General

Q: What is Amazon Lightsail?

Amazon Lightsail is the easiest way to get started with AWS for developers, small businesses, students, and other users who need a simple virtual private server (VPS) solution. Lightsail provides developers compute, storage, and networking capacity and capabilities to deploy and manage websites and web applications in the cloud. Lightsail includes everything you need to launch your project quickly – a virtual machine, SSD-based storage, data transfer, DNS management, and a static IP – for a low, predictable monthly price.

Q: What can I do with Lightsail?

You can get preconfigured virtual private server plans that include everything to easily deploy and manage your application. Lightsail is best suited to projects that require a few virtual private servers and users who prefer a simple management interface. Common use cases for Lightsail include running websites, web applications, blogs, e-commerce sites, simple software, and more.

Q: What is a Lightsail plan?

Also referred to as a bundle, a Lightsail plan includes a virtual server with a fixed amount of memory (RAM) and compute (vCPUs), SSD-based storage (disks), and a free data transfer allowance. Lightsail plans also offer static IP addresses (5 per account) and DNS management (3 domain zones per account). Lightsail plans are charged on an hourly, on-demand basis, so you only pay for a plan when you're using it.

Q: What is a Lightsail instance?

A Lightsail instance is a virtual private server (VPS) that lives in the AWS Cloud. Use your Lightsail instances to store your data, run your code, and build web-based applications or websites. Your instances can connect to each other and to other AWS resources through both public (Internet) and private (VPC) networking. You can create, manage, and connect easily to instances right from the Lightsail console.

Q: What software can I run on my instance?

Lightsail offers a range of operating system and application templates that are automatically installed when you create a new Lightsail instance. Application templates include WordPress, Drupal, Joomla!, Ghost, Magento, Redmine, LAMP, Nginx (LEMP), MEAN, Node.js, Django, and more.

You can install additional software on your instances by using the in-browser SSH or your own SSH client.

Q: How do I create a Lightsail instance?

After logging in to Lightsail, you can use the Lightsail console, command line interface (CLI), or API to create and manage instances.

The first time you log in to the console, choose **Create Instance**. The create instance page is where you can choose the software, location, and name for your instance. Once you choose **Create**, your new instance will spin up automatically within minutes.

Q: Does Lightsail offer an API?

Yes. Everything you do in the Lightsail console is backed by a publically available API. Learn how to install and use the Lightsail CLI and API.

Q: How do I sign up for Lightsail?

To start using Lightsail, choose Get Started and log in. You use your Amazon Web Services account to access Lightsail; if you don't already have one, you'll be prompted to create one.

Lightsail resources

Q: How do Lightsail instances perform?

Lightsail instances are specifically engineered by AWS for web servers, developer environments, and small database use cases. Such workloads don't use the full CPU often or consistently, but occasionally need a performance burst. Lightsail uses burstable performance instances that provide a baseline level of CPU performance with the additional ability to burst above the baseline. This design enables you to get the performance you need, when you need it, while protecting you from the variable performance or other common side effects that you might typically experience from over-subscription in other environments. For more information on burstable performance, go here.

If you need highly configurable environments and instances with consistently high CPU performance for applications such as video encoding or HPC applications, we recommend you use Amazon EC2.

Q: How do I connect to a Lightsail instance?

Lightsail offers a 1-click secure connection to your instance's terminal right from your browser, supporting SSH access for Linux/Unix-based instances and RDP access for Windows-based instances. To use 1-click connections, launch your instance management screens, click **Connect using SSH** or **Connect using RDP**, and a new browser window opens and automatically connects to your instance.

If you prefer to connect to your Linux/Unix-based instance using your own client, Lightsail will do the SSH key storing and management work for you, and provide you with a secure key to use in your SSH client.

Q: How do I use IPs in Lightsail?

Each Lightsail instance automatically gets a private IP address and a public IP address. You can use the private IP to transmit data between Lightsail instances and AWS resources privately, for free. You can use the public IP to connect to your instance from the Internet, such as through a registered domain name or through an SSH or RDP connection from your local computer. You can also attach a static IP to the instance, which substitutes the public IP with an IP address that doesn't change even if the instance is stopped and started.

Q: What is a static IP?

A static IP is a fixed, public IP that is dedicated to your Lightsail account. You can assign a static IP to an instance, replacing its public IP. If you decide to replace your instance with another one, you can reassign the static IP to the new instance. In this way, you don't have to reconfigure any external systems (like DNS records) to point to a new IP every time you want to replace your instance.

Q: What are DNS records?

DNS is a globally distributed service that translates human readable names like www.example.com into numeric IP addresses, like 192.0.2.1 that computers use to connect to each other. With Lightsail, you can easily map your registered domain names such as photos.example.com to the public IPs of your Lightsail instances. In this way, when users type human readable names like example.com into their browsers, Lightsail automatically translates the address into the IP of the instance you want to direct your users to. Each of these translations is referred to as a DNS query.

It's important to know that in order to use a domain in Lightsail, you must first register it. You can register new domains using Amazon Route 53, or your preferred DNS registrar.

Q: Can I manage firewall settings for my instance?

Yes. You can control the data traffic for your instances by using the Lightsail firewall. From the Lightsail console, you can set rules about which ports of your instance are publicly accessible for different types of traffic.

Q: What are snapshots?

Snapshots are point-in-time backups of instances, databases, or block storage disks. You can create a snapshot of your resources at any time, or you can enable automatic snapshots on instances and disks to have Lightsail create snapshots for you. You can use snapshots as baselines to create new resources or to back up your data. A snapshot contains all of the data that is needed to restore your resource (from the moment when the snapshot was taken). When you restore a resource by creating it from a snapshot, the new resource begins as an exact replica of the original resource that was used to create the snapshot.

You can manually take snapshots of your Lightsail instances, disks, and databases, or you can use automatic snapshots to instruct Lightsail to take daily snapshots of your instances and disks automatically. For more information, see Snapshots in Amazon Lightsail.

Q: What are automatic snapshots?

Automatic snapshots are a way to schedule daily snapshots of your Linux/Unix instances in Amazon Lightsail. You can pick a time of the day, and Lightsail will automatically take a snapshot for you each day at the time you chose and always keep your seven most recent automatic snapshots. Enabling snapshots is free – you only pay for the actual storage used by your snapshots.

Q: What are the differences between manual snapshots and automatic snapshots?

Automatic snapshots cannot be tagged or exported directly to Amazon EC2. However, automatic snapshots can be copied and converted into manual snapshots. To copy an automatic snapshot into a manual one, choose Keep from the automatic snapshot’s context menu to copy it as a manual snapshot.

Q: How can I back up my instances?

If you want to back up your data, you can use the Lightsail console or API to take a snapshot of your instance. If there is a failure or bad code deployment, you can later use your instance snapshot to create a brand new instance. We recommend stopping your instance temporarily when taking a snapshot, to ensure your data is complete and not corrupted in any way.

Q: What is the difference between stopping and deleting my instance?

When you stop your instance, it is powered down at its current state and is available for you to start again at any time. Stopping your instance will release its public IP, so it is recommended that you use static IPs for instances that must retain the same IP after they stop.

When you delete your instance, you are performing a destructive action. Unless you have taken an instance snapshot, all of your instance data will be lost and you cannot recover it again. The instance's public and private IPs will also be released. If you were using a static IP with that instance, the static IP is detached, but remains in your account.

Q: Can I upgrade my plan?

Yes. You can take a snapshot of your instance, and use the API to launch a new, larger size instance. You can launch new instances from snapshot using the Lightsail console or the CLI. Find instructions on how to use the CLI here.

Q: How can I connect Lightsail instances to other resources in my AWS account?

You can connect your Lightsail instances to VPC resources in your AWS account privately, by using VPC peering. Just choose Enable VPC peering on your Lightsail account page, and Lightsail does the work for you. Once VPC peering is enabled, you can address other AWS resources in your default AWS VPC by using their private IPs. Find instructions here.

Note that you need to have a default VPC set up in your AWS account in order for VPC peering with Lightsail to work. AWS accounts created before December 2013 do not have a default VPC, and you will need to set one up. Find out more about setting up your default VPC here.

Q: In which regions is Lightsail available?

Lightsail is currently available in all the Availability Zones in the following AWS Regions:

US East (N. Virginia)
US East (Ohio)
US West (Oregon)
Canada (Central)
Europe (Frankfurt)

Europe (Ireland)
Europe (London)
Europe (Paris)
Asia Pacific (Mumbai)
Asia Pacific (Singapore)
Asia Pacific (Sydney)
Asia Pacific (Tokyo)
Asia Pacific (Seoul)

Q: What are Availability Zones?

Availability Zones are collections of data centers that run on physically distinct, independent infrastructure and are engineered to be highly reliable. Common points of failure such as generators and cooling equipment are not shared between Availability Zones. Additionally, Availability Zones are physically separate, so that even extremely uncommon disasters such as fires, tornados, or flooding can affect only a single Availability Zone.

Q: What are the Lightsail service limits?

You can currently create up to 20 Lightsail instances, 5 static IPs, 3 DNS zones, 20 TB of attached block storage, and 5 load balancers in a Lightsail account. You can also generate up to 20 certificates during each calendar year. If you need to increase your account limit for instances, static IPs, block storage, or certificates in your account, please open a case with customer service. We do not currently support increases for DNS zones or load balancers.

Q: How can I get more help?

We're here for you. Quick Assist in Lightsail offers immediate helpful tips about your actions in the console. From the Lightsail console, you can also access a library of getting started guides, overviews, and how-to topics. And if you want to use the API or CLI, Lightsail has a full API reference for all supported programming languages. You can also use Lightsail support resources:

- If you have an issue with your account or billing, contact customer service online. You get free 24x7 access with your Lightsail account.

- If you have a general question about how to use Lightsail, search the Lightsail documentation and support forums.

Additionally, AWS Support offers an array of paid plans to cover your individual needs.

Q: What operating systems can I use with Amazon Lightsail?

Lightsail currently supports 6 Linux or Unix-like distributions – Amazon Linux, Debian, FreeBSD, OpenSUSE, and Ubuntu – and 2 Windows Server versions – 2012 R2 and 2016.

Q: What are tags?

A tag is a label that you assign to a Lightsail resource. Each tag consists of a key and a value, both of which you define. A tag value is optional, so you can choose to create “key-only” tags for filtering resources in the Lightsail console.

Q: What can I use tags to do in Lightsail?

Tags have multiple use cases - they enable you to group and filter your resources in the Lightsail console and API, track and organize your costs in your bill, and regulate who can see or modify your resources through access management rules. By tagging your resources you can:

- Organize - use the Lightsail console and API filters to view and manage resources based on their tags you have assigned them. This is useful when you have many resources of the same type—you can quickly identify a specific resource based on the tags you've assigned to it.
- Cost-allocate - track and allocate costs across different projects or users by tagging your resources and creating “cost allocation tags” in the billing console. For instance, you can split out your bill and understand your costs by project or by client.
- Manage access - control how users with access to your AWS account can edit, create, and delete Lightsail resources by using AWS Identity and Access Management policies. This allows you to more easily collaborate with others without needing to give them full access to your Lightsail resources.

Learn more about using tags in Lightsail >>

Q: What resources can be tagged?

Lightsail current supports tagging for the following resources:

- Instances (Linux and Windows)
- Block storage disks
- Load balancers
- Databases
- DNS zones
- Instance, disk, and database snapshots

Manual snapshots also support tags and are automatically given the same tags as the source resource. You can edit these tags when you use a snapshot to create a new instance, disk, or database.

Q: How can I tag my Lightsail snapshots?

The Lightsail console automatically tags manual snapshots with the same tags as its parent resource. However, tags are not automatically copied over from a resource to its automatic snapshots. If you use the Lightsail API to create a snapshot, you can choose the tags for the snapshot yourself.

Important: Database snapshots tags are not currently included in billing reports (cost allocation tags).

Q: What is the difference between key-value and key-only Tags?

Lightsail tags are key-value pairs, allowing you to organize resources such as instances across different categories, e.g. project:Blog, project:Game, project:Test. This allows you full conttol across all use cases such as resource organization, bill reporting, and access management. The Lightsail console also allows you to tag your resources with key-only tags for quick filtering in the console.

Q: Does Amazon Lightsail support monitoring and alerting?

Yes. With Amazon Lightsail you can collect metrics on various resources, including instances, load balancers and databases. For any individual resource you can setup to two alarms thresholds for each metric. If the alarm threshold is breached you will receive a notification in the Lightsail console, and, optionally, you can choose to receive an email message and/or SMS message. There is no additional charge for the alerting and monitoring feature in Lightsail, however, you may incur charges from your mobile carrier for SMS messaging.

Billing and account management

Q: What do Lightsail plans cost?

Lightsail plans are billed on an on-demand hourly rate, so you pay only for what you use. For every Lightsail plan you use, we charge you the fixed hourly price, up to the maximum monthly plan cost. The least expensive Lightsail plan starts at $0.0047 USD/hour ($3.50 USD/month). Lightsail plans that include a Windows Server license start at $0.01075 USD/hour ($8 USD/month).

Q: When am I getting charged for a plan?

Your Lightsail instances are charged only when they're in the running or stopped state. If you delete your Lightsail instance before the end of the month, we only charge you a prorated cost, based on the total number of hours that you used your Lightsail instance. For example, if you use the least expensive Lightsail plan for 100 hours in a month, you will be charged 47 cents (100*0.0047).

Q: What do Lightsail static IPs cost?

They're free in Lightsail, as long as you are using them! You don't pay for a static IP if it is attached to an instance. Public IPs are a scarce resource and Lightsail is committed to helping to use them efficiently, so we charge a small $0.005 USD/hour fee for static IPs not attached to an instance for more than 1 hour.

Q: What does data transfer cost?

Your plan includes a free data transfer allowance. Both data transfer in and data transfer out of your instance count toward your data transfer allowance.

If you exceed your data transfer allowance, you will only get charged for data transfer OUT from a Lightsail instance to the Internet or to AWS resources using the public IP address of the instance. Both data transfer IN to Lightsail instances and data transfer OUT from a Lightsail instance when using the instance's private IP address are free beyond your data transfer allowance.

Q: What is my data transfer plan allowance?

Every single Lightsail plan also includes a healthy amount of free IN and OUT data transfer. For example, using the cheapest Lightsail bundle you can send and receive up to 1 TB of data to the Internet within the month, at no extra charge.

Q: How does my data transfer allowance work?

Any type of data transfer you consume is covered by your Lightsail plan and counted towards your data transfer allowance. As long as your instance's data transfer is below the plan allowance, you do not incur any data transfer charges. Your data transfer allowance will reset every month, and you can consume it whenever you need within the month. If you delete your instance before the month ends and create another one, data transfer allowance is shared between the two instances.

Q: What if I exceed my data transfer plan allowance?

We have designed our data transfer plans so that the vast majority of our customers will be fully covered by their allowance and not incur any additional charges. Even if you exceed your data transfer allowance, some types of data transfer are free. Data transfer IN to Lightsail instances is always free. Data transfer OUT from a Lightsail instance to another Lightsail instance or AWS resource in the same Region is also free if private IP addresses are used.

Q: What types of data transfer do I get charged for?

When you exceed the monthly free data transfer allowance of your plan, you will get charged for data transfer OUT from a Lightsail instance to the Internet or to another AWS Region or to AWS resources in the same Region when using public IP addresses. The charge for these types of data transfer above the free allowance is as follows:

- US East (N. Virginia): $0.09 USD/GB
- US East (Ohio): $0.09 USD/GB
- US West (Oregon): $0.09 USD/GB
- Canada (Central): $0.09 USD/GB
- Europe (Frankfurt): $0.09 USD/GB
- Europe (Ireland): $0.09 USD/GB
- Europe (London): $0.09 USD/GB
- Europe (Paris): $0.09 USD/GB
- Asia Pacific (Mumbai): $0.13 USD/GB
- Asia Pacific (Singapore): $0.12 USD/GB
- Asia Pacific (Sydney): $0.17 USD/GB
- Asia Pacific (Tokyo): $0.14 USD/GB
- Asia Pacific (Seoul): $0.13 USD/GB

Instances created in different Availability Zones can communicate between zones privately and for free, and are much less likely to be impaired concurrently. Availability Zones enable you to build highly available applications and websites without increasing the cost of data transfer or compromising your application's security.

Q: How do my data transfer plan allowances vary by Region?

All AWS Regions have the same data transfer plan allowance as listed on amazonlightsail.com and amazonlightsail.com/pricing, with the exception of the Asia Pacific (Mumbai) and the Asia Pacific (Sydney) Regions. In these two AWS Regions, the data transfer plan allowance is as follows:

- 512 MB plan: 500 GB
- 1 GB plan: 1 TB
- 2 GB plan: 1.5 TB
- 4 GB plan: 2 TB
- 8 GB plan: 2.5 TB
- 16 GB plan: 3 TB
- 32 GB plan: 3.5 TB

Q: How does my data transfer allowance work with my load balancers?

Your load balancer does not consume your data transfer allowance. Traffic between the load balancer and the target instances is metered and counts toward your data transfer allowance for your instances, in the same way that traffic in from and out to the internet is counted toward your data transfer allowance for Lightsail instances that are not behind a load balancer. Traffic into and out of your load balancer to the internet is not calculated toward the data transfer allowance for your instances.

Q: What does Lightsail DNS management cost?

DNS management is free within Lightsail. You can create up to 3 DNS zones and as many records as you want for each DNS zone. You also get a monthly allowance of 3 million DNS queries per month to your zones. Beyond your first 3 million queries in a month, you are charged $0.40 USD/million DNS queries.

Q: What do Lightsail snapshots cost?

Lightsail snapshots cost $0.05 USD/GB-month for both instance snapshots and for disk snapshots. This means if you take a snapshot of your 30 GB SSD instance and keep it for a month, you pay $1.50 USD at the end of the month.

When you take multiple successive snapshots of the same instance, Lightsail automatically cost-optimizes your snapshots. For each new snapshot you take, you're changed only for the part of the instance that changed. In the example above, if your instance only changes by 2 GB, your second instance snapshot costs only $0.10 USD per month.

Q: What does Lightsail block storage cost?

Lightsail Block storage costs $0.10 USD per GB per month.

Q: What do Lightsail load balancers cost?

Lightsail load balancers cost $18 per month.

Q: What does certificate management cost?

Lightsail certificates and certificate management are free with use of a Lightsail load balancer.

Q: Can I try Lightsail for free?

Yes! Whether you're an existing or new AWS customer, you get 750 hours of free usage of the 512 MB Lightsail plan for free. You also can try Lightsail plans that include a Windows Server license for free using the same plan size.

You can use your 750 hours of usage across as many instances as you like. For example, you can run a single Lightsail instance for a whole month, or 10 Lightsail instances for 75 hours. The free trial offer is only applicable to usage within the first calendar month from when you sign up to use Lightsail.

Q: How can I manage my AWS account?

Lightsail is an AWS service and runs on the AWS trusted and proven cloud infrastructure. You use the same AWS account and credentials to log in to Lightsail and the AWS Management Console.

You can manage your AWS account, including changing your AWS account password, user name, contact information, or billing information from the AWS Billing and Cost Management console.

Q: What are the Lightsail legal terms of use?

Lightsail is an Amazon web service, so to use Lightsail, you first agree to the AWS Customer Agreement and Service Terms. When creating Lightsail instances, you also agree that your use of software is also subject to the end user license agreement of the seller, available for your review on the create instance page.

Q: How can I pay my Lightsail bill?

You can pay and manage your bill through the AWS Billing and Cost Management console. AWS accepts most major credit cards. Learn more about managing your payment methods.

Lightsail block storage

Q: What can I do with Lightsail block storage?

Lightsail block storage provides additional storage volumes (called “attached disks” in Lightsail) that you can attach to your Lightsail instance, similar to an individual hard drive. Attached disks are useful for applications or software that need to separate out specific data from their core service and to protect application data in case of a failure or other issue with your instance and system disk. Attached disks offers consistent performance and low latency needed for applications or software that frequently access their stored data.

Lightsail block storage uses solid-state drives (SSD). This type of block storage balances a low price and good performance and is intended to support the vast majority of workloads that run on Lightsail. For customers with applications that require sustained IOPS performance, high amounts of throughput per disk, or that are running large databases like MongoDB, Cassandra, etc., we recommend using EC2 with GP2 or Provisioned IOPS SSD storage instead of Lightsail.

Q: How large can I make my attached disk?

Each attached disk can be up to 16 TB.

Q: How many disks can I attach per Lightsail instance?

You can attach up to 15 disks per Lightsail instance.

Q: Can I attach a disk to more than 1 instance?

No, disks can only be attached to one instance at a time.

Q: Does my disk need to be attached to an instance?

No, you can choose not to attach a disk to an instance. The disk will remain in your account in an unattached state. There is no difference in price if your disk is not attached to an instance.

Q: Can I increase the size of my attached disk?

Yes, you can increase the size of a disk by taking a disk snapshot and then creating a new, larger disk from snapshot.

Q: Does Lightsail block storage offer encryption?

Yes, to help keep your data secure, all Lightsail attached disks and disk snapshots are encrypted at rest by default, using keys that Lightsail manages on your behalf. Lightsail also provides encryption of data as it moves between Lightsail instances and attached disks.

Q: What availability can I expect from Lightsail block storage?

Lightsail block storage is designed to be highly available and reliable. Each attached disk is automatically replicated within its Availability Zone to protect you from component failure. Although Lightsail does not guarantee SLAs, Lightsail block storage disks are designed for 99.99% availability and for an annual failure rate of less than 0.2%. Lightsail also supports disk snapshots to allow regular backups of your data.

Q: How do I back up my attached disk?

You can back up your disk by taking a disk snapshot. You can also backup your entire instance and any attached disks by taking an instance snapshot.

Q: How are attached disks different than the storage included in my Lightsail plan?

The system disk included with your Lightsail plan is your instance’s root device. If you terminate your instance, the system disk will be deleted as well. If you experience an instance failure, the system disk could be impacted. You also cannot detach your system disk or back it up separately from your instance. Data stored on an attached disk persists independently of the instance. Attached disks can be detached and moved between instances and can be backed up independently from an instance using disk snapshot. To protect your data, we recommend that you use your Lightsail instance’s system disk only for temporary data. For data requiring a higher level of durability, we recommend using attached disks and regularly backing up your disk using disk or instance snapshots.

Lightsail load balancer

Q: What can I do with Lightsail load balancers?

Lightsail load balancers allow you to build highly available websites and applications. By distributing traffic across instances in different Availability Zones and pointing traffic to only healthy target instances, Lightsail load balancers reduce the risk of your application going down due to an issue with your instance or to a datacenter outage. With Lightsail load balancers and multiple target instances, your website or application can also accommodate increases in web traffic and maintain good performance for your visitors during peak load times.

In addition, you can use Lightsail load balancers to build secure applications and accept HTTPS traffic. Lightsail takes the complexity out of requesting, provisioning, and maintaining SSL/TLS certificates. The built-in certificate management requests and renews certificates on your behalf and adds the certificate to your load balancer automatically.

Q: Can I use load balancers with instances in different AWS Regions or different Availability Zones?

You cannot use load balancers with instances running in different AWS Regions. You can, however, use target instances across different Availability Zones with your load balancer. In fact, we recommend that you distribute your target instances across Availability Zones to maximize the availability of your application.

Q: How does my Lightsail load balancer deal with traffic spikes?

Lightsail load balancers scale automatically to handle traffic spikes to your application without you having to manually adjust them. If your application experiences a transient spike in traffic, your Lightsail load balancer will automatically scale and continue to efficiently direct traffic to your Lightsail instances. While your Lightsail load balancer is designed to easily manage traffic spikes, applications that consistently experience very high volume levels of traffic may experience performance degradation or throttling. If you expect your application consistently to manage more than 5 GB/hour of data or consistently to have a large number of connections (>400k new connections/hour, >15k active, concurrent connections), we recommend using Amazon EC2 with Application Load Balancing instead.

Q: How do Lightsail load balancers route traffic to my target instances?

Lightsail load balancers direct traffic to your healthy target instances based on a round robin algorithm.

Q: How does Lightsail know if my target instances are healthy?

During load balancer creation, you will be asked to specify a path (a common file or webpage URL) for Lightsail to ping. If the target instance can be reached using this path, then Lightsail will route traffic there. If one of your target instances is unresponsive, Lightsail will not route traffic to that instance. You can update the Health check path if needed in the load balancer management screens.

Q: What is the difference between key-value and key-only Tags?

Lightsail tags are key-value pairs, allowing you to organize resources such as instances across different categories, e.g. project:Blog, project:Game, project:Test. This allows you full control across all use cases such as resource organization, bill reporting, and access management. The Lightsail console also allows you to tag your resources with key-only tags for quick filtering in the console.

Q: Can I assign one instance to multiple load balancers?

Yes, Lightsail supports adding instances as target instances for more than one load balancer, if desired.

Q: What happens to my target instances when I delete my load balancer?

If you delete your load balancer, the attached target instances will continue to run normally and will appear in the Lightsail console as regular Lightsail instances. Please note that you will likely need to update your DNS records to direct traffic to one of your former target instances after you delete the load balancer.

Q: What is session persistence?

Session persistence enables the load balancer to bind a visitor's session to a specific target instance. This ensures that all requests from the user during the session are sent to the same target instance. Lightsail supports session persistence for applications that require visitors to hit the same target instances for data consistency. For example, many applications that require user authentication can benefit from using session persistence. You can turn on session persistence for specific load balancer from the load balancer management screens after creation.

Q: What kind of connections do Lightsail load balancers support?

Lightsail load balancers support HTTP and HTTPS connections.

Certificate management

Q: How can I use Lightsail-provisioned certificates?

SSL/TLS certificates are used to establish the identity of your website or application and secure connections between browsers and your website.

Lightsail provides a signed certificate to use with your load balancer, and the load balancer provides SSL/TLS termination before routing verified traffic to your target instances over the secure AWS network. Lightsail certificates can only be used with Lightsail load balancers, not with individual Lightsail instances.

Q: How do I validate my certificate?

Lightsail certificates are domain validated, meaning that you need to provide proof of identity by validating that you own or have access to your website's domain before the certificate can be provisioned by the certificate authority. When you request a new certificate, Lightsail will prompt you to add a CNAME to the DNS zone(s) of the domain or domains you are validating. You will add this CNAME wherever you currently manage your DNS zones – either Lightsail DNS management or an external DNS hosting provider (e.g., Route 53, GoDaddy, Namecheap, etc.). Once your certificate is validated, you can remove the CNAME record from your DNS zone, if desired.

Q: What happens if I cannot validate my domain?

You must be able to validate that you own a domain for security purposes. This means if you or someone in your organization cannot add a DNS record to validate your certificate for any reason, you will not be able to use an HTTPS-enabled load balancer with Lightsail.

Q: How many domains and subdomains can I add to my certificate?

You can add up to 10 domains or subdomains per certificate. Lightsail does not currently support wild card domains.

Q: How can I change the domains associated with my certificate?

To change the domains (add/delete) associated with your certificate, you will need to resubmit the certificate and revalidate your ownership of the domain(s). Follow the steps in the certificate management screens to regenerate your certificate and add or remove domains when prompted.

Q: How do I renew my certificate?

Lightsail provides managed renewal for your SSL/TLS certificates. This means that Lightsail tries to renew the certificates automatically before they expire with no action required from you. Your Lightsail certificate must be actively associated load balancer before it can be automatically renewed.

Q: What happens to my certificate when I delete my load balancer?

If your load balancer is deleted, your certificate is deleted as well. If you need to use a certificate for the same domain(s) in the future, you will need to request and validate a new certificate.

Q: Can I download my certificate provided by Lightsail?

No, Lightsail certificates are bound to your Lightsail account and cannot be removed and used outside of Lightsail.

Upgrade to EC2

Q: What is Upgrade to EC2?

Upgrade to EC2 is a feature that allows you to create a copy of your Lightsail instance in Amazon EC2. When you upgrade to EC2, you can pick among the wide set of instance types, configurations, and pricing models that EC2 offers, and have even more fine-tuned control over your networking, storage, and compute environment.

Q: Why would I want to upgrade to EC2?

Lightsail offers you an easy way to run and scale a wide set of cloud-based applications, at a bundled, predictable, and low price. Lightsail also automatically sets up your cloud environment configurations such as networking and access management.

Upgrading to EC2 allows you to run your application on a wider set of instance types, ranging from virtual machines with more CPU power, memory, and networking capabilities, to specialized or accelerated instances with FPGAs and GPUs. In addition, EC2 performs less automatic management and set-up, allowing you more control over how you configure your cloud environment, such as your VPC.

Q: How does it work?

To get started, you need to export your Lightsail instance manual snapshot. You'll then use the Upgrade to EC2 wizard to create an instance in EC2.

Customers who are comfortable with EC2 can then use the EC2 creation wizard or API to create a new EC2 instance as they would from an existing EC2 AMI. Alternatively, Lightsail also provides a guided Lightsail console experience to help you easily create a new EC2 instance.

Note: Exporting Ghost and Django instance manual snapshots to Amazon EC2 is not supported at this time. We apologize for the inconvenience.

Q: How am I billed?

Using the Upgrade to EC2 feature is free. Once you have exported your snapshots to EC2, you will be charged for the EC2 image separately and in addition to your Lightsail snapshot. Any new EC2 instances you launch will also be billed by EC2, including their EBS storage volume(s) and data transfer. Refer to the EC2 pricing page for details on the pricing for your new instance and resources. Lightsail resources that continue to run in your Lightsail account will continue to be billed at their regular rates until they are deleted.

Q: Can I export managed database or disk snapshots?

The upgrade feature allows you to export Lightsail disk manual snapshots but doesn't currently support managed database snapshots. Disk snapshots can be rehydrated as EBS volumes from the EC2 console or API.

Q: What Lightsail resources can I upgrade?

Lightsail's upgrade to EC2 feature is designed to support the export of Linux and Windows instances and their attached block storage (if applicable) to EC2. It also supports the export of unattached block storage disks to EBS. It does not currently support the export of load balancers, databases, static IPs or DNS records.

Managed databases

Q: What are Lightsail's managed databases?

Lightsail's managed databases are instances that are dedicated to running databases, instead of other workloads like webserver, mail server, etc. A Lightsail database can contain multiple user-created databases, and you can access it by using the same tools and applications that you use with a stand-alone database. Lightsail maintains the security and health of your database's underlying infrastructure and operating system, so that you can run a database without deep expertise in infrastructure management.

Like regular Lightsail instances, Lightsail databases come with a fixed amount of memory, computing power, and SSD based storage in their plans that you can scale up over time. Lightsail will automatically install and configure your chosen database for you upon creation.

Q: What can I do with Lightsail's managed databases?

Lightsail managed databases provide an easy, low maintenance way to store your data in the cloud. You can run Lightsail databases either as a new databases or by migrating from an existing on-premises or hosted database to Lightsail.

They can also allow you to scale your application to accept larger amounts of traffic and more intensive loads, by separating out your database into a dedicated instance. Lightsail databases are especially useful for stateful applications – like WordPress and most common CMSs – that need data to be kept in sync when you scale beyond a single instance. Lightsail databases can be paired with a Lightsail load balancer and two or more Lightsail instances to create a powerful, scaled application. By using Lightsail High Availability plans, you can also create add redundancy to your database, helping to ensure high uptime for your application.

Q: What does Lightsail manage for me on my managed database?

Lightsail manages a range of maintenance activities and security for your database and its underlying infrastructure. Lightsail automatically backs up your database and allows point in time restore from the past 7 days using the database restore tool, to help protect against data loss or component failure. Lightsail also automatically encrypts your data at rest and in motion for increased security and stores your database password for easy and secure connections to your database. On the maintenance side, Lightsail runs maintenance on your database during your set maintenance window. This maintenance include automatic upgrades to the latest minor database version and all management of the underlying infrastructure and operating system.

Q: What managed database plans does Lightsail offer?

Lightsail offers 4 sizes of databases in Standard and High Availability plans. Each plan comes with a fixed amount of storage and a monthly allowance of data transfer. You can also scale up to larger plans over time, as needed, and switch between Standard and High Availability plans. High Availability plans mirror the same resources as Standard plans and additionally include a standby database running in a separate Availability Zones from your primary database for redundancy.

Q: What is a High Availability plan?

Managed databases are available in Standard and High Availability plans. Standard and High availability plans have identical plan resources, including memory, storage, and data transfer allowance. High Availability plans add redundancy and durability to your database, by automatically creating standby database in a separate Availability Zone from your primary database, synchronously replicating data to the standby database, and providing failover to the standby database in case of infrastructure failure and during maintenance so that you ensure uptime even when databases is being automatically upgraded/maintained by Lightsail. Use High Availability plans for running production applications or software where high uptime is required.

Q: How do I scale up or down my managed database?

You can scale up your database by taking a snapshot of your database and creating a new, larger database plan from snapshot or by creating a new, larger database using the emergency restore feature. You can also switch from Standard to High Availability plans and vice versa using either method. You cannot scale down your database. To learn more, see Creating a database from a snapshot in Amazon Lightsail.

Q: How can I back up my managed database?

Lightsail backs up your data automatically and allows restore of this data from a specific point in time to a new database. Automatic backup is a free service for your database but only saves the last 7 days of data. If you delete your database, all automatic backup records are deleted and point in time restore is no longer possible. To retain backups of data after deleting your database or to retain a backup for more than 7 days in the past, use manual snapshots.

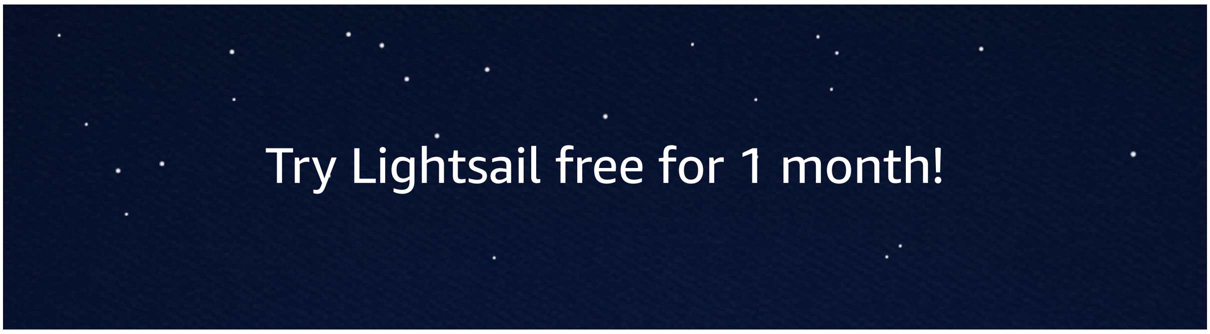
You can take manual snapshots of your Lightsail managed databases from the database management pages. Manual snapshots contain all the data from your database and can be used as backups for data that you want to store permanently. You can also use manual snapshots to create a new, larger database or to switch between Standard and High Availability plans. Manual snapshots are stored until you delete them and are billed at $0.05 USD/GB-month.

Q: What happens to my data if I delete my managed database?

If you delete your managed database, both your database itself and all automatic backups will be deleted. There is no way to recover this data unless you take a manual snapshot before deleting your database. During deletion of your database, Lightsail provides a one-click option to take a manual snapshot, if desired, to help protect against accidental loss of data. Taking a manual snapshot before deletion is optional but highly recommended. You can delete your manual snapshot in the future when you no longer need the stored data.

Q: How do managed databases work with my Lightsail instances?

After you create your managed database, you can start using your database with your application immediately, using your Lightsail instances as web servers or other dedicated workloads for your app. To connect your Lightsail instance to a database, use your database endpoint and reference your securely stored password to configure the database as your data store in the code of your application. You can find connection data in the database management screens. The file name and location for your database configuration file will vary by application. Note that you can connect many instances to one database, either using the same tables or using different ones.



Amazon Athena FAQs

General

Q: What is Amazon Athena?
Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to setup or manage, and you can start analyzing data immediately. You don't even need to load your data into Athena, it works directly with data stored in S3. To get started, just log into the Athena Management Console, define your schema, and start querying. Amazon Athena uses Presto with full standard SQL support and works with a variety of standard data formats, including CSV, JSON, ORC, Apache Parquet and Avro. While Amazon Athena is ideal for quick, ad-hoc querying and integrates with Amazon QuickSight for easy visualization, it can also handle complex analysis, including large joins, window functions, and arrays.

Q: What can I do with Amazon Athena?
Amazon Athena helps you analyze data stored in Amazon S3. You can use Athena to run ad-hoc queries using ANSI SQL, without the need to aggregate or load the data into Athena. Amazon Athena can process unstructured, semi-structured, and structured data sets. Examples include CSV, JSON, Avro or columnar data formats such as Apache Parquet and Apache ORC. Amazon Athena integrates with Amazon QuickSight for easy visualization. You can also use Amazon Athena to generate reports or to explore data with business intelligence tools or SQL clients, connected via an ODBC or JDBC driver.

Q: How do I get started with Amazon Athena?
To get started with Amazon Athena, simply log into the AWS Management Console for Athena and create your schema by writing DDL statements on the console or by using a create table wizard. You can then start querying data using a built-in query editor. Athena queries data directly from Amazon S3 so there's no loading required.

Q: How do you access Amazon Athena?
Amazon Athena can be accessed via the AWS Management Console, an API, or an ODBC or JDBC driver. You can programmatically run queries, add tables or partitions using the ODBC or JDBC driver.

Q: What are the service limits associated with Amazon Athena?

Please click here to learn more about service limits.

Q: What is the underlying technology behind Amazon Athena?
Amazon Athena uses Presto with full standard SQL support and works with a variety of standard data formats, including CSV, JSON, ORC, Avro, and Parquet. Athena can handle complex analysis, including large joins, window functions, and arrays. Because Amazon Athena uses Amazon S3 as the underlying data store, it is highly available and durable with data redundantly stored across multiple facilities and multiple devices in each facility. Learn more about Presto here.

Q: How does Amazon Athena store table definitions and schema?
Amazon Athena uses a managed Data Catalog to store information and schemas about the databases and tables that you create for your data stored in Amazon S3. In regions where AWS Glue is available, you can upgrade to using the AWS Glue Data Catalog with Amazon Athena. In regions where AWS Glue is not available, Athena uses an internal Catalog.
You can modify the catalog using DDL statements or via the AWS Management Console. Any schemas you define are automatically saved unless you explicitly delete them. Athena uses schema-on-read technology, which means that your table definitions applied to your data in S3 when queries are being executed. There's no data loading or transformation required. You can delete table definitions and schema without impacting the underlying data stored on Amazon S3.

Q: Why should I upgrade to AWS Glue Data Catalog?
AWS Glue is a fully managed ETL service. Glue has three main components: 1) a crawler that automatically scans your data sources, identifies data formats and infers schemas, 2) a fully managed ETL service that allows you to transform and move data to various destinations, and 3) a Data Catalog that stores metadata information about databases & tables either stored in S3 or an ODBC- or JDBC-compliant data store. To use the benefits of Glue, you must upgrade from using Athena's internal Data Catalog to the Glue Data Catalog.
The benefits of upgrading to the Glue Data Catalog are:

1. **Unified Metadata Repository:** AWS Glue is integrated across a wide range of AWS services. AWS Glue supports data stored in Amazon Aurora, Amazon RDS MySQL, Amazon RDS PostreSQL, Amazon Redshift, and Amazon S3, as well as MySQL and PostgreSQL databases in your Virtual Private Cloud (Amazon VPC) running on Amazon EC2. AWS Glue provides out-of-the-box integration with Amazon Athena, Amazon EMR, Amazon Redshift Spectrum, and any Apache Hive Metastore-compatible application.

2. **Automatic schema and partition recognition:** AWS Glue automatically crawls your data sources, identifies data formats, and suggests schemas and transformations. Crawlers can help automate table creation and automatic loading of partitions.

3. **Easy to build pipelines:** AWS Glue’s ETL engine generates Python code that is customizable, reusable, and portable. You can edit the code using your favorite IDE or notebook and share it with others using GitHub. Once your ETL job is ready, you can schedule it to run on AWS Glue's fully managed, scale-out Spark infrastructure. AWS Glue is serverless, so it handles provisioning, configuration, and scaling of the resources required to run your ETL jobs, allowing you to tightly integrate ETL in your workflow.

Click here to learn more about the Glue Data Catalog.

Q: Is there a step-by-step to upgrade to the AWS Data Catalog?
Yes. Step-by-Step guide can be found here.

Q: What regions is Amazon Athena available in?
Please refer to Regional Products and Services for details of Amazon Athena service availability by region.

Preview features

Q: What preview features are available in Athena?
You can now invoke your SageMaker machine learning (ML) models in an Athena SQL query to run inference. The ability to use ML models in SQL queries makes complex tasks such anomaly detection, customer cohort analysis and sales predictions as simple as writing a SQL query. Learn more.

With federated query, you can now analyze data stored across a variety of data stores, either on-premises or hosted in AWS, within the same query. Athena supports federated query in relational, non-relational, object, or custom data sources. You can also write your own data source connector using our Query Federation SDK. Learn more.

With User-Defined Functions (UDFs), you can now write your own functions in Java and invoke them in your Athena SQL query. Learn more.

You can connect Athena to your external Apache Hive Metastore. If your dataset is stored in Amazon S3, in addition to using the AWS Glue Data Catalog as your metadata store, you can connect Athena to your Hive Metastore using an AWS Lambda-based data source connector. Learn more.

Q: How do I test the preview features?
All Athena queries originating from the Workgroup *AmazonAthenaPreviewFunctionality* will be considered preview test queries. You can create and set up a new Workgroup

AmazonAthenaPreviewFunctionality using Athena APIs or Athena UX. To create the new Workgroup, follow the steps here.

The following notes are important for using preview features. Please do not edit the Workgroup name. You can edit other Workgroup properties such as Enable CloudWatch metrics and Enable Requester Pays. You can use the Athena Console, JDBC/ODBC drivers or APIs to submit your test queries. Please make sure to specify the Workgroup: *AmazonAthenaPreviewFunctionality* when you submit test queries. Preview functionality is available only in us-east-1 region. If you use Athena in any other region and submit queries using Workgroup:A *mazonAthenaPreviewFunctionality*, your query will fail. Cross AWS region calls are not supported in the preview mode.

Q: Is it safe to use Athena preview features in my production account?
We recommend to not onboard your production workload to the preview Workgroup *AmazonAthenaPreviewFunctionality.* Query performance may vary between the preview Workgroup and the other workgroups in your account. Additionally, we may add new features and bug fixes to the preview Workgroup that may not be backwards compatible.

Q: How can I submit my queries?
You can submit your queries using the Athena Console, Athena APIs, or using the Athena preview JDBC driver with any off-the-shelf query and result visualization tools such as SQL WorkBench.

Q: How do I provide feedback on preview feature functionality?
Your feedback is important to us. Please email us your feedback to athena-feedback@amazon.com.

Q: Are there any charges to test preview features?
During the preview, you are not charged for the data scanned from federated data sources. However, you are charged standard Athena rates for data scanned from Amazon S3. Additionally, you are charged standard rates for the AWS services that you use with Athena, such as Amazon S3, AWS Lambda, AWS Glue, Amazon SageMaker, and AWS Serverless Application Repository. For example, you are charged S3 rates for storage, requests, and inter-region data transfer. By default, query results are stored in an S3 bucket of your choice and are also billed at standard Amazon S3 rates. If you use AWS Lambda, you are charged based on the number of requests for your functions and the duration, the time it takes for your code to execute.

Q: What happens when the preview ends?
All queries submitted using the Workgroup *AmazonAthenaPreviewFunctionality* will fail. You can continue to submit queries from other Workgroups. If you do not specify any

workgroup, the query automatically executes using the default Primary Workgroup. Please note that the preview of any feature may end at any time.

When to use Athena vs other big data services

Q: What is the difference between Amazon Athena, Amazon EMR, and Amazon Redshift?
Query services like Amazon Athena, data warehouses like Amazon Redshift, and sophisticated data processing frameworks like Amazon EMR, all address different needs and use cases. You just need to choose the right tool for the job. Amazon Redshift provides the fastest query performance for enterprise reporting and business intelligence workloads, particularly those involving extremely complex SQL with multiple joins and sub-queries. Amazon EMR makes it simple and cost effective to run highly distributed processing frameworks such as Hadoop, Spark, and Presto when compared to on-premises deployments. Amazon EMR is flexible - you can run custom applications and code, and define specific compute, memory, storage, and application parameters to optimize your analytic requirements. Amazon Athena provides the easiest way to run ad-hoc queries for data in S3 without the need to setup or manage any servers.

Q: When should you use a full featured enterprise data warehouse, like Amazon Redshift vs. a query service like Amazon Athena?
A data warehouse like Amazon Redshift is your best choice when you need to pull together data from many different sources – like inventory systems, financial systems, and retail sales systems – into a common format, and store it for long periods of time, to build sophisticated business reports from historical data; then a data warehouse like Amazon Redshift is the best choice.

Data warehouses collect data from across the company and act as the “single source of truth” for report generation and analysis. Data warehouses pull data from many sources, format and organize it, store it, and support complex, high speed queries that produce business reports. The query engine in Amazon Redshift has been optimized to perform especially well on this use case - where you need to run complex queries that join large numbers of very large database tables. TPC-DS is a standard benchmark designed to replicate this use case, and Redshift runs these queries up to 20x faster than query services that are optimized for unstructured data. When you need to run queries against highly structured data with lots of joins across lots of very large tables, you should choose Amazon Redshift.

By comparison, query services like Amazon Athena make it easy to run interactive queries against data directly in Amazon S3 without worrying about formatting data or managing infrastructure. For example, Athena is great if you just need to run a quick query on some web logs to troubleshoot a performance issue on your site. With query services, you can get started fast. You just define a table for your data and start querying using standard SQL.

You can also use both services together. If you stage your data on Amazon S3 before loading it into Amazon Redshift, that data can also be registered with and queried by Amazon Athena.

Q: When should I use Amazon EMR vs. Amazon Athena?
Amazon EMR goes far beyond just running SQL queries. With EMR you can run a wide variety of scale-out data processing tasks for applications such as machine learning, graph analytics, data transformation, streaming data, and virtually anything you can code. You should use Amazon EMR if you use custom code to process and analyze extremely large datasets with the latest big data processing frameworks such as Spark, Hadoop, Presto, or Hbase. Amazon EMR gives you full control over the configuration of your clusters and the software installed on them.

You should use Amazon Athena if you want to run interactive ad hoc SQL queries against data on Amazon S3, without having to manage any infrastructure or clusters.

Q: Can I use Amazon Athena to query data that I process using Amazon EMR?
Yes, Amazon Athena supports many of the same data formats as Amazon EMR. Athena’s data catalog is Hive metastore compatible. If you're using EMR and already have a Hive metastore, you simply execute your DDL statements on Amazon Athena, and then you can start querying your data right away without impacting your Amazon EMR jobs.

Q: How does federated query in Athena relate to other AWS services? [preview]
Federated query in Athena allows you to run SQL queries across variety of relational, non-relational, and custom data sources. You get a unified way to run SQL queries across various data stores.

Q: How does machine learning in Athena relate to other AWS services? [preview]
Athena SQL queries can invoke ML models deployed on Amazon SageMaker. You can specify the Amazon S3 location where they want to store results of these Athena SQL queries.

Creating tables, data formats and partitions

Q: How do I create tables and schemas for my data on Amazon S3?

Amazon Athena uses Apache Hive DDL to define tables. You can run DDL statements using the Athena console, via an ODBC or JDBC driver, via the API, or using the Athena create table wizard. If you use the AWS Glue Data Catalog with Athena, you can also use Glue crawlers to automatically infer schemas and partitions. An AWS Glue crawler connects to a data store, progresses through a prioritized list of classifiers to extract the schema of your data and other statistics, and then populates the Glue Data Catalog with this metadata. Crawlers can run periodically to detect the availability of new data as well as changes to existing data, including table definition changes. Crawlers automatically add new tables, new partitions to existing table, and new versions of table definitions. You can customize Glue crawlers to classify your own file types.

When you create a new table schema in Amazon Athena the schema is stored in the Data Catalog and used when executing queries, but it does not modify your data in S3. Athena uses an approach known as schema-on-read, which allows you to project your schema onto your data at the time you execute a query. This eliminates the need for any data loading or transformation. Learn more about creating tables.

Q: What data formats does Amazon Athena support?
Amazon Athena supports a wide variety of data formats like CSV, TSV, JSON, or Textfiles and also supports open source columnar formats such as Apache ORC and Apache Parquet. Athena also supports compressed data in Snappy, Zlib, LZO, and GZIP formats. By compressing, partitioning, and using columnar formats you can improve performance and reduce your costs.

Q: What kind of data types does Amazon Athena support?
Amazon Athena supports both simple data types such as INTEGER, DOUBLE, VARCHAR and complex data types such as MAPS, ARRAY and STRUCT.

Q: Can I run any Hive Query on Athena?
Amazon Athena uses Hive only for DDL (Data Definition Language) and for creation/modification and deletion of tables and/or partitions. Please click here for a complete list of statements supported. Athena uses Presto when you run SQL queries on Amazon S3. You can run ANSI-Compliant SQL SELECT statements to query your data in Amazon S3.

Q: What is a SerDe?
SerDe stands for Serializer/Deserializer, which are libraries that tell Hive how to interpret data formats. Hive DLL statements require you to specify a SerDe, so that the system knows how to interpret the data that you’re pointing to. Amazon Athena uses SerDes to interpret the data read from Amazon S3. The concept of SerDes in Athena is the same as the concept used in Hive. Amazon Athena supports the following SerDes:

1. Apache Web Logs: "org.apache.hadoop.hive.serde2.RegexSerDe"

2. CSV: "org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe"
3. TSV: "org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe"
4. Custom Delimiters: "org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe"
5. Parquet: "org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe"
6. Orc: "org.apache.hadoop.hive.ql.io.orc.OrcSerde"
7. JSON: “org.apache.hive.hcatalog.data.JsonSerDe” OR org.openx.data.jsonserde.JsonSerDe

Q: Can I add my own SerDe (Serializer/Deserializer) to Amazon Athena?
Currently, you cannot add your own SerDe to Amazon Athena. We appreciate your feedback, so if there are any SerDes you would like to see added, please contact the Athena team at Athena-feedback@amazon.com

Q: I created Parquet/ORC files using Spark/Hive. Will I be able to query them via Athena?
Yes, Parquet and ORC files created via Spark can be read in Athena.

Q: I have data coming from Kinesis Firehose. How can I query it using Athena?
If your Kinesis Firehose data is stored in Amazon S3, you can query it using Amazon Athena. Simply create a schema for your data in Athena and start querying. We recommend that you organize the data into partitions to optimize performance. You can add partitions created by Kinesis Firehose using ALTER TABLE DDL statements. Learn more about partitions.

Q: Does Amazon Athena support data partitioning?
Yes. Amazon Athena allows you to partition your data on any column. Partitions allow you to limit the amount of data each query scans, leading to cost savings and faster performance. You can specify your partitioning scheme using the PARTITIONED BY clause in the CREATE TABLE statement. Learn more about partitioning data.

Q: How do I add new data to an existing table in Amazon Athena?
If your data is partitioned, you will need to run a metadata query (ALTER TABLE ADD PARTITION) to add the partition to Athena once new data becomes available on Amazon S3. If your data is not partitioned, just adding the new data (or files) to the existing prefix automatically adds the data to Athena. Learn more about partitioning data.

Q: I already have large quantities of log data in Amazon S3. Can I use Amazon Athena to query it?
Yes, Amazon Athena makes it easy to run standard SQL queries on your existing log data. Athena queries data directly from Amazon S3 so there’s no data movement or loading

required. Simply define your schema using DDL statements and start querying your data right away.

Querying and data formats

Q: What kinds of queries does Amazon Athena support?
Amazon Athena supports ANSI SQL queries. Amazon Athena uses Presto, an open source, in-memory, distributed SQL engine, and can handle complex analysis, including large joins, window functions, and arrays.

Q: Can I use Amazon QuickSight with Amazon Athena?
Yes. Amazon Athena integrates with Amazon QuickSight, allowing you to easily visualize your data stored in Amazon S3.

Q: Does Athena support other BI Tools and SQL Clients?
Yes. Amazon Athena comes with an ODBC and JDBC driver that you can use with other business intelligence tools and SQL clients. Learn more about using an ODBC or JDBC driver with Athena.

Q: How do I access the functions supported by Amazon Athena?
Click here to learn more about functions supported by Amazon Athena.

Q: How do I improve the performance of my query?
You can improve the performance of your query by compressing, partitioning, or converting your data into columnar formats. Amazon Athena supports open source columnar data formats such as Apache Parquet and Apache ORC. Converting your data into a compressed, columnar format lowers your cost and improves query performance by enabling Athena to scan less data from S3 when executing your query.

Q: Does Athena support User Defined Functions (UDFs)? [preview]
Amazon Athena now supports user-defined functions (UDFs) to enable you to write custom scalar functions and invoke them in SQL queries. While Athena provides built-in functions, UDFs enables you to perform custom processing such as compressing and decompressing data, redacting sensitive data, or applying customized decryption.

You can write their UDFs in Java using the Athena Query Federation SDK. When a UDF is used in a SQL query submitted to Athena, it is invoked and executed on AWS Lambda. UDFs can be used in both SELECT and FILTER clauses of a SQL query. You can invoke multiple UDFs in the same query.

Q: What is the user experience when writing a UDF? [preview]
You can use the Athena Query Federation SDK to write your UDF. UDF examples are provided here. You can upload your function to AWS Lambda and then invoke it in your Athena query. Click here to get started.

Athena will invoke your UDF on a batch of dataset rows to optimize performance.

Federated query [in preview]

Q: Why should you use federated queries in Athena? [preview]
Developers often pick relational, key-value, document, in-memory, search, graph, time-series and ledger databases along with storing their data on S3. Running analytics on data spread across wide variety of data sources can be complex and time consuming. Analysts often have to learn new programming languages and database constructs, and build complex pipelines that extract, transform, and create copies of data before they can analyze them. Similarly, data scientists often need to extract data from multiple data sources to create a data set fit for feature extraction and training. This process is time consuming and inhibits building self-service platforms where analysts and data scientists can easily build pipelines that can extract data from multiple source. Analysts typically have to depend on data engineering teams to build such pipelines, introducing delays and complexity. Federated query eliminates this complexity by providing a simple to use, pay-per-query, serverless service that allows you to run SQL queries across a variety of such data stores. You can use well-known SQL constructs to query data across multiple data sources for quick analysis, or use scheduled SQL queries to extract and transform data from multiple data sources, and store them in S3 for further analysis.

In addition, you may also have proprietary or custom databases and catalogs. Athena federated queries are extensible because they allow you to write your own or use community-developed connectors to run SQL queries against to any data source or custom catalog of your choice. There are open source reference implementations for several such data sources that can be used as baselines for developing new ones.

Q: What use cases do Athena federated queries support? [preview]
Athena federated queries supports a wide variety of use-cases. One example is ad-hoc analysis, where you often have data stored in variety of data stores. Consider an e-commerce company that uses Amazon ElasticCache Redis to store active orders, Amazon DocumentDB or MongoDB to store customer-specific information such as email address, shipping address, Amazon CloudWatch Logs (e.g. of custom data store) to store order processing application log events. You may want to understand what happened with a

specific order that was reported as delayed. You can use a simple query to join data across the multiple data stores to quickly run analysis.

Another example is ETL from multiple data sources. Running analytics often requires assembling data from multiple data sources, so that it can be further published in a data warehouse or queried using engines such as Athena, Apache Spark, or Apache Presto. Such assembly requires building data pipelines that can extract and transform data from multiple sources on a schedule. Building data pipelines often requires learning new programming languages such as Python and Java, or using large-scale distributed systems such as Apache Spark. Analysts often have to rely on data engineering teams to build such pipelines. With Athena federated queries, anyone can express their pipelines as SQL statements and schedule them to run on schedule.

A third example is machine learning extracts: data scientists often need to extract data from multiple data sources to create a data set fit for feature extraction and training. This process is time consuming and inhibits building self-service platforms.

Q: How do Athena data source connectors work? [preview]
You can run SQL queries against new data stores by registering the data store with Athena. To register a data source, you use an Athena Data Source Connector specific to the data source. A connector can be used to extend Athena’s querying capability to new data sources. You can use AWS provided open source connectors, build your own or contribute to existing connectors, or use community or marketplace-built connectors. Depending on the type of data source, a connector manages metadata information, identifies specific parts of the tables that need to be scanned, read or filtered, and manages parallelism.

Connectors run as AWS Lambda functions in the customer account. Each connector is composed of two Lambda functions specific to a data source – one for metadata, and one for record reading. You can deploy Lambda functions using code in the Github repository or can use pre-deployed Lambda functions from AWS Serverless Application Repository. Once the Lambda functions are deployed, they produce a unique Amazon Resource Name or ARN. You must register these ARNs with Athena. Registering an ARN allows Athena to understand which Lambda function to talk to during query execution. Once both the ARN is registered, you can query the registered data source. The process needs to be repeated for each data source.

When a query runs on a federated data source, Athena will fan out the Lambda invocations reading metadata and data in parallel. The number of parallel invocations depend on the Lambda concurrency limits in your account. For example, if you have a limit of 300 concurrent Lambda invocations, Athena can run invoke 300 parallel Lambda functions for record reading. For two queries running in parallel, Athena will invoke twice the number of concurrent executions. You can define their own limit allowing them to control cost and throughput to data source.

Q: What connectors are available for Athena federated query? [preview]
Athena has open sourced data source connectors to Apache HBase, Amazon DocumentDB, Amazon DynamoDB, and Amazon CloudWatch Logs and CloudWatch Metrics. Athena also has a generic JDBC connector that connect to any JDBC-compliant data source and an AWS Configuration Management Database (CMDB) connector that allows customers to run queries on AWS resource metadata.

Q: How do I use the Query Federation SDK? [preview]
You can use the Query Federation SDK to create your own connector to use when querying a data source using Athena. Template implementations are provided for each of the connectors. You can use the templates as baseline. Get started by visiting our documentation.

Q: Can I use federated query capabilities for ETL? What is the workflow? [preview]
All Athena query results are stored in an Amazon S3 location that you set. You can use Athena's federated query capabilities to execute a query that scans data sources of your choice and store the result in S3 in one SQL query. Common SQL constructs such as JOINs, Filter clauses, etc. are supported. Additionally, you can also define your own functions using Athena's UDF functionality to pre- or post-process your result dataset.

Q: Are you going to release SDK support in programming languages other than JAVA? [preview]
Please let us know what programming languages you want support for by emailing athena-feedback@amazon.com

Q: What are the known limitations of the Query Federation SDK? [preview]
At the Preview launch, Query Federation SDK only supports Reads and JAVA-based Lambda functions.

Machine learning [in preview]

Q: What use cases does Athena support for embedded ML? [preview]
Athena use-cases for ML span across different industries, as in the following examples. Financial risk data analysts can run what-if analysis and Monte Carlo simulations. Business analysts might run linear regression or forecasting models to predict future values to help them create richer and forward-looking business dashboards that forecast revenues. Marketing analysts could use k-means clustering models to help determine their different

customer segments. Security analysts could use logistic regression models (bivariant and multivariant) to find anomalies and detect security incidents from various logs.

Q: What ML models can be used with Athena? [preview]
Athena can invoke any ML model that is deployed on Amazon SageMaker. You have the flexibility to train your own model using your proprietary data, or use a model that is pre-trained and deployed on SageMaker. For example, cluster analysis would likely be trained on your own data, because you want to categorize new records into the same categories you used for previous records. On the other hand, for predicting real world sports events, you could use a publicly available model, since the training data used would be in the public domain already. Domain-specific or industry-specific predictions will typically be trained on your own data in SageMaker, while undifferentiated ML needs may use external models.

Q: Can I train my ML model using Athena? [preview]
You cannot train and deploy your ML models on SageMaker using Athena. You can train your ML model, or use an existing pre-trained model that is deployed on SageMaker using Athena. Documentation detailing training steps on SageMaker is here.

Q: Can I run inference on models deployed on other services such as Comprehend, Forecasting or Models deployed on my own EC2 cluster? [preview]
Athena only supports invoking ML models deployed on SageMaker. We welcome feedback on what other services you want to use with Athena. Please email us your feedback to: athena-feedback@amazon.com.

Q: What are the performance implications of using Athena queries for SageMaker inference? [preview]
We are constantly adding operational performance improvements to our features and services. To optimize performance of your Athena ML queries, we batch rows when invoking your SageMaker ML model for inference. At this time, we do not support user provided row batch size overrides.

Q: What features does Athena ML support? [preview]
Athena offers ML inference (prediction) capabilities wrapped by a SQL interface. You can also call an Athena user-defined function (UDFs, also included in the Preview) to invoke pre- or post-processing logic on your result set. Inputs can include any column, record or table, and multiple calls can be batched together for higher scalability. You can run inference in the Select phase or in the Filter phase. To learn more, please visit our documentation.

Q: What ML models can I use? [preview]

Amazon SageMaker supports a variety of ML algorithms. You can also create your proprietary ML model and deploy it on Amazon SageMaker. For example, cluster analysis would likely be trained on your own data, because you want to categorize new records into the same categories you used for previous records. On the other hand, for predicting real world sports events, you could use a publicly available model, since the training data used would be in the public domain.

We expect that domain-specific or industry-specific predictions will typically be trained on your own data in SageMaker, while undifferentiated ML needs such as machine translation will use external models.

Security & availability

Q: How do I control access to my data?
Amazon Athena allows you to control access to your data by using AWS Identity and Access Management (IAM) policies, Access Control Lists (ACLs), and Amazon S3 bucket policies. With IAM policies, you can grant IAM users fine-grained control to your S3 buckets. By controlling access to data in S3, you can restrict users from querying it using Athena.

Q: Can Athena query encrypted data in Amazon S3?
Yes, you can query data that's encrypted using Server-Side Encryption with Amazon S3-Managed Encryption Keys, Server-Side Encryption with AWS Key Management Service (KMS) – Managed Keys, and Client-Side Encryption with keys managed by KMS. Amazon Athena also integrates with KMS and provides you an option to encrypt your result sets.

Q: Is Athena highly available?
Yes. Amazon Athena is highly available and executes queries using compute resources across multiple facilities, automatically routing queries appropriately if a particular facility is unreachable. Athena uses Amazon S3 as its underlying data store, making your data highly available and durable. Amazon S3 provides durable infrastructure to store important data and is designed for durability of 99.999999999% of objects. Your data is redundantly stored across multiple facilities and multiple devices in each facility.

Q: Can I provide cross-account access to someone else's S3 bucket?
Yes, you can provide cross-account access to Amazon S3.

Pricing & billing

Q: How is Amazon Athena priced?

Amazon Athena is priced per query and charges based on the amount of data scanned by the query. You can store data in a variety of formats on Amazon S3. If you compress your data, partition, or convert it to columnar storage formats, you pay less because you scan less data. Converting data to the columnar format allows Athena to read only the columns it needs to process the query. Please see the Athena pricing page for more details

Q: Why do I get charged less when I use a columnar format?

Amazon Athena charges you for the amount of data scanned per query. Compressing your data allows Amazon Athena to scan less data. Converting your data to columnar formats allows Athena to selectively read only required columns to process the data. Partitioning your data also allows Athena to restrict the amount of data scanned. This leads to cost savings and improved performance. See pricing example for details.

Q: How do I lower my costs?

You can save 30%-90% on your query costs and get better performance by compressing, partitioning, and converting your data into columnar formats. Each of these operations reduces the amount of data Amazon Athena needs to scan to execute a query. Amazon Athena supports Apache Parquet and ORC, two of the most popular open-source columnar formats. You can see the amount of data scanned per query, on the Athena console.

Q: Does Amazon Athena charge me for failed queries?

No, you are not charged for failed queries.

Q: Does Amazon Athena charge me for cancelled queries?

Yes, if you cancel a query manually, you are charged for the amount of data scanned up to the point at which you cancelled the query.

Q: Are there any additional charges associated with Amazon Athena?

Amazon Athena queries data directly from Amazon S3, so your source data is billed at S3 rates. When Amazon Athena runs a query, it stores the results in an S3 bucket of your choice and you are billed at standard S3 rates for these result sets. We recommend you monitor these buckets and use lifecycle policies to control how much data gets retained.

Q. Will I be charged for using AWS Glue Data Catalog?

Yes, you are charged separately for using the AWS Glue Data Catalog. Click here to learn more about Glue Data Catalog pricing.

Amazon CloudSearch FAQs

General

Q: What is Amazon CloudSearch?

Amazon CloudSearch is a fully-managed service in the AWS Cloud that makes it easy to set up, manage, and scale a search solution for your website or application.

Q: What are the benefits of running a managed search service like Amazon CloudSearch over running my own search service on EC2?

Amazon CloudSearch provides several benefits over running your own self-managed search service including easy configuration, auto scaling for data and traffic, self-healing clusters, and high availability with Multi-AZ. With a few clicks in the AWS Management Console, you can create a search domain and upload the data you want to make searchable, and Amazon CloudSearch automatically provisions the required resources and deploys a highly tuned search index.

Q: What is a search engine?

A search engine makes it possible to search large collections of mostly textual data items (called documents) to quickly find the best matching results. Search requests are usually a few words of unstructured text, such as "matt damon movies". The returned results are usually ranked with the best matching, or most relevant, items listed first (the ones that are most "about" the search words).

Documents may be completely unstructured, or they can contain multiple fields that can optionally be searched individually. For example, a search service for movies might have documents with fields for title, director, actor, description, and reviews. Results returned by a search engine are typically proxies for the underlying documents, such as URLs that reference particular web pages. However, the search service can also return the actual contents of individual fields.

Q: What benefits does Amazon CloudSearch offer?

Amazon CloudSearch is a fully managed search service that automatically scales with the volume of data and complexity of search requests to deliver fast and accurate results. Amazon CloudSearch lets customers add search capability without needing to manage hosts, traffic and data scaling, redundancy, or software packages. Users pay low hourly rates only for the resources consumed. Amazon CloudSearch can offer significantly lower total cost of ownership compared to operating and managing your own search environment.

Q: Can Amazon CloudSearch be used with a storage service?

A search service and a storage service are complementary. A search service requires that your documents already be stored somewhere, whether it's in files of a file system, data in Amazon S3, or records in an Amazon DynamoDB or Amazon RDS instance. The search service is a rapid retrieval system that makes those items searchable with sub-second latencies through a process called indexing.

Q: Can Amazon CloudSearch be used with a database?

Search engines and databases are not mutually exclusive - in fact, they are often used together. If you already have a database that contains structured data, you might want to use a search engine to intelligently filter and rank the database contents using search keywords as relevance criteria.

A search service can be used to index and search both structured and unstructured data. Content can come from multiple sources and can include database fields along with files in a variety of formats, web pages, and so on. A search service can support customizable result ranking as well as special search features such as using facets for filtering that are not available in databases.

Q: What regions is Amazon CloudSearch available in?

Amazon CloudSearch is available in the following AWS Regions: US East (Northern Virgina), US West (Oregon), US West (N. California), EU (Ireland), EU (Frankfurt), South America (Sao Paulo) and Asia Pacific (Singapore, Tokyo, Sydney, and Seoul).

About the 2013-01-01 API

Q: What new features does Amazon CloudSearch support?

With this latest release Amazon CloudSearch supports several new search and administration features. The key new features include:

- Language support:
  - 34 languages, plus "multiple" to handle mixed language fields
  - Per-field language configuration
  - Language-specific text analysis
  - Multiple levels of algorithmic stemming are available for many languages, including "none"
- Enhanced search features:
  - Suggestions
  - Highlighting
  - Geospatial search
  - New data types: date, double, 64 bit signed int, latlon
  - Sloppy phrase search

  - Term boosting
  - Enhanced range searching for all field types
  - Support for multiple query parsers: simple, structured, lucene, dismax
  - Query parser configuration options
- Administration features:
  - High availability option
  - IAM integration
  - User configurable scaling
- Available in additional AWS Regions: Asia Pacific (Tokyo), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Seoul), and South America (Sao Paulo)

Q: Does Amazon CloudSearch still support dictionary stemming?

Yes. The new version of Amazon CloudSearch supports dictionary stemming in addition to algorithmic stemming.

Q: Does the new version of Amazon CloudSearch use Apache Solr?

Yes. The latest version of Amazon CloudSearch has been modified to use Apache Solr as the underlying text search engine. Amazon CloudSearch now provides several popular search engine features available with Apache Solr in addition to the managed search service experience that makes it easy to set up, operate, and scale a search domain.

Q: Can I access the new version of Amazon CloudSearch through the console?

Yes. You can access the new version of Amazon CloudSearch through the console. If you are a current Amazon CloudSearch customer with existing search domains, you have the option to select which version of Amazon CloudSearch you want to use when creating new search domains. New customers will use the new version of Amazon CloudSearch by default and will not have access to the 2011-01-01 version.

Q: What data types does the new version of Amazon CloudSearch support?

Amazon CloudSearch supports two types of text fields, text and literal. Text fields are processed according to the language configured for the field to determine individual words that can serve as matches for queries. Literal fields are not processed and must match exactly, including case. CloudSearch also supports four numeric types: int, double, date, and latlon. Int fields hold 64-bit, signed integer values. Double fields hold double-width floating point values. Date fields hold dates specified in UTC (Coordinated Universal Time) according to IETF RFC3339: yyyy-mm-ddT00:00:00Z. Latlon fields contain a location stored as a latitude and longitude value pair.

Q: Will my existing search domains created with the 2011-02-01 version of Amazon CloudSearch continue to work?

Yes. Existing search domains created with the 2011-02-01 version of Amazon CloudSearch will continue to work.

Q: Will I be able to use the new features on my existing search domains created with the 2011-01-01 version of Amazon CloudSearch?

No. Existing search domains created with the 2011-01-01 version of Amazon CloudSearch will not have access to the features available in the new version. To access the new features you will have to create a new search domain using the 2013-01-01 version of Amazon CloudSearch.

Q: How can I migrate my applications built using the 2011-01-01 version of Amazon CloudSearch to the new version of Amazon CloudSearch?

To use the new version of Amazon CloudSearch you need to recreate existing domains using the new version of Amazon CloudSearch and re-upload your data. For more information, see Migrating to the 2013-01-01 API in the Amazon CloudSearch Developer Guide.

Q: Will AWS continue to support the 2011-02-01 version of Amazon CloudSearch?

Yes. AWS will continue support for the 2011-02-01 version of Amazon CloudSearch.

Q: Can I create new search domains using the 2011-02-01 version of Amazon CloudSearch?

Current Amazon CloudSearch customers who have existing 2011-02-01 domains will be able to choose whether their new domains use the 2011-02-01 API or the new 2013-01-01 API. Search domains created by new customers will automatically be created with the 2013-01-01 API.

Q: Can I take advantage of the free trial offer with the new version of Amazon CloudSearch?

New customers will still be able to take advantage of the free trial offer available with Amazon CloudSearch. See the Amazon CloudSearch Free Trial page for details.

Getting Started

Q: How do I get started with Amazon CloudSearch?

To sign up for Amazon CloudSearch, click the **Create Free Account** button on the Amazon CloudSearch detail page and complete the sign-up process. You must have an Amazon Web Services account. If you do not already have one, you will be prompted to create an AWS account when you begin the Amazon CloudSearch sign-up process.

After you have signed up, select **Amazon CloudSearch** from the AWS Management Console. Using the Amazon CloudSearch console you can quickly create a search domain, configure your search fields, upload sample data, and send search queries to your search domain. You can also use the AWS SDKs and the CLI to perform these operations.

For more information, see the Getting Started tutorial in the Amazon CloudSearch Developer Guide.

Q: Do the AWS SDKs support Amazon CloudSearch?

Yes, the AWS SDKs for Java, Ruby, Python, .Net, PHP, and Node.js provide support for CloudSearch. Using the AWS SDKs you can quickly create a search domain, configure your search fields, upload data, and send search queries to your search domain.

Q: Does the AWS CLI support Amazon CloudSearch?

Yes, the AWS CLI provides support for CloudSearch. Using the AWS CLI you can quickly create a search domain, configure your search fields, upload data, and send search queries to your search domain.

Q: Can I still use the Amazon CloudSearch CLTs?

Yes, the Amazon CloudSearch CLTs will continue to work.

Search Domains, Data, and Indexing

Q: What is a search domain and how do I create one?

A search domain is a data container and a set of services that make the data searchable. These services include:

- A document service that allows you upload data to your domain for indexing.
- A search service that allows you to perform search requests against your indexed data.
- A configuration service for controlling your domain's behavior (including relevance ranking).

You can create, manage, and delete search domains using the AWS Management Console, AWS SDKs, or AWS CLI.

Q: How do I upload documents to my search domain?

You upload documents to your domain using the AWS Management Console, AWS SDKs, or AWS CLI.

Q: Do my documents need to be in a particular format?

To make your data searchable, you need to format your data in JSON or XML. Each item that you want to be able to receive as a search result is represented as a document. Every document has a unique document ID and one or more fields that contain the data that you want to search and return in results. Amazon CloudSearch generates a search index from your document data according to the index fields configured for the domain. As your data changes, you submit updates to add or delete documents from your index.

Q: How do I create document batches formatted for Amazon CloudSearch?

To create document batches that describe your data, you create JSON or XML text files that specify:

- The operation type: add or delete
- A unique identifier
- The actual fields and their data

The following example shows a single document batch formatted in JSON:

```
[

    {

       "fields" : {

         "directors" : [

             "Francis Lawrence"

          ],

          "release_date" : "2013-11-11T00:00:00Z",

          "genres" : [
             "Action",

             "Adventure",

             "Sci-Fi",

             "Thriller"

          ],

          "image_url" : "http://ia.media-imdb.com/images/M/MV5xMzzAx._V1_SX400_.jpg",

          "plot" : "Katniss Everdeen and Peeta Mellark become targets of the Capitol after their victory in the 74th Hunger Games sparks a rebellion in the Districts of Panem.",

          "title" : "The Hunger Games: Catching Fire",

          "rank" : 4,

          "running_time_secs" : 8760,

          "actors" : [

             "Jennifer Lawrence",

             "Josh Hutcherson",

             "Liam Hemsworth"

           ],

          "year" : 2013

       },

       "id" : "tt1951264",

       "type" : "add"
```

```
    }

]
```

Note that numeric values such as the year are not enclosed in quotes, and that values in a multi-value field such as genres are listed in a JSON array.

To make this data available to Amazon CloudSearch, you can save it to a file and upload it using the AWS Management Console, AWS SDKs, or AWS CLI.

Q: How do my documents get indexed?

Documents are automatically indexed when you upload them to your search domain. You can also explicitly re-index your documents when you make configuration changes by sending an IndexDocuments request.

Q: When do I need to re-index my domain?

Certain configuration options, such as adding a new index field or updating your stemming or stopword dictionaries, are not available until your domain is re-indexed. When you have made changes that require indexing, the domain's status will indicate that it needs to be indexed. You can initiate indexing from the AWS Management Console, AWS SDKs, or AWS CLI.

Q: How do I send search requests to my search domain?

Every search domain has a REST-based search service with a unique URL (search endpoint) that accepts search requests for its document set. You can send search requests from the AWS Management Console, AWS SDKs, or AWS CLI.

Q: Can a search domain span multiple Availability Zones?

Yes. If you enable the Multi-AZ option, Amazon CloudSearch deploys additional instances in a second availability zone in the same Region. For more information, see Configuring Availability Options in the Amazon CloudSearch Developer Guide.

Q: Can I move a search domain from one region to another?

At this time, there is no way to automatically migrate a search domain from one region to another. You will need to create a new domain in the target region, configure the domain and upload your data, then delete the original domain.

Q: How do I delete my search domain?

To delete a search domain, click on Delete Domain button in the Amazon CloudSearch console. You can also delete domains through the AWS SDKs or AWS CLI.

Q: How do I delete documents from my search domain?

To delete documents you specify a delete operation in your batch upload that contains the ID of the document you want to remove.

You can submit data updates through the AWS Management Console, AWS SDKs, or AWS CLI.

Q: How do I empty my search domain?

If you wish to maintain your domain's endpoints, you can send a delete for each document that is in your domain.

Q: Why is my domain in the "Processing" state?

A domain can be in one of three different states: "processing," "active," or "reindexing." Normally, your domain will be in the "active" state, which indicates that no changes are currently being made, that the domain can be queried and updated, and that all previous changes are currently visible in the search results.

When a domain needs to be re-indexed, Amazon CloudSearch needs to rebuild the index entirely. However, the domain does not enter the "processing" state until you initiate reindexing. During this stage, the domain can still be queried and updated, but the configuration changes won't be visible in search results until indexing is completed, and the domain's status changes back to "active."

You can also continue to upload document batches to your domain. However, if you submit a large volume of updates while your domain is in the "processing" state, it can increase the amount of time it takes for the updates to be applied to your search index. If this becomes an issue, slow down your update rate until the domain returns to the "active" state.

Best Practices

Q: What are the best practices for bootstrapping data into CloudSearch?

After you've launched your domain, the next step is loading your data into Amazon CloudSearch. You'll likely need to upload a single large dataset, and then make smaller updates or additions as new data comes in. The following guidelines will help make bootstrapping your initial data into CloudSearch quick and easy.

1. Use the curl-v command line tool when preparing your script

During the upload of a dataset, the script you've written reads your data and uses it to create JSON or XML documents. We recommend preparing this script in advance, and using curl or another simple command line tool to see if you're able to upload the documents that the script creates. The “-v” option in curl often provides more detailed information about syntax problems than the AWS SDK or Boto, which both suppress errors for production purposes. Curl displays more detailed error messages, which helps identify the sources of any issues.

2. Use the UTF-8 character code

Make sure that all data is formatted in the UTF-8 character code format, and that any bad Unicode characters have been removed before uploading to CloudSearch. Illegal characters will cause the document upload to fail.

3. Batch your documents

Batching your documents is perhaps the most important step in data bootstrapping. Submitting documents to CloudSearch individually is not only inefficient, but also leads to preventable errors.

A document batch is simply a collection of add and delete operations that represent the documents you want to add, update, or delete from your domain. Batches are described in either JSON or XML, and when you upload them to a domain, the data is indexed automatically, according to the domain's indexing options. Since you’re billed for the total number of document batches uploaded to your search domain, it’s more cost-effective to upload your data in batches of 5 MB, the maximum allowed per upload. You can also upload batches in parallel to reduce the amount of time it takes to upload your data.

4. Pre-scale

It’s also important to pre-scale your data before uploading it to CloudSearch. Pre-scaling involves selecting the appropriate instance type for the amount of data you wish to upload.

Choosing an instance with enough capacity to handle the size of your upload can help prevent errors and a high replication count. Although replication can help decrease search response time, it doesn’t increase the size of the data pipe or address core problems in data uploads.

CloudSearch will automatically scale up to larger instances as you send more data. Still, pre-selecting the appropriate instance type saves time later in the bootstrapping process, as scaling from one instance to another tends to be a slower process. Below is a sample script to pre-scale the domain for boostrapping and to restore the instance type after data is loaded.

Pre-scale before bootstrapping:

```
aws cloudsearch update-scaling-parameters --domain-name foo --scaling-parameters DesiredInstanceType=search.m3.2xlarge

aws cloudsearch index-documents --domain-name foo
```

Restore after data loading:

```
aws cloudsearch update-scaling-parameters --domain-name foo --scaling-parameters DesiredInstanceType=search.m1.small

aws cloudsearch index-documents --domain-name foo
```

Q: What are some ways to avoid 504 errors?

If you’re seeing 504 errors or high replication counts, try moving to larger instance type. For example, if you’re having problems with m3.large, move up to m3.xlarge. If you continue to get 504 errors even after pre-scaling, start batching the data and increase the delay between retries.

Q: What are the best practices to accelerate domain configuration and re-indexing?

When you change the configuration options of your search domain, you must rebuild your search index for those changes to take effect in search results. Rebuilding the index can take 30 to 60 minutes whether you make one configuration change at a time or several configuration changes at once. Even if your domain has only a small number of documents, re-indexing takes this time because of the processing and provisioning necessary to build the index and distribute it. Therefore, you should plan your configuration changes ahead of time, make all of your changes at once, and then re-index your domain. The same applies when setting up a new domain - plan your configuration before you set it up so that you can index only once and get up and running in the shortest time possible.

Some domain changes require re-indexing while others just require re-deploying the existing index. Redeploying the domain takes 10 to 15 minutes compared to 30-60 minutes for re-indexing. During re-deployment, CloudSearch creates new nodes, deploys the index on them, and shuts down the old nodes. Your domain status changes to “Processing” during re-deployment. When re-indexing is needed, your domain status changes to “Needs Indexing,” followed by “Processing” once you have initiated indexing. Once the new index is created, your domain is re-deployed. The following table summarizes which changes require re-indexing followed by re-deployment and which changes require just re-deployment. Understanding this will help you better plan your configuration changes.

Change	Needs re-indexing	Needs re-deployment
Multi-AZ	No	Yes
Index fields	Yes	Yes
Index field options	Yes	Yes
Instance type	Yes	Yes
Partition count	Yes	Yes
Replication count	No	Yes
Suggesters	Yes	Yes
Expressions	No	Yes
Analysis schemes	Yes	Yes

Search Features

Q: What search features does Amazon CloudSearch provide?

Amazon CloudSearch provides features to index and search both structured data and plain text, including faceted search, free text search, Boolean search expressions, customizable relevance ranking, query time rank expressions, field weighting, searching and sorting of results using any field, and text processing options including tokenization, stopwords, stemming and synonyms. It also provides near real-time indexing for document updates. New features include:

- Autocomplete suggestions
- Highlighting
- Geospatial search
- New data types: date, double, 64 bit signed int, LatLon
- Dynamic fields

- Index field statistics
- Sloppy phrase search
- Term boosting
- Enhanced range searching for all field types
- Search filters that don't affect relevance
- Support for multiple query parsers: simple, structured, lucene, dismax
- Query parser configuration options

Q: What is faceting?

Faceting allows you to categorize your search results into refinements on which the user can further search. For example, a user might search for "umbrellas", and facets allow you to group the results by price, such as $0-$10, $10-$20, $20-$40, and so on. Amazon CloudSearch also allows for result counts to be included in facets, so that each refinement has a count of the number of documents in that group. The example could then be: $0-$10 (4 items), $10-$20 (123 items), $20-$40 (57 items), and so on.

Q: What languages does Amazon CloudSearch support?

Amazon CloudSearch currently supports 34 languages: Arabic (ar), Armenian (hy), Basque (eu), Bulgarian (bg), Catalan (ca), simplified Chinese (zh-Simp), traditional Chinese (zh-Trad), Czech (cs), Danish (da), Dutch (nl), English (en), Finnish (fi), French (fr), Galician (gl), German (de), Greek (el), Hebrew (he), Hindi (hi), Hungarian (hu), Indonesian (id), Irish (ga), Italian (it), Japanese (ja), Korean (ko), Latvian (la), Norwegian (no), Persian (fa), Portuguese (pt), Romanian (ro), Russian (ru), Spanish (es), Swedish (sv), Thai (th), and Turkish (tr). In addition, Amazon CloudSearch supports a Multiple (mul) option for fields that contain mixed languages.

Q: Does Amazon CloudSearch support geospatial search?

Yes, Amazon CloudSearch has a native type to support latitude and longitude (latlon), so that you can easily implement geographically-based searching and sorting. For more information, see Searching and Ranking Results by Geographic Location in the Amazon CloudSearch Developer Guide.

Performance

Q: How quickly will my uploaded documents become searchable?

Documents uploaded to a search domain typically become searchable within seconds to a few minutes.

Q: How many search requests can I send to my search domain?

There is no intrinsic limit on the number of search requests that can be sent to a search domain.

Q: What factors affect the latency of my search requests?

Your search requests are typically processed within a few hundred milliseconds, frequently much faster. Latency is affected by many factors including the time it takes for your request and responses to travel between your own application and your search domain, the complexity of your search request, and how heavily you are using your search domain.

Q: What makes one search request more complex than another?

Amazon CloudSearch is designed to efficiently process a wide range of search requests very quickly. Search requests vary in complexity depending on the expressions that determine which documents match and additional criteria that determine how closely each document matches. Search requests that match a large number of documents take longer to process than those that match very few documents. Search requests that compute complex expressions take longer to process than those that rank using a simple criteria such as a single field. To help you understand the difference in complexity between Search requests, the time it took to process the request is returned as part of the response.

Q: Where should I run my search application to minimize communication time with my search domain?

Applications hosted in the same AWS Region as your search domain will experience the fastest communication times.

Scaling

Q: What is a search instance?

A search instance is a single search engine in the cloud that indexes documents and responds to search requests. It has a finite amount of RAM and CPU resources for indexing data and processing requests.

Q: What is a search partition?

A search partition is the portion of your data which fits on a single search instance. A search domain can have one or more search partitions, and the number of search partitions can change as your documents are indexed.

Q: How does my search domain scale to meet my application needs?

Search domains scale in two dimensions: data and traffic. As your data volume grows, you need more (or larger) Search instances to contain your indexed data, and your index is partitioned among the search instances. As your request volume or request complexity increases, each Search Partition must be replicated to provide additional CPU for that Search Partition. For example, if your data requires three search partitions, you will have 3 search instances in your search domain. As your traffic increases beyond the capacity of a single search instance, each partition is replicated to provide additional CPU capacity, adding an additional three search instances to your search domain. Further increases in traffic will result in additional replicas, to a maximum of 5, for each search partition.

Q: How much data can I upload to my search domain?

The number of partitions you need depends on your data and configuration, so the maximum data you can upload is the data set that when your search configuration is applied results in 10 search partitions. When you exceed your search partition limit, your domain will stop accepting uploads until you delete documents and re-index your domain. If you need more than 10 search partitions, please contact us.

Q: Do I need to select the number and type of search instances for my search domain?

CloudSearch is a fully managed search service that automatically scales your search domain and selects the number and type of search instances. All search instances in a given search domain are of the same type and this type can change over time as your data or traffic grows.

You can also configure scaling options for an Amazon CloudSearch domain to:

- Increase the upload capacity
- Speed up search requests
- Increase the search capacity
- Improve fault tolerance

Q: What instance types does Amazon CloudSearch support?

Amazon CloudSearch supports the following instance types:

- Small Search Instance
- Large Search Instance
- Extra Large Search Instance
- Double Extra Large Search Instance

Q: How do I find out the number and type of search instances in my search domain?

You can find out the number and type of search instances in your search domain by using the AWS Management Console, AWS SDKs, or AWS CLI. The number and type of search instances change over time and automatically scale up and down according to your indexable data and search traffic.

Q: How quickly does my search domain scale to accommodate changes in data and traffic?

Search domains typically react to increases in traffic changes within minutes. Changes in data volume or a reduction in traffic might take longer but you can accelerate this process by invoking an IndexDocuments operation. If you are about to upload a large amount of data or expect a surge in query traffic, you can prescale your domain by setting the desired instance type and replication count. For more information, see Configuring Scaling Options in the Amazon CloudSearch Developer Guide.

Q: Does Amazon CloudSearch support Multi-AZ deployments?

Yes. Amazon CloudSearch supports Multi-AZ deployments. When you enable the Multi-AZ option, Amazon CloudSearch provisions and maintains extra instances for your search domain in a second Availability Zone to ensure high availability. Updates are automatically applied to the instances in both Availability Zones. Search traffic is distributed across all of the instances and the instances in either zone are capable of handling the full load in the event of a failure.

Q: How does the new Multi-AZ feature work? Will my system experience any downtime in the event of a failure?

When the Multi-AZ option is enabled, Amazon CloudSearch instances in either zone are capable of handling the full load in the event of a failure. If there's service disruption or the instances in one zone become degraded, Amazon CloudSearch routes all traffic to the other Availability Zone. Redundant instances are restored in a separate Availability Zone without any administrative intervention or disruption in service.

Some inflight queries might fail and will need to be retried. Updates sent to the search domain are stored durably and will not be lost in the event of a failure.

Q: Can a search domain be deployed in more than 2 Availability Zones?

No. The maximum number of Availability Zones a domain can be deployed in is two.

Q: Can I modify the Multi-AZ configuration on my search domain?

Yes. You can turn the Multi-AZ configuration on and off for your search domains. The service is not interrupted when this setting is changed.

Q: Can I choose which Availability Zones my search domain is deployed in?

No. At this time Amazon CloudSearch automatically chooses an alternate Availability Zone in the same Region.

Q: Can I choose the instance type my domain uses?

Yes. With the latest release, Amazon CloudSearch enables you to specify the desired instance type for your domain. If necessary, Amazon CloudSearch will scale your domain up to a larger instance type, but will never scale back to a smaller instance type.

Q: What is the fastest way to get my data into CloudSearch?

By default, all domains start out on a small search instance. If you need to upload a large amount of data, you should prescale your domain to a larger instance type. For more information, see Bulk Uploads in the Amazon CloudSearch Developer Guide.

Q: How do I know which instance type I should choose for my initial setup?

For datasets of less than 1 GB of data or fewer than one million 1 KB documents, start with the default settings of a single small search instance. For larger data sets consider pre-warming the domain by setting the desired instance type. For data sets up to 8 GB, start with a large search instance. For datasets between 8 GB and 16 GB, start with an extra large search instance. For datasets between 16 GB and 32 GB, start with a double extra large search instance. Contact us if you need more upload capacity or have more than 500 GB to index.

Security

Q: What additional security features are available with the new version of Amazon CloudSearch?

With the latest release, Amazon CloudSearch now provides IAM integration for the configuration service and all search domain services. You can control access to specific Amazon CloudSearch actions and require request authentication for all requests. Requests are authenticated using Signature Version 4 signing.

Q: How do I upload my data to Amazon CloudSearch securely?

You send us your data using a secure and encrypted SSL connection by using HTTPS instead of HTTP when you connect to Amazon CloudSearch.

Q: My data is already encrypted. Can I just send you the encrypted data and the encryption key?

We do not support user-generated encryption keys. You will need to decrypt the data and upload it using HTTPS.

Q: Do you support encrypted search results?

Yes. We support HTTPS for all Amazon CloudSearch requests.

Q: How can I prevent specific users from accessing my search domain?

Amazon CloudSearch supports IAM integration for the configuration service and all search domain services. You can grant users full access to Amazon CloudSearch, restrict their access to specific domains, and allow or deny access to specific actions.

Pricing

Q: How will I be charged and billed for my use of Amazon CloudSearch?

There are no set-up fees or commitments to begin using the service. Following the end of the month, your credit card will automatically be charged for that month's usage. You can view your charges for the current billing period at any time on the AWS web site by logging into your Amazon Web Services account and clicking **Account Activity** under Your Web Services Account.

Q: How much does it cost to use Amazon CloudSearch?

There are no changes to the pricing structure for Amazon CloudSearch at this time. For detailed pricing information, see Amazon CloudSearch Pricing.

Q: Is a free trial available for Amazon CloudSearch?

Yes, a free trial is available for new CloudSearch customers. For more information, see Amazon CloudSearch 30 Day Free Trial.

Q: How much does it cost to use the new version of Amazon CloudSearch?

There are no changes to the pricing structure for Amazon CloudSearch at this time. See the Pricing page for more information.

Q: Are there any cost savings to using the new version of Amazon CloudSearch?

The latest version of Amazon CloudSearch features advanced index compression and supports larger indexes on each instance type. This makes the new version of Amazon CloudSearch more efficient than the previous version and can result in significant cost savings.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

- Amazon CloudSearch
- Pricing
- Getting Started
- Free Trial
- Developer Resources
- Testimonials
- What's New
- FAQs
- Product Details

RELATED LINKS

- Documentation
- Management Console
- Release Notes
- Discussion Forum

Amazon Elasticsearch Service FAQs

General

Q: What is Amazon Elasticsearch Service?

Amazon Elasticsearch Service is a managed service that makes it easy to deploy, operate, and scale Elasticsearch clusters in the AWS Cloud.

Q: Which Elasticsearch version does Amazon Elasticsearch Service support?

Amazon Elasticsearch Service currently supports Elasticsearch versions 7.1, 6.8, 6.7, 6.5, 6.4, 6.3, 6.2, 6.0, 5.6, 5.5, 5.3, 5.1, 2.3, and 1.5.

Q: What is an Amazon Elasticsearch Service domain?

Amazon Elasticsearch Service domains are Elasticsearch clusters created using the Amazon Elasticsearch Service console, CLI, or API. Each domain is an Elasticsearch cluster in the cloud with the compute and storage resources you specify. You can create and delete domains, define infrastructure attributes, and control access and security. You can run one or more Amazon Elasticsearch Service domains.

Q: What does Amazon Elasticsearch Service manage on my behalf?

Amazon Elasticsearch Service manages the work involved in setting up a domain, from provisioning infrastructure capacity in the network environment you request to installing the Elasticsearch software. Once your domain is running, Amazon Elasticsearch Service automates common administrative tasks, such as performing backups, monitoring instances and patching software. Amazon Elasticsearch Service integrates with Amazon CloudWatch to produce metrics that provide information about the state of the domains. Amazon

Elasticsearch Service also offers options to modify your domain instance and storage settings to simplify the task of tailoring your domain based to your application needs.

Q: Does Amazon Elasticsearch Service support the open source Elasticsearch APIs?

Amazon Elasticsearch Service supports most of the commonly used Elasticsearch APIs, so the code, applications, and popular tools that you're already using with your current Elasticsearch environments work seamlessly. For a full list of supported Elasticsearch operations, see our documentation.

Q: What are the Availability Zone (AZ) deployment options available on Amazon Elasticsearch Service?

Amazon Elasticsearch Service offers customers the option to deploy their instances across one, two, or three AZs. Customers running development or test workloads can pick the single AZ option. Those running production-grade workloads should use two or three AZs. Three AZ deployments are strongly recommended for workloads with higher availability requirements.

Note: The three AZ option is only available in regions where there are three or more AZs.

Q: In which regions does Amazon Elasticsearch Service offer three AZ deployments?

Amazon Elasticsearch Service supports three AZ deployments in the following regions: US East (N. Virginia, Ohio), US West (Oregon), EU (Ireland, Frankfurt, London, Paris), China (Ningxia), and Asia Pacific (Singapore, Sydney, Tokyo).

Setup and configuration

Q: Can I create and modify my Amazon Elasticsearch Service domain through the Amazon Elasticsearch Service console?

Yes. You can create a new Amazon Elasticsearch Service domain with the Domain Creation Wizard in the console with just a few clicks. While creating a new domain you can specify the number of instances, instance types, and EBS volumes you want allocated to your domain. You can also modify or delete existing Amazon Elasticsearch Service domains using the console.

Q: Does Amazon Elasticsearch Service support Amazon VPC?

Yes, Amazon Elasticsearch Service is integrated with Amazon VPC. When choosing VPC access, IP addresses from your VPC are attached to your Amazon Elasticsearch Service domain and all network traffic stays within the AWS network and is not accessible to the Internet. Moreover, you can use security groups and IAM policies to restrict access to your Amazon Elasticsearch Service domains.

Q: Can I use CloudFormation Templates to provision Amazon Elasticsearch Service domains?

Yes. AWS CloudFormation supports Amazon Elasticsearch Service. For more information, see the CloudFormation Template Reference documentation.

Q: Does Amazon Elasticsearch Service support configuring dedicated master nodes?

Yes. You can configure dedicated master nodes for your domains. When choosing a dedicated master configuration, you can specify the instance type and instance count.

Q: Can I create multiple Elasticsearch indices within a single Amazon Elasticsearch Service domain?

Yes. You can create multiple Elasticsearch indices within the same Amazon Elasticsearch Service domain. Elasticsearch automatically distributes the indices and any associated replicas between the instances allocated to the domain.

Q: How do I ingest data into my Amazon Elasticsearch Service domain?

Amazon Elasticsearch Service supports three options for data ingestion:

- For large data volumes, we recommend Amazon Kinesis Firehose, a fully managed service that automatically scales to match the throughput of your data and requires no ongoing administration. It can also transform, batch and compress the data before loading it.
- Amazon Elasticsearch Service supports integration with Logstash. You can configure your Amazon Elasticsearch Service domain as the data store for all logs arriving from your Logstash implementation.
- You can use native Elasticsearch APIs, such as the index and bulk APIs, to load data into your domain.

Q: Does Amazon Elasticsearch Service support integration with Logstash?

Yes. Amazon Elasticsearch Service supports integration with Logstash. You can set up your Amazon Elasticsearch Service domain as the backend store for all logs coming through your Logstash implementation. You can set up access control on your Amazon Elasticsearch Service domain to either use request signing to authenticate calls from your Logstash implementation, or use resource based IAM policies to include IP addresses of instances running your Logstash implementation.

Q: Does Amazon Elasticsearch Service support integration with Kibana?

Yes. Amazon Elasticsearch Service includes a built-in Kibana install that is deployed with your Amazon Elasticsearch Service domain.

Q: Can I create custom reports with the Kibana installation included with Amazon Elasticsearch Service?

Yes. Kibana supports creating and saving custom reports through the user interface. For more information on using Kibana, refer to Kibana documentation.

Q: What storage options are available with Amazon Elasticsearch Service?

You can choose between local on-instance storage or EBS volumes. During domain creation, if you select EBS storage, you can increase and decrease the size of the storage volume as necessary.

Q: What types of EBS volumes does Amazon Elasticsearch Service support?

You can choose between Magnetic, General Purpose, and Provisioned IOPS EBS volumes.

Q: Is there a limit on the amount of EBS storage that can be allocated to an Amazon Elasticsearch Service domain?

Yes. Amazon Elasticsearch Service supports one EBS volume (max size of 1.5 TB) per instance associated with a domain. With the default maximum of 20 data nodes allowed per Amazon Elasticsearch Service domain, you can allocate about 30 TB of EBS storage to a single domain. You can request a service limit increase up to 200 instances per domain by creating a case with the AWS Support Center. With 200 instances, you can allocate about 300 TB of EBS storage to a single domain.

Q: How are dedicated master instances distributed across AZs?

If you deploy your data instances in a single AZ, your dedicated master instances are also deployed in the same AZ. However, if you deploy your data instances across two or three AZs, Amazon Elasticsearch Service automatically distributes the dedicated master instances across three AZs. The exception to this rule occurs if a region only has two AZs or if you select an older-generation instance type for the master instances that is not available in all AZs. For more details, refer our documentation.

Q: What is the recommended AZ configuration for production workloads?

For production workloads, we recommend deploying your data instances across three AZs since it offers better availability. Also, we recommend provisioning instances in multiples of three for equal distribution across AZs. In regions where three AZs are not available, we recommend using a two AZ deployment with an even number of data instances. In all cases, we recommend provisioning three dedicated master instances.

Q: How can I configure my domain for three AZ deployment?

You can enable three AZ deployment for both existing and new domains using the AWS console, CLI or SDKs. For more details, refer our documentation.

Q: Is there a fee for enabling three AZ deployment?

No. Amazon Elasticsearch Service does not charge anything for enabling three AZ deployment. You only pay for the number of instances in your domain, not the number of AZs to which they are deployed.

Q: I no longer see the "zone awareness" option in my console. Is my domain no longer zone aware?

All domains configured for multiple AZs will have zone awareness enabled to ensure shards are distributed across Availability Zones. In the console, you can now explicitly choose two or three AZ deployments. Domains previously configured with "Zone Awareness" will continue to be deployed across two AZs unless they are reconfigured. For more details, refer our documentation.

Q: How does Amazon Elasticsearch Service handle instance failures and AZ disruptions?

If one or more instances in an AZ are unreachable or not functional, Amazon Elasticsearch Service automatically tries to bring up new instances in the same AZ to replace the affected instances. In the rare event that new instances cannot be brought up in the AZ, Amazon Elasticsearch Service brings up new instances in the other available AZs if the domain has been configured to deploy instances across multiple AZs. Once the AZ issue resolves, Amazon Elasticsearch Service rebalances the instances such that they are equally distributed across the AZs configured for the domain. For more details refer our documentation.

Q: If I have only one replica for the indices in my domain, should I use two or three AZs?

Even when you configure one replica, we recommend three AZs. If an AZ disruption occurs in a three AZ domain, you only lose one-third of your capacity but if the disruption occurs in a two AZ domain, you lose half your capacity, which can be more disruptive. Also, in a three AZ domain, when an AZ is disrupted, Amazon Elasticsearch Service can fall back to the two remaining AZs, and still support cross-AZ replication . In a two AZ domain, you lose cross-AZ replication if one AZ is disrupted, which can further reduce availability. For more details refer our documentation.

Q: How do I leverage three AZ deployment for my VPC domain?

The number of AZs your domain is deployed to corresponds to the number of subnets you have configured for your VPC domain. You need to configure at least three subnets in your VPC domain to enable three AZ deployment. For more details on configuring VPC, refer our documentation.

Administration

Q: Can programs running on servers in my own data center access my Amazon Elasticsearch Service domains?

Yes. The programs with public Internet access can access Amazon Elasticsearch Service domains through a public endpoint. If your data center is already connected to Amazon VPC through Direct Connect or SSH tunneling, you can also use VPC access. In both cases, you can configure IAM policies and security groups to allow programs running on servers outside of AWS to access your Amazon Elasticsearch Service domains. Click here for more information about signed requests.

Q: How can I migrate data from my existing Elasticsearch cluster to a new Amazon Elasticsearch Service domain?

To migrate data from an existing Elasticsearch cluster you should create a snapshot of an existing Elasticsearch cluster, and store the snapshot in your Amazon S3 bucket. Then you can create a new Amazon Elasticsearch Service domain and load data from the snapshot into the newly created Amazon Elasticsearch Service domain using the Elasticsearch restore API.

Q: How can I scale an Amazon Elasticsearch Service domain?

Amazon Elasticsearch Service allows you to control the scaling of your Amazon Elasticsearch Service domains using the console, API, and CLI. You can scale your Amazon Elasticsearch Service domain by adding, removing, or modifying instances or storage volumes depending on your application needs. Amazon Elasticsearch Service is integrated with Amazon CloudWatch to provide metrics about the state of your Amazon Elasticsearch Service domains to enable you to make appropriate scaling decisions for your domains.

Q: Does scaling my Amazon Elasticsearch Service domain require downtime?

No. Scaling your Amazon Elasticsearch Service domain by adding or modifying instances, and storage volumes is an online operation that does not require any downtime.

Q: Does Amazon Elasticsearch Service support cross-zone replication?

Yes. If you enable replicas for your Elasticsearch indices and use multiple Availability Zones, Amazon Elasticsearch Service automatically distributes your primary and replica shards across instances in different AZs.

Q: Does Amazon Elasticsearch Service expose any performance metrics through Amazon CloudWatch?

Yes. Amazon Elasticsearch Service exposes several performance metrics through Amazon CloudWatch including number of nodes, cluster health, searchable documents, EBS metrics (if applicable), CPU, memory and disk utilization for data and master nodes. Please refer to the service documentation for a full listing of available CloudWatch metrics.

Q: I wish to perform security analysis or operational troubleshooting of my Amazon Elasticsearch Service deployment. Can I get a history of all the Amazon Elasticsearch Service API calls made on my account?

Yes. AWS CloudTrail is a web service that records AWS API calls for your account and delivers log files to you. The AWS API call history produced by AWS CloudTrail enables security analysis, resource change tracking, and compliance auditing. Learn more about AWS CloudTrail at the AWS CloudTrail detail page, and turn it on via CloudTrail's AWS Management Console home page.

Q: What is a snapshot?

A snapshot is a copy of your Amazon Elasticsearch Service domain at a moment in time.

Q: Why would I need snapshots?

Creating snapshots can be useful in case of data loss caused by node failure, as well as the unlikely event of a hardware failure. You can use snapshots to recover your Amazon Elasticsearch Service domain with preloaded data or to create a new Amazon Elasticsearch Service domain with preloaded data. Another common reason to use backups is for archiving purposes. Snapshots are stored in Amazon S3.

Q: Does Amazon Elasticsearch Service provide automated snapshots?

Yes. By default, Amazon Elasticsearch Service automatically creates hourly snapshots of each Amazon Elasticsearch Service domain and retains them for 14 days.

Q: How long are the automated daily hourly snapshots stored by Amazon Elasticsearch Service?

Amazon Elasticsearch Service will retain the last 14 days' worth of automated hourly snapshots.

Q: Is there a charge for the automated daily hourly snapshots?

There is no additional charge for the automated hourly snapshots. The snapshots are stored for free in an Amazon Elasticsearch Service S3 bucket and will be made available for node recovery purposes.

Q: Can I create additional snapshots of my Amazon Elasticsearch Service domains as needed?

Yes. You can use the Elasticsearch snapshot API to create additional manual snapshots in addition to the daily-automated snapshots created by Amazon Elasticsearch Service. The manual snapshots are stored in your S3 bucket and will incur relevant Amazon S3 usage charges.

Q: Can snapshots created by the manual snapshot process be used to recover a domain in the event of a failure?

Yes. Customers can create a new Amazon Elasticsearch Service domain and load data from the snapshot into the newly created Amazon Elasticsearch Service domain using the Elasticsearch restore API.

Q: What happens to my snapshots when I delete my Amazon Elasticsearch Service domain?

The daily snapshots retained by Amazon Elasticsearch Service will be deleted as part of domain deletion. Before deleting a domain, you should consider creating a snapshot of the domain in your own S3 buckets using the manual snapshot process. The snapshots stored in your S3 bucket will not be affected if you delete your Amazon Elasticsearch Service domain.

Q: What types of Elasticsearch logs are exposed by Amazon Elasticsearch Service?

Amazon Elasticsearch Service exposes three Elasticsearch logs through Amazon CloudWatch Logs: error logs, search slow logs, and index slow logs. These logs are useful for troubleshooting performance and stability issues with one's domain.

Q: What are slow logs?

Slow logs are log files that help track the performance of various stages in an operation. Elasticsearch exposes two kinds of slow logs:

- Index Slow Logs – These logs provide insights into the indexing process and can be used to fine-tune the index setup.
- Search Slow Logs – These logs provide insights into how fast or slow queries and fetches are performing. These logs help fine tune the performance of any kind of search operation on Elasticsearch.

For complete details on Elasticsearch slow logs, please refer to Elasticsearch documentation.

Q: How can I enable slow logs on Amazon Elasticsearch Service?

Slows logs can be enabled via the click of a button from the Console or via our CLI and APIs. For more details please refer to our documentation.

Q: Can I only enable slow logs for specific indices?

Yes. You can update the settings for a specific index to enable or disable slow logs for it. For more details refer to our documentation.

Q: Does turning on slow logs in Amazon Elasticsearch Service automatically enable logging for all indexes?

No. Turning on slow logs in Amazon Elasticsearch Service enables the option to publish the generated logs to Amazon CloudWatch Logs for indices in the given domain. However, in order to generate the logs you have to update the settings for one or more indices to start the logging process. For more details on setting the index configuration for enabling slow logs, please refer to our documentation.

Q: If I turn off the slow logs in Amazon Elasticsearch Service, does it mean that log files are no longer being generated?

No. The generation of log files are dependent on the index settings. To turn off generation of the log files you have to update the index configuration. For more details on setting the index configuration for enabling slow logs, see our documentation.

Q: Can I change the granularity of logging?

You can only change the granularity of logging for slow logs. Elasticsearch exposes multiple levels of logging for slow logs. You need to set the appropriate level in the configuration of your index. For more details on setting the index configuration for enabling slow logs, please refer to Elasticsearch documentation.

Q: Will enabling slow logs or error logs cost me anything?

When slow logs or error logs are enabled, Amazon Elasticsearch Service starts publishing the generated logs to CloudWatch Logs. Amazon Elasticsearch

Service does not charge anything for enabling the logs. However, standard CloudWatch charges will apply.

Q: What kinds of error logs are exposed by Amazon Elasticsearch Service?

Elasticsearch uses Apache Log4j 2 and its built-in log levels (from least to most severe) of TRACE, DEBUG, INFO, WARN, ERROR, and FATAL. If you enable error logs, Amazon Elasticsearch Service publishes log lines of WARN, ERROR, and FATAL, and select errors from the DEBUG level to CloudWatch. For more details, refer our documentation.

Q: How can I enable error logs on Amazon Elasticsearch Service?

Error logs can be enabled via the click of a button from the AWS Console or via our CLI and APIs. For more details please refer to our documentation.

Q: Can I enable error logs for only specific indices?

No, error logs are exposed for the entire domain. That is, once enabled, log entries from all indices in the domain will be made available.

Q: Are error logs available for all versions of Elasticsearch supported by Amazon Elasticsearch Service?

No, error logs are available only for Elasticsearch versions 5.x and above.

Q: Is there any limit on the size of each log entry?

Yes. Each log entry made into CloudWatch will be limited to 255,000 characters. If your log entry is bigger than that, it will be truncated to 255,000 characters.

Q: What is the recommended best practice for using slow logs?

Slow logs are only needed when you want to troubleshoot your indexes or fine-tune performance. The recommended approach is to only enable logging for those indexes for which you need additional performance insights. Also, once the investigation is done, you should turn off logging so that you don't incur any additional costs on account of it. For more details, see our documentation.

Q: How can I consume logs from CloudWatch Logs?

CloudWatch offers multiple ways to consume logs. You can view log data, export it to S3, or process it in real time. To learn more, see the CloudWatch Logs developer guide.

Q: Are slow logs available for all versions of Elasticsearch supported by Amazon Elasticsearch Service?

Yes. slow logs can be enabled for all versions of Elasticsearch supported by Amazon Elasticsearch Service. However, there are slight differences in the way log settings can be specified for each version of Elasticsearch. Please refer to our documentation for more details.

Q: Will the cluster have any down time when logging is turned on or off?

No. There will not be any down-time. Every time the log status is updated, we will deploy a new cluster in the background and replace the existing cluster with the new one. This process will not cause any down time. However, since a new cluster is deployed the update to the log status will not be instantaneous.

Q: Which Elasticsearch versions does the in-place upgrade feature support?

Amazon Elasticsearch Service currently supports in-place version upgrade for domains with Elasticsearch versions 5.x and above. The target versions that we support for the upgrade are 5.6, 6.3, 6.4, 6.5, 6.7, 6.8, and 7.1. For more details refer our documentation.

Q: My domain runs a version of Elasticsearch older than 5.x. How do I upgrade those domains?

Please refer to our documentation for details on migrating from various Elasticsearch versions.

Q: Will my domain be offline while the in-place upgrade is in progress?

No. Your domain remains available throughout the upgrade process. However, part of the upgrade process involves relocating shards, which can impact domain performance. We recommend upgrading when the load on your domain is low.

Q: How can I check if my domain’s Elasticsearch version can be upgraded?

In-place version upgrade is available only for domains running Elasticsearch 5.x and above. If your domain is of version 5.x or above, you can run the upgrade eligibility check to validate whether your domain can be upgraded to the desired version. Please refer to our documentation to learn more.

Q: What are the tests done by Amazon Elasticsearch Service to validate my domains upgrade eligibility?

For detailed list of the tests we run to validate upgrade eligibility, please refer to our documentation.

Q: Can I update my domain configuration while the version upgrade is in progress?

No. Once the in-place version upgrade has been triggered, you cannot make changes to your domain configuration until the upgrade completes or fails. You can continue reading and writing data while the upgrade is in progress. Also, you can delete the domain, in which case the upgrade is terminated and the domain deleted.

Q: What happens to the automated system snapshot when the in-place version upgrade is in progress?

The version upgrade process automatically takes a snapshot of the system and only starts the actual upgrade if the snapshot succeeds. If the upgrade is in progress when the automated snapshot’s start time is reached, the automated snapshot is skipped for that day and continued on the next day.

Q: How does Amazon Elasticsearch Service safeguard against issues that can crop up during version upgrades?

Amazon Elasticsearch Service runs a set of tests before triggering the upgrade to check for known issues that can block the upgrade. If no issues are encountered, the service takes a snapshot of the domain and starts the upgrade process if the snapshot is successful. The upgrade is not triggered if there are any issues encountered with any of the steps.

Q: What happens if the system encounters issues while performing the in-place version upgrade?

If encountered issues are minor and fixable, Amazon Elasticsearch Service automatically tries to address them and unblock the upgrade. However, if an issue blocks the upgrade, the service reverts back to the snapshot that was taken before the upgrade and logs the error. For more details on viewing the logs from the upgrade progress, please refer to our documentation.

Q: Can I view the history of upgrades on my domain?

Yes. You can view the upgrade logs from the AWS console or request them using the CLI or SDKs. Please refer to our documentation for more details.

Q: Can I pause or cancel the version upgrade after it has been triggered?

No. After the upgrade has been triggered, it cannot be paused or cancelled until it either completes or fails.

Q: Can I run in-place version upgrade on multiple domains in parallel?

Yes. However, if you want to keep all of your domains on the same version, we recommend running the upgrade eligibility check on all domains before upgrading them. This extra step can help catch issues with one domain that might not be present on others.

Q: How long does the in-place version upgrade take?

Depending on the amount of data and the size of the cluster, upgrades can take anywhere from a few minutes to a few hours to complete.

Q: Can I just upgrade the domain quickly without retaining any of the data?

No. With in-place version upgrade, all the data in your cluster is also restored as part of the upgrade process. If you only wish to upgrade the domain alone, you can take a snapshot of your data, delete all your indexes from the domain and then trigger an in-place version upgrade. Alternatively, you can create a separate domain with the newer version and then restore your data to that domain.

Q: Can I downgrade to previous version if I’m not comfortable with the new version?

No. If you need to downgrade to an older version, you must take a snapshot of your upgraded domain and restore it to a domain that uses the older Elasticsearch version.

Security

Q: How can I secure my Amazon Elasticsearch Service domain?

Amazon Elasticsearch Service provides multiple security features and is HIPAA eligible and compliant with PCI DSS, SOC, ISO, and FedRamp standards, so that you can meet your security and compliance needs. Access to Amazon Elasticsearch Service management APIs for operations such as creating and scaling domains are controlled with AWS Identity and Access Management (IAM) policies.

Amazon Elasticsearch Service domains can be configured to be accessible with an endpoint within your VPC or a public endpoint accessible to the internet. Network access for VPC endpoints is controlled with security groups and for public endpoints access can be granted or restricted by IP address.

In addition to network-based access control, Amazon Elasticsearch Service provides user authentication via IAM and basic authentication using username and password. Authorization can be granted at the domain level (via Domain Access Policies) as well as at the index, document, and field level (via the fine-grained access control feature powered by Open Distro for Elasticsearch). Additionally the fine-grained access control feature extends Kibana with read-only views and secure multi-tenant support.

Amazon Elasticsearch Service also supports an integration with Amazon Cognito, to allow your end-users to log-in to Kibana through enterprise identity providers such as Microsoft Active Directory using SAML 2.0, Amazon Cognito User Pools, and more. Once you sign-in, Amazon Cognito establishes a session using the appropriate IAM principal, which provides access to the Amazon Elasticsearch Service domain. These IAM principals are then available to be used with the fine-grained access control feature powered by Open Distro for Elasticsearch.

Q: How does security authentication and authorization work in Amazon Elasticsearch Service?

Amazon Elasticsearch Service security has three main layers: Network, Domain access policies, and fine-grained access control. The first security layer is the network, which determines whether requests reach a domain. We support public access via the internet or VPC access limited to specific security groups in your VPC. The domain access policy is the second security layer. After a request reaches a domain endpoint, the Domain Access Policy allows or denies the request access to a given URL. The Domain Access Policy accepts or rejects requests at the edge of the domain, before they reach Elasticsearch itself. The third and final security layer is fine-grained access control. After a Domain Access Policy allows a request to reach a domain endpoint, fine-grained access control evaluates the user credentials and either authenticates the user or denies the request. If fine-grained access control authenticates the user, it fetches all roles mapped to that user and uses the complete set of permissions to determine what data the user has access to.

Q: Does Amazon Elasticsearch Service support encryption?

Yes, Amazon Elasticsearch Service supports encryption at rest through AWS Key Management Service (KMS), node-to-node encryption over TLS, and the ability to require clients to communicate of HTTPS. Encryption at rest encrypts shards, log files, swap files, and automated S3 snapshots. You can use AWS-managed keys or choose one of your own. Node-to-node encryption enables TLS for all communications between nodes. Amazon Elasticsearch Service automatically deploys and rotates certificates throughout the life of the domain. If you require you clients to communicate over HTTPS, you also have the ability to specify the minimum TLS version.

Q: If I set up VPC access for my Amazon Elasticsearch Service domain, how can I access Kibana?

When VPC access is enabled, the endpoint for Amazon Elasticsearch Service is only accessible within the customer VPC. To use your laptop to access Kibana from outside the VPC, you need to connect the laptop to the VPC using VPN or VPC Direct Connect.

Pricing

On-Demand Instance pricing

Q: How will I be charged and billed for my use of Amazon Elasticsearch Service?

You pay only for what you use, and there are no minimum or setup fees. You are billed based on:

- Amazon Elasticsearch Service instance hours – Based on the class (e.g. Standard Small, Large, Extra Large) of the Amazon Elasticsearch Service instance consumed. Partial Amazon Elasticsearch Service instance hours consumed are billed as full hours.
- Storage (per GB per month) – Amazon EBS Storage capacity you have provisioned to your Amazon Elasticsearch Service instance. If you scale your provisioned storage capacity within the month, your bill will be pro-rated.
- Provisioned IOPS per month – Amazon EBS Provisioned IOPS rate, regardless of IOPS consumed (for Amazon Elasticsearch Service Provisioned IOPS (SSD) Storage only).
- Data transfer – Regular AWS data transfer charges apply.

Please refer to the Amazon Elasticsearch Service pricing page for detailed pricing information.

Q: When does billing of my Amazon Elasticsearch Service domain begin and end?

Billing commences for an Amazon Elasticsearch Service instance as soon as the instance is available. Billing continues until the Amazon Elasticsearch Service instance terminates, which would occur upon deletion or in the event of instance failure.

Q: What defines billable instance hours for Amazon Elasticsearch Service?

Amazon Elasticsearch Service instance hours are billed for each hour your instance is running in an available state. If you no longer wish to be charged for your Amazon Elasticsearch Service instance, you must delete the domain to avoid being billed for additional instance hours. Partial Amazon Elasticsearch Service instance hours consumed are billed as full hours.

Reserved Instance pricing

Q: What is a Reserved Instance (RI)?

Amazon Elasticsearch Service Reserved Instances give you the option to reserve an instance for a one- or three-year term, and in turn receive significant savings compared to the On-Demand Instance pricing.

Q: How are Reserved Instances different from On-Demand Instances?

Functionally, Reserved Instances and On-Demand Instances are exactly the same. The only difference is how your instance(s) are billed. With Reserved Instances, you purchase a one- or three-year reservation and receive a lower effective hourly usage rate (compared to On-Demand Instances) for the duration of the term. Unless you purchase Reserved Instances in a Region, all instances in that Region are billed at On-Demand Instance hourly rates.

Q: What are the payment options for Reserved Instances?

Three options are available:

- No Upfront Reserved Instances (NURI) – NURIs offer significant savings compared to On-Demand Instance prices. You pay nothing upfront, but commit to paying for the Reserved Instance over the course of the one- or three-year term.

- Partial Upfront Reserved Instances (PURI) – PURIs offer higher savings than NURIs. You pay for a portion of the total cost upfront, and the remainder over the course of the term. This option balances payments between upfront and hourly.
- All Upfront Reserved Instances (AURI) – AURIs offer the highest savings of all of the Reserved Instance payment options. You pay for the entire reservation with one upfront payment, and pay nothing on an hourly basis.

Q: How do I purchase Reserved Instances?

You purchase Reserved Instances in the "Reserved Instance" section of the AWS Management Console for Amazon Elasticsearch Service. Alternatively, you can use the Amazon Elasticsearch Service API or AWS Command Line Interface to list and purchase Reserved Instances.

Once you purchase a Reserved Instance, you can use it just like an On-Demand Instance. As long as the purchased reservation is active, Amazon Elasticsearch Service applies the reduced hourly rate to it.

Q: Are Reserved Instances specific to an Availability Zone?

Amazon Elasticsearch Service Reserved Instances are purchased for a Region rather than for a specific Availability Zone. After you purchase a Reserved Instance for a Region, the discount applies to matching usage in any Availability Zone within that Region.

Q: How many Reserved Instances can I purchase?

You can procure up to 100 Reserved Instances in a single purchase. If you need more Reserved Instances, you need to place more purchase requests.

Q: Do Reserved Instances include a capacity reservation?

Amazon Elasticsearch Service Reserved Instances are purchased for a Region rather than for a specific Availability Zone. Hence, they are not capacity reservations. Even if capacity is limited in one Availability Zone, Reserved Instances can still be purchased in the Region. The discount applies to matching usage in any Availability Zone within that Region.

Q: What if I have an existing On-Demand Instance that I'd like to convert to a Reserved Instance?

Simply purchase a Reserved Instance of the same type as the existing On-Demand Instance. If the Reserved Instance purchase succeeds, Amazon Elasticsearch Service automatically applies the new hourly usage charge for the duration of your reservation.

Q: If I sign up for a Reserved Instance, when does the term begin? What happens to my Reserved Instance when the term ends?

Pricing changes and the reservation term associated with your Reserved Instance become active after your request is received and the payment authorization is processed. If the one-time payment (if applicable) or new hourly rate (if applicable) cannot be successfully authorized by the next billing period, the discounted price does not take effect and your term does not begin. You can follow the status of your reservation using the console, API, or CLI. For more details, refer our documentation.

When your Reserved Instance term expires, your Reserved Instance reverts to the appropriate On-Demand Instance hourly usage rate for your instance class and Region.

Q: How do I control which instances are billed at the Reserved Instance rate?

When computing your bill, our system automatically applies your reservation(s) such that all eligible instances are charged at the lower hourly Reserved Instance rate. Amazon Elasticsearch Service does not distinguish between On-Demand and Reserved Instances while operating Elasticsearch Service domains.

Q: If I scale my Reserved Instance up or down, what happens to my reservation?

Each Reserved Instance is associated with the instance type and Region that you picked for it. If you change the instance type in the Region where you have the Reserved Instance, you will not receive discounted pricing. You must ensure that your reservation matches the instance type you plan to use. For more details, please refer to Amazon Elasticsearch Service Reserved Instance Documentation.

Q: Can I move a Reserved Instance from one Region or Availability Zone to another?

Each Reserved Instance is associated with a specific Region, which is fixed for the lifetime of the reservation and cannot be changed. Each Reserved Instance can, however, be used in any of the Availability Zones within the associated Region.

Q: Are Reserved Instances applicable if use multiple Availability Zones?

A Reserved Instance is for an AWS Region and can be used in any of the Availability Zones in that Region.

Q: Are Reserved Instances available for both Master nodes and Data nodes?

Yes. Amazon Elasticsearch Service does not differentiate between Master and Data nodes when applying Reserved Instance discounts.

Q: Can I cancel a Reserved Instance?

No, you cannot cancel your Reserved Instances, and the one-time payment (if applicable) and discounted hourly usage rate (if applicable) are not refundable. Also, you cannot transfer the Reserved Instance to another account. You must pay for every hour during your Reserved Instance's term, regardless of your usage.

Q: If I purchase a Reserved Instance from a payer (master) account, is it accessible to all the member accounts?

Yes. Reserved Instance pricing and application follows the policies defined for consolidated billing on AWS. More details can be found here.

Q: If AWS reduces prices of On-Demand Instances for Amazon Elasticsearch Service, will the amount I pay for my current Reserved Instances change?

No. The price you pay for already-purchased Reserved Instances does not change for the term of the reservation.

Q: Can I sell my Reserved Instances on the Reserved Instance Marketplace?

No. Reserved Instances purchased on Amazon Elasticsearch Service cannot be sold on the Reserved Instance Marketplace.

Q: Are volume discounts available for Reserved Instance purchase?

No. We do not offer volume discounts for Amazon Elasticsearch Service Reserved Instances.

Service Level Agreement

Q: What does the Amazon Elasticsearch Service SLA guarantee?

Our Amazon Elasticsearch Service SLA guarantees a Monthly Uptime Percentage of at least 99.9% for Amazon Elasticsearch Service.

Q: How do I know if I qualify for a SLA Service Credit?

You are eligible for a SLA credit for Amazon Elasticsearch Service under the Amazon Elasticsearch Service SLA if multi-AZ domains on Amazon Elasticsearch Service have a Monthly Uptime Percentage of less than 99.9% during any monthly billing cycle.

For full details on all of the terms and conditions of the SLA, as well as details on how to submit a claim, please see the Amazon Elasticsearch Service SLA details page.

UltraWarm (preview)

Q. What is UltraWarm?

UltraWarm is a fully-managed, low-cost, warm storage tier for Amazon Elasticsearch Service. It is compatible with Elasticsearch and Kibana, enabling you to analyze data using the same tools that Amazon Elasticsearch Service provides today. UltraWarm seamlessly integrates with Amazon Elasticsearch

Service's existing features such as integrated alerting, SQL querying, and more. UltraWarm is currently available in preview.

Q. Why should I use UltraWarm?

UltraWarm enables you to cost effectively expand the data you want to analyze on Amazon Elasticsearch Service gaining valuable insights on data that previously may have been deleted or archived. With UltraWarm, you can now economically retain more of your data to interactively analyze it whenever you want.

Q. How does UltraWarm relate to/work with Amazon Elasticsearch Service?

Amazon Elasticsearch Service supports two integrated storage tiers, hot and UltraWarm. The hot tier is powered by data nodes which are used for indexing, updating, and providing the fastest access to data. UltraWarm nodes complement the hot tier by providing low cost, read-only tier for older and less-frequently accessed data.

Q. Why does UltraWarm only need primary data for durability?

UltraWarm uses Amazon Simple Storage Service (Amazon S3) for storage, which is designed for 99.999999999 percent durability, and removes the need to configure an Elasticsearch replica for your warm data. Additionally, if you have more than one UltraWarm node, in the event of a node failure, the other UltraWarm nodes will automatically access the data as needed.

Q. How much data can I store in UltraWarm?

UltraWarm (preview) supports up to 900 TB of primary data. UltraWarm is designed to allow you to fully utilize 100% of this storage and because UltraWarm stores data on S3 for durability, you do not need to use additional storage for Elasticsearch replicas.

Q. What are the performance characteristics of UltraWarm?

UltraWarm delivers an interactive experience in Kibana by implementing granular Elasticsearch I/O caching, prefetching, and query engine optimizations to provide similar performance to high-density instances using local storage.

Q. How can I start using UltraWarm?

To get started with UltraWarm, create a new Amazon Elasticsearch Service domain with UltraWarm enabled via the console, CLI, or APIs. Once your domain is created you can move data from hot to UltraWarm using the Elasticsearch APIs. Learn more.

Open Distro for Elasticsearch

Q: How does Open Distro for Elasticsearch relate to the Amazon Elasticsearch Service?

Open Distro for Elasticsearch is committed to ensuring that there is an innovative, 100% open source distribution of Elasticsearch that is available to everyone. We use Open Distro for Elasticsearch to involve the community in much of the new feature development for the Amazon Elasticsearch Service. This gives customers the ability to provide feedback and contribute ahead of the open source features being made available on the service. Open Distro for Elasticsearch features supported on the service include security, alerting, SQL, and more. To learn more about Open Distro for Elasticsearch or get involved please visit the Open Distro website.

Q: Should I consider Open Distro for Elasticsearch “upstream” from the software running on Amazon Elasticsearch Service?

Yes. We are developing the value added features of Open Distro for Elasticsearch in the open. Once they are generally available in Open Distro for Elasticsearch we will then integrate and roll them out on Amazon Elasticsearch Service.

Q: Are you going to offer Open Distro for Elasticsearch on Amazon Elasticsearch Service?

Amazon Elasticsearch Service runs on Open Distro for Elasticsearch. Open Distro for Elasticsearch is a stack of value added plugins and extensions installed with open source Elasticsearch and Kibana. When customers select a version of

Elasticsearch on the service, they will get the corresponding version of the supported Open Distro for Elasticsearch features. Some example features include security, alerting, and SQL. To learn more see our supported features page.

Amazon EMR FAQs

General

Q: What is Amazon EMR?

Amazon EMR is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data. It utilizes a hosted Hadoop framework running on the web-scale infrastructure of Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3).

Q: What is Apache Hadoop?

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly. You can learn more about Apache Hadoop here.

Q: What can I do with Amazon EMR?

Using Amazon EMR, you can instantly provision as much or as little capacity as you like to perform data-intensive tasks for applications such as web indexing, data mining, log file analysis, machine learning, financial analysis, scientific simulation, and bioinformatics research. Amazon EMR lets you focus on crunching or analyzing your data without having to worry about time-consuming set-up, management or tuning of Hadoop clusters or the compute capacity upon which they sit.

Amazon EMR is ideal for problems that necessitate the fast and efficient processing of large amounts of data. The web service interfaces allow you to build processing workflows, and programmatically monitor progress of running clusters. In addition, you can use the simple web interface of the AWS

Management Console to launch your clusters and monitor processing-intensive computation on clusters of Amazon EC2 instances.

Q: Who can use Amazon EMR?

Anyone who requires simple access to powerful data analysis can use Amazon EMR. You don’t need any software development experience to experiment with several sample applications available in the Developer Guide and on the AWS Big Data Blog.

Q: What can I do with Amazon EMR that I could not do before?

Amazon EMR significantly reduces the complexity of the time-consuming set-up, management and tuning of Hadoop clusters or the compute capacity upon which they sit. You can instantly spin up large Hadoop clusters which will start processing within minutes, not hours or days. When your cluster finishes its processing, unless you specify otherwise, it will be automatically terminated so you are not paying for resources you no longer need.

Using this service you can quickly perform data-intensive tasks for applications such as web indexing, data mining, log file analysis, machine learning, financial analysis, scientific simulation, and bioinformatics research.

As a software developer, you can also develop and run your own more sophisticated applications, allowing you to add functionality such as scheduling, workflows, monitoring, or other features.

Q: What is the data processing engine behind Amazon EMR?

Amazon EMR uses Apache Hadoop as its distributed data processing engine. Hadoop is an open source, Java software framework that supports data-intensive distributed applications running on large clusters of commodity hardware. Hadoop implements a programming model named “MapReduce,” where the data is divided into many small fragments of work, each of which may be executed on any node in the cluster. This framework has been widely used by developers, enterprises and startups and has proven to be a reliable software platform for processing up to petabytes of data on clusters of thousands of commodity machines.

Q: What is Apache Spark?

Apache Spark^{TM} is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. Amazon EMR is the best place to deploy Apache Spark in the cloud, because it combines the integration and testing rigor of commercial Hadoop & Spark distributions with the scale, simplicity, and cost effectiveness of the cloud. It allows you to launch Spark clusters in minutes without needing to do node provisioning, cluster setup, Spark configuration, or cluster tuning. Learn more about Spark and Spark on Amazon EMR.

Q: What is Presto?

Presto (or PrestoDB) is an open source, distributed SQL query engine, designed from the ground up for fast analytic queries against data of any size. With Amazon EMR, you can launch Presto clusters in minutes without needing to do node provisioning, cluster setup, Presto configuration, or cluster tuning. EMR enables you to provision one, hundreds, or thousands of compute instances in minutes. Learn more about Presto and Presto on Amazon EMR.

Q: What is an Amazon EMR cluster?

Amazon EMR historically referred to an Amazon EMR cluster (and all processing steps assigned to it) as a "cluster". Every cluster has a unique identifier that starts with "j-".

Q: What is a cluster step?

A cluster step is a user-defined unit of processing, mapping roughly to one algorithm that manipulates the data. A step is a Hadoop MapReduce application implemented as a Java jar or a streaming program written in Java, Ruby, Perl, Python, PHP, R, or C++. For example, to count the frequency with which words appear in a document, and output them sorted by the count, the first step would be a MapReduce application which counts the occurrences of each word, and the second step would be a MapReduce application which sorts the output from the first step based on the counts.

Q: What are different cluster states?

STARTING – The cluster provisions, starts, and configures EC2 instances.
BOOTSTRAPPING – Bootstrap actions are being executed on the cluster.
RUNNING – A step for the cluster is currently being run.
WAITING – The cluster is currently active, but has no steps to run.
TERMINATING - The cluster is in the process of shutting down.
TERMINATED - The cluster was shut down without error.
TERMINATED_WITH_ERRORS - The cluster was shut down with errors.

Q: What are different step states?

PENDING – The step is waiting to be run.
RUNNING – The step is currently running.
COMPLETED – The step completed successfully.
CANCELLED – The step was cancelled before running because an earlier step failed or cluster was terminated before it could run.
FAILED – The step failed while running.

Launching a cluster

Q: How can I access Amazon EMR?

You can access Amazon EMR by using the AWS Management Console, Command Line Tools, SDKS, or the EMR API.

Q: How can I launch a cluster?

You can launch a cluster through the AWS Management Console by filling out a simple cluster request form. In the request form, you specify the name of your cluster, the location in Amazon S3 of your input data, your processing application, your desired data output location, and the number and type of Amazon EC2 instances you'd like to use. Optionally, you can specify a location to store your cluster log files and SSH Key to login to your cluster while it is running. Alternatively, you can launch a cluster using the RunJobFlow API or using the 'create' command in the Command Line Tools.

Q: How can I get started with Amazon EMR?

To sign up for Amazon EMR, click the “Sign Up Now” button on the Amazon EMR detail page http://aws.amazon.com/emr. You must be signed up for Amazon EC2 and Amazon S3 to access Amazon EMR; if you are not already signed up for these services, you will be prompted to do so during the Amazon EMR sign-up process. After signing up, please refer to the Amazon EMR documentation, which includes our Getting Started Guide – the best place to get going with the service.

Q: How can I terminate a cluster?

At any time, you can terminate a cluster via the AWS Management Console by selecting a cluster and clicking the “Terminate” button. Alternatively, you can use the TerminateJobFlows API. If you terminate a running cluster, any results that have not been persisted to Amazon S3 will be lost and all Amazon EC2 instances will be shut down.

Q: Does Amazon EMR support multiple simultaneous cluster?

Yes. At any time, you can create a new cluster, even if you’re already running one or more clusters.

Q: How many clusters can I run simultaneously?

You can start as many clusters as you like. You are limited to 20 instances across all your clusters. If you need more instances, complete the Amazon EC2 instance request form and your use case and instance increase will be considered. If your Amazon EC2 limit has been already raised, the new limit will be applied to your Amazon EMR clusters.

Developing

Q: Where can I find code samples?

Check out the sample code in these Articles and Tutorials.

Q: How do I develop a data processing application?

You can develop a data processing job on your desktop, for example, using Eclipse or NetBeans plug-ins such as IBM MapReduce Tools for Eclipse (http://www.alphaworks.ibm.com/tech/mapreducetools). These tools make it easy to develop and debug MapReduce jobs and test them locally on your machine. Additionally, you can develop your cluster directly on Amazon EMR using one or more instances.

Q: What is the benefit of using the Command Line Tools or APIs vs. AWS Management Console?

The Command Line Tools or APIs provide the ability to programmatically launch and monitor progress of running clusters, to create additional custom functionality around clusters (such as sequences with multiple processing steps, scheduling, workflow, or monitoring), or to build value-added tools or applications for other Amazon EMR customers. In contrast, the AWS Management Console provides an easy-to-use graphical interface for launching and monitoring your clusters directly from a web browser.

Q: Can I add steps to a cluster that is already running?

Yes. Once the job is running, you can optionally add more steps to it via the AddJobFlowSteps API. The AddJobFlowSteps API will add new steps to the end of the current step sequence. You may want to use this API to implement conditional logic in your cluster or for debugging.

Q: Can I run a persistent cluster?

Yes. Amazon EMR clusters that are started with the –alive flag will continue until explicitly terminated. This allows customers to add steps to a cluster on demand. You may want to use this to debug your application without having to repeatedly wait for cluster startup. You may also use a persistent cluster to run a long-running data warehouse cluster. This can be combined with data warehouse and analytics packages that runs on top of Hadoop such as Hive and Pig.

Q: Can I be notified when my cluster is finished?

You can sign up for up Amazon SNS and have the cluster post to your SNS topic when it is finished. You can also view your cluster progress on the AWS Management Console or you can use the Command Line, SDK, or APIs get a status on the cluster.

Q: What programming languages does Amazon EMR support?

You can use Java to implement Hadoop custom jars. Alternatively, you may use other languages including Perl, Python, Ruby, C++, PHP, and R via Hadoop Streaming. Please refer to the Developer's Guide for instructions on using Hadoop Streaming.

Q: What OS versions are supported with Amazon EMR?

Amazon EMR creates cluster instances using an Amazon Linux Amazon Machine Image (AMI) that is optimized for Amazon EMR. The AMI contains the Amazon Linux operating system, other software, and configurations required for each instance to host your cluster applications. As an alternative, you can specify a custom AMI that you create based on the Amazon Linux AMI. This allows you to perform sophisticated pre-configuration for virtually any application. For more information, see Using a Custom AMI.

Q: Can I view the Hadoop UI while my cluster is running?

Yes. Please refer to the Hadoop UI section in the Developer's Guide for instructions on how to access the Hadoop UI.

Q: Does Amazon EMR support third-party software packages?

Yes. The recommended way to install third-party software packages on your cluster is to use Bootstrap Actions. Alternatively you can package any third party libraries directly into your Mapper or Reducer executable. You can also upload statically compiled executables using the Hadoop distributed cache mechanism.

Q: Which Hadoop versions does Amazon EMR support?

For the latest versions supported by Amazon EMR, please reference the documentation.

Q: Does Amazon contribute Hadoop improvements to the open source community?

Yes. Amazon EMR is active with the open source community and contributes many fixes back to the Hadoop source.

Q: Does Amazon EMR update the version of Hadoop it supports?

Amazon EMR periodically updates its supported version of Hadoop based on the Hadoop releases by the community. Amazon EMR may choose to skip some Hadoop releases.

Q: How quickly does Amazon EMR retire support for old Hadoop versions?

Amazon EMR service retires support for old Hadoop versions several months after deprecation. However, Amazon EMR APIs are backward compatible, so if you build tools on top of these APIs, they will work even when Amazon EMR updates the Hadoop version it's using.

EMR Notebooks

Q: What are EMR Notebooks?

EMR Notebooks provide a managed environment, based on Jupyter Notebook, that allows data scientists, analysts, and developers to prepare and visualize data, collaborate with peers, build applications, and perform interactive analysis using EMR clusters.

Q: What can I do with EMR Notebooks?

You can use EMR Notebooks to build Apache Spark applications and run interactive queries on your EMR cluster with minimal effort. Multiple users can create serverless notebooks directly from the console, attach them to an existing shared EMR cluster, or provision a minimum 1-node cluster directly from the console and immediately start experimenting with Spark. You can detach notebooks and re-attach them to new clusters. Notebooks are auto-saved to S3 buckets, and you can retrieve saved notebooks from the console to resume work. EMR Notebooks are prepackaged with the libraries found in the Anaconda repository, allowing you to import and use these libraries in your notebooks code and use them to manipulate data and visualize results. Further, EMR notebooks have integrated Spark monitoring capabilities that you can use to monitor the progress of your Spark jobs and debug code from within the notebook.

Q: How do I get started with EMR Notebooks?

To get started with EMR Notebooks, open the EMR console and choose **Notebooks** in the navigation pane. From there, just choose **Create Notebook**, enter a name for your notebook, choose an EMR cluster or instantly create a new one, provide a service role for the notebook to use, and choose an S3 bucket where you want to save your notebook files and then click on **Create Notebook**. After the notebook shows a **Ready** status, choose **Open** to start the notebook editor.

Q: Can I open EMR Notebooks without logging into the AWS Management Console?

No, to create or open a notebook and run queries on your EMR cluster you need to log into the AWS Management Console. The notebook files are saved to your S3 bucket in ipynb format and can be downloaded and opened locally from your machine.

Q: What programming languages does EMR Notebooks support?

EMR notebook supports PySpark, SparkR, SparkSQL, Spark (Scala), and Python kernels.

Q: What libraries are available with EMR Notebooks?

Libraries found in the open-source Anaconda repositories are available to import in your code. You can import these libraries and use it locally within notebooks.

Q: Can I install custom libraries to use in my notebook code?

All Spark queries run on your EMR cluster, so you need to install all runtime libraries that your Spark application uses on the cluster. You can use a bootstrap action or a custom AMI to install required libraries when you create a cluster. For more information, see Create Bootstrap Actions to Install Additional Software and Using a Custom AMI in the *Amazon EMR Management Guide*. Installing libraries from within the notebook editor is not supported.

Q: What are the service limits associated with EMR Notebooks?

Notebooks use the master node of your EMR cluster to run queries. The size of the master instance limits the number of notebooks that you can attach to a cluster. Once you exceed the limit, you must stop an active notebook before you can start another.

Q: How do I stop my notebook?

You can use the EMR console. Choose **Notebooks**, select the notebook from the list, and choose **Stop**. This terminates the notebook session and makes it unavailable to open in the notebook editor. You can choose **Start** to restart the notebook.

Q: How do I delete my notebook?

You can use the EMR console. Choose **Notebooks**, select the notebook from the list, and choose **Delete**. Deleting a notebook only removes it from the listing in the console. The notebook file persists in the Amazon S3 location that you specified when you created the notebook.

Q: How do I run queries and execute code from a notebook?

Spark queries that you run within a notebook execute on the EMR cluster that you choose when you create the notebook. The programming language kernel that you select from within the notebook editor interacts with the Livy server installed on your EMR cluster to create a Spark session, and all your queries run on the cluster. The output from the Spark application are communicated back to the kernel using Livy and can be seen within the notebook.

Before executing code within the notebook editor, you need to be sure that the notebook has a status of **Ready**. This status means that the interface between the applications on the cluster and the notebook editor are prepared to run queries and execute code. To open the editor, select the notebook from the Notebooks list, and then choose **Open** to start the notebook editor in a new browser tab. In the notebook editor, from the **Kernel** list, select the programming language kernel for your queries. After the kernel starts and is ready, you can run code as you usually would in a Jupyter notebook - for example, clicking the **Run** button in a single cell, choosing **Run All** from the **Cell** menu, and so on.

Q: What EMR release versions are supported with EMR Notebooks?

EMR Notebooks can be attached to EMR clusters running EMR release 5.18.0 or later.

Q: Can I create a notebook or open the notebook editor without an EMR cluster?

No, to create or open your EMR Notebook from the console you need to attach it to a running EMR cluster for the duration of your notebook session. You can quickly create a compatible EMR cluster when you create the notebook or before you restart it. You can always down a previously created notebook file in the ipynb format from the S3 location you chose when you created the notebook.

Q: Can I leave my notebook session running indefinitely?

No. If a notebook is idle for an extended time, the notebook is stopped. If the notebook editor is still open, code that you run in the editor fails. You can start a notebook again from the EMR console and then re-open the notebook editor.

Q: What happens if I close the notebook editor while it‘s running some code on the cluster?

Closing the notebook editor will not impact any code running on the cluster, but if you don’t re-open the notebook editor for an extended time, the notebook will be stopped, and you will not get any output back into the notebook. You can re-start this notebook and resume your work by clicking on the notebook link.

Q: Does the EMR cluster shut down if it is no longer attached to a notebook?

No. You need to terminate the cluster to shut it down.

Q: What other Apache Hadoop applications can I use with EMR Notebooks?

EMR Notebooks currently supports Spark in the Hadoop ecosystem.

Q: Can I use a notebook with different EMR clusters?

Yes, you can change EMR clusters. Notebooks must be stopped before you can change clusters. You can then select the cluster from the **Notebooks** list, choose **View Details**, choose **Change cluster** to select a running cluster or create a new one, and then choose **Change cluster and start the notebook**.

Q: Where are the notebooks saved?

Notebook files are saved automatically at regular intervals to the ipynb file format in the Amazon S3 location that you specify when you create the notebook. The notebook file has the same name as your notebook in the EMR console. You can also save the notebook manually at any time using the **Save and Checkpoint** feature within the notebook editor. This creates an ipynb file with the same name in a sub-folder named **checkpoint**. The most recent checkpoint file overwrites previous checkpoint files. The **Save as** feature in the notebook editor is not available.

Q: How do I use version control with my notebook? Can I use repositories like GitHub?

EMR Notebooks currently do not integrate with repositories for version control.

Q: How do I use my saved notebooks?

To work with a saved notebook, click on the notebook from the **Notebooks** list using the EMR console.

Q: Can I integrate my corporate Active Directory with EMR Notebooka?

EMR Notebooks can be accessed only through the AWS Management Console for EMR. You can federate users from your Active Directory (AD) to the AWS Management for a single sign-on experience. Please visit Enabling federation to AWS using Active Directory, ADFS and SAML 2.0 for more information

Q: What are the IAM policies needed to use the notebooks?

Users must have an identity-based policy statement that gives permissions to create and use EMR Notebooks. In addition to the user policy, EMR Notebooks uses a service role to access other AWS resources and perform actions. For more information, see Security for EMR Notebooks in the *Amazon EMR Release Guide*.

Q: How does the notebook communicate with the master node of my EMR cluster and what security is available for that?

The EMR master node uses Livy to interact with the notebook editor. Each EMR Notebook uses Amazon EC2 security groups to control the network traffic between the Livy server on the master node and an EMR notebook. The default security group rules limit network traffic so that only Livy traffic can pass between notebook editors and master nodes on clusters that notebooks use. You can provide custom security groups with customized inbound and outbound rules for each notebook, and each cluster, to further restrict allowed communication between specific notebooks and clusters from the notebook console page or provide permissions in the notebook service role to have the notebook service create the security groups on your behalf. For more information, see Specifying EC2 Security Groups in the *Amazon EMR Release Guide*.

Q: As an Admin, how do I control access to the EMR cluster for notebook users?

You can limit the Amazon EMR clusters that a user can query with a notebook by using tags on the cluster. If a user has the permissions to create a notebook, they can attach to any Amazon EMR cluster unless access is restricted through the use of tags. For more information, see EMR Notebook tags in the *Amazon EMR Release Guide*.

Q: Can multiple users open the same notebook concurrently?

No, only one user can open a notebook at a time. To view the current user, select the notebook from the Notebooks list, choose View details, and you can see the user name and IAM Amazon Resource Name (ARN) of the user who last modified the notebook as Last modified by. For more information on ARNs see Amazon Resource Names in *AWS General Reference.*

Q: How do I restrict the ability of users to edit or delete my notebook?

You can control access to your notebooks using notebook tags together with identity-based IAM policies. By default, a tag associated with the user creating the notebook is added automatically to the notebook. For more information, see Using Notebook Tags to Control IAM User Access in the *EMR Management Guide*.

Q: Can I attach my notebook to a Kerberos enabled EMR cluster?

No, kerberized EMR clusters are currently not supported.

Q: Can I terminate a cluster if a notebook is using it?

Yes. If the notebook editor is still open, code that you run in the editor fails and the notebook will be stopped after sometime.

Q: What is the cost of using EMR Notebooks?

EMR notebooks are provided at no additional charge to Amazon EMR customers. You will be charged as usual for the attached EMR clusters in your account. You and find out more about the pricing for your cluster by visiting https://aws.amazon.com/emr/pricing/

Debugging

Q: How can I debug my cluster?

You first select the cluster you want to debug, then click on the “Debug” button to access the debug a cluster window in the AWS Management Console. This will enable you to track progress and identify issues in steps, jobs, tasks, or task attempts of your clusters. Alternatively you can SSH directly into the Amazon Elastic Compute Cloud (Amazon EC2) instances that are running your cluster and use your favorite command-line debugger to troubleshoot the cluster.

Q: What is the cluster debug tool?

The cluster debug tool is a part of the AWS Management Console where you can track progress and identify issues in steps, jobs, tasks, or task attempts of your clusters. To access the cluster debug tool, first select the cluster you want to debug and then click on the “Debug” button.

Q: How can I enable debugging of my cluster?

To enable debugging you need to set “Enable Debugging” flag when you create a cluster in the AWS Management Console. Alternatively, you can pass the --enable-debugging and --log-uri flags in the Command Line Client when creating a cluster.

Q: Where can I find instructions on how to use the debug a cluster window?

Please reference the AWS Management Console section of the Developer’s Guide for instructions on how to access and use the debug a cluster window.

Q: What types of clusters can I debug with the debug a cluster window?

You can debug all types of clusters currently supported by Amazon EMR including custom jar, streaming, Hive, and Pig.

Q: Why do I have to sign-up for Amazon SimpleDB to use cluster debugging?

Amazon EMR stores state information about Hadoop jobs, tasks and task attempts under your account in Amazon SimpleDB. You can subscribe to Amazon SimpleDB here.

Q: Can I use the cluster debugging feature without Amazon SimpleDB subscription?

You will be able to browse cluster steps and step logs but will not be able to browse Hadoop jobs, tasks, or task attempts if you are not subscribed to Amazon SimpleDB.

Q: Can I delete historical cluster data from Amazon SimpleDB?

Yes. You can delete Amazon SimpleDB domains that Amazon EMR created on your behalf. Please reference the Amazon SimpleDB documentation for instructions.

Managing data

Q: How do I get my data into Amazon S3?

You can use Amazon S3 APIs to upload data to Amazon S3. Alternatively, you can use many open source or commercial clients to easily upload data to Amazon S3.

Q: How do I get logs for completed clusters?

Hadoop system logs as well as user logs will be placed in the Amazon S3 bucket which you specify when creating a cluster.

Q: Do you compress logs?

No. At this time Amazon EMR does not compress logs as it moves them to Amazon S3.

Q: Can I load my data from the internet or somewhere other than Amazon S3?

Yes. Your Hadoop application can load the data from anywhere on the internet or from other AWS services. Note that if you load data from the internet, EC2 bandwidth charges will apply. Amazon EMR also provides Hive-based access to data in DynamoDB.

Billing

Q: Can Amazon EMR estimate how long it will take to process my input data?

No. As each cluster and input data is different, we cannot estimate your job duration.

Q: How much does Amazon EMR cost?

As with the rest of AWS, you pay only for what you use. There is no minimum fee and there are no up-front commitments or long-term contracts. Amazon EMR pricing is in addition to normal Amazon EC2 and Amazon S3 pricing.

For Amazon EMR pricing information, please visit EMR's pricing page.

Amazon EC2, Amazon S3 and Amazon SimpleDB charges are billed separately. Pricing for Amazon EMR is per-second consumed for each instance type (with a one-minute minimum), from the time cluster is requested until it is terminated. For additional details on Amazon EC2 Instance Types, Amazon EC2 Spot Pricing, Amazon EC2 Reserved Instances Pricing, Amazon S3 Pricing, or Amazon SimpleDB Pricing, follow the links below:

Amazon EC2 Instance Types

Amazon EC2 Reserved Instances Pricing

Amazon EC2 Spot Instances Pricing

Amazon S3 Pricing

Amazon SimpleDB Pricing

Q: When does billing of my Amazon EMR cluster begin and end?

Billing commences when Amazon EMR starts running your cluster. You are only charged for the resources actually consumed. For example, let’s say you launched 100 Amazon EC2 Standard Small instances for an Amazon EMR cluster, where the Amazon EMR cost is an incremental $0.015 per hour. The Amazon EC2 instances will begin booting immediately, but they won’t necessarily all start at the same moment. Amazon EMR will track when each instance starts and will check it into the cluster so that it can accept processing tasks.

In the first 10 minutes after your launch request, Amazon EMR either starts your cluster (if all of your instances are available) or checks in as many instances as possible. Once the 10 minute mark has passed, Amazon EMR will start processing (and charging for) your cluster as soon as 90% of your requested instances are available. As the remaining 10% of your requested instances check in, Amazon EMR starts charging for those instances as well.

So, in the above example, if all 100 of your requested instances are available 10 minutes after you kick off a launch request, you'll be charged $1.50 per hour (100 * $0.015) for as long as the cluster takes to complete. If only 90 of your requested instances were available at the 10 minute mark, you'd be charged $1.35 per hour (90 * $0.015) for as long as this was the number of instances running your cluster. When the remaining 10 instances checked in, you'd be charged $1.50 per hour (100 * $0.015) for as long as the balance of the cluster takes to complete.

Each cluster will run until one of the following occurs: you terminate the cluster with the TerminateJobFlows API call (or an equivalent tool), the cluster shuts itself down, or the cluster is terminated due to software or hardware failure.

Q: Where can I track my Amazon EMR, Amazon EC2 and Amazon S3 usage?

You can track your usage in the Billing & Cost Management Console.

Q: How do you calculate the Normalized Instance Hours displayed on the console ?

On the AWS Management Console, every cluster has a Normalized Instance Hours column that displays the approximate number of compute hours the cluster has used, rounded up to the nearest hour. Normalized Instance Hours are hours of compute time based on the standard of 1 hour of m1.small usage = 1 hour normalized compute time. The following table outlines the normalization factor used to calculate normalized instance hours for the various instance sizes:

Instance Size	Normalization Factor
Small	1
Medium	2
Large	4
Xlarge	8
2xlarge	16
4xlarge	32
8xlarge	64

For example, if you run a 10-node r3.8xlarge cluster for an hour, the total number of Normalized Instance Hours displayed on the console will be 640 (10 (number of nodes) x 64 (normalization factor) x 1 (number of hours that the cluster ran) = 640).

This is an approximate number and should not be used for billing purposes. Please refer to the Billing & Cost Management Console for billable Amazon EMR usage. Note that we recently changed the normalization factor to accurately reflect the weights of the instances, and the normalization factor does not affect your monthly bill.

Q: Does Amazon EMR support Amazon EC2 On-Demand, Spot, and Reserved Instances?

Yes. Amazon EMR seamlessly supports On-Demand, Spot, and Reserved Instances. Click here to learn more about Amazon EC2 Reserved Instances. Click here to learn more about Amazon EC2 Spot Instances.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Security

Q: How do I prevent other people from viewing my data during cluster execution?

Amazon EMR starts your instances in two Amazon EC2 security groups, one for the master and another for the other cluster nodes. The master security group has a port open for communication with the service. It also has the SSH port open to allow you to SSH into the instances, using the key specified at startup. The other nodes start in a separate security group, which only allows interaction with the master instance. By default both security groups are set up to not allow access from external sources including Amazon EC2 instances belonging to other customers. Since these are security groups within your account, you can reconfigure them using the standard EC2 tools or dashboard. Click here to learn more about EC2 security groups.

Q: How secure is my data?

Amazon S3 provides authentication mechanisms to ensure that stored data is secured against unauthorized access. Unless the customer who is uploading the data specifies otherwise, only that customer can access the data. Amazon EMR customers can also choose to send data to Amazon S3 using the HTTPS protocol for secure transmission. In addition, Amazon EMR always uses HTTPS to send data between Amazon S3 and Amazon EC2. For added security, customers may encrypt the input data before they upload it to Amazon S3 (using any common data encryption tool); they then need to add a decryption step to the beginning of their cluster when Amazon EMR fetches the data from Amazon S3.

Q: Can I get a history of all EMR API calls made on my account for security or compliance auditing?

Yes. AWS CloudTrail is a web service that records AWS API calls for your account and delivers log files to you. The AWS API call history produced by CloudTrail enables security analysis, resource change tracking, and compliance auditing. Learn more about CloudTrail at the AWS CloudTrail detail page, and turn it on via CloudTrail's AWS Management Console.

Regions and Availability Zones

Q: How does Amazon EMR make use of Availability Zones?

Amazon EMR launches all nodes for a given cluster in the same Amazon EC2 Availability Zone. Running a cluster in the same zone improves performance of the jobs flows because it provides a higher data access rate. By default, Amazon EMR chooses the Availability Zone with the most available resources in which to run your cluster. However, you can specify another Availability Zone if required.

Q: In what Regions is this Amazon EMR available?

For a list of the supported Amazon EMR AWS regions, please visit the AWS Region Table for all AWS global infrastructure.

Q: Which Region should I select to run my clusters?

When creating a cluster, typically you should select the Region where your data is located.

Q: Can I use EU data in a cluster running in the US region and vice versa?

Yes you can. If you transfer data from one region to the other you will be charged bandwidth charges. For bandwidth pricing information, please visit the pricing section on the EC2 detail page.

Q: What is different about the AWS GovCloud (US) region?

The AWS GovCloud (US) region is designed for US government agencies and customers. It adheres to US ITAR requirements. In GovCloud, EMR does not support spot instances or the enable-debugging feature. The EMR Management Console is not yet available in GovCloud.

Q: Is Amazon EMR supported in AWS Local Zones?

A: EMR does not yet support AWS Local Zones, but will in the coming months.

EMR on AWS Outposts

Q: What is AWS Outposts?

AWS Outposts brings native AWS services, infrastructure, and operating models to virtually any data center, co-location space, or on-premises facility. Using EMR on Outposts, you can deploy, manage, and scale EMR clusters on-premises, just as you would in the cloud.

Q: When should I use EMR on Outposts?

If you have existing on-premises Apache Hadoop deployments and are struggling to meet capacity demands during peak utilization, you can use EMR on Outposts to augment your processing capacity without having to move data to the cloud. EMR on Outposts enables you to launch a new EMR cluster on-premises in minutes and connect to existing datasets in on-premises HDFS storage to meet this demand and maintain SLAs.

If you need to process data that needs to remain on-premises for governance, compliance, or other reasons, you can use EMR on Outposts to deploy and run applications like Apache Hadoop and Apache Spark on-premises, close to your data. This reduces the need to move large amounts of on-premises data to the cloud, reducing the overall time needed to process that data.

If you're in the process of migrating data and Apache Hadoop workloads to the cloud and want to start using EMR before your migration is complete, you can use AWS Outposts to launch EMR clusters that connect to your existing on-premises HDFS storage. You can then gradually migrate your data to Amazon S3 as part of an evolution to a cloud architecture.

Q: What EMR versions are supported with EMR on Outposts?

The minimum supported Amazon EMR release is 5.28.0.

Q: What EMR applications are available when using Outposts?

All applications in EMR release 5.28.0 and above are supported. See our release notes for a full list of EMR applications.

Q: What EMR features are not supported with EMR on Outposts?

- EC2 Spot instances are not available in AWS Outposts. When creating a cluster, you must choose EC2 On-Demand instances.

- A subset of EC2 instance types are available in AWS Outposts. For a list of supported instance types with EMR and Outposts, please see our documentation.
- When adding Amazon EBS volumes to instances, only the General Purpose SSD (GP2) storage type is supported in AWS Outposts.

Q: Can I use EMR clusters in an Outpost to read data from my existing on-premises Apache Hadoop clusters?

Workloads running on EMR in an Outpost can read and write data in existing HDFS storage, allowing you to easily integrate with existing on-premises Apache Hadoop deployments. This gives you the ability to augment your data processing needs using EMR without the need to migrate data.

Q: Can I choose where to store my data?

When an EMR cluster is launched in an Outpost, all of the compute and data storage resources are deployed in your Outpost. Data written locally to the EMR cluster is stored on local EBS volumes in your Outpost. Tools such as Apache Hive, Apache Spark, Presto, and other EMR applications can each be configured to write data locally in an Outpost, to external file system such as an existing HDFS installation, or to Amazon S3. Using EMR on Outposts, you have full control over storing your data in Amazon S3 or locally in your Outpost.

Q: Do any EMR features need uploaded data to S3?

When launching an EMR cluster in an Outpost, you have the option to enable logging. When logging is enabled, cluster logs will be uploaded to the S3 bucket that you specify. These logs are used to simplify debugging clusters after they have been terminated. When disabled, no logs will be uploaded to S3.

Q: What happens if my Outpost is out of capacity?

When launching a cluster in an Outpost, EMR will attempt to launch the number and type of EC2 On-Demand instances you've requested. If there is no capacity available on the Outpost, EMR will receive an insufficient capacity notice. EMR will retry for a period of time, and if no capacity becomes available, the cluster will fail to launch. The same process applies when resizing a cluster.

If there is insufficient capacity on the Outpost for the requested instance types, EMR will be unable to scale up the cluster. You can easily set up Amazon CloudWatch alerts to monitor your capacity utilization on Outposts and receive alerts when instance capacity is lower than a desired threshold.

Q: What happens if network connectivity is interrupted between my Outpost and AWS?

If network connectivity between your Outpost and its AWS Region is lost, your clusters in Outposts will continue to run, but there will be actions you will be unable to take until connectivity is restored. Until connectivity is restored, you cannot create new clusters or take new actions on existing clusters. In case of instance failures, the instance will not be automatically replaced. Also, actions such as adding steps to a running cluster, checking step execution status, and sending CloudWatch metrics and events will be delayed until connectivity is restored.

We recommend that you provide reliable and highly available network connectivity between your Outpost and the AWS Region. If network connectivity between your Outpost and its AWS Region is lost for more than a few hours, clusters that have terminate protection enabled will continue to run, and clusters that have terminate protection disabled may be terminated. If network connectivity will be impacted due to routine maintenance, we recommend proactively enabling terminate protection.

Managing your cluster

Q: How does Amazon EMR use Amazon EC2 and Amazon S3?

Customers upload their input data and a data processing application into Amazon S3. Amazon EMR then launches a number of Amazon EC2 instances as specified by the customer. The service begins the cluster execution while pulling the input data from Amazon S3 using S3N protocol into the launched Amazon EC2 instances. Once the cluster is finished, Amazon EMR transfers the output data to Amazon S3, where customers can then retrieve it or use as input in another cluster.

Q: How is a computation done in Amazon EMR?

Amazon EMR uses the Hadoop data processing engine to conduct computations implemented in the MapReduce programming model. The customer implements their algorithm in terms of map() and reduce() functions. The service starts a customer-specified number of Amazon EC2 instances, comprised of one master and multiple other nodes. Amazon EMR runs Hadoop software on these instances. The master node divides input data into blocks, and distributes the processing of the blocks to the other nodes. Each node then runs the map function on the data it has been allocated, generating intermediate data. The intermediate data is then sorted and partitioned and sent to processes which apply the reducer function to it locally on the nodes. Finally, the output from the reducer tasks is collected in files. A single "cluster" may involve a sequence of such MapReduce steps.

Q: How reliable is Amazon EMR?

Amazon EMR manages an Amazon EC2 cluster of compute instances using Amazon's highly available, proven network infrastructure and datacenters. Amazon EMR uses industry proven, fault-tolerant Hadoop software as its data processing engine. Hadoop splits the data into multiple subsets and assigns each subset to more than one Amazon EC2 instance. So, if an Amazon EC2 instance fails to process one subset of data, the results of another Amazon EC2 instance can be used.

Q: How quickly will my cluster be up and running and processing my input data?

Amazon EMR starts resource provisioning of Amazon EC2 On-Demand instances almost immediately. If the instances are not available, Amazon EMR will keep trying to provision the resources for your cluster until they are provisioned or you cancel your request. The instance provisioning is done on a best-efforts basis and depends on the number of instances requested, time when the cluster is created, and total number of requests in the system. After resources have been provisioned, it typically takes fewer than 15 minutes to start processing.

In order to guarantee capacity for your clusters at the time you need it, you may pay a one-time fee for Amazon EC2 Reserved Instances to reserve instance capacity in the cloud at a discounted hourly rate. Like On-Demand Instances, customers pay usage charges only for the time when their instances are running. In this way, Reserved Instances enable businesses with known instance requirements to maintain the elasticity and flexibility of On-Demand Instances, while also reducing their predictable usage costs even further.

Q: Which Amazon EC2 instance types does Amazon EMR support?

Amazon EMR supports 12 EC2 instance types including Standard, High CPU, High Memory, Cluster Compute, High I/O, and High Storage. Standard Instances have memory to CPU ratios suitable for most general-purpose applications. High CPU instances have proportionally more CPU resources than memory (RAM) and are well suited for compute-intensive applications. High Memory instances offer large memory sizes for high throughput applications. Cluster Compute instances have proportionally high CPU with increased network performance and are well suited for High Performance Compute (HPC) applications and other demanding network-bound applications. High Storage instances offer 48 TB of storage across 24 disks and are ideal for applications that require sequential access to very large data sets such as data warehousing and log processing. See the EMR pricing page for details on available instance types and pricing per region.

Q: How do I select the right Amazon EC2 instance type?

When choosing instance types, you should consider the characteristics of your application with regards to resource utilization and select the optimal instance family. One of the advantages of Amazon EMR with Amazon EC2 is that you pay only for what you use, which makes it convenient and inexpensive to test the performance of your clusters on different instance types and quantity. One effective way to determine the most appropriate instance type is to launch several small clusters and benchmark your clusters.

Q: How do I select the right number of instances for my cluster?

The number of instances to use in your cluster is application-dependent and should be based on both the amount of resources required to store and process your data and the acceptable amount of time for your job to complete. As a general guideline, we recommend that you limit 60% of your disk space to storing the data you will be processing, leaving the rest for intermediate output. Hence, given 3x replication on HDFS, if you were looking to process 5 TB on m1.xlarge instances, which have 1,690 GB of disk space, we recommend your cluster contains at least (5 TB * 3) / (1,690 GB * .6) = 15 m1.xlarge core nodes. You may want to increase this number if your job generates a high amount of intermediate data or has significant I/O requirements. You may also want to include additional task nodes to improve processing performance. See Amazon EC2 Instance Types for details on local instance storage for each instance type configuration.

Q: How long will it take to run my cluster?

The time to run your cluster will depend on several factors including the type of your cluster, the amount of input data, and the number and type of Amazon EC2 instances you choose for your cluster.

Q: If the master node in a cluster goes down, can Amazon EMR recover it?

No. If the master node goes down, your cluster will be terminated and you'll have to rerun your job. Amazon EMR currently does not support automatic failover of the master nodes or master node state recovery. In case of master node failure, the AWS Management console displays "The master node was terminated" message which is an indicator for you to start a new cluster. Customers can back-up their intermediate data on EMR cluster to Amazon S3.

Q: If another node goes down in a cluster, can Amazon EMR recover from it?

Yes. Amazon EMR is fault tolerant for node failures and continues job execution if a node goes down. Amazon EMR will also provision a new node when a core node fails. However, Amazon EMR will not replace nodes if all nodes in the cluster are lost.

Q: Can I SSH onto my cluster nodes?

Yes. You can SSH onto your cluster nodes and execute Hadoop commands directly from there. If you need to SSH into a specific node, you have to first SSH to the master node, and then SSH into the desired node.

Q: What is Amazon EMR Bootstrap Actions?

Bootstrap Actions is a feature in Amazon EMR that provides users a way to run custom set-up prior to the execution of their cluster. Bootstrap Actions can be used to install software or configure instances before running your cluster. You can read more about bootstrap actions in EMR's Developer Guide.

Q: How can I use Bootstrap Actions?

You can write a Bootstrap Action script in any language already installed on the cluster instance including Bash, Perl, Python, Ruby, C++, or Java. There are several pre-defined Bootstrap Actions available. Once the script is written, you need to upload it to Amazon S3 and reference its location when you start a cluster. Please refer to the “Developer’s Guide”: http://docs.amazonwebservices.com/ElasticMapReduce/latest/DeveloperGuide/ for details on how to use Bootstrap Actions.

Q: How do I configure Hadoop settings for my cluster?

The EMR default Hadoop configuration is appropriate for most workloads. However, based on your cluster’s specific memory and processing requirements, it may be appropriate to tune these settings. For example, if your cluster tasks are memory-intensive, you may choose to use fewer tasks per core and reduce your job tracker heap size. For this situation, a pre-defined Bootstrap Action is available to configure your cluster on startup. See the Configure Memory Intensive Bootstrap Action in the Developer’s Guide for configuration details and usage instructions. An additional predefined bootstrap action is available that allows you to customize your cluster settings to any value of your choice. See the Configure Hadoop Bootstrap Action in the Developer’s Guide for usage instructions.

Q: Can I modify the number of nodes in a running cluster?

Yes. Nodes can be of two types: (1) core nodes, which both host persistent data using Hadoop Distributed File System (HDFS) and run Hadoop tasks and (2) task nodes, which only run Hadoop tasks. While a cluster is running you may increase the number of core nodes and you may either increase or decrease the number of task nodes. This can be done through the API, Java SDK, or though the command line client. Please refer to the Resizing Running clusters section in the Developer's Guide for details on how to modify the size of your running cluster.

Q: When would I want to use core nodes versus task nodes?

As core nodes host persistent data in HDFS and cannot be removed, core nodes should be reserved for the capacity that is required until your cluster completes. As task nodes can be added or removed and do not contain HDFS, they are ideal for capacity that is only needed on a temporary basis.

Q: Why would I want to modify the number of nodes in my running cluster?

There are several scenarios where you may want to modify the number of nodes in a running cluster. If your cluster is running slower than expected, or timing requirements change, you can increase the number of core nodes to increase cluster performance. If different phases of your cluster have different capacity needs, you can start with a small number of core nodes and increase or decrease the number of task nodes to meet your cluster's varying capacity requirements.

Q: Can I automatically modify the number of nodes between cluster steps?

Yes. You may include a predefined step in your workflow that automatically resizes a cluster between steps that are known to have different capacity needs. As all steps are guaranteed to run sequentially, this allows you to set the number of nodes that will execute a given cluster step.

Q: How can I allow other IAM users to access my cluster?

To create a new cluster that is visible to all IAM users within the EMR CLI: Add the --visible-to-all-users flag when you create the cluster. For example: elastic-mapreduce --create --visible-to-all-users. Within the Management Console, simply select "Visible to all IAM Users" on the Advanced Options pane of the Create cluster Wizard.

To make an existing cluster visible to all IAM users you must use the EMR CLI. Use --set-visible-to-all-users and specify the cluster identifier. For example: elastic-mapreduce --set-visible-to-all-users true --jobflow j-xxxxxxx. This can only be done by the creator of the cluster.

To learn more, see the Configuring User Permissions section of the EMR Developer Guide.

Tagging your cluster

Q: What Amazon EMR resources can I tag?

You can add tags to an active Amazon EMR cluster. An Amazon EMR cluster consists of Amazon EC2 instances, and a tag added to an Amazon EMR cluster will be propagated to each active Amazon EC2 instance in that cluster. You cannot add, edit, or remove tags from terminated clusters or terminated Amazon EC2 instances which were part of an active cluster.

Q: Does Amazon EMR tagging support resource-based permissions with IAM Users?

No, Amazon EMR does not support resource-based permissions by tag. However, it is important to note that propagated tags to Amazon EC2 instances behave as normal Amazon EC2 tags. Therefore, an IAM Policy for Amazon EC2 will act on tags propagated from Amazon EMR if they match conditions in that policy.

Q: How many tags can I add to a resource?

You can add up to ten tags on an Amazon EMR cluster.

Q: Do my Amazon EMR tags on a cluster show up on each Amazon EC2 instance in that cluster? If I remove a tag on my Amazon EMR cluster, will that tag automatically be removed from each associated EC2 instance?

Yes, Amazon EMR propagates the tags added to a cluster to that cluster's underlying EC2 instances. If you add a tag to an Amazon EMR cluster, it will also appear on the related Amazon EC2 instances. Likewise, if you remove a tag from an Amazon EMR cluster, it will also be removed from its associated Amazon EC2 instances. However, if you are using IAM policies for Amazon EC2 and plan to use Amazon EMR's tagging functionality, you should make sure that permission to use the Amazon EC2 tagging APIs CreateTags and DeleteTags is granted.

Q: How do I get my tags to show up in my billing statement to segment costs?

Select the tags you would like to use in your AWS billing report here. Then, to see the cost of your combined resources, you can organize your billing information based on resources that have the same tag key values.

Q: How do I tell which Amazon EC2 instances are part of an Amazon EMR cluster?

An Amazon EC2 instance associated with an Amazon EMR cluster will have two system tags:

- aws:elasticmapreduce:instance-group-role=CORE
  - Key = instance-group role ; Value = [CORE or TASK];
- aws:elasticmapreduce:job-flow-id=j-12345678
  - Key = job-flow-id ; Value = [JobFlowID]

Q: Can I edit tags directly on the Amazon EC2 instances?

Yes, you can add or remove tags directly on Amazon EC2 instances that are part of an Amazon EMR cluster. However, we do not recommend doing this, because Amazon EMR's tagging system will not sync the changes you make to an associated Amazon EC2 instance directly. We recommend that tags for Amazon EMR clusters be added and removed from the Amazon EMR console, CLI, or API to ensure that the cluster and its associated Amazon EC2 instances have the correct tags.

Using EBS volumes

Q: What can I do now that I could not do before?

Most EC2 instances have fixed storage capacity attached to an instance, known as an "instance store". You can now add EBS volumes to the instances in your Amazon EMR cluster, allowing you to customize the storage on an instance. The feature also allows you to run Amazon EMR clusters on EBS-Only instance families such as the M4 and C4.

Q: What are the benefits of adding EBS volumes to an instance running on Amazon EMR?

You will benefit by adding EBS volumes to an instance in the following scenarios:

1. Your processing requirements are such that you need a large amount of HDFS (or local) storage that what is available today on an instance. With support for EBS volumes, you will be able to customize the storage capacity on an instance relative to the compute capacity that the instance provides. Optimizing the storage on an instance will allow you to save costs.
2. You are running on an older generation instance family (such as the M1 and M2 family) and want to move to latest generation instance family but are constrained by the storage available per node on the next generation instance types. Now you can use any of the new generation instance type and add EBS volumes to optimize the storage. Internal benchmarks indicate that you can save cost and improve performance by moving from an older generation instance family (M1 or M2) to a new generation one (M4, C4 & R3). The Amazon EMR team recommends that you run your application to arrive at the right conclusion.
3. You want to use or migrate to the next-generation EBS-Only M4 and C4 family.

Q: Can I persist my data on an EBS volume after a cluster is terminated?

Currently, Amazon EMR will delete volumes once the cluster is terminated. If you want to persist data outside the lifecycle of a cluster, consider using Amazon S3 as your data store.

Q: What kind of EBS volumes can I attach to an instance?

Amazon EMR allows you to use different EBS Volume Types: General Purpose SSD (GP2), Magnetic and Provisioned IOPS (SSD).

Q: What happens to the EBS volumes once I terminate my cluster?

Amazon EMR will delete the volumes once the EMR cluster is terminated.

Q: Can I use an EBS with instances that already have an instance store?

Yes, You can add EBS volumes to instances that have an instance store.

Q: Can I attach an EBS volume to a running cluster?

No, currently you can only add EBS volumes when launching a cluster.

Q: Can I snapshot volumes from a cluster?

The EBS API allows you to Snapshot a cluster. However, Amazon EMR currently does not allow you to restore from a snapshot.

Q: Can I use encrypted EBS volumes?

No, encrypted volumes are not supported in the current release.

Q: What happens when I remove an attached volume from a running cluster?

Removing an attached volume from a running cluster will be treated as a node failure. Amazon EMR will replace the node and the EBS volume with each of the same.

Using Hive

Q: What is Apache Hive?

Hive is an open source datawarehouse and analytics package that runs on top of Hadoop. Hive is operated by a SQL-based language called Hive QL that allows users to structure, summarize, and query data sources stored in Amazon S3. Hive QL goes beyond standard SQL, adding first-class support for map/reduce functions and complex extensible user-defined data types like Json and Thrift. This capability allows processing of complex and even unstructured data sources such as text documents and log files. Hive allows user extensions via user-defined functions written in Java and deployed via storage in Amazon S3. You can learn more about Apache Hive here.

Q: What can I do with Hive running on Amazon EMR?

Using Hive with Amazon EMR, you can implement sophisticated data-processing applications with a familiar SQL-like language and easy to use tools available with Amazon EMR. With Amazon EMR, you can turn your Hive applications into a reliable data warehouse to execute tasks such as data analytics, monitoring, and business intelligence tasks.

Q: How is Hive different than traditional RDBMS systems?

Traditional RDBMS systems provide transaction semantics and ACID properties. They also allow tables to be indexed and cached so that small amounts of data can be retrieved very quickly. They provide for fast update of small amounts of data and for enforcement of referential integrity constraints. Typically they run on a single large machine and do not provide support for executing map and reduce functions on the table, nor do they typically support acting over complex user defined data types.

In contrast, Hive executes SQL-like queries using MapReduce. Consequently, it is optimized for doing full table scans while running on a cluster of machines and is therefore able to process very large amounts of data. Hive provides partitioned tables, which allow it to scan a partition of a table rather than the whole table if that is appropriate for the query it is executing.

Traditional RDMS systems are best for when transactional semantics and referential integrity are required and frequent small updates are performed. Hive is best for offline reporting, transformation, and analysis of large data sets; for example, performing click stream analysis of a large website or collection of websites.

One of the common practices is to export data from RDBMS systems into Amazon S3 where offline analysis can be performed using Amazon EMR clusters running Hive.

Q: How can I get started with Hive running on Amazon EMR?

The best place to start is to review our written documentation located here.

Q: Are there new features in Hive specific to Amazon EMR?

Yes. There are four new features which make Hive even more powerful when used with Amazon EMR, including:

a/ The ability to load table partitions automatically from Amazon S3. Previously, to import a partitioned table you needed a separate alter table statement for each individual partition in the table. Amazon EMR a now includes a new statement type for the Hive language: “alter table recover partitions.” This statement allows you to easily import tables concurrently into many clusters without having to maintain a shared meta-data store. Use this functionality to read from tables into which external processes are depositing data, for example log files.

b/ The ability to specify an off-instance metadata store. By default, the metadata store where Hive stores its schema information is located on the master node and ceases to exist when the cluster terminates. This feature allows you to override the location of the metadata store to use, for example a MySQL instance that you already have running in EC2.

c/ Writing data directly to Amazon S3. When writing data to tables in Amazon S3, the version of Hive installed in Amazon EMR writes directly to Amazon S3 without the use of temporary files. This produces a significant performance improvement but it means that HDFS and S3 from a Hive perspective behave differently. You cannot read and write within the same statement to the same table if that table is located in Amazon S3. If you want to update a table located in S3, then create a temporary table in the cluster’s local HDFS filesystem, write the results to that table, and then copy them to Amazon S3.

d/ Accessing resources located in Amazon S3. The version of Hive installed in Amazon EMR allows you to reference resources such as scripts for custom map and reduce operations or additional libraries located in Amazon S3 directly from within your Hive script (e.g., add jar s3://elasticmapreduce/samples/hive-ads/libs/jsonserde.jar).

Q: What types of Hive clusters are supported?

There are two types of clusters supported with Hive: interactive and batch. In an interactive mode a customer can start a cluster and run Hive scripts interactively directly on the master node. Typically, this mode is used to do ad hoc data analyses and for application development. In batch mode, the Hive script is stored in Amazon S3 and is referenced at the start of the cluster. Typically, batch mode is used for repeatable runs such as report generation.

Q: How can I launch a Hive cluster?

Both batch and interactive clusters can be started from AWS Management Console, EMR command line client, or APIs. Please refer to the Hive section in the Release Guide for more details on launching a Hive cluster.

Q: When should I use Hive vs. PIG?

Hive and PIG both provide high level data-processing languages with support for complex data types for operating on large datasets. The Hive language is a variant of SQL and so is more accessible to people already familiar with SQL and relational databases. Hive has support for partitioned tables which allow Amazon EMR clusters to pull down only the table partition relevant to the query being executed rather than doing a full table scan. Both PIG and Hive have query plan optimization. PIG is able to optimize across an entire scripts while Hive queries are optimized at the statement level.

Ultimately the choice of whether to use Hive or PIG will depend on the exact requirements of the application domain and the preferences of the implementers and those writing queries.

Q: What version of Hive does Amazon EMR support?

Amazon EMR supports multiple versions of Hive, including version 0.11.0.

Q: Can I write to a table from two clusters concurrently

No. Hive does not support concurrently writing to tables. You should avoid concurrently writing to the same table or reading from a table while you are writing to it. Hive has non-deterministic behavior when reading and writing at the same time or writing and writing at the same time.

Q: Can I share data between clusters?

Yes. You can read data in Amazon S3 within a Hive script by having ‘create external table’ statements at the top of your script. You need a create table statement for each external resource that you access.

Q: Should I run one large cluster, and share it amongst many users or many smaller clusters?

Amazon EMR provides a unique capability for you to use both methods. On the one hand one large cluster may be more efficient for processing regular batch workloads. On the other hand, if you require ad-hoc querying or workloads that vary with time, you may choose to create several separate cluster tuned to the specific task sharing data sources stored in Amazon S3.

Q: Can I access a script or jar resource which is on my local file system?

No. You must upload the script or jar to Amazon S3 or to the cluster’s master node before it can be referenced. For uploading to Amazon S3 you can use tools including s3cmd, jets3t or S3Organizer.

Q: Can I run a persistent cluster executing multiple Hive queries?

Yes. You run a cluster in a manual termination mode so it will not terminate between Hive steps. To reduce the risk of data loss we recommend periodically persisting all of your important data in Amazon S3. It is good practice to regularly transfer your work to a new cluster to test your process for recovering from master node failure.

Q: Can multiple users execute Hive steps on the same source data?

Yes. Hive scripts executed by multiple users on separate clusters may contain create external table statements to concurrently import source data residing in Amazon S3.

Q: Can multiple users run queries on the same cluster?

Yes. In the batch mode, steps are serialized. Multiple users can add Hive steps to the same cluster, however, the steps will be executed serially. In interactive mode, several users can be logged on to the same cluster and execute Hive statements concurrently.

Q: Can data be shared between multiple AWS users?

Yes. Data can be shared using standard Amazon S3 sharing mechanism described here.

Q: Does Hive support access from JDBC?

Yes. Hive provides JDBC drive, which can be used to programmatically execute Hive statements. To start a JDBC service in your cluster you need to pass an optional parameter in the Amazon EMR command line client. You also need to establish an SSH tunnel because the security group does not permit external connections.

Q: What is your procedure for updating packages on EMR AMIs?

On first boot, the Amazon Linux AMIs for EMR connect to the Amazon Linux AMI yum repositories to install security updates. When you use a custom AMI, you can disable this feature, but we don’t recommend this for security reasons.

Q: Can I update my own packages on EMR clusters?

Yes. You can use Bootstrap Actions to install updates to packages on your clusters.

Q: Can I process DynamoDB data using Hive?

Yes. Simply define an external Hive table based on your DynamoDB table. You can then use Hive to analyze the data stored in DynamoDB and either load the results back into DynamoDB or archive them in Amazon S3. For more information please visit our Developer Guide.

Using Hudi

Q: What is Apache Hudi?

Apache Hudi (Incubating) is an open-source data management framework used to simplify incremental data processing and data pipeline development. Apache Hudi enables you to manage data at the record-level in Amazon S3 to simplify Change Data Capture (CDC) and streaming data ingestion, and provides a framework to handle data privacy use cases requiring record level updates and deletes. Data sets managed by Apache Hudi are stored in S3 using open storage formats, and integrations with Presto, Apache Hive, Apache Spark, and AWS Glue Data Catalog give you near real-time access to updated data using familiar tools.

Q: When should I use Apache Hudi?

Apache Hudi helps you with uses cases requiring record-level data management on S3. There are five common use cases that benefit from these abilities:

1. Complying with data privacy laws that require organizations to remove user data, or update user preferences when user's choose to change their preferences as to how their data can be used. Apache Hudi gives you the ability to perform record-level insert, update, and delete operations on your data stored in S3, using open source data formats such as Apache Parquet, and Apache Avro.
2. Consuming real time data streams and applying change data capture logs from enterprise systems. Many organizations require Enterprise Data Warehouses (EDW) and Operational Data Stores (ODS) data to be available in Amazon S3 so it’s accessible to SQL engines like Apache Hive and Presto for data processing and analytics. Apache Hudi simplifies applying change logs, and gives users near real-time access to data.
3. Reinstating late arriving, or incorrect data. Late arriving, or incorrect data requires the data to be restated, and existing data sets updated to incorporate new, or updated records. Apache Hudi allows you to “upsert” records into an existing data set, relying on the framework to insert or update records based on their presence in the data set.
4. Tracking change to data sets and providing the ability to rollback changes. With Apache Hudi, each change to a data set is tracked as a commit, and can be easily rolled back, allowing you to find specific changes to a data set and “undo” them.

5. Simplifying file management on S3. To make sure data files are efficiently sized, customers have to build custom solutions that monitor and re-write many small files into fewer large files. With Apache Hudi, data files on S3 are managed, and users can simply configure an optimal file size to store their data and Hudi will merge files to create efficiently sized files.

Q: How do I create an Apache Hudi data set?

Apache Hudi data sets are created using Apache Spark. Creating a data set is as simple as writing an Apache Spark DataFrame. The metadata for Apache Hudi data sets can optionally be stored in the AWS Glue Data Catalog or the Hive metastore to simplify data discovery and for integrating with Apache Hive and Presto.

Q: How does Apache Hudi manage data sets?

When creating a data set with Apache Hudi, you can choose what type of data access pattern the data set should be optimized for. For read-heavy use cases, Apache Hudi will choose the “Copy on Write” data management strategy to optimize for frequent reads of the data set. This strategy organizes data using columnar storage formats, and merges existing data and new updates when the updates are written. For write-heavy workloads, Apache Hudi uses the “Merge on Read” data management strategy which organizes data using a combination of columnar and row storage formats, and existing data and new updates are merged when the data is read.

Q: How do I write to an Apache Hudi data set?

Changes to Apache Hudi data sets are made using Apache Spark. With Apache Spark, Apache Hudi data sets are operated on using the Spark DataSource API, enabling you to read and write data using a Spark DataFrame. New data can be added and updates can be applied to the DataFrame and when saved. Apache Hudi will manage writing the changes to the individual files on S3.

Q: How do I read from an Apache Hudi data set?

You can read data using either Apache Spark, Apache Hive, or Presto. When you create a data set, you have the option to publish the metadata of that data set in either the AWS Glue Data Catalog, or the Hive metastore. If you choose to publish the metadata in a metastore, your data set will look just like an ordinary table, and you can query that table using Apache Hive and Presto.

Q: What considerations or limitations should I be aware of when using Apache Hudi?

For a full list of consideration and limitations when using Apache Hudi on Amazon EMR, please refer to our Amazon EMR documentation.

Q: How does my existing data work with Apache Hudi?

If you have existing data that you want to now manage with Apache Hudi, you can easily convert your Apache Parquet data to Apache Hudi data sets using an import tool provided with Apache Hudi on Amazon EMR, or you can use Apache Spark to rewrite your existing data as an Apache Hudi data set.

Using Impala

Q: What is Impala?

Impala is an open source tool in the Hadoop ecosystem for interactive, ad hoc querying using SQL syntax. Instead of using MapReduce, it leverages a massively parallel processing (MPP) engine similar to that found in traditional relational database management systems (RDBMS). With this architecture, you can query your data in HDFS or HBase tables very quickly, and leverage Hadoop's ability to process diverse data types and provide schema at runtime. This lends Impala to interactive, low-latency analytics. In addition, Impala uses the Hive metastore to hold information about the input data, including the partition names and data types. Also, Impala on Amazon EMR requires AMIs running Hadoop 2.x or greater. Click here to learn more about Impala.

Q: What can I do with Impala running on Amazon EMR?

Similar to using Hive with Amazon EMR, leveraging Impala with Amazon EMR can implement sophisticated data-processing applications with SQL syntax.

However, Impala is built to perform faster in certain use cases (see below). With Amazon EMR, you can use Impala as a reliable data warehouse to execute tasks such as data analytics, monitoring, and business intelligence. Here are three use cases:

- Use Impala instead of Hive on long-running clusters to perform ad hoc queries. Impala reduces interactive queries to seconds, making it an excellent tool for fast investigation. You could run Impala on the same cluster as your batch MapReduce workflows, use Impala on a long-running analytics cluster with Hive and Pig, or create a cluster specifically tuned for Impala queries.
- Use Impala instead of Hive for batch ETL jobs on transient Amazon EMR clusters. Impala is faster than Hive for many queries, which provides better performance for these workloads. Like Hive, Impala uses SQL, so queries can easily be modified from Hive to Impala.
- Use Impala in conjunction with a third party business intelligence tool. Connect a client ODBC or JDBC driver with your cluster to use Impala as an engine for powerful visualization tools and dashboards.

Both batch and interactive Impala clusters can be created in Amazon EMR. For instance, you can have a long-running Amazon EMR cluster running Impala for ad hoc, interactive querying or use transient Impala clusters for quick ETL workflows.

Q: How is Impala different than traditional RDBMSs?

Traditional relational database systems provide transaction semantics and database atomicity, consistency, isolation, and durability (ACID) properties. They also allow tables to be indexed and cached so that small amounts of data can be retrieved very quickly, provide for fast updates of small amounts of data, and for enforcement of referential integrity constraints. Typically, they run on a single large machine and do not provide support for acting over complex user defined data types. Impala uses a similar distributed query system to that found in RDBMSs, but queries data stored in HDFS and uses the Hive metastore to hold information about the input data. As with Hive, the schema for a query is provided at runtime, allowing for easier schema changes. Also, Impala can query a variety of complex data types and execute user defined functions. However, because Impala processes data in-memory, it is important to understand the hardware limitations of your cluster and optimize your queries for the best performance.

Q: How is Impala different than Hive?

Impala executes SQL queries using a massively parallel processing (MPP) engine, while Hive executes SQL queries using MapReduce. Impala avoids Hive's overhead from creating MapReduce jobs, giving it faster query times than Hive. However, Impala uses significant memory resources and the cluster's available memory places a constraint on how much memory any query can consume. Hive is not limited in the same way, and can successfully process larger data sets with the same hardware. Generally, you should use Impala for fast, interactive queries, while Hive is better for ETL workloads on large datasets. Impala is built for speed and is great for ad hoc investigation, but requires a significant amount of memory to execute expensive queries or process very large datasets. Because of these limitations, Hive is recommended for workloads where speed is not as crucial as completion. Click here to view some performance benchmarks between Impala and Hive.

Q: Can I use Hadoop 1?

No, Impala requires Hadoop 2, and will not run on a cluster with an AMI running Hadoop 1.x.

Q: What instance types should I use for my Impala cluster?

For the best experience with Impala, we recommend using memory-optimized instances for your cluster. However, we have shown that there are performance gains over Hive when using standard instance types as well. We suggest reading our Performance Testing and Query Optimization section in the Amazon EMR Developer's Guide to better estimate the memory resources your cluster will need with regards to your dataset and query types. The compression type, partitions, and the actual query (number of joins, result size, etc.) all play a role in the memory required. You can use the EXPLAIN statement to estimate the memory and other resources needed for an Impala query.

Q: What happens if I run out of memory on a query?

If you run out of memory, queries fail and the Impala daemon installed on the affected node shuts down. Amazon EMR then restarts the daemon on that node so that Impala will be ready to run another query. Your data in HDFS on the node remains available, because only the daemon running on the node shuts down, rather than the entire node itself. For ad hoc analysis with Impala, the query time can often be measured in seconds; therefore, if a query fails, you can discover the problem quickly and be able to submit a new query in quick succession.

Q: Does Impala support user defined functions?

Yes, Impala supports user defined functions (UDFs). You can write Impala specific UDFs in Java or C++. Also, you can modify UDFs or user-defined aggregate functions created for Hive for use with Impala. For information about Hive UDFs, click here.

Q: Where is the data stored for Impala to query?

Impala queries data in HDFS or in HBase tables.

Q: Can I run Impala and MapReduce at the same time on a cluster?

Yes, you can set up a multitenant cluster with Impala and MapReduce. However, you should be sure to allot resources (memory, disk, and CPU) to each application using YARN on Hadoop 2.x. The resources allocated should be dependent on the needs for the jobs you plan to run on each application.

Q: Does Impala support ODBC and JDBC drivers?

While you can use ODBC drivers, Impala is also a great engine for third-party tools connected through JDBC. You can download and install the Impala client JDBC driver from http://elasticmapreduce.s3.amazonaws.com/libs/impala/1.2.1/impala-jdbc-1.2.1.zip. From the client computer where you have your business intelligence tool installed, connect the JDBC driver to the master node of an Impala cluster using SSH or a VPN on port 21050. For more information, see Open an SSH Tunnel to the Master Node.

Using Pig

Q: What is Apache Pig?

Pig is an open source analytics package that runs on top of Hadoop. Pig is operated by a SQL-like language called Pig Latin, which allows users to structure, summarize, and query data sources stored in Amazon S3. As well as SQL-like operations, Pig Latin also adds first-class support for map/reduce functions and complex extensible user defined data types. This capability allows processing of complex and even unstructured data sources such as text documents and log files. Pig allows user extensions via user-defined functions written in Java and deployed via storage in Amazon S3.

Q: What can I do with Pig running on Amazon EMR?

Using Pig with Amazon EMR, you can implement sophisticated data-processing applications with a familiar SQL-like language and easy to use tools available with Amazon EMR. With Amazon EMR, you can turn your Pig applications into a reliable data warehouse to execute tasks such as data analytics, monitoring, and business intelligence tasks.

Q: How can I get started with Pig running on Amazon EMR?

The best place to start is to review our written documentation located here.

Q: Are there new features in Pig specific to Amazon EMR?

Yes. There are three new features which make Pig even more powerful when used with Amazon EMR, including:

a/ Accessing multiple filesystems. By default a Pig job can only access one remote file system, be it an HDFS store or S3 bucket, for input, output and temporary data. EMR has extended Pig so that any job can access as many file systems as it wishes. An advantage of this is that temporary intra-job data is always stored on the local HDFS, leading to improved perfomance.

b/ Loading resources from S3. EMR has extended Pig so that custom JARs and scripts can come from the S3 file system, for example “REGISTER s3:///my-bucket/piggybank.jar” c/ Additional Piggybank function for String and DateTime processing.

Q: What types of Pig clusters are supported?

There are two types of clusters supported with Pig: interactive and batch. In an interactive mode a customer can start a cluster and run Pig scripts interactively directly on the master node. Typically, this mode is used to do ad hoc data analyses and for application development. In batch mode, the Pig script is stored in Amazon S3 and is referenced at the start of the cluster. Typically, batch mode is used for repeatable runs such as report generation.

Q: How can I launch a Pig cluster?

Both batch and interactive clusters can be started from AWS Management Console, EMR command line client, or APIs.

Q: What version of Pig does Amazon EMR support?

Amazon EMR supports multiple versions of Pig.

Q: Can I write to a S3 bucket from two clusters concurrently

Yes, you are able to write to the same bucket from two concurrent clusters.

Q: Can I share input data in S3 between clusters?

Yes, you are able to read the same data in S3 from two concurrent clusters.

Q: Can data be shared between multiple AWS users?

Yes. Data can be shared using standard Amazon S3 sharing mechanism described here http://docs.amazonwebservices.com/AmazonS3/latest/index.html?S3_ACLs.html

Q: Should I run one large cluster, and share it amongst many users or many smaller clusters?

Amazon EMR provides a unique capability for you to use both methods. On the one hand one large cluster may be more efficient for processing regular batch workloads. On the other hand, if you require ad-hoc querying or workloads that vary with time, you may choose to create several separate cluster tuned to the specific task sharing data sources stored in Amazon S3.

Q: Can I access a script or jar resource which is on my local file system?

No. You must upload the script or jar to Amazon S3 or to the cluster's master node before it can be referenced. For uploading to Amazon S3 you can use tools including s3cmd, jets3t or S3Organizer.

Q: Can I run a persistent cluster executing multiple Pig queries?

Yes. You run a cluster in a manual termination mode so it will not terminate between Pig steps. To reduce the risk of data loss we recommend periodically persisting all important data in Amazon S3. It is good practice to regularly transfer your work to a new cluster to test you process for recovering from master node failure.

Q: Does Pig support access from JDBC?

No. Pig does not support access through JDBC.

Using HBase

Q: What is Apache HBase?

HBase is an open source, non-relational, distributed database modeled after Google's BigTable. It was developed as part of Apache Software Foundation's Hadoop project and runs on top of Hadoop Distributed File System(HDFS) to provide BigTable-like capabilities for Hadoop. HBase provides you a fault-tolerant, efficient way of storing large quantities of sparse data using column-based compression and storage. In addition, HBase provides fast lookup of data because data is stored in-memory instead of on disk. HBase is optimized for sequential write operations, and it is highly efficient for batch inserts, updates, and deletes. HBase works seamlessly with Hadoop, sharing its file system and serving as a direct input and output to Hadoop jobs. HBase also integrates with Apache Hive, enabling SQL-like queries over HBase tables, joins with Hive-based tables, and support for Java Database Connectivity (JDBC). You can learn more about Apache HBase here.

Q: Are there new features in HBase specific to Amazon EMR?

With Amazon EMR you can back up HBase to Amazon S3 (full or incremental, manual or automated) and you can restore from a previously created backup. Learn more about HBase and EMR.

Q: Which versions of HBase are supported on Amazon EMR?

Amazon EMR supports HBase 0.94.7 and HBase 0.92.0. To use HBase 0.94.7 you must specify AMI version 3.0.0. If you are using the CLI you must use version 2013-10-07 or later.

Kinesis connector

Q: What does EMR Connector to Kinesis enable?

The connector enables EMR to directly read and query data from Kinesis streams. You can now perform batch processing of Kinesis streams using existing Hadoop ecosystem tools such as Hive, Pig, MapReduce, Hadoop Streaming, and Cascading.

Q: What does the EMR connector to Kinesis enable that I couldn’t have done before?

Reading and processing data from a Kinesis stream would require you to write, deploy and maintain independent stream processing applications. These take time and effort. However, with this connector, you can start reading and analyzing a Kinesis stream by writing a simple Hive or Pig script. This means you can analyze Kinesis streams using SQL! Of course, other Hadoop ecosystem tools could be used as well. You don’t need to developed or maintain a new set of processing applications.

Q: Who will find this functionality useful?

The following types of users will find this integration useful:

- Hadoop users who are interested in utilizing the extensive set of Hadoop ecosystem tools to analyze Kinesis streams.
- Kinesis users who are looking for an easy way to get up and running with stream processing and ETL of Kinesis data.
- Business analysts and IT professionals who would like to perform ad-hoc analysis of data in Kinesis streams using familiar tools like SQL (via Hive) or scripting languages like Pig.

Q: What are some use cases for this integration?

The following are representative use cases are enabled by this integration:

- Streaming Log Analysis: You can analyze streaming web logs to generate a list of top 10 error type every few minutes by region, browser, and access domains.
- Complex Data Processing Workflows: You can join Kinesis stream with data stored in S3, Dynamo DB tables, and HDFS. You can write queries that join clickstream data from Kinesis with advertising campaign information stored in a DynamoDB table to identify the most effective categories of ads that are displayed on particular websites.
- Ad-hoc Queries: You can periodically load data from Kinesis into HDFS and make it available as a local Impala table for fast, interactive, analytic queries.

Q: What EMR AMI version do I need to be able to use the connector?

You need to use EMR’s AMI version 3.0.4 and later.

Q: Is this connector a stand-alone tool?

No, it is a built in component of the Amazon distribution of Hadoop and is present on EMR AMI versions 3.0.4 and later. Customer simply needs to spin up a cluster with AMI version 3.0.4 or later to start using this feature.

Q: What data format is required to allow EMR to read from a Kinesis stream?

The EMR Kinesis integration is not data format-specific. You can read data in any format. Individual Kinesis records are presented to Hadoop as standard records that can be read using any Hadoop MapReduce framework. Individual frameworks like Hive, Pig and Cascading have built in components that help with serialization and deserialization, making it easy for developers to query data from many formats without having to implement custom code. For example, in Hive users can read data from JSON files, XML files and SEQ files by specifying the appropriate Hive SerDe when they define a table. Pig has a similar component called Loadfunc/Evalfunc and Cascading has a similar component called a Tap. Hadoop users can leverage the extensive ecosystem of Hadoop adapters without having to write format-specific code. You can also implement custom deserialization formats to read domain specific data in any of these tools.

Q: How do I analyze a Kinesis stream using Hive in EMR?

Create a table that references a Kinesis stream. You can then analyze the table like any other table in Hive. Please see our tutorials for page more details.

Q: Using Hive, how do I create queries that combine Kinesis stream data with other data source?

First create a table that references a Kinesis stream. Once a Hive table has been created, you can join it with tables mapping to other data sources such as Amazon S3, Amazon Dynamo DB, and HDFS. This effectively results in joining data from Kinesis stream to other data sources.

Q: Is this integration only available for Hive?

No, you can use Hive, Pig, MapReduce, Hadoop Streaming, and Cascading.

Q: How do I setup scheduled jobs to run on a Kinesis stream?

The EMR Kinesis input connector provides features that help you configure and manage scheduled periodic jobs in traditional scheduling engines such as Cron. For example, you can develop a Hive script that runs every N minutes. In the configuration parameters for a job, you can specify a **Logical Name** for the job. The Logical Name is a label that will inform the EMR Kinesis input connector that individual instances of the job are members of the same periodic schedule. The Logical Name allows the process to take advantage of iterations, which are explained next.

Since MapReduce is a batch processing framework, to analyze a Kinesis stream using EMR, the continuous stream is divided in to batches. Each batch is called an **Iteration**. Each Iteration is assigned a number, starting with 0. Each Iteration's boundaries are defined by a start sequence number and end sequence number. Iterations are then processed sequentially by EMR.

In the event of an attempt's failure, the EMR Kinesis input connector will re-try the iteration within the Logical Name from the known start sequence number of the iteration. This functionality ensures that successive attempts on the same iteration will have precisely the same input records from the Kinesis stream as the previous attempts. This guarantees idempotent (consistent) processing of a Kinesis stream.

You can specify Logical Names and Iterations as runtime parameters in your respective Hadoop tools. For example, in the tutorial section "Running queries with checkpoints", the code sample shows a scheduled Hive query that designates a Logical Name for the query and increments the iteration with each successive run of the job.

Additionally, a sample cron scheduling script is provided in the tutorials.

Q: Where is the metadata for Logical Names and Iterations stored?

The metadata that allows the EMR Kinesis input connector to work in scheduled periodic workflows is stored in Amazon DynamoDB. You must provision an Amazon Dynamo DB table and specify it as an input parameter to the Hadoop Job. It is important that you configure appropriate IOPS for the table to enable this integration. Please refer to the getting started tutorial for more information on setting up your Amazon Dynamo DB table.

Q: What happens when an iteration processing fails?

Iterations identifiers are user-provided values that map to specific boundary (start and end sequence numbers) in a Kinesis stream. Data corresponding to these boundaries is loaded in the Map phase of the MapReduce job. This phase is managed by the framework and will be automatically re-run (three times by default) in case of job failure. If all the retries fail, you would still have options to retry the processing starting from last successful data boundary or past data boundaries. This behavior is controlled by providing kinesis.checkpoint.iteration.no parameter during processing. Please refer to the getting started tutorial for more information on how this value is configured for different tools in the Hadoop ecosystem.

Q: Can I run multiple queries on the same iteration?

Yes, you can specify a previously run iteration by setting the kinesis.checkpoint.iteration.no parameter in successive processing. The implementation ensures that successive runs on the same iteration will have precisely the same input records from the Kinesis stream as the previous runs.

Q: What happens if records in an Iteration expire from the Kinesis stream?

In the event that the beginning sequence number and/or end sequence number of an iteration belong to records that have expired from the Kinesis steam, the Hadoop job will fail. You would need to use a different Logical Name to process data from the beginning of the Kinesis stream.

Q: Can I push data from EMR into Kinesis stream?

No. The EMR Kinesis connector currently does not support writing data back into a Kinesis stream.

Q: Does the EMR Hadoop input connector for Kinesis enable continuous stream processing?

The Hadoop MapReduce framework is a batch processing system. As such, it does not support continuous queries. However there is an emerging set of Hadoop ecosystem frameworks like Twitter Storm and Spark Streaming that enable to developers build applications for continuous stream processing. A Storm connector for Kinesis is available at on GitHub here and you can find a tutorial explaining how to setup Spark Streaming on EMR and run continuous queries here.

Additionally, developers can utilize the Kinesis client library to develop real-time stream processing applications. You can find more information on developing custom Kinesis applications in the Kinesis documentation here.

Q: Can I specify access credential to read a Kinesis stream that is managed in another AWS account?

Yes. You can read streams from another AWS account by specifying the appropriate access credentials of the account that owns the Kinesis stream. By default, the Kinesis connector utilizes the user-supplied access credentials that are specified when the cluster is created. You can override these credentials to access streams from other AWS Accounts by setting the kinesis.accessKey and kinesis.secretKey parameters. The following examples show how to set the kinesis.accessKey and kinesis.secretKey parameters in Hive and Pig.

Code sample for Hive:

```
...
STORED BY
'com.amazon.emr.kinesis.hive.KinesisStorageHandler'
TBLPROPERTIES(
"kinesis.accessKey"="AwsAccessKey",
"kinesis.secretKey"="AwsSecretKey",
);
```

Code sample for Pig:

```
…
raw_logs = LOAD 'AccessLogStream' USING com.amazon.emr.kinesis.pig.KinesisStreamLoader('kinesis.accessKey=AwsAccessKey',
'kinesis.secretKey=AwsSecretKey'
) AS (line:chararray);
```

Q: Can I run multiple parallel queries on a single Kinesis Stream? Is there a performance impact?

Yes, a customer can run multiple parallel queries on the same stream by using separate logical names for each query. However, reading from a shard within a Kinesis stream is subjected to a rate limit of of 2MB/sec. Thus, if there are N parallel queries running on the same stream, each one would get roughly (2/N)

MB/sec egress rate per shard on the stream. This may slow down the processing and in some cases fail the queries as well.

Q: Can I join and analyze multiple Kinesis streams in EMR?

Yes, for example in Hive, you can create two tables mapping to two different Kinesis streams and create joins between the tables.

Q: Does the EMR Kinesis connector handle Kinesis scaling events, such as merge and split events?

Yes. The implementation handles split and merge events. The Kinesis connector ties individual Kinesis shards (the logical unit of scale within a Kinesis stream) to Hadoop MapReduce map tasks. Each unique shard that exists within a stream in the logical period of an Iteration will result in exactly one map task. In the event of a shard split or merge event, Kinesis will provision new unique shard Ids. As a result, the MapReduce framework will provision more map tasks to read from Kinesis. All of this is transparent to the user.

Q: What happens if there are periods of “silence” in my stream?

The implementation allows you to configure a parameter called kinesis.nodata.timeout. For example, consider a scenario where kinesis.nodata.timeout is set to 2 minutes and you want to run a Hive query every 10 minutes. Additionally, consider some data has been written to the stream since the last iteration (10 minutes ago). However, currently no new records are arriving, i.e. there is a silence in the stream. In this case, when the current iteration of the query launches, the Kinesis connector would find that no new records are arriving. The connector will keep polling the stream for 2 minutes and if no records arrive for that interval then it will stop and process only those records that were already read in the current batch of stream. However, if new records start arriving before kinesis.nodata.timeout interval is up, then the connector will wait for an additional interval corresponding to a parameter called kinesis.iteration.timeout. Please look at the tutorials to see how to define these parameters.

Q: How do I debug a query that continues to fail in each iteration?

In the event of a processing failure, you can utilize the same tools they currently do when debugging Hadoop Jobs. Including the Amazon EMR web console, which helps identify and access error logs. More details on debugging an EMR job can be found here.

Q: What happens if I specify a DynamoDB table that I don't have access to?

The job would fail and the exception would show up in error logs for the job.

Q: What happens if job doesn't fail but checkpointing to DynamoDB fails?

The job would fail and the exception would show up in error logs for the job.

Q: How do I maximize the read throughput from Kinesis stream to EMR?

Throughput from Kinesis stream increases with instance size used and record size in the Kinesis stream. We recommend that you use m1.xlarge and above for both master and core nodes for this feature.

Service Level Agreement

Q: What does the Amazon EMR SLA guarantee?

Our Amazon EMR SLA guarantees a Monthly Uptime Percentage of at least 99.9% for Amazon EMR.

Q: How do I know if I qualify for a SLA Service Credit?

You are eligible for a SLA credit for Amazon EMR under the Amazon EMR SLA if more than one Availability Zone in which you are running a task, within the same region has a Monthly Uptime Percentage of less than 99.9% during any monthly billing cycle.

For full details on all of the terms and conditions of the SLA, as well as details on how to submit a claim, please see the Amazon EMR SLA details page.

Amazon Kinesis Data Streams FAQs

General

Q: What is Amazon Kinesis Data Streams?

Amazon Kinesis Data Streams enables you to build custom applications that process or analyze streaming data for specialized needs. You can continuously add various types of data such as clickstreams, application logs, and social media to an Amazon Kinesis data stream from hundreds of thousands of sources. Within seconds, the data will be available for your Amazon Kinesis Applications to read and process from the stream.

Q: What does Amazon Kinesis Data Streams manage on my behalf?

Amazon Kinesis Data Streams manages the infrastructure, storage, networking, and configuration needed to stream your data at the level of your data throughput. You do not have to worry about provisioning, deployment, ongoing-maintenance of hardware, software, or other services for your data streams. In addition, Amazon Kinesis Data Streams synchronously replicates data across three availability zones, providing high availability and data durability.

Q: What can I do with Amazon Kinesis Data Streams?

Amazon Kinesis Data Streams is useful for rapidly moving data off data producers and then continuously processing the data, be it to transform the data before emitting to a data store, run real-time metrics and analytics, or derive more complex data streams for further processing. The following are typical scenarios for using Amazon Kinesis Data Streams:

- Accelerated log and data feed intake: Instead of waiting to batch up the data, you can have your data producers push data to an Amazon Kinesis data stream as soon as the data is produced, preventing data loss in case of data producer failures. For example, system and application logs can be continuously added to a data stream and be available for processing within seconds.
- Real-time metrics and reporting: You can extract metrics and generate reports from Amazon Kinesis data stream data in real-time. For example, your Amazon Kinesis Application can work on metrics and reporting for system and application logs as the data is streaming in, rather than wait to receive data batches.
- Real-time data analytics: With Amazon Kinesis Data Streams, you can run real-time streaming data analytics. For example, you can add clickstreams to your Amazon Kinesis data stream and have your Amazon Kinesis Application run analytics in real-time, enabling you to gain insights out of your data at a scale of minutes instead of hours or days.
- Complex stream processing: You can create Directed Acyclic Graphs (DAGs) of Amazon Kinesis Applications and data streams. In this scenario, one or more Amazon Kinesis Applications can add data to another Amazon Kinesis data stream for further processing, enabling successive stages of stream processing.

Q: How do I use Amazon Kinesis Data Streams?

After you sign up for Amazon Web Services, you can start using Amazon Kinesis Data Streams by:

- Creating an Amazon Kinesis data stream through either AWS Management Console or CreateStream operation.
- Configuring your data producers to continuously add data to your data stream.
- Building your Amazon Kinesis Applications to read and process data from your data stream, using either Amazon Kinesis API or Amazon Kinesis Client Library (KCL).

Q: What are the limits of Amazon Kinesis Data Streams?

The throughput of an Amazon Kinesis data stream is designed to scale without limits via increasing the number of shards within a data stream. However, there are certain limits you should keep in mind while using Amazon Kinesis Data Streams:

- By default, Records of a stream are accessible for up to 24 hours from the time they are added to the stream. You can raise this limit to up to 7 days by enabling extended data retention.
- The maximum size of a data blob (the data payload before Base64-encoding) within one record is 1 megabyte (MB).
- Each shard can support up to 1000 PUT records per second.

For more information about other API level limits, see Amazon Kinesis Data Streams Limits.

Q: How does Amazon Kinesis Data Streams differ from Amazon SQS?

Amazon Kinesis Data Streams enables real-time processing of streaming big data. It provides ordering of records, as well as the ability to read and/or replay records in the same order to multiple Amazon Kinesis Applications. The Amazon Kinesis Client Library (KCL) delivers all records for a given partition key to the same record processor, making it easier to build multiple applications reading from the same Amazon Kinesis data stream (for example, to perform counting, aggregation, and filtering).

Amazon Simple Queue Service (Amazon SQS) offers a reliable, highly scalable hosted queue for storing messages as they travel between computers. Amazon SQS lets you easily move data between distributed application components and helps you build applications in which messages are processed independently (with message-level ack/fail semantics), such as automated workflows.

Q: When should I use Amazon Kinesis Data Streams, and when should I use Amazon SQS?

We recommend Amazon Kinesis Data Streams for use cases with requirements that are similar to the following:

- Routing related records to the same record processor (as in streaming MapReduce). For example, counting and aggregation are simpler when all records for a given key are routed to the same record processor.
- Ordering of records. For example, you want to transfer log data from the application host to the processing/archival host while maintaining the order of log statements.
- Ability for multiple applications to consume the same stream concurrently. For example, you have one application that updates a real-time dashboard and another that archives data to Amazon Redshift. You want both applications to consume data from the same stream concurrently and independently.
- Ability to consume records in the same order a few hours later. For example, you have a billing application and an audit application that runs a few hours behind the billing application. Because Amazon Kinesis Data Streams stores data for up to 7 days, you can run the audit application up to 7 days behind the billing application.

We recommend Amazon SQS for use cases with requirements that are similar to the following:

- Messaging semantics (such as message-level ack/fail) and visibility timeout. For example, you have a queue of work items and want to track the successful completion of each item independently. Amazon SQS tracks the ack/fail, so the application does not have to maintain a persistent checkpoint/cursor. Amazon SQS will delete acked messages and redeliver failed messages after a configured visibility timeout.
- Individual message delay. For example, you have a job queue and need to schedule individual jobs with a delay. With Amazon SQS, you can configure individual messages to have a delay of up to 15 minutes.
- Dynamically increasing concurrency/throughput at read time. For example, you have a work queue and want to add more readers until the backlog is cleared. With Amazon Kinesis Data Streams, you can scale up to a sufficient number of shards (note, however, that you'll need to provision enough shards ahead of time).

- Leveraging Amazon SQS's ability to scale transparently. For example, you buffer requests and the load changes as a result of occasional load spikes or the natural growth of your business. Because each buffered request can be processed independently, Amazon SQS can scale transparently to handle the load without any provisioning instructions from you.

Key concepts

Q: What is a shard?

Shard is the base throughput unit of an Amazon Kinesis data stream. One shard provides a capacity of 1MB/sec data input and 2MB/sec data output. One shard can support up to 1000 PUT records per second. You will specify the number of shards needed when you create a data stream. For example, you can create a data stream with two shards. This data stream has a throughput of 2MB/sec data input and 4MB/sec data output, and allows up to 2000 PUT records per second. You can monitor shard-level metrics in Amazon Kinesis Data Streams and add or remove shards from your data stream dynamically as your data throughput changes by resharding the data stream.

Q: What is a record?

A record is the unit of data stored in an Amazon Kinesis data stream. A record is composed of a sequence number, partition key, and data blob. Data blob is the data of interest your data producer adds to a data stream. The maximum size of a data blob (the data payload before Base64-encoding) is 1 megabyte (MB).

Q: What is a partition key?

Partition key is used to segregate and route records to different shards of a data stream. A partition key is specified by your data producer while adding data to an Amazon Kinesis data stream. For example, assuming you have a data stream with two shards (shard 1 and shard 2). You can configure your data producer to use two partition keys (key A and key B) so that all records with key A are added to shard 1 and all records with key B are added to shard 2.

Q: What is a sequence number?

A sequence number is a unique identifier for each record. Sequence number is assigned by Amazon Kinesis when a data producer calls PutRecord or PutRecords operation to add data to an Amazon Kinesis data stream. Sequence numbers for the same partition key generally increase over time; the longer the time period between PutRecord or PutRecords requests, the larger the sequence numbers become.

Creating data streams

Q: How do I create an Amazon Kinesis data stream?

After you sign up for Amazon Web Services, you can create an Amazon Kinesis data stream through either Amazon Kinesis Management Console or CreateStream operation.

Q: How do I decide the throughput of my Amazon Kinesis data stream?

The throughput of an Amazon Kinesis data stream is determined by the number of shards within the data stream. Follow the steps below to estimate the initial number of shards your data stream needs. Note that you can dynamically adjust the number of shards within your data stream via resharding.

1. Estimate the average size of the record written to the data stream in kilobytes (KB), rounded up to the nearest 1 KB. (average_data_size_in_KB)
2. Estimate the number of records written to the data stream per second. (number_of_records_per_second)
3. Decide the number of Amazon Kinesis Applications consuming data concurrently and independently from the data stream. (number_of_consumers)
4. Calculate the incoming write bandwidth in KB (incoming_write_bandwidth_in_KB), which is equal to the average_data_size_in_KB multiplied by the number_of_records_per_seconds.
5. Calculate the outgoing read bandwidth in KB (outgoing_read_bandwidth_in_KB), which is equal to the incoming_write_bandwidth_in_KB multiplied by the number_of_consumers.

You can then calculate the initial number of shards (number_of_shards) your data stream needs using the following formula:

number_of_shards = max (incoming_write_bandwidth_in_KB/1000, outgoing_read_bandwidth_in_KB/2000)

Q: What is the minimum throughput I can request for my Amazon Kinesis data stream?

The throughput of an Amazon Kinesis data stream scales by unit of shard. One single shard is the smallest throughput of a data stream, which provides 1MB/sec data input and 2MB/sec data output.

Q: What is the maximum throughput I can request for my Amazon Kinesis data stream?

The throughput of an Amazon Kinesis data stream is designed to scale without limits. By default, each account can provision 10 shards per region. You can use the Amazon Kinesis Data Streams Limits form to request more than 10 shards within a single region.

Q: How can record size affect the throughput of my Amazon Kinesis data stream?

A shard provides 1MB/sec data input rate and supports up to 1000 PUT records per sec. Therefore, if the record size is less than 1KB, the actual data input rate of a shard will be less than 1MB/sec, limited by the maximum number of PUT records per second.

Adding data to Kinesis data streams

Q: How do I add data to my Amazon Kinesis data stream?

You can add data to an Amazon Kinesis data stream via PutRecord and PutRecords operations, Amazon Kinesis Producer Library (KPL), or Amazon Kinesis Agent.

Q: What is the difference between PutRecord and PutRecords?

PutRecord operation allows a single data record within an API call and PutRecords operation allows multiple data records within an API call. For more information about PutRecord and PutRecords operations, see PutRecord and PutRecords.

Q: What is Amazon Kinesis Producer Library (KPL)?

Amazon Kinesis Producer Library (KPL) is an easy to use and highly configurable library that helps you put data into an Amazon Kinesis data stream. KPL presents a simple, asynchronous, and reliable interface that enables you to quickly achieve high producer throughput with minimal client resources.

Q: What programming languages or platforms can I use to access Amazon Kinesis API?

Amazon Kinesis API is available in Amazon Web Services SDKs. For a list of programming languages or platforms for Amazon Web Services SDKs, see Tools for Amazon Web Services.

Q: What programming language is Amazon Kinesis Producer Library (KPL) available in?

Amazon Kinesis Producer Library (KPL)'s core is built with C++ module and can be compiled to work on any platform with a recent C++ compiler. The library is currently available in a Java interface. We are looking to add support for other programming languages.

Q: What is Amazon Kinesis Agent?

Amazon Kinesis Agent is a pre-built Java application that offers an easy way to collect and send data to your Amazon Kinesis data stream. You can install the agent on Linux-based server environments such as web servers, log servers, and database servers. The agent monitors certain files and continuously sends data to your data stream. For more information, see Writing with Agents.

Q: What platforms do Amazon Kinesis Agent support?

Amazon Kinesis Agent currently supports Amazon Linux or Red Hat Enterprise Linux.

Q: Where do I get Amazon Kinesis Agent?

You can download and install Amazon Kinesis Agent using the following command and link:

On Amazon Linux: sudo yum install –y aws-kinesis-agent

On Red Hat Enterprise Linux: sudo yum install –y https://s3.amazonaws.com/streaming-data-agent/aws-kinesis-agent-latest.amzn1.noarch.rpm

From GitHub: awlabs/amazon-kinesis-agent

Q: How do I use Amazon Kinesis Agent?

After installing Amazon Kinesis Agent on your servers, you configure it to monitor certain files on the disk and then continuously send new data to your Amazon Kinesis data stream. For more information, see Writing with Agents.

Q: What happens if the capacity limits of an Amazon Kinesis data stream are exceeded while the data producer adds data to the data stream?

The capacity limits of an Amazon Kinesis data stream are defined by the number of shards within the data stream. The limits can be exceeded by either data throughput or the number of PUT records. While the capacity limits are exceeded, the put data call will be rejected with a ProvisionedThroughputExceeded exception. If this is due to a temporary rise of the data stream's input data rate, retry by the data producer will eventually lead to completion of the requests. If this is due to a sustained rise of the data stream's input data rate, you should increase the number of shards within your data stream to provide enough capacity for the put data calls to consistently succeed. In both cases, Amazon CloudWatch metrics allow you to learn about the change of the data stream's input data rate and the occurrence of ProvisionedThroughputExceeded exceptions.

Q: What data is counted against the data throughput of an Amazon Kinesis data stream during a PutRecord or PutRecords call?

Your data blob, partition key, and data stream name are required parameters of a PutRecord or PutRecords call. The size of your data blob (before Base64 encoding) and partition key will be counted against the data throughput of your Amazon Kinesis data stream, which is determined by the number of shards within the data stream.

Enhanced fan-out

Q: What is enhanced fan-out?

Enhanced fan-out is an optional feature for Kinesis Data Streams consumers that provides logical 2 MB/sec throughput pipes between consumers and shards. This allows customers to scale the number of consumers reading from a data stream in parallel, while maintaining high performance.

Q: How do consumers use enhanced fan-out?

Consumers must first register themselves with the Kinesis Data Streams service. By default, consumer registration activates enhanced fan-out. If you are using the KCL, KCL version 2.x takes care of registering your consumers automatically, and uses the name of the KCL application as the consumer name. Once registered, all registered consumers will have their own logical enhanced fan-out throughput pipes provisioned for them. Then, consumers use the HTTP/2 SubscribeToShard API to retrieve data inside of these throughput pipes. The HTTP/1 GetRecords API does not currently support enhanced fan-out, so you will need to upgrade to KCL 2.x, or alternatively register your consumer and have the consumer call the SubscribeToShard API.

Q: How is enhanced fan-out utilized by a consumer?

Consumers utilize enhanced fan-out by retrieving data with the SubscribeToShard API. The name of the registered consumer is used within the SubscribeToShard API, which leads to utilization of the enhanced fan-out benefit provided to the registered consumer.

Q: When should I use enhanced fan-out?
You should use enhanced fan-out if you have, or expect to have, multiple consumers retrieving data from a stream in parallel, or if you have at least one consumer that requires

the use of the SubscribeToShard API to provide sub-200ms data delivery speeds between producers and consumers.

Q: Can I have consumers using enhanced fan-out, and others not?

Yes, you can have multiple consumers using enhanced fan-out and others not using enhanced fan-out at the same time. The use of enhanced fan-out does not impact the limits of shards for traditional GetRecords usage.

Q: Is there a limit on the number of consumers using enhanced fan-out on a given stream?

There is a default limit of 20 consumers using enhanced fan-out per data stream. If you need more than 20, please submit a limit increase request though AWS support. Keep in mind that you can have more than 20 total consumers reading from a stream by having 20 consumers using enhanced fan-out and other consumers not using enhanced fan-out at the same time.

Q: How do consumers register to use enhanced fan-out and the HTTP/2 SubscribeToShard API?

We recommend using KCL 2.x, which will automatically register your consumer and use both enhanced fan-out and the HTTP/2 SubscribeToShard API. Otherwise, you can manually register a consumer using the RegisterStreamConsumer API and then you can use the SubscribeToShard API with the name of the consumer you registered.

Q: Is there a cost associated with the use of enhanced fan-out?

Yes, there is an on-demand hourly cost for every combination of shard in a stream and consumer (a consumer-shard hour) registered to use enhanced fan-out, in addition to a data retrieval cost for every GB retrieved. See the Kinesis Data Streams pricing page for more details.

Q: How is a consumer-shard hour calculated?

A consumer-shard hour is calculated by multiplying the number of registered stream consumers with the number of shards in the stream. For example, if a consumer-shard hour costs $0.015, for a 10 shard data stream, this consumer using enhanced fan-out would be able to read from 10 shards, and thus incur a consumer-shard hour charge of $0.15 per hour (1 consumer x 10 shards x $0.015 per consumers-shard hour). If there were two consumers registered for enhanced fan-out simultaneously, the total consumer-shard hour charge would be $0.30 per hour (2 consumers x 10 shards x $0.015).

Q: Does consumer-shard hour billing for enhanced fan-out automatically prorate if I terminate or start a consumer within the hour?

Yes, you will only pay for the prorated portion of the hour the consumer was registered to use enhanced fan-out.

Q: How does billing for enhanced fan-out data retrievals work?

You pay a low per GB rate that is metered per byte of data retrieved by consumers using enhanced fan-out. There is no payload roundup or delivery minimum.

Q: Do I need to change my producers or my data stream to use enhanced fan-out?

No, enhance fan-out can be activated without impacting data producers or data streams.

Reading and processing data from Kinesis data streams

Q: What is an Amazon Kinesis Application?

An Amazon Kinesis Application is a data consumer that reads and processes data from an Amazon Kinesis data stream. You can build your applications using either Amazon Kinesis Data Analytics, Amazon Kinesis API or Amazon Kinesis Client Library (KCL).

Q: What is Amazon Kinesis Client Library (KCL)?

Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js | .NET is a pre-built library that helps you easily build Amazon Kinesis Applications for reading and processing data from an Amazon Kinesis data stream.

KCL handles complex issues such as adapting to changes in data stream volume, load-balancing streaming data, coordinating distributed services, and processing data with fault-tolerance. KCL enables you to focus on business logic while building applications. KCL 2.x supports both the HTTP/1 GetRecords and HTTP/2 SubscribeToShard APIs with enhanced fan-out for retrieving data from a stream. KCL 1.x does not support the SubscribeToShard API or enhanced fan-out.

Q: How do I upgrade from KCL 1.x to 2.x to use SubscribeToShard and enhanced fan-out?

Visit the Kinesis Data Streams user documentation to learn how to upgrade from KCL 1.x to KCL 2.x.

Q: What is the SubscribeToShard API?

The SubscribeToShard API is a high performance streaming API that pushes data from shards to consumers over a persistent connection without a request cycle from the client. The SubscribeToShard API uses the HTTP/2 protocol to deliver data to registered consumers whenever new data arrives on the shard, typically within 70ms, offering ~65% faster delivery compared to the GetRecords API.. The consumers will enjoy fast delivery even when multiple registered consumers are reading from the same shard.

Q: Can I use SubscribeToShard without using enhanced fan-out?

No, SubscribeToShard requires the use of enhanced fan-out, which means you also need to register your consumer with the Kinesis Data Streams service before you can use SubscribeToShard.

Q: How long does the SubscribeToShard persistent connection last?

The persistent connection can last up to 5 minutes.

Q: Does the Kinesis Client Library (KCL) support SubscribeToShard?

Yes, version 2.x of the KCL uses SubscribeToShard and enhanced fan-out to retrieve data with high performance from a Kinesis data stream.

Q: Is there a cost associated with using SubscribeToShard?

No, there is no additional cost associated with SubscribeToShard, but you must use SubscribeToShard with enhanced fan-out which does have an additional hourly cost for each consumer-shard combination and per GB of data delivered by enhanced fan-out.

Q: Do I need to use enhanced fan-out if I want to use SubscribeToShard?

Yes, to use SubscribeToShard you need to register your consumers, and registration activates enhanced fan-out. By default, your consumer will utilize enhanced fan-out automatically when data is retrieved via SubscribeToShard.

Q: What is Amazon Kinesis Connector Library?

Amazon Kinesis Connector Library is a pre-built library that helps you easily integrate Amazon Kinesis Data Streams with other AWS services and third-party tools. Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js | .NET is required for using Amazon Kinesis Connector Library. The current version of this library provides connectors to Amazon DynamoDB, Amazon Redshift, Amazon S3, and Elasticsearch. The library also includes sample connectors of each type, plus Apache Ant build files for running the samples.

Q: What is Amazon Kinesis Storm Spout?

Amazon Kinesis Storm Spout is a pre-built library that helps you easily integrate Amazon Kinesis Data Streams with Apache Storm. The current version of Amazon Kinesis Storm Spout fetches data from Amazon Kinesis data stream and emits it as tuples. You will add the spout to your Storm topology to leverage Amazon Kinesis Data Streams as a reliable, scalable, stream capture, storage, and replay service.

Q: What programming language are Amazon Kinesis Client Library (KCL), Amazon Kinesis Connector Library, and Amazon Kinesis Storm Spout available in?

Amazon Kinesis Client Library (KCL) is currently available in Java, Python, Ruby, Node.js, and .NET. Amazon Kinesis Connector Library and Amazon Kinesis Storm Spout are currently available in Java. We are looking to add support for other programming languages.

Q: Do I have to use Amazon Kinesis Client Library (KCL) for my Amazon Kinesis Application?

No, you can also use Amazon Kinesis API to build your Amazon Kinesis Application. However, we recommend using Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js | .NET if applicable because it performs heavy-lifting tasks associated with distributed stream processing, making it more productive to develop applications.

Q: How does Amazon Kinesis Client Library (KCL) interact with an Amazon Kinesis Application?

Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js | .NET acts as an intermediary between Amazon Kinesis Data Streams and your Amazon Kinesis Application. KCL uses the IRecordProcessor interface to communicate with your application. Your application implements this interface, and KCL calls into your application code using the methods in this interface.

For more information about building application with KCL, see Developing Consumer Applications for Amazon Kinesis Using the Amazon Kinesis Client Library.

Q: What is a worker and a record processor generated by Amazon Kinesis Client Library (KCL)?

An Amazon Kinesis Application can have multiple application instances and a worker is the processing unit that maps to each application instance. A record processor is the processing unit that processes data from a shard of an Amazon Kinesis data stream. One worker maps to one or more record processors. One record processor maps to one shard and processes records from that shard.

At startup, an application calls into Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js | .NET to instantiate a worker. This call provides KCL with configuration information for the application, such as the data stream name and AWS credentials. This call also passes a reference to an IRecordProcessorFactory implementation. KCL uses this factory to create new record processors as needed to process data from the data stream. KCL communicates with these record processors using the IRecordProcessor interface.

Q: How does Amazon Kinesis Client Library (KCL) keep tracking data records being processed by an Amazon Kinesis Application?

Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js | .NET automatically creates an Amazon DynamoDB table for each Amazon Kinesis Application to track and maintain state information such as resharding events and sequence number checkpoints. The DynamoDB table shares the same name with the application so that you need to make sure your application name doesn't conflict with any existing DynamoDB tables under the same account within the same region.

All workers associated with the same application name are assumed to be working together on the same Amazon Kinesis data stream. If you run an additional instance of the same application code, but with a different application name, KCL treats the second instance as an entirely separate application also operating on the same data stream.

Please note that your account will be charged for the costs associated with the Amazon DynamoDB table in addition to the costs associated with Amazon Kinesis Data Streams.

For more information about how KCL tracks application state, see Tracking Amazon Kinesis Application state.

Q: How can I automatically scale up the processing capacity of my Amazon Kinesis Application using Amazon Kinesis Client Library (KCL)?

You can create multiple instances of your Amazon Kinesis Application and have these application instances run across a set of Amazon EC2 instances that are part of an Auto Scaling group. While the processing demand increases, an Amazon EC2 instance running your application instance will be automatically instantiated. Amazon Kinesis Client Library (KCL) for Java | Python | Ruby | Node.js | .NET will generate a worker for this new instance and automatically move record processors from overloaded existing instances to this new instance.

Q: Why does GetRecords call return empty result while there is data within my Amazon Kinesis data stream?

One possible reason is that there is no record at the position specified by the current shard iterator. This could happen even if you are using TRIM_HORIZON as shard iterator type. An Amazon Kinesis data stream represents a continuous stream of data. You should call GetRecords operation in a loop and the record will be returned when the shard iterator advances to the position where the record is stored.

Q: What is ApproximateArrivalTimestamp returned in GetRecords operation?

Each record includes a value called ApproximateArrivalTimestamp. It is set when the record is successfully received and stored by Amazon Kinesis. This timestamp has millisecond precision and there are no guarantees about the timestamp accuracy. For example, records in a shard or across a data stream might have timestamps that are out of order.

Q: What happens if the capacity limits of an Amazon Kinesis data stream are exceeded while Amazon Kinesis Application reads data from the data stream?

The capacity limits of an Amazon Kinesis data stream are defined by the number of shards within the data stream. The limits can be exceeded by either data throughput or the number of read data calls. While the capacity limits are exceeded, the read data call will be rejected with a ProvisionedThroughputExceeded exception. If this is due to a temporary rise of the data stream's output data rate, retry by the Amazon Kinesis Application will eventually lead to completions of the requests. If this is due to a sustained rise of the data stream's output data rate, you should increase the number of shards within your data stream to provide enough capacity for the read data calls to consistently succeed. In both cases, Amazon CloudWatch metrics allow you to learn about the change of the data stream's output data rate and the occurrence of ProvisionedThroughputExceeded exceptions.

Managing Kinesis data streams

Q: How do I change the throughput of my Amazon Kinesis data stream?

There are two ways to change the throughput of your data stream. You can use the UpdateShardCount API or the AWS Management Console to scale the number of shards in a data stream, or you can change the throughput of an

Amazon Kinesis data stream by adjusting the number of shards within the data stream (resharding).

Q: How long does it take to change the throughput of my Amazon Kinesis data stream using UpdateShardCount or the AWS Management Console?

Typical scaling requests should take a few minutes to complete. Larger scaling requests will take longer than smaller ones.

Q: What are the limitations of UpdateShardCount?

For information about limitations of UpdateShardCount, see the Amazon Kinesis Data Streams Service API Reference.

Q: Does Amazon Kinesis Data Streams remain available when I change the throughput of my Amazon Kinesis data stream using UpdateShardCount or via resharding?

Yes. You can continue adding data to and reading data from your Amazon Kinesis data stream while you use UpdateShardCount or reshard to change the throughput of the data stream.

Q: What is resharding?

Resharding is the process used to scale your data stream using a series of shard splits or merges. In a shard split, a single shard is divided into two shards, which increases the throughput of the data stream. In a shard merge, two shards are merged into a single shard, which decreases the throughput of the data stream. For more information, see Resharding a Data Stream in the Amazon Kinesis Data Streams developer guide.

Q: How often can I and how long does it take to change the throughput of my Amazon Kinesis data stream by resharding it?

A resharding operation such as shard split or shard merge takes a few seconds. You can only perform one resharding operation at a time. Therefore, for an Amazon Kinesis data stream with only one shard, it takes a few seconds to double the throughput by splitting one shard. For a data stream with 1000 shards, it takes 30K seconds (8.3 hours) to double the throughput by splitting

1000 shards. We recommend increasing the throughput of your data stream ahead of the time when extra throughput is needed.

Q: How do I change the data retention period of my Amazon Kinesis data stream?

Amazon Kinesis stores your data for up to 24 hours by default. You can raise data retention period to up to 7 days by enabling extended data retention.

For more information about changing data retention period, see Changing Data Retention Period.

Q: How do I monitor the operations and performance of my Amazon Kinesis data stream?

Amazon Kinesis Data Streams Management Console displays key operational and performance metrics such as throughput of data input and output of your Amazon Kinesis data streams. Amazon Kinesis Data Streams also integrates with Amazon CloudWatch so that you can collect, view, and analyze CloudWatch metrics for your data streams and shards within those data streams. For more information about Amazon Kinesis Data Streams metrics, see Monitoring Amazon Kinesis Data Streams with Amazon CloudWatch.

Please note that all stream-level metrics are free of charge. All enabled shard-level metrics are charged at Amazon CloudWatch Pricing.

Q: How do I manage and control access to my Amazon Kinesis data stream?

Amazon Kinesis Data Streams integrates with AWS Identity and Access Management (IAM), a service that enables you to securely control access to your AWS services and resources for your users. For example, you can create a policy that only allows a specific user or group to add data to your Amazon Kinesis data stream. For more information about access management and control of your data stream, see Controlling Access to Amazon Kinesis Data Streams Resources using IAM.

Q: How do I log API calls made to my Amazon Kinesis data stream for security analysis and operational troubleshooting?

Amazon Kinesis integrates with Amazon CloudTrail, a service that records AWS API calls for your account and delivers log files to you. For more information about API call logging and a list of supported Amazon Kinesis API operations, see Logging Amazon Kinesis API calls Using Amazon CloudTrail.

Q: How do I effectively manage my Amazon Kinesis data streams and the costs associated with these data streams?

Amazon Kinesis Data Streams allows you to tag your Amazon Kinesis data streams for easier resource and cost management. A tag is a user-defined label expressed as a key-value pair that helps organize AWS resources. For example, you can tag your data streams by cost centers so that you can categorize and track your Amazon Kinesis Data Streams costs based on cost centers. For more information about Amazon Kinesis Data Streams tagging, see Tagging Your Amazon Kinesis Data Streams.

Q: How can I describe how I'm utilizing my shard limit?

You can understand how you're utilizing your shard limit for an account using the DescribeLimits API. The DescribeLimits API will return the shard limit and the number of open shards in your account. If you need to raise your shard limit, please request a limit increase.

Security

Q: When I use Kinesis Data Streams, how secure is my data?

Kinesis is secure by default. Only the account and data stream owners have access to the Kinesis resources they create. Kinesis supports user authentication to control access to data. You can use AWS IAM policies to selectively grant permissions to users and groups of users. You can securely put and get your data from Kinesis through SSL endpoints using the HTTPS protocol. If you need extra security you can use server-side encryption with AWS KMS master keys to encrypt data stored in your data stream. AWS KMS allows you to use AWS generated KMS master keys for encryption, or if you prefer you can bring your own master key into AWS KMS. Lastly, you can use your own encryption libraries to encrypt data on the client-side before putting the data into Kinesis.

Q: Can I privately access Kinesis Data Streams APIs from my Amazon Virtual Private Cloud (VPC) without using public IPs?

Yes, you can privately access Kinesis Data Streams APIs from your Amazon Virtual Private Cloud (VPC) by creating VPC Endpoints. With VPC Endpoints, the routing between the VPC and Kinesis Data Streams is handled by the AWS network without the need for an Internet gateway, NAT gateway, or VPN connection. The latest generation of VPC Endpoints used by Kinesis Data Streams are powered by AWS PrivateLink, a technology that enables private connectivity between AWS services using Elastic Network Interfaces (ENI) with private IPs in your VPCs. To learn more about PrivateLink, visit the PrivateLink documentation.

Encryption

Q: Can I encrypt the data I put into a Kinesis data stream?

Yes, and there are two options for encrypting the data you put into a Kinesis data stream. You can use server-side encryption , which is a fully managed feature that automatically encrypts and decrypts data as you put and get it from a data stream. Or you can write encrypted data to a data stream by encrypting and decrypting on the client side.

Q: Why should I use server-side encryption instead of client-side encryption?

Customers often choose server-side encryption over client-side encryption for one of the following reasons:

- It hard to enforce client-side encryption
- They want a second layer of security on top of client-side encryption
- It is hard to implement client-side key management schemes

Q: What is server-side encryption?

Server-side encryption for Kinesis Data Streams automatically encrypts data using a user specified AWS KMS master key (CMK) before it is written to the data stream storage layer, and decrypts the data after it is retrieved from storage. Encryption makes writes impossible and the payload and the partition key unreadable unless the user writing or reading from the data stream has the permission to use the key selected for encryption on the data stream. As a result, server-side encryption can make it easier to meet internal security and compliance requirements governing your data.

With server-side encryption your client-side applications (producers and consumers) do not need to be aware of encryption, they do not need to manage CMKs or cryptographic operations, and your data is encrypted when it is at rest and in motion within the Kinesis Data Streams service. All CMKs used by the server-side encryption feature are provided by the AWS KMS. AWS KMS makes it easy to use an AWS-managed CMK for Kinesis(a “one-click” encryption method), your own AWS KMS generated CMK, or a CMK that you imported for encryption.

Q: Is there a server-side encryption getting started guide?

Yes, there is a getting started guide in the user documentation.

Q: Does server-side encryption interfere with how my applications interact with Kinesis Data Streams?

Possibly. This depends on the key you use for encryption and the permissions governing access to the key.

- If you use the AWS-managed CMK for Kinesis (key alias = aws/kinesis) your applications will not be impacted by enabling or disabling encryption with this key.
- If you use a different master key, like a custom AWS KMS master key or one you imported into the AWS KMS service, and if your producers and consumers of a data stream do not have permission to use the AWS KMS CMK used for encryption, then your PUT and GET requests will fail. Before you can use server-side encryption you must configure AWS KMS key policies to allow encryption and decryption of messages. For examples and more information about AWS KMS permissions, see AWS KMS API Permissions: Actions and Resources Reference in the AWS Key Management Service

Developer Guide or the permissions guidelines in the Kinesis Data Streams server-side encryption user documentation.

Q: Is there an additional cost associated with the use of server-side encryption?

Yes, however if you are using the AWS-managed CMK for Kinesis and are not exceeding the free tier KMS API usage costs, then your use of server-side encryption is free. The following describes the costs by resource:

Keys:

- The AWS-managed CMK for Kinesis (alias = “aws/kinesis”) is free.
- User generated KMS keys are subject to KMS key costs. Learn more.

KMS API Usage:

- API usage costs apply for every CMK, including custom ones. Kinesis Data Streams calls KMS approximately every 5 minutes when it is rotating the data key. In a 30-day month, the total cost of KMS API calls initiated by a Kinesis data stream should be less than a few dollars. Please note that this cost scales with the number of user credentials you use on your data producers and consumers because each user credential requires a unique API call to AWS KMS. When you use IAM role for authentication, each assume-role-call will result in unique user credentials and you might want to cache user credentials returned by the assume-role-call to save KMS costs.

Q: Which AWS regions offer server-side encryption for Kinesis Data Streams?

Kinesis Data Streams server-side encryption is available in the AWS GovCloud region and all public regions except the China (Beijing) region.

Q: How do I start, update, or remove server-side encryption from a data stream?

All of these operations can be completed using the AWS management console or using the AWS SDK. To learn more, see the Kinesis Data Streams server-side encryption getting started guide.

Q: What encryption algorithm is used for server-side encryption?

Kinesis Data Streams uses an AES-GCM 256 algorithm for encryption.

Q: If I encrypt a data stream that already has data written to it, either in plain text or ciphertext, will all of the data in the data stream be encrypted or decrypted if I update encryption?

No, only new data written into the data stream will be encrypted (or left decrypted) by the new application of encryption.

Q: What does server-side encryption for Kinesis Data Streams encrypt?

Server-side encryption encrypts the payload of the message along with the partition key, which is specified by the data stream producer applications.

Q: Is server-side encryption a shard specific feature or a stream specific feature?

Server-side encryption is a stream specific feature.

Q: Can I change the CMK that is used to encrypt a specific data stream?

Yes, using the AWS management console or the AWS SDK you can choose a new master key to apply to a specific data stream.

Q: Can you walk me through the encryption lifecycle of my data from the point in time when I send it to a Kinesis data stream with server-side encryption enabled, and when I retrieve it?

The following walks you through how Kinesis Data Streams uses AWS KMS CMKs to encrypt a message before it is stored in the PUT path, and to decrypt it after it is retrieved in the GET path. Kinesis and AWS KMS perform the following actions (including decryption) when you call putRecord(s) or getRecords on a data stream with server-side encryption enabled.

1. Data is sent from a customer's Kinesis producer application (client) to Kinesis using SSL via HTTPS.
2. Data is received by Kinesis, stored in RAM, and encryption is applied to the payload and partition key of a record.

3. Kinesis requests a plaintext input keying material (IKM) and a copy of the IKM is encrypted by using the customer’s selected KMS master key.
4. AWS KMS creates an IKM, encrypts it by using the master key, and sends both the plaintext IKM and the encrypted IKM to Kinesis.
5. Kinesis uses the plaintext IKM to derive data keys that are unique per-record.
6. Kinesis encrypts the payload and partition key using the data keys and removes the plaintext key from memory.
7. Kinesis appends the encrypted IKM to the encrypted data.
8. The plaintext IKM is cached in memory for reuse until it expires after 5 minutes.
9. Kinesis delivers the encrypted message to a backend store where it is stored at rest and fetch-able by a getRecords call.

Kinesis and AWS KMS perform the following actions (including decryption) when you call getRecords.

1. When a getRecords call is made, the frontend of Kinesis retrieves the encrypted record from the backend service.
2. Kinesis makes a request to KMS using a token generated by the customer’s request. AWS KMS authorizes it.
3. Kinesis decrypts the encrypted IKM stored with the record.
4. Kinesis recreates the per-record data keys from the decrypted IKM.
5. If authorized, Kinesis decrypts the payload and removes the plaintext data key from memory.
6. Kinesis delivers the payload over SSL and HTTPS to the Kinesis consumer (client) requesting the records.

Pricing and billing

Q: Is Amazon Kinesis Data Streams available in AWS Free Tier?

No. Amazon Kinesis Data Streams is not currently available in AWS Free Tier. AWS Free Tier is a program that offers free trial for a group of AWS services. For more details about AWS Free Tier, see AWS Free Tier.

Q: How much does Amazon Kinesis Data Streams cost?

Amazon Kinesis Data Streams uses simple pay as you go pricing. There is neither upfront cost nor minimum fees, and you only pay for the resources you use. The cost of Amazon Kinesis Data Streams has two core dimensions and three optional dimensions:

- Hourly Shard cost determined by the number of shards within your Amazon Kinesis data stream.
- PUT Payload Unit cost determined by the number of 25KB payload units that your data producers add to your data stream.

Optional:

- Extended data retention is an optional cost determined by the number of shard hours incurred by your data stream. When extended data retention is enabled, you pay the extended retention rate for each shard in your stream.
- Enhanced fan-out is an optional cost with two cost dimensions: consumer-shard hours and data retrievals. Consumer-shard hours reflect the number of shards in a stream multiplied by the number of consumers using enhanced fan-out. Data retrievals are determined by the number of GBs delivered to consumers using enhanced fan-out.

For more information about Amazon Kinesis Data Streams costs, see Amazon Kinesis Data Streams Pricing.

Q: Does my PUT Payload Unit cost change by using PutRecords operation instead of PutRecord operation?

PUT Payload Unit charge is calculated based on the number of 25KB payload units added to your Amazon Kinesis data stream. PUT Payload Unit cost is consistent when using PutRecords operation or PutRecord operation.

Q: Am I charged for shards in "CLOSED" state?

A shard could be in "CLOSED" state after resharding. You will not be charged for shards in "CLOSED" state.

Q: Other than Amazon Kinesis Data Streams costs, are there any other costs that might incur to my Amazon Kinesis Data Streams usage?

If you use Amazon EC2 for running your Amazon Kinesis Applications, you will be charged for Amazon EC2 resources in addition to Amazon Kinesis Data Streams costs.

Amazon Kinesis Client Library (KCL) uses Amazon DynamoDB tables to track state information of record processing. If you use KCL for your Amazon Kinesis Applications, you will be charged for Amazon DynamoDB resources in addition to Amazon Kinesis Data Streams costs.

If you enable Enhanced Shard-Level Metrics, you will be charged for Amazon CloudWatch costs associated with enabled shard-level metrics in addition to Amazon Kinesis Data Streams costs.

Please note that the above are three common, but not exhaustive, cases.

Service Level Agreement

Q: What does the Amazon Kinesis Data Streams SLA guarantee?

Our Amazon Kinesis Data Streams SLA guarantees a Monthly Uptime Percentage of at least 99.9% for Amazon Kinesis Data Streams.

Q: How do I know if I qualify for a SLA Service Credit?

You are eligible for a SLA credit for Amazon Kinesis Data Streams under the Amazon Kinesis Data Streams SLA if more than one Availability Zone in which you are running a task, within the same region has a Monthly Uptime Percentage of less than 99.9% during any monthly billing cycle.

For full details on all of the terms and conditions of the SLA, as well as details on how to submit a claim, please see the Amazon Kinesis Data Streams SLA details page.

Amazon MSK FAQs

General

Q: What is Amazon MSK?
Amazon MSK is a new AWS streaming data service that manages Apache Kafka infrastructure and operations, making it easy for developers and DevOps managers to run Apache Kafka applications on AWS without the need to become experts in operating Apache Kafka clusters. Amazon MSK is an ideal place to run existing or new Apache Kafka applications in AWS. Amazon MSK operates and maintains Apache Kafka clusters, provides enterprise-grade security features out of the box, and has built-in AWS integrations that accelerate development of streaming data applications. To get started, you can migrate existing Apache Kafka workloads into Amazon MSK, or with a few clicks, you can build new ones from scratch in minutes. There are no data transfer charges for in-cluster traffic, and no commitments or upfront payments required. You only pay for the resources that you use.

Q: What is Apache Kafka?
Apache Kafka is an open-source, high performance, fault-tolerant, and scalable platform for building real-time streaming data pipelines and applications. Apache Kafka is a streaming data store that decouples applications producing streaming data (producers) into its data store from applications consuming streaming data (consumers) from its data store. Organizations use Apache Kafka as a data source for applications that continuously analyze and react to streaming data. Learn more about Apache Kafka.

Q: What is streaming data?
Streaming data is a continuous stream of small records or events (a record or event is typically a few kilobytes) generated by thousands of machines, devices, websites, and applications. Streaming data includes a wide variety of data such as log files generated by customers using your mobile or web applications, ecommerce purchases, in-game player activity, information from social

networks, financial trading floors, geospatial services, and telemetry from connected devices or instrumentation in data centers. Streaming data services like Amazon MSK and Amazon Kinesis Data Streams make it easy for you to continuously collect, process, and deliver streaming data. Learn more about streaming data.

Q: What are Apache Kafka's primary capabilities?
Apache Kafka has three key capabilities:

- Apache Kafka stores streaming data in a fault-tolerant way as a continuous series of records and preserves the order in which the records were produced.
- Apache Kafka acts as a buffer between data producers and data consumers. Apache Kafka allows many data producers (e.g. websites, IoT devices, Amazon EC2 instances) to continuously publish streaming data and categorize this data using Apache Kafka topics. Multiple data consumers (e.g. machine learning applications, Lambda functions) read from these topics at their own rate, similar to a message queue or enterprise messaging system.
- Data consumers process data from Apache Kafka topics on a first-in-first-out basis, preserving the order data was produced.

Q: What are the key concepts of Apache Kafka?
Apache Kafka stores records in topics. Data producers write records to topics and consumers read records from topics. Each record in Apache Kafka consists of a key, a value, and a timestamp. Apache Kafka partitions topics and replicates these partitions across multiple nodes called brokers. Apache Kafka runs as a cluster on one or more brokers, and brokers can be located in multiple AWS availability zones to create a highly available cluster. Apache Kafka relies on Apache ZooKeeper to coordinate cluster tasks and can maintain state for resources interacting with an Apache Kafka cluster.

Q: When should I use Apache Kafka?
Apache Kafka is used to support real-time applications that transform, deliver, and react to streaming data, and for building real-time streaming data pipelines that reliably get data between multiple systems or applications.

Q: What does Amazon MSK do?
Amazon MSK makes it easy to get started and run open-source versions of

Apache Kafka in AWS with high availability and security while providing integration with AWS services without the operational overhead of running an Apache Kafka cluster. Amazon MSK allows you to use and configure open-source versions of Apache Kafka while the service manages the setup, provisioning, AWS integrations, and on-going maintenance of Apache Kafka clusters.

With a few clicks in the console, you can provision an Amazon MSK cluster. From there, Amazon MSK replaces unhealthy brokers, automatically replicates data for high availability, manages Apache ZooKeeper nodes, automatically deploys hardware patches as needed, manages the integrations with AWS services, makes important metrics visible through the console, and will support Apache Kafka version upgrades when more than one version is supported so you can take advantage of improvements to the open-source version of Apache Kafka.

Q: What Apache Kafka versions does Amazon MSK support?
For supported Kafka versions, see the Amazon MSK documentation.

Q: Are Apache Kafka APIs compatible with Amazon MSK?
Yes, all data plane and admin APIs are natively supported by Amazon MSK.

Q: Is the Apache Kafka AdminClient supported by Amazon MSK?
Yes.

Data production and consumption

Q: Can I use Apache Kafka APIs to get data in and out of Apache Kafka?
Yes, Amazon MSK supports the native Apache Kafka producer and consumer APIs. Your application code does not need to change when clients begin to work with clusters within Amazon MSK.

Q: Can I use Apache Kafka Connect, Apache Kafka Streams, or any other ecosystem component of Apache Kafka with Amazon MSK?

Yes, you can use any component that leverages the Apache Kafka producer and consumer APIs, and the Apache Kafka AdminClient. Tools that upload .jar files into Apache Kafka clusters are currently not compatible with Amazon MSK, including Confluent Control Center, Confluent Auto Data Balancer, Uber uReplicator, and LinkedIn Cruise Control.

Migrating to Amazon MSK

Q: Can I migrate data within my existing Apache Kafka cluster to Amazon MSK?
Yes, you can use third-party tools or open source tools like MirrorMaker that come with open source Apache Kafka to replicate data from clusters into an Amazon MSK cluster.

Version upgrades

Q: Are Apache Kafka version upgrades supported?
Cluster version upgrades are not currently supported but we plan to support version upgrades soon.

Q: How will the upgrade process work under the hood?
When you deploy a new Apache Kafka version, Amazon MSK will use a rolling upgrade process that upgrades one broker or Apache ZooKeeper node at a time before moving on to the next resource. Throughout the upgrade process your cluster will be in an ‘Updating’ state and will transition to an ‘Active’ state when finished. It’s important to note that if you chose to not replicate data to multiple brokers within a cluster that is being upgraded, your cluster will experience downtime.

Clusters

Q: How do I create my first Amazon MSK cluster?

You can create your first cluster with a few clicks in the AWS management console or using the AWS SDKs. First, in the Amazon MSK console select an AWS region to create an Amazon MSK cluster in. Choose a name for your cluster, the VPC you want to run the cluster with, a data replication strategy for the cluster, and the subnets for each AZ. Next, pick a broker instance type and quantity of brokers per AZ, and click create.

Q: What resources are within a cluster?
Each cluster contains broker instances, provisioned storage, and Apache ZooKeeper nodes.

Q: What types of broker instances can I provision within an Amazon MSK cluster?
You can choose instances within the EC2 M5 instance family.

Q: Does Amazon MSK offer Reserved Instance pricing?
No, not at this time.

Q: Do I need to provision and pay for broker boot volumes?
No, each broker you provision includes boot volume storage managed by the Amazon MSK service.

Q: When I create an Apache Kafka cluster, do the underlying resources (e.g. Amazon EC2 instances) show up in my EC2 console?
Some resources, like elastic network interfaces (ENIs), will show up in your Amazon EC2 account. Other Amazon MSK resources will not show up in your EC2 account as these are managed by the Amazon MSK service.

Q: What do I need to provision within an Amazon MSK cluster?
You need to provision broker instances and broker storage with every cluster you create. You do not provision Apache ZooKeeper nodes as these resources are included at no additional charge with each cluster you create.

Q: What is the default broker configuration for a cluster?
Unless otherwise specified, Amazon MSK uses the same defaults specified by the open-source version of Apache Kafka. The default settings are documented here.

Q: Can I provision brokers such that they are imbalanced across AZs (e.g. 3 in us-east-1a, 2 in us-east-1b, 1 in us-east-1c)?
No, Amazon MSK enforces the best practice of balancing broker quantities across AZs within a cluster.

Q: How does data replication work in Amazon MSK?
Amazon MSK uses Apache Kafka's leader-follower replication to replicate data between brokers. Amazon MSK makes it easy to deploy clusters with multi-AZ replication and gives you the option to use a custom replication strategy by topic. By default with each of the replication options, leader and follower brokers will be deployed and isolated using the replication strategy specified. For example, if you select a 3 AZ broker replication strategy with 1 broker per AZ cluster, Amazon MSK will create a cluster of three brokers (1 broker in three AZs in a region), and by default (unless you choose to override the topic replication factor) the topic replication factor will also be 3.

Q: Can I change the default broker configurations or upload a cluster configuration to Amazon MSK?
Yes, Amazon MSK allows you create custom configurations and apply them to new clusters. Custom configurations can be created using the AWS CLI or Console. Custom configurations support for existing clusters will be available June 2019.

Q: What configuration properties am I able to customize?
The configurations properties that you can customize are documented here.

Q: What is the default configuration of a new topic?
Amazon MSK uses Apache Kafka's default configuration unless otherwise specified here:

Replication factor	Cluster default
Min.Insync.Replicas	2

Topics

Q: How do I create topics?
Once your Apache Kafka cluster has been created, you can create topics using the Apache Kafka APIs. All topic and partition level actions and configurations are performed using Apache Kafka APIs. The following command is an example of creating a topic using Apache Kafka APIs:

```
bin/kafka-topics.sh --create —bootstrap-server ConnectionString:9092 --replication-factor 3 --partitions 1 --topic TopicName
```

Networking

Q: Does Amazon MSK run in an Amazon VPC?
Yes, Amazon MSK always runs within an Amazon VPC managed by the Amazon MSK service. Amazon MSK resources will be available to your own Amazon VPC, subnet, and security group you select when the cluster is setup. IP addresses from your VPC are attached to your Amazon MSK resources through elastic network interfaces (ENIs), and all network traffic stays within the AWS network and is not accessible to the Internet.

Q: Is the connection between my clients and an Amazon MSK cluster always private?
Yes, the only way data can be produced and consumed from an Amazon MSK cluster is over a private connection between your clients in your VPC and the Amazon MSK cluster. Amazon MSK does not support public endpoints.

Q: How will the brokers in my Amazon MSK cluster be made accessible to clients within my VPC?
The brokers in your cluster will be made accessible to clients in your VPC through elastic network interfaces (ENIs) which will appear in your account. The Security Groups on the ENIs will dictate the source and type of ingress and egress traffic allowed on your brokers.

Q: How can I give clients running in different AWS accounts access to my cluster?

You can use VPC peering to give clients running in different AWS accounts access to your cluster.

Connecting to the VPC

Q: How do I connect to my AWS MSK cluster outside of the VPC?

There are several methods to connect to your AWS MSK clusters outside of your VPC.

- VPN: https://docs.aws.amazon.com/vpc/latest/userguide/vpn-connections.html
- VPC Peering: https://docs.aws.amazon.com/vpc/latest/peering/what-is-vpc-peering.html
- VPC Transit Gateway: https://docs.aws.amazon.com/vpc/latest/tgw/what-is-transit-gateway.html
- AWS Direct Connect: https://aws.amazon.com/directconnect/
- REST Proxy: A REST proxy can be installed on an instance running within your VPC. REST proxies allow your producers and consumers to communicate to the cluster through HTTP API requests.

Encryption

Q: Can I encrypt data in my Amazon MSK cluster?
Yes, Amazon MSK uses Amazon EBS server-side encryption and AWS KMS keys to encrypt storage volumes.

Q: Is data encrypted in-transit between brokers within an Amazon MSK cluster?
Yes, by default new clusters have encryption in-transit enabled via TLS for inter-broker communication. You can opt-out of using encryption in-transit when a cluster is created.

Q: Is data encrypted in-transit between my Apache Kafka clients and the Amazon MSK service?
Yes, by default in-transit encryption is set to TLS only for clusters created from the CLI or AWS Console. Additional configuration is required for clients to communicate with clusters using TLS encryption. You can change the default encryption setting by selecting the TLS/plaintext or plaintext settings. Read More: MSK Encryption

Q: Is data encrypted in-transit as it moves between brokers and Apache ZooKeeper nodes in an Amazon MSK cluster?
No, the default version of Apache Zookeeper bundled with Apache Kafka does not support encryption. However it is important to note that communications between Apache Zookeeper and Apache Kafka brokers is limited to broker, topic, and partition state information.

Authentication and Authorization

Q: How can I restrict the scope of connectivity to an Amazon MSK cluster across multiple clients in my VPC?
Amazon MSK supports TLS based authentication and you can use this feature to authenticate client connections to an Amazon MSK cluster. Amazon MSK allows you to deploy private CAs within the AWS Certificate Manager service to an MSK cluster. When TLS client authentication is enabled, only clients presenting TLS certificates generated from the previously loaded private CAs can authenticate with the cluster.

Q: How does authorization work in Amazon MSK?
Apache Kafka uses access control lists (ACLs) for authorization and Amazon MSK supports the use of ACLs. To enable ACLs you must enable client authentication using TLS.

Q: How can I authenticate and authorize a client at the same time?
Amazon MSK customers who use client TLS authentication can use the Dname of clients TLS certificates as the principal of the ACL to authenticate and authorize client requests.

Monitoring, Metrics, Logging, Tagging

Q: How do I monitor the performance of my clusters or topics?
You can monitor the performance of your clusters using the Amazon MSK console, Amazon CloudWatch console, or you can access JMX and host metrics using Open Monitoring with Prometheus, an open source monitoring solution.

Q: What is the cost for the different CloudWatch monitoring levels?
The cost of monitoring your cluster using Amazon CloudWatch is dependent on the monitoring level and the size of your Apache Kafka cluster. Amazon CloudWatch charges per metric per month and includes a free tier; see Amazon CloudWatch pricing for more information. For details on the number of metrics exposed for each monitoring level, see Amazon MSK monitoring documentation.

Q: What monitoring tools are compatible with Open Monitoring with Prometheus?
Tools that are designed to read from Prometheus exporters are compatible with Open Monitoring, like: Datadog, Lenses, New Relic, Sumologic, or a Prometheus server. For details on Open Monitoring, see Amazon MSK Open Monitoring documentation.

Q: How do I monitor the health and performance of clients?
You can use any client-side monitoring supported by the Apache Kafka version you are using.

Q: Can I tag Amazon MSK resources?
Yes, you can tag Amazon MSK clusters from the AWS CLI or Console.

Q: How do I monitor consumer lag?
Consumer lag within your Amazon MSK cluster can be monitored using consumer lag tools like Linkedin's Burrow: https://github.com/linkedin/Burrow

Apache ZooKeeper

Q: What is Apache ZooKeeper?

From https://zookeeper.apache.org/: “Apache ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications,” including Apache Kafka.

Q: Does Amazon MSK use Apache ZooKeeper?
Yes, Amazon MSK uses Apache ZooKeeper and manages Apache ZooKeeper within each cluster as a part of the Amazon MSK service. Apache ZooKeeper nodes are included with each cluster at no additional cost.

Q: How do my clients interact with Apache ZooKeeper?
Your clients can interact with Apache ZooKeeper through an Apache ZooKeeper endpoint provided by the service. This endpoint is provided in the AWS management console or using the DescribeCluster API.

Integrations

Q: What AWS services does Amazon MSK integrate with?
Amazon MSK integrates with:

- Amazon VPC for network isolation and security
- Amazon CloudWatch for metrics
- Amazon KMS for storage volume encryption
- Amazon IAM for authentication of cluster APIs
- AWS CloudTrail for AWS API logs
- AWS Certificate Manager for Private CAs used for client TLS authentication
- AWS CloudFormation for describing and provisioning Amazon MSK clusters using code
- Amazon Kinesis Data Analytics for full managed Apache Flink applications

Scaling

Q: How can I scale up storage in my cluster?
You can scale up storage in your cluster using the AWS Management Console or the AWS CLI.

Q. Can I scale the number of brokers in an existing cluster?
Yes. You can scale out, or increase, the number of brokers for existing Amazon MSK clusters.

Q. Can I scale a broker instance size in an existing cluster?
No. Scaling the instance size of brokers in an existing cluster is not currently supported by Amazon MSK, but is on our roadmap.

Pricing and Availability

Q: How does Amazon MSK pricing work?
Pricing is based is per Apache Kafka broker-hour, and per provisioned storage-hour. AWS data transfer rates apply for data transfer in and out of Amazon MSK. For more information, visit our pricing page.

Q: Do I pay for data transfer as a result of data replication?
No, all in-cluster data transfer is included with the service at no additional charge.

Q: What AWS regions offer Amazon MSK?
Amazon MSK region availability is documented here.

Q: How does data transfer pricing work?
You will pay standard AWS data transfer charges for data transferred in and out of an Amazon MSK cluster. You will not be charged for data transfer within the cluster in a region, including data transfer between brokers and data transfer between brokers and Apache ZooKeeper nodes.

Compliance

Q: What compliance programs are in scope for Amazon MSK?
Amazon MSK is compliant or eligible for the following programs:

- HIPAA eligible
- PCI
- ISO
- SOC 1,2,3

For a complete list of AWS services and compliance programs, please see AWS Services in Scope by Compliance Program.

Service Level Agreement

Q: What does the Amazon MSK SLA guarantee?
Our Amazon MSK SLA guarantees a Monthly Uptime Percentage of at least 99.9% for Amazon MSK.

Q: How do I know if I qualify for a SLA Service Credit?
You are eligible for a SLA credit for Amazon MSK under the Amazon MSK SLA if Multi-AZ deployments on Amazon MSK have a Monthly Uptime Percentage of less than 99.9% during any monthly billing cycle.

For full details on all of the terms and conditions of the SLA, as well as details on how to submit a claim, please see the Amazon MSK SLA details page.

Get started with Amazon MSK

Calculate your costs

Visit the Amazon MSK pricing page.

Review the getting-started guide

Learn how to set up your Apache Kafka cluster on Amazon MSK in this step-by-step guide.

Run your Apache Kafka cluster

Start running your Apache Kafka cluster on Amazon MSK. Log in to the Amazon MSK console.

Amazon Redshift FAQs

General

Learn about what's new with Amazon Redshift on the What's New page.

View more detailed information and usage guidance in the Documentation.

Q: What is Amazon Redshift?
Amazon Redshift is a fast, fully managed data warehouse that makes it simple and cost-effective to analyze all your data using standard SQL and your existing Business Intelligence (BI) tools. It allows you to run complex analytic queries against petabytes of structured data using sophisticated query optimization, columnar storage on high-performance storage, and massively parallel query execution. Most results come back in seconds. With Redshift, you can start small for just $0.25 per hour with no commitments and scale out to petabytes of data for $1,000 per terabyte per year, less than a tenth the cost of traditional solutions. Amazon Redshift also includes Amazon Redshift Spectrum, allowing you to directly run SQL queries against exabytes of unstructured data in Amazon S3 data lakes. No loading or transformation is required, and you can use open data formats, including Avro, CSV, Ion, JSON, ORC, Parquet, and more. Redshift Spectrum automatically scales query compute capacity based on the data being retrieved, so queries against Amazon S3 run fast, regardless of data set size.

Amazon Redshift gives you fast querying capabilities over structured data using familiar SQL-based clients and business intelligence (BI) tools using standard ODBC and JDBC connections. Queries are distributed and parallelized across multiple physical resources. You can easily scale an Amazon Redshift data warehouse up or down with a few clicks in the AWS Management Console or with a single API call. Amazon Redshift automatically patches and backs up your data warehouse, storing the backups for a user-defined retention period.

Amazon Redshift uses replication and continuous backups to enhance availability and improve data durability and can automatically recover from component and node failures. In addition, Amazon Redshift supports Amazon Virtual Private Cloud (Amazon VPC), SSL, AES-256 encryption, and Hardware Security Modules (HSMs) to protect your data in transit and at rest.

As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources you use. Amazon Redshift lets you pay as you go. You can even try Amazon Redshift for free.

For information about Amazon Redshift regional availability, see the AWS Region Table.

Q: Why would I use Amazon Redshift over an on-premises data warehouse?

On-premises data warehouses require significant time and resource to administer, especially for large datasets. In addition, the financial cost associated with building, maintaining, and growing self-managed, on-premise data warehouses is very high. As your data grows, you have to constantly trade-off what data to load into your data warehouse and what data to archive in storage so you can manage costs, keep ETL complexity low, and deliver good performance. Amazon Redshift not only significantly lowers the cost and operational overhead of a data warehouse, but with Redshift Spectrum, it also makes it easy to analyze large amounts of data in its native format without requiring you to load the data.

Q: How do I use Amazon Redshift features that are in preview?

When creating a Amazon Redshift cluster you can pick three tracks for maintenance: Current, Trailing, or Preview. Within the Preview track, PREVIEW_FEATURES should be selected to use Redshift features that are available in preview.

Q: What is AQUA (Advanced Query Accelerator) for Amazon Redshift?

AQUA is a new distributed and hardware-accelerated cache for Redshift. Learn more and sign up to be considered for the preview.

Q: What is Redshift Spectrum?

Redshift Spectrum is a feature of Amazon Redshift that enables you to run queries against exabytes of unstructured data in Amazon S3, with no loading or ETL required. When you issue a query, it goes to the Amazon Redshift SQL endpoint, which generates and optimizes a query plan. Amazon Redshift determines what data is local and what is in Amazon S3, generates a plan to minimize the amount of Amazon S3 data that needs to be read, requests Redshift Spectrum workers out of a shared resource pool to read and process data from Amazon S3.

Redshift Spectrum scales out to thousands of instances if needed, so queries run quickly regardless of data size. In addition, you can use the exact same SQL for Amazon S3 data as you do for your Amazon Redshift queries today and connect to the same Amazon Redshift endpoint using your same BI tools. Redshift Spectrum lets you separate storage and compute, allowing you to scale each independently. You can setup as many Amazon Redshift clusters as you need to query your Amazon S3 data lake, providing high availability and limitless concurrency. Redshift Spectrum gives you the freedom to store your data where you want, in the format you want, and have it available for processing when you need it.

For information about Redshift Spectrum regional availability, please visit the Amazon Redshift pricing page.

Q: What does Amazon Redshift manage on my behalf?

Amazon Redshift manages the work needed to set up, operate, and scale a data warehouse, from provisioning the infrastructure capacity to automating ongoing administrative tasks such as backups, and patching. Amazon Redshift automatically monitors your nodes and drives to help you recover from failures. For Redshift Spectrum, Amazon Redshift manages all the computing infrastructure, load balancing, planning, scheduling and execution of your queries on data stored in Amazon S3.

Q: How does the performance of Amazon Redshift compare to most traditional databases for data warehousing and analytics?

Amazon Redshift uses a variety of innovations to achieve up to ten times higher performance than traditional databases for data warehousing and analytics workloads:

- *Columnar Data Storage:* Instead of storing data as a series of rows, Amazon Redshift organizes the data by column. Unlike row-based systems, which are ideal for transaction processing, column-based systems are ideal for data warehousing and analytics, where queries often involve aggregates performed over large data sets. Since only the columns involved in the queries are processed and columnar data is stored sequentially on the storage media, column-based systems require far fewer I/Os, greatly improving query performance.
- *Advanced Compression:* Columnar data stores can be compressed much more than row-based data stores because similar data is stored sequentially on disk. Amazon Redshift employs multiple compression techniques and can often achieve significant compression relative to traditional relational data stores. When loading data into an empty table, Amazon Redshift automatically samples your data and selects the most appropriate compression scheme.
- *Massively Parallel Processing (MPP):* Amazon Redshift automatically distributes data and query load across all nodes. Amazon Redshift makes it easy to add nodes to your data warehouse and enables you to maintain fast query performance as your data warehouse grows.
- *Redshift Spectrum:* Redshift Spectrum enables you to run queries against exabytes of data in Amazon S3. There is no loading or ETL required. Even if you don’t store any of your data in Amazon Redshift, you can still use Redshift Spectrum to query datasets as large as an exabyte in Amazon S3. When you issue a query, it goes to the Amazon Redshift SQL endpoint, which generates the query plan. Amazon Redshift determines what data is local and what is in Amazon S3, generates a plan to minimize the amount of Amazon S3 data that needs to be read, requests Redshift Spectrum workers out of a shared resource pool to read and process data from Amazon S3, and pulls results back into your Amazon Redshift cluster for any remaining processing.

Q: How do I get started with Amazon Redshift?

You can sign up and get started within minutes from the Amazon Redshift detail page or via the AWS Management Console. If you don't already have an AWS account, you'll be prompted to create one.

To use the Redshift Spectrum feature, you need to first store your data in Amazon S3. You can then define the metadata about that data in your Amazon Redshift cluster or register the metadata you may already have in your Hive metastore with your cluster. You can issue a CREATE EXTERNAL SCHEMA SQL command in your Amazon Redshift cluster to define or register a database in your catalog as an external schema within Amazon Redshift. You can then issue queries against Amazon S3 using the same SQL you use for local tables and any BI tool that supports Amazon Redshift today. The external database definition you create using Amazon Redshift SQL is registered in the same data catalog that Amazon Athena uses. You can optionally manage the external database definition from the Amazon Athena Catalog as well.

Visit our Getting Started page to see how to try Amazon Redshift for free.

Q: How do I create and access an Amazon Redshift data warehouse cluster?

You can easily create an Amazon Redshift data warehouse cluster by using the AWS Management Console or the Amazon Redshift APIs. You can start with a single node, 160GB data warehouse and scale all the way to petabytes or more with a few clicks in the AWS Console or a single API call.

The single node configuration enables you to get started with Amazon Redshift quickly and cost-effectively and scale up to a multi-node configuration as your needs grow. A Redshift data warehouse cluster can contain from 1-128 compute nodes, depending on the node type. For details, please see our documentation.

The multi-node configuration requires a leader node that manages client connections and receives queries, and two compute nodes that store data and perform queries and computations. The leader node is provisioned for you automatically and you are not charged for it.

Simply specify your preferred Availability Zone (optional), the number of nodes, node types, a master name and password, security groups, your preferences for backup retention, and other system settings. Once you've chosen your desired configuration, Amazon Redshift will provision the required resources and set up your data warehouse cluster.

Once your data warehouse cluster is available, you can retrieve its endpoint and JDBC and ODBC connection string from the AWS Management Console or by using the Redshift APIs. You can then use this connection string with your favorite database tool, programming language, or Business Intelligence (BI) tool. You will need to authorize network requests to your running data warehouse cluster. For a detailed explanation, please refer to our Getting Started Guide.

Q: What is the maximum storage capacity per compute node? What is the recommended amount of data per compute node for optimal performance?

You can create a cluster using either RA3, DC, or DS node types. RA3 node types enable you to scale and pay for compute and storage independently. You choose the number of instances you need based on performance requirements, and only pay for the managed storage that you use.

RA3 is available now as RA3.16XL, which allows building a cluster with up to 8 petabytes in managed storage. With RA3, customers pay for the actual storage they use. The RA3 clusters run with minimum 2 nodes and the minimum sized RA3.16XL cluster can support up-to 128 TB. Redshift managed storage uses large, high-performance SSDs in each Amazon Redshift RA3 instance for fast local storage and Amazon S3 for longer-term durable storage. If the data in an instance grows beyond the size of the large local SSDs, Redshift managed storage automatically offloads that data to Amazon S3. Customers pay the same low rate for Redshift managed storage regardless of whether the data sits in high performance SSDs or in Amazon S3. For workloads that require a lot of storage, but not as much compute capacity, this lets customers automatically scale their data warehouse storage capacity without adding and paying for additional instances

DC node types are available in two sizes. The Large has 160GB of SSD storage, 2 Intel Xeon E5-2670v2 (Ivy Bridge) virtual cores and 15GiB of RAM. The Eight Extra Large is sixteen times bigger with 2.56TB of SSD storage, 32 Intel Xeon E5-2670v2 virtual cores, and 244GiB of RAM. You can get started with a single

DC2.Large node for $0.25 per hour and scale all the way up to 128 8XL nodes with 326TB of SSD storage, 3,200 virtual cores, and 24TiB of RAM.

DS node types are available in two sizes, Extra Large and Eight Extra Large. The Extra Large (XL) has 3 HDDs with a total of 2TB of magnetic storage, whereas Eight Extra Large (8XL) has 24 HDDs with a total of 16TB of magnetic storage. DS2.8XLarge has 36 Intel Xeon E5-2676 v3 (Haswell) virtual cores and 244GiB of RAM, and DS2.XL has 4 Intel Xeon E5-2676 v3 (Haswell) virtual cores, and 31GiB of RAM.

Please see our pricing page for more detail.

Q: When would I use Amazon Redshift vs. Amazon RDS?

Both Amazon Redshift and Amazon RDS enable you to run traditional relational databases in the cloud while offloading database administration. Customers use Amazon RDS databases primarily for online-transaction processing (OLTP) workload while Redshift is used primarily for reporting and analytics. Amazon Redshift harnesses the scale and resources of multiple nodes and uses a variety of optimizations to provide order of magnitude improvements over traditional databases for analytic and reporting workloads against very large data sets. Amazon Redshift provides an excellent scale-out option as your data and query complexity grows if you want to prevent your reporting and analytic processing from interfering with the performance of your OLTP workload. Now, with the new Federated Query feature (preview), you can easily query data across your Amazon RDS or Aurora database services with Redshift.

Q: When would I use Amazon Redshift or Redshift Spectrum vs. Amazon EMR?

You should use Amazon EMR if you use custom code to process and analyze extremely large datasets with big data processing frameworks such as Apache Spark, Hadoop, Presto, or Hbase. Amazon EMR gives you full control over the configuration of your clusters and the software you install on them.

Data warehouses like Amazon Redshift are designed for a different type of analytics altogether. Data warehouses are designed to pull together data from lots of different sources, like inventory, financial, and retail sales systems. In order to ensure that reporting is consistently accurate across the entire company, data warehouses store data in a highly structured fashion. This structure builds data consistency rules directly into the tables of the database. Amazon Redshift is the best service to use when you need to perform complex queries on massive collections of structured and semi-structured data and get super fast performance.

While the Redshift Spectrum feature is great for running queries against data in Amazon Redshift and S3, it really isn't a fit for the types of use cases that enterprises typically ask from processing frameworks like Amazon EMR. Amazon EMR goes far beyond just running SQL queries. Amazon EMR is a managed service that lets you process and analyze extremely large data sets using the latest versions of popular big data processing frameworks, such as Spark, Hadoop, and Presto, on fully customizable clusters. With Amazon EMR, you can run a wide variety of scale-out data processing tasks for applications such as machine learning, graph analytics, data transformation, streaming data, and virtually anything you can code.

You can use Redshift Spectrum together with EMR. Redshift Spectrum uses the same approach to store table definitions as Amazon EMR. Redshift Spectrum can support the same Apache Hive Metastore used by Amazon EMR to locate data and table definitions. If you're using Amazon EMR and have a Hive Metastore already, you just have to configure your Amazon Redshift cluster to use it. You can then start querying that data right away along with your Amazon EMR jobs. Therefore, if you're already using EMR to process a large data store, you can use Redshift Spectrum to query that data right at the same time without interfering with your Amazon EMR jobs.

Query services, data warehouses, and complex data processing frameworks all have their place, and they are used for different things. You just need to choose the right tool for the job.

Q: When should I use Amazon Athena vs. Redshift Spectrum?

Amazon Athena is the simplest way to give any employee the ability to run ad-hoc queries on data in Amazon S3. Athena is serverless, so there is no infrastructure to setup or manage, and you can start analyzing your data immediately.

If you have frequently accessed data, that needs to be stored in a consistent, highly structured format, then you should use a data warehouse like Amazon Redshift. This gives you the flexibility to store your structured, frequently accessed data in Amazon Redshift, and use Redshift Spectrum to extend your Amazon Redshift queries out to the entire universe of data in your Amazon S3 data lake. This gives you the freedom to store your data where you want, in the format you want, and have it available for processing when you need.

Q: Why should I use Amazon Redshift instead of running my own MPP data warehouse cluster on Amazon EC2?

Amazon Redshift automatically handles many of the time-consuming tasks associated with managing your own data warehouse including:

- *Setup:* With Amazon Redshift, you simply create a data warehouse cluster, define your schema, and begin loading and querying your data. Provisioning, configuration and patching are all managed for you.
- *Data Durability:* Amazon Redshift replicates your data within your data warehouse cluster and continuously backs up your data to Amazon S3, which is designed for eleven nines of durability. Amazon Redshift mirrors each drive's data to other nodes within your cluster. If a drive fails, your queries will continue with a slight latency increase while Redshift rebuilds your drive from replicas. In case of node failure(s), Amazon Redshift automatically provisions new node(s) and begins restoring data from other drives within the cluster or from Amazon S3. It prioritizes restoring your most frequently queried data so your most frequently executed queries will become performant quickly.
- *Scaling:* You can add or remove nodes from your Amazon Redshift data warehouse cluster with a single API call or via a few clicks in the AWS Management Console as your capacity and performance needs change. You can also schedule your scaling and resize operations by using the scheduler capability in Redshift.
- *Automatic Updates and Patching:* Amazon Redshift automatically applies upgrades and patches your data warehouse so you can focus on your application and not on its administration.
- *Exabyte Scale Query Capability:* Redshift Spectrum enables you to run queries against exabytes of data in Amazon S3. There is no loading or ETL

required. Even if you don’t store any of your data in Amazon Redshift, you can still use Redshift Spectrum to query datasets as large as an exabyte in Amazon S3.

Billing

Q: How will I be charged and billed for my use of Amazon Redshift?

You pay only for what you use, and there are no minimum or setup fees. Billing commences for a data warehouse cluster as soon as the data warehouse cluster is available. Billing continues until the data warehouse cluster terminates, which would occur upon deletion or in the event of instance failure. You are billed based on:

- *Compute node hours:* Compute node hours are the total number of hours you run across all your compute nodes for the billing period. Node usage hours are billed for each hour your data warehouse cluster is running in an available state. If you no longer wish to be charged for your data warehouse cluster, you must terminate it to avoid being billed for additional node hours. Partial node hours consumed are billed as full hours. You are billed for 1 unit per node per hour, so a 3-node data warehouse cluster running persistently for an entire month would incur 2,160 instance hours. You will not be charged for leader node hours; only compute nodes will incur charges.
- *Backup Storage:* Backup storage is the storage associated with your automated and manual snapshots for your data warehouse. Increasing your backup retention period or taking additional snapshots increases the backup storage consumed by your data warehouse. There is no additional charge for backup storage up to 100% of your provisioned storage for an active data warehouse cluster. For example, if you have an active Single Node XL data warehouse cluster with 2TB of local instance storage, we will provide up to 2TB-Month of backup storage at no additional charge. Backup storage beyond the provisioned storage size and backups stored after your cluster is terminated are billed at standard Amazon S3 rates.
- *Data transfer:* There is no data transfer charge for data transferred to or from Amazon Redshift and Amazon S3 within the same AWS Region. For all

other data transfers into and out of Amazon Redshift, you will be billed at standard AWS data transfer rates.

- *Data scanned:* With Redshift Spectrum, you are charged for the amount of Amazon S3 data scanned to execute your query. There are no charges for Redshift Spectrum when you're not running queries. If you store data in a columnar format, such as Parquet or RC, your charges will go down as Redshift Spectrum will only scan the columns needed by the query, rather than processing entire rows. Similarly, if you compress your data, using one of Redshift Spectrum's supported formats, your costs will also go down. You pay the standard Amazon S3 rates for data storage and Amazon Redshift instance rates for the cluster used.

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

For Amazon Redshift pricing information, please visit the Amazon Redshift pricing page.

Data Integration and Loading

Q: How do I load data into my Amazon Redshift data warehouse?

You can load data into Amazon Redshift from a range of data sources including Amazon S3, Amazon DynamoDB, Amazon EMR, AWS Glue, AWS Data Pipeline and or any SSH-enabled host on Amazon EC2 or on-premises. Amazon Redshift attempts to load your data in parallel into each compute node to maximize the rate at which you can ingest data into your data warehouse cluster. Clients can connect to Amazon Redshift using ODBC or JDBC and issue 'insert' SQL commands to insert the data. Please note this is slower than using S3 or DynamoDB since those methods load data in parallel to each compute node while SQL insert statements load via the single leader node. For more details on loading data into Amazon Redshift, please view our Getting Started Guide.

Q: How do I load data from my existing Amazon RDS, Amazon EMR, Amazon DynamoDB, and Amazon EC2 data sources to Amazon Redshift?

You can use our COPY command to load data in parallel directly to Amazon Redshift from Amazon EMR, Amazon DynamoDB, or any SSH-enabled host. Redshift Spectrum also enables you to load data from Amazon S3 into your cluster with a simple INSERT INTO command. This could enable you to load data from various formats such as Parquet and RC into your cluster. Note that if you use this approach, you will accrue Redshift Spectrum charges for the data scanned from Amazon S3. The Redshift Federated Query (Preview) feature enables you to combine data from your Amazon RDS and Aurora (PostgreSQL).

In addition, many ETL companies have certified Amazon Redshift for use with their tools, and a number are offering free trials to help you get started loading your data. AWS Data Pipeline provides a high performance, reliable, fault tolerant solution to load data from a variety of AWS data sources. You can use AWS Data Pipeline to specify the data source, desired data transformations, and then execute a pre-written import script to load your data into Amazon Redshift. In addition, AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy to prepare and load data for analytics. You can create and run an AWS Glue ETL job with a few clicks in the AWS Management Console.

Q: I have a lot of data for initial loading into Amazon Redshift. Transferring via the Internet would take a long time. How do I load this data?

You can use AWS Import/Export to transfer the data to Amazon S3 using portable storage devices. In addition, you can use AWS Direct Connect to establish a private network connection between your network or datacenter and AWS. You can choose 1Gbit/sec or 10Gbit/sec connection ports to transfer your data.

Security

Q: How does Amazon Redshift keep my data secure?

Amazon Redshift encrypts and keeps your data secure in transit and at rest using industry-standard encryption techniques. To keep data secure in transit, Amazon Redshift supports SSL-enabled connections between your client application and your Redshift data warehouse cluster. To keep your data secure at rest, Amazon Redshift encrypts each block using hardware-accelerated AES-256 as it is written to disk. This takes place at a low level in the I/O subsystem, which encrypts everything written to disk, including intermediate query results. The blocks are backed up as is, which means that backups are encrypted as well. By default, Amazon Redshift takes care of key management but you can choose to manage your keys using your own hardware security modules (HSMs) or manage your keys through AWS Key Management Service.

Redshift Spectrum supports Amazon S3's Server Side Encryption (SSE) using your account's default key managed used by the AWS Key Management Service (KMS).

Q: Can I use Amazon Redshift in Amazon Virtual Private Cloud (Amazon VPC)?

Yes, you can use Amazon Redshift as part of your VPC configuration. With Amazon VPC, you can define a virtual network topology that closely resembles a traditional network that you might operate in your own datacenter. This gives you complete control over who can access your Amazon Redshift data warehouse cluster. You can use Redshift Spectrum with an Amazon Redshift cluster that is part of your VPC.

Q: Can I access my Amazon Redshift compute nodes directly?

No. Your Amazon Redshift compute nodes are in a private network space and can only be accessed from your data warehouse cluster's leader node. This provides an additional layer of security for your data.

Availability and Durability

Q: What happens to my data warehouse cluster availability and data durability if a drive on one of my nodes fails?

Your Amazon Redshift data warehouse cluster will remain available in the event of a drive failure however you may see a slight decline in performance for certain queries. In the event of a drive failure, Amazon Redshift will transparently use a replica of the data on that drive which is stored on other drives within that node. In addition, Amazon Redshift will attempt to move your data to a healthy drive or will replace your node if it is unable to do so. Single node clusters do not support data replication. In the event of a drive failure, you will need to restore the cluster from snapshot on S3. We recommend using at least two nodes for production.

Q: What happens to my data warehouse cluster availability and data durability in the event of individual node failure?

Amazon Redshift will automatically detect and replace a failed node in your data warehouse cluster. The data warehouse cluster will be unavailable for queries and updates until a replacement node is provisioned and added to the DB. Amazon Redshift makes your replacement node available immediately and loads your most frequently accessed data from S3 first to allow you to resume querying your data as quickly as possible. Single node clusters do not support data replication. In the event of a drive failure, you will need to restore the cluster from snapshot on S3. We recommend using at least two nodes for production.

Q: What happens to my data warehouse cluster availability and data durability if my data warehouse cluster's Availability Zone (AZ) has an outage?

If your Amazon Redshift data warehouse cluster's Availability Zone becomes unavailable, you will not be able to use your cluster until power and network access to the AZ are restored. Your data warehouse cluster's data is preserved so you can start using your Amazon Redshift data warehouse as soon as the AZ becomes available again. In addition, you can also choose to restore any existing snapshots to a new AZ in the same Region. Amazon Redshift will restore your most frequently accessed data first so you can resume queries as quickly as possible.

Q: Does Amazon Redshift support Multi-AZ Deployments?

Currently, Amazon Redshift only supports Single-AZ deployments. You can run data warehouse clusters in multiple AZ's by loading data into two Amazon Redshift data warehouse clusters in separate AZs from the same set of Amazon S3 input files. With Redshift Spectrum, you can spin up multiple clusters across AZs and access data in Amazon S3 without having to load it into your cluster. In addition, you can also restore a data warehouse cluster to a different AZ from your data warehouse cluster snapshots.

Backup and Restore

Q: How does Amazon Redshift back up my data? How do I restore my cluster from a backup?

Amazon Redshift replicates all your data within your data warehouse cluster when it is loaded and also continuously backs up your data to S3. Amazon Redshift always attempts to maintain at least three copies of your data (the original and replica on the compute nodes and a backup in Amazon S3). Redshift can also asynchronously replicate your snapshots to S3 in another region for disaster recovery.

By default, Amazon Redshift enables automated backups of your data warehouse cluster with a 1-day retention period. You can configure this to be as long as 35 days.

Free backup storage is limited to the total size of storage on the nodes in the data warehouse cluster and only applies to active data warehouse clusters. For example, if you have total data warehouse storage of 8TB, we will provide at most 8TB of backup storage at no additional charge. If you would like to extend your backup retention period beyond one day, you can do so using the AWS Management Console or the Amazon Redshift APIs. For more information on automated snapshots, please refer to the Amazon Redshift Management Guide. Amazon Redshift only backs up data that has changed so most snapshots only use up a small amount of your free backup storage.

When you need to restore a backup, you have access to all the automated backups within your backup retention window. Once you choose a backup from which to restore, we will provision a new data warehouse cluster and restore your data to it.

Q: How do I manage the retention of my automated backups and snapshots?

You can use the AWS Management Console or ModifyCluster API to manage the period of time your automated backups are retained by modifying the RetentionPeriod parameter. If you wish to turn off automated backups altogether, you can set up the retention period to 0 (not recommended).

Q: What happens to my backups if I delete my data warehouse cluster?

When you delete a data warehouse cluster you have the ability to specify whether a final snapshot is created upon deletion. This enables a restore of the deleted data warehouse cluster at a later date. All previously created manual snapshots of your data warehouse cluster will be retained and billed at standard Amazon S3 rates, unless you choose to delete them.

Scalability

Q: How do I scale the size and performance of my Amazon Redshift data warehouse cluster?

If you would like to increase query performance or respond to CPU, memory or I/O over-utilization, you can increase the number of nodes within your data warehouse cluster using Elastic Resize via the AWS Management Console or the ModifyCluster API. When you modify your data warehouse cluster, your requested changes will be applied immediately. Metrics for compute utilization, storage utilization, and read/write traffic to your Amazon Redshift data warehouse cluster are available free of charge via the AWS Management Console or Amazon CloudWatch APIs. You can also add additional, user-defined metrics via Amazon Cloudwatch custom metric functionality.

With the Concurrency Scaling feature, you can support virtually unlimited concurrent users and concurrent queries, with consistently fast query performance. When concurrency scaling is enabled, Amazon Redshift automatically adds additional cluster capacity when you need it to process an increase in concurrent read queries.

With Redshift Spectrum, you can run multiple Amazon Redshift clusters accessing the same data in Amazon S3. You can use different clusters for different use cases. For example, you can use one cluster for standard reporting and another for data science queries. Your marketing team can use their own clusters different from your operations team. Depending on the type and number of nodes in your local cluster, and the number of files need to be processed for your query, Redshift Spectrum automatically distributes the execution of your query to several Redshift Spectrum workers out of a shared resource pool to read and process data from Amazon S3, and pulls results back into your Amazon Redshift cluster for any remaining processing.

Q: Will my data warehouse cluster remain available during scaling?

It depends. When you using the Concurrency Scaling feature, the cluster is fully available for read and write during concurrency scaling. With Elastic Resize, the cluster is unavailable for 4 to 8 minutes of the resize period. With the Redshift RA3 storage elasticity in managed storage, the cluster is fully available and data is automatically moved between managed storage and compute nodes.

Concurrency

Q: How do I manage resources to ensure that my Redshift cluster can provide consistently fast performance during periods of high concurrency?

A typical data warehouse has significant variance in concurrent query usage over the course of a day. It is more cost-effective to add resources just for the period during which they are required rather than provisioning to peak demand. Amazon Redshift handles this automatically on your behalf.

Concurrency Scaling is a feature in Amazon Redshift that provides consistently fast query performance, even with thousands of concurrent queries. With this feature, Amazon Redshift automatically adds transient capacity when needed to handle heavy demand. Amazon Redshift automatically routes queries to scaling clusters, which are provisioned in seconds and begin processing queries immediately.

This feature is free for most customers. Each Amazon Redshift cluster earns up to one hour of free Concurrency Scaling credits per day. This gives you predictability in your month-to-month cost, even during periods of fluctuating analytical demand.

Q: What is Elastic Resize and how is it different from Concurrency Scaling?

Elastic Resize adds or removes nodes from a single Redshift cluster within minutes to manage its query throughput. For example, an ETL workload for certain hours in a day or month-end reporting may need additional Redshift resources to complete on time. Concurrency Scaling adds additional cluster resources to increase the overall query concurrency.

Q: Can I access the Concurrency Scaling clusters directly?

No. Concurrency Scaling is a massively-scalable pool of Redshift resources, to which customers do not have direct access.

Querying and Analytics

Q: Are Amazon Redshift and Redshift Spectrum compatible with my preferred business intelligence software package and ETL tools?

Amazon Redshift uses industry-standard SQL and is accessed using standard JDBC and ODBC drivers. You can download Amazon Redshift custom JDBC and ODBC drivers from the Connect Client tab of the Redshift Console. We have validated integrations with popular BI and ETL vendors, a number of which are offering free trials to help you get started loading and analyzing your data. You can also go to the AWS Marketplace to deploy and configure solutions designed to work with Amazon Redshift in minutes.

Redshift Spectrum supports all Amazon Redshift client tools. The client tools can continue to connect to the Amazon Redshift cluster endpoint using ODBC or JDBC connections. No changes are required.

You use exactly the same query syntax and have the same query capabilities to access tables in Redshift Spectrum as you have for tables in the local storage of your Redshift cluster. External tables are referenced using the schema name defined in the CREATE EXTERNAL SCHEMA command where they were registered.

Q: What data formats and compression formats does Redshift Spectrum support?

Redshift Spectrum currently supports many open source data formats, including Avro, CSV, Grok, Ion, JSON, ORC, Parquet, RCFile, RegexSerDe, SequenceFile, TextFile, and TSV.

Redshift Spectrum currently supports Gzip and Snappy compression.

Q: What happens if a table in my local storage has the same name as an external table?

Just like with local tables, you can use the schema name to pick exactly which one you mean by using schema_name.table_name in your query.

Q: I use a Hive Metastore to store metadata about my S3 data lake. Can I use Redshift Spectrum?

Yes. The CREATE EXTERNAL SCHEMA command supports Hive Metastores. We do not currently support DDL against the Hive Metastore.

Q: How do I get a list of all external database tables created in my cluster?

You can query the system table SVV_EXTERNAL_TABLES to get that information.

Monitoring

Q: How do I monitor the performance of my Amazon Redshift data warehouse cluster?

Metrics for compute utilization, storage utilization, and read/write traffic to your Amazon Redshift data warehouse cluster are available free of charge via the AWS Management Console or Amazon CloudWatch APIs. You can also add additional, user-defined metrics via Amazon Cloudwatch's custom metric functionality. The AWS Management Console provides a monitoring dashboard that helps you monitor the health and performance of all your clusters. Amazon Redshift also provides information on query and cluster performance via the AWS Management Console. This information enables you to see which users and queries are consuming the most system resources and diagnose performance issues by viewing query plans and execution statistics. In addition, you can see the resource utilization on each of your compute nodes to ensure that you have data and queries that are well balanced across all nodes.

Maintenance

Q: What is a maintenance window? Will my data warehouse cluster be available during software maintenance?

Amazon Redshift periodically performs maintenance to apply fixes, enhancements and new features to your cluster. You can change the scheduled maintenance windows by modifying the cluster, either programmatically or by using the Redshift Console. During these maintenance windows, your Amazon Redshift cluster is not available for normal operations. For more information about maintenance windows and schedules by region, see Maintenance Windows in the Amazon Redshift Management Guide.

Amazon QuickSight FAQs

General

Q: What is Amazon QuickSight? >>

Q: How is Amazon QuickSight different from traditional Business Intelligence (BI) solutions?>>

Q: What is SPICE? >>

Q: How can I get started with Amazon QuickSight? >>

Q: Are there any pre-requisites to use SageMaker inferencing in QuickSight? >>

Authors and Readers

Q: Who is a QuickSight “Author”? >>

Q: Who is a QuickSight “Reader”? >>

Q: Can I upgrade a Reader to an Author? >>

Q: I have a Standard Edition account. Can I add Readers? >>

Q: Can I use the QuickSight Reader account for programmatic access to QuickSight? >>

Q: Who is a QuickSight “Admin”? >>

Q: Can I upgrade a Reader or Author to Admin? >>

Q: How long is a Reader session? >>

Q: When does a Reader session start and end? >>

Q: Will a Reader be logged out after a 30-minute session? >>

Q: Will a Reader be charged if QuickSight is open in a browser in a background tab? >>

Q: What does up to $5/mo. on reader charges mean? >>

Q: Can Qucksight “Authors” or “Readers” invite more users? >>

Q: Can I use the QuickSight Reader account for display and scripted refresh of QuickSight dashboards on monitors or large screens? >>

Mobile and web access

Q: Can I use Amazon QuickSight on my mobile device? >>

Q: On which browsers is Amazon QuickSight supported? >>

Data management

Q: Which data sources does Amazon QuickSight support? >>

Q: Can I connect Amazon QuickSight to my Amazon EC2 or on-premises database? >>

Q: How do I upload my data files into Amazon QuickSight? >>

Q: How do I access my data in AWS data sources? >>

Q: My source data is not in a clean format. How do I format and transform the data before visualizing? >>

Q: How much data can I analyze with Amazon QuickSight? >>

Q: How does QuickSight's integration with SageMaker work? >>

Q: Does QuickSight leverage SageMaker models to perform inference on incremental data or the full data every time it runs? >>

User management

Q: How do I manage user access for Amazon QuickSight? >>

Visualization and analysis

Q: How do I create an analysis with Amazon QuickSight? >>

Q: What is AutoGraph? >>

Q: How do I create a dashboard? >>

Q: What types of visualizations are supported in Amazon QuickSight? >>

Q: What is a suggested visualization? How does Amazon QuickSight generate suggestions? >>

Q: What are Stories? >>

Q: What type of calculations does Amazon QuickSight enable? >>

Q: How can I get sample data to explore in QuickSight? >>

Security and access

Q: How is data transmitted to Amazon QuickSight? >>

Q: Can I choose the AWS region to connect to hosted or on-premises databases over JDBC/ODBC? >>

Q: Does Amazon QuickSight support multi-factor authentication?

Q: How do I connect my VPC to Amazon QuickSight? >>

Q: What is row-level security? >>

Q: What does private VPC access in the context of Amazon QuickSight mean? >>

Sharing

Q: How do I share an analysis, dashboard, or story in Amazon QuickSight? >>

Upgrades and Downgrades

Q: Can I upgrade from Standard Edition to Enterprise Edition? >>

Q: Can I downgrade from Enterprise Edition to Standard Edition? >>

Keep in touch

To stay in touch with all of the new and innovative QuickSight features in the coming months, sign up for the QuickSight update email.

AWS Data Exchange FAQs

General

Who are the primary users of AWS Data Exchange?

AWS Data Exchange makes it easy for AWS customers to securely exchange and use third-party data in AWS. Data analysts, product managers, portfolio managers, data scientists, quants, clinical trial technicians, and developers in nearly every industry would like access to more data to drive analytics, train machine-learning models, and make data-driven decisions. But there is no one place to find data from multiple providers and no consistency in how providers deliver data, leaving them to deal with a mix of shipped physical media, FTP credentials, and bespoke API calls. Conversely, many organizations would like to make their data available for either research or commercial purposes, but it’s too hard and expensive to build and maintain data delivery, entitlement, and billing technology, which further depresses the supply of valuable data.

What AWS regions is AWS Data Exchange available in?

AWS Data Exchange has a single, globally available product catalog offered by providers. You can see the same catalog regardless of which region you are using. The resources underlying the product (data sets, revisions, assets) are regional resources that you manage programmatically or through the AWS Data Exchange console in specific AWS Regions. See the AWS Regional Availability Table for a list of AWS Regions in which AWS Data Exchange is available today.

Subscriber

What type of data can I subscribe to in AWS Data Exchange?

Today, AWS Data Exchange contains more than 1,000 data products from a broad range of domains, including financial services (e.g. top US businesses by revenue), healthcare and life sciences (e.g. population health management), geospatial (e.g. satellite imagery), weather (e.g. historical and future trajectories of temperature), and mapping (e.g. street level imagery and foot traffic patterns). For a complete list of data providers, see the AWS Data Exchange catalog. If the data you are looking for is not available and you would like us to source it, please contact customer support.

How can I see the catalog of AWS Data Exchange products?

Anyone can browse the AWS Data Exchange catalog in AWS Marketplace under the "Data" category, or by searching AWS Marketplace for keywords of interest. Authenticated AWS customers can also browse the AWS Data Exchange catalog alongside existing subscriptions in the AWS Data Exchange console. For more details, see Getting Started as a Subscriber.

How will I know when new revisions to data sets I'm subscribed to become available?

As a subscriber with an active subscription to a product, you will receive a CloudWatch event from AWS Data Exchange everytime new revision(s) are published by the provider. You can use this CloudWatch Event to automate consumption of new data. To learn more, see Logging and monitoring on AWS Data Exchange.

Can I migrate pre-existing data subscriptions to be delivered by AWS Data Exchange?

Yes. AWS Data Exchange allows qualified data providers to fulfill to existing subscribers using a "Bring-Your-Own-Subscription (BYOS)" entitlement at no additional cost, as long as the data provider makes the same product publicly available to other AWS customers. Using BYOS, the existing billing relationship will continue between you and the data provider. Talk to your data provider to leverage this capability.

What are the subscription durations that are available for AWS Data Exchange products?

Data providers list products subscriptions that range from 1 to 36 months. Subscription duration options can be found on each product's detail page.

Can I set my subscription to auto-renew?

Subscriptions to public offers auto-renew by default, though you can disable auto-renew at the time of subscribing or change auto-renewal settings at any time in the "Subscriptions" section of AWS Data Exchange's console. Private offers made exclusively to an individual subscriber do not auto-renew.

Can data providers change the terms of the offer that I am subscribed to? How would it affect my subscription and renewal?

Yes. Data providers can update the terms of the offer at any time but doing so will not affect existing subscriptions. For subscriptions set to auto-renew, AWS Data Exchange will automatically renew the subscription at the latest terms that the provider specified on or by renewal date, **which may be different from the original subscription terms**. For more information, see Product Subscriptions.

How are refunds handled?

AWS Data Exchange requires data providers to specify their refund policy, which you can see on the subscription details page. For any refund requests, you will need to contact the provider directly. Once a provider approves a refund request, AWS will process and issue the approved refund.

How do I know that data I subscribe to is free of any malware?

Security and Compliance is a shared responsibility between AWS and the customer. To promote a safe, secure, and trustworthy service for everyone, AWS Data Exchange scans data published by providers before making it available to subscribers. If AWS detects malware, AWS will remove the affected file(s). AWS Data Exchange does not guarantee that the data you consume as a subscriber is free of any potential malware. Customers are encouraged to conduct their own additional due-diligence to ensure compliance with their internal security controls. You can explore many third-party anti-malware and security products in AWS Marketplace.

How do I remain compliant with applicable data privacy laws when subscribing to AWS Data Exchange products?

Security and Compliance is a shared responsibility among AWS, the data provider, and the subscriber. AWS Data Exchange is responsible for facilitating data transactions and promoting transparency around data category restrictions. Detailed restrictions around eligible data sets and other related legal compliance matters are set forth in Terms and Conditions for AWS Marketplace Providers which every data provider must agree to before listing any data products. If AWS learns that these terms are breached in any way, AWS will remove such content from AWS Data Exchange and the data provider may be suspended from the service.

Providers and subscribers are responsible for conducting their own additional due-diligence to ensure compliance with any data privacy laws. If you suspect that a data product or AWS Data Exchange resources are being used for abusive or illegal purposes, you can report it using AWS's abuse report form.

Are there are any restrictions for how AWS Data Exchange and any data obtained through AWS Data Exchange can be used?

Yes, AWS explicitly prohibits the use of AWS Data Exchange for any illegal or fraudulent activities. Data may not be used for any activities that result in the violation of an individual's rights or unlawfully discriminate against others based on race, ethnicity, sexual orientation, gender identity, or other related groups. Subscribers may not use any content obtained through AWS Data Exchange that was anonymized and/or aggregated by the data provider to create, generate, or infer any information relating to a person's identity (e.g. attempting to triangulate with other data sources).

If AWS learns that these terms are breached in any way, AWS may remove the subscriber's access to the data product and the subscriber may be suspended from future use of AWS Data Exchange. If you suspect that a data product or AWS Data Exchange resources are being used for abusive or illegal purposes, you can report it using AWS's abuse report form and we will investigate.

Further details on these restrictions can be found in the AWS Service Terms.

How do I report abusive content and/or request information be removed from a product suspected of abuse?

If you suspect that a data product or AWS Data Exchange resources are being used for abusive or illegal purposes, you can report it using AWS's abuse report form and we will investigate. If AWS learns that our terms are breached in any way, AWS may remove the subscriber's access to the data product and the data provider or the subscriber may be suspended or terminated from future use of AWS Data Exchange.

Provider

How do I qualify to become a data provider on AWS Data Exchange?

To become a data provider on AWS Data Exchange, data providers must agree to the Terms and Conditions for AWS Marketplace Providers ("AWS Marketplace Terms & Conditions"). Data providers must use a valid legal entity domiciled in the United States or a member state of the EU, supply valid banking and taxation identification, and be qualified by the AWS Data Exchange business operations team. Each data provider will also undergo a detailed review by the AWS Data Exchange team prior to being granted permission to list data products on the catalog.

How is data organized in AWS Data Exchange?

Data in AWS Data Exchange is organized using three building blocks – data sets, revisions, and assets. A data set is container for data that belongs together (e.g. end of data pricing for equities trading in the U.S.). Data sets contain a series of revisions, which data providers publish as needed to make new assets available. Revisions can represent changes or new data (e.g. today's end of day prices), corrections to previous revisions, or entirely new snapshots. Assets are any file that can be stored in Amazon S3 (e.g. CSV, parquet, images). For more details, see Working with Data Sets. Using these building blocks, you can organize the data any way you choose, whether hierarchically to build complex data models or as single data files.

After I create data sets, how do I publish and make them available to my subscribers?

Data sets are made available to subscribers as part of a product. A product is a collection of one or more data sets, metadata that makes the product discoverable in the AWS Data Exchange catalog, pricing, and a Data Subscription Agreement that subscribers must agree to before subscribing. For more information see Publishing Products.

Can I choose which customers can subscribe to my data?

Yes. You have an option to enable subscription verification on any product, which will require prospective subscribers to fill out a subscription request form including their identity and intended use-case details before subscribing. For these products, you will have up to 45 days to either approve or decline the subscription request. For more details, see Subscription Verification for Providers.

Do I have to package files in a specific format?

AWS Data Exchange allows you to package files in any file format, though you should consider organizing the file and format in a way that allows subscribers to gain insight from the data in a more convenient manner. Parquet formatted files, for example, will allow subscribers to instantly run ad-hoc queries using Amazon Athena in a cost-effective way. Binary or other proprietary file formats will require the subscriber understands how to parse the information, which AWS recommends explaining in each product description.

Who owns the data I am distributing as a provider through AWS Data Exchange?

You retain ownership of the data you distribute as a data provider on AWS Data Exchange. The AWS Marketplace Terms and Conditions require each data provider to attest that they have the legal right to distribute the data they publish. Subscribers must legally agree to the Data Subscription Agreement specified by the data provider before gaining access to data sets contained in a product, which remains available for both data providers and subscribers. Consistent with Amazon Web Services' acceptable use policy, AWS Data

Exchange may suggest corrective action where there is evidence of abuse, but it is the data provider's responsibility to enforce and govern the terms of use.

How do I specify the Data Subscription Agreement (DSA)?
AWS Data Exchange provides an optional Data Subscription Agreement (DSA) template that incorporates inputs from multiple AWS customers and data providers. You can choose to use this DSA template, copy and edit it with their own terms and conditions, or specify custom terms by uploading a DSA of their choice. AWS Data Exchange will associate the DSA specified for the product without any further modifications. See Publishing Products for further information.

What ways can I price my data sets?

AWS Data Exchange currently supports subscription-based pricing from 1 to 36-month duration terms.

Is AWS Data Exchange suitable for data providers who want to distribute their data for free?

Yes. Many data providers make their data products available for free for research, scientific, or other non-commercial use cases.

Can I customize pricing or terms for select customers?

Yes. "Private offers" allows you to make public products available to select AWS customers for a custom price, duration, and/or Data Subscription Agreement (DSA). To learn more, see Create Private Offers.

Are there any restrictions on what data can be made available on AWS Data Exchange?

Yes. AWS Data Exchange's Product Publishing Guidelines and Terms and Conditions for AWS Marketplace Providers restrict certain categories of data. Data products listed on AWS Data Exchange may not include information that can be used to identify any person, unless that information is already legally available to the public. Examples of this include newspaper articles, open court records, public company filings, or public online profiles.

The following categories of information must be aggregated and/or anonymized so that no person in your data product can be identified: biometric or genetic data, health, racial or ethnic origin, political opinions, religious or philosophical beliefs, sex or sexual orientation, trade union membership, personal payment or financial information (e.g., credit history), or other similar categories of sensitive information. Providers must also include an acknowledgement in their product description that their product contains aggregated and/or anonymized data when applicable.

Examples of data sets that can be included on AWS Data Exchange: (1) Historic stock prices for public companies, (2) Names of judges and their court opinions, and (3) Aggregated and/or anonymized research findings from pharmaceutical drug studies.

Examples of data sets that are prohibited on AWS Data Exchange: (1) Lists of names organized by race, (2) Geo-location data that can be used to identify a person, and (3) protected health information under HIPAA.

Can I remove a product that I published from the catalog?

Yes. You can un-publish a product at any time. Un-publishing a product ensures that no new subscribers are able to view and subscribe to your product, including auto-renewal cancellation for existing subscribers. You will need to keep data current for any existing subscribers until each subscription expires.

What happens if I need to remove data from AWS Data Exchange?
You can remove or change the price or Data Subscription Agreement (DSA) of a product at any time, although existing subscriptions will remain in effect until their next renewal. If a data provider erroneously publishes data, you can open a support case here to have the data un-published.

How do I know who is subscribing to the data I have listed on AWS Data Exchange?

AWS Data Exchange provides daily, weekly, and monthly reports detailing subscription activity.

When/how often will I receive payments?

Data providers will receive a disbursement for subscriptions less fulfillment fees once a month. AWS will disburse all funds that AWS has received from subscribers by that date to the bank account linked to the data provider's AWS account used at registration.

How will AWS handle collection and remittances of U.S. Sales and Use Tax?

When listing your data sets, you can enable collection and remittance of U.S. Sales and Use tax. Data providers can also configure their tax nexus to account for places where you have a physical presence to direct AWS to collect appropriate taxes. Please review the AWS Marketplace U.S Tax Collection Support Terms and Conditions. For details on sales tax collection in other geographies, see AWS Marketplace Sellers & Tax Collection.

Is Amazon.com or AWS providing customers' data on AWS Data Exchange?

No. Neither Amazon.com nor AWS are providing customers' data on AWS Data Exchange.

Can Amazon access data products listed on AWS Data Exchange?

Amazon.com and AWS can subscribe to a data product like any other customer and must agree to the restrictions in a provider's Data Subscription Agreement. Otherwise, AWS will only access data products as needed to provide the AWS Data Exchange service.

AWS Data Pipeline FAQs

General

Q: What is AWS Data Pipeline?

AWS Data Pipeline is a web service that makes it easy to schedule regular data movement and data processing activities in the AWS cloud. AWS Data Pipeline integrates with on-premise and cloud-based storage systems to allow developers to use their data when they need it, where they want it, and in the required format. AWS Data Pipeline allows you to quickly define a dependent chain of data sources, destinations, and predefined or custom data processing activities called a pipeline. Based on a schedule you define, your pipeline regularly performs processing activities such as distributed data copy, SQL transforms, MapReduce applications, or custom scripts against destinations such as Amazon S3, Amazon RDS, or Amazon DynamoDB. By executing the scheduling, retry, and failure logic for these workflows as a highly scalable and fully managed service, Data Pipeline ensures that your pipelines are robust and highly available.

Q: What can I do with AWS Data Pipeline?

Using AWS Data Pipeline, you can quickly and easily provision pipelines that remove the development and maintenance effort required to manage your daily data operations, letting you focus on generating insights from that data. Simply specify the data sources, schedule, and processing activities required for your data pipeline. AWS Data Pipeline handles running and monitoring your processing activities on a highly reliable, fault-tolerant infrastructure. Additionally, to further ease your development process, AWS Data Pipeline provides built-in activities for common actions such as copying data between Amazon Amazon S3 and Amazon RDS, or running a query against Amazon S3 log data.

Q: How is AWS Data Pipeline different from Amazon Simple Workflow Service?

While both services provide execution tracking, handling retries and exceptions, and running arbitrary actions, AWS Data Pipeline is specifically designed to facilitate the specific steps that are common across a majority of data-driven workflows. For example: executing activities after their input data meets specific readiness criteria, easily copying data between different data stores, and scheduling chained transforms. This highly specific focus means that Data Pipeline workflow definitions can be created rapidly, and with no code or programming knowledge.

Q: What is a pipeline?

A pipeline is the AWS Data Pipeline resource that contains the definition of the dependent chain of data sources, destinations, and predefined or custom data processing activities required to execute your business logic.

Q: What is a data node?

A data node is a representation of your business data. For example, a data node can reference a specific Amazon S3 path. AWS Data Pipeline supports an expression language that makes it easy to reference data which is generated on a regular basis. For example, you could specify that your Amazon S3 data format is s3://example-bucket/my-logs/logdata-#{scheduledStartTime('YYYY-MM-dd-HH')}.tgz.

Q: What is an activity?

An activity is an action that AWS Data Pipeline initiates on your behalf as part of a pipeline. Example activities are EMR or Hive jobs, copies, SQL queries, or command-line scripts.

Q: What is a precondition?

A precondition is a readiness check that can be optionally associated with a data source or activity. If a data source has a precondition check, then that check must complete successfully before any activities consuming the data source are launched. If an activity has a precondition, then the precondition check must complete successfully before the activity is run. This can be useful if you are running an activity that is expensive to compute, and should not run until specific criteria are met.

Q: What is a schedule?

Schedules define when your pipeline activities run and the frequency with which the service expects your data to be available. All schedules must have a start date and a frequency; for example, every day starting Jan 1, 2013, at 3pm. Schedules can optionally have an end date, after which time the AWS Data Pipeline service does not execute any activities. When you associate a schedule with an activity, the activity runs on it. When you associate a schedule with a data source, you are telling the AWS Data Pipeline service that you expect the data to be updated on that schedule. For example, if you define an Amazon S3 data source with an hourly schedule, the service expects that the data source contains new files every hour.

Back to top »

Functionality

Q: Does Data Pipeline supply any standard Activities?

Yes, AWS Data Pipeline provides built-in support for the following activities:

- CopyActivity: This activity can copy data between Amazon S3 and JDBC data sources, or run a SQL query and copy its output into Amazon S3.
- HiveActivity: This activity allows you to execute Hive queries easily.
- EMRActivity: This activity allows you to run arbitrary Amazon EMR jobs.
- ShellCommandActivity: This activity allows you to run arbitrary Linux shell commands or programs.

Q: Does AWS Data Pipeline supply any standard preconditions?

Yes, AWS Data Pipeline provides built-in support for the following preconditions:

- DynamoDBDataExists: This precondition checks for the existence of data inside a DynamoDB table.
- DynamoDBTableExists: This precondition checks for the existence of a DynamoDB table.
- S3KeyExists: This precondition checks for the existence of a specific AmazonS3 path.
- S3PrefixExists: This precondition checks for at least one file existing within a specific path.

- ShellCommandPrecondition: This precondition runs an arbitrary script on your resources and checks that the script succeeds.

Q: Can I supply my own custom activities?

Yes, you can use the ShellCommandActivity to run arbitrary Activity logic.

Q: Can I supply my own custom preconditions?

Yes, you can use the ShellCommandPrecondition to run arbitrary precondition logic.

Q: Can you define multiple schedules for different activities in the same pipeline?

Yes, simply define multiple schedule objects in your pipeline definition file and associate the desired schedule to the correct activity via its schedule field. This allows you to define a pipeline in which, for example, log files are stored in Amazon S3 each hour to drive generation of an aggregate report one time per day.

Q: What happens if an activity fails?

An activity fails if all of its activity attempts return with a failed state. By default, an activity retries three times before entering a hard failure state. You can increase the number of automatic retries to 10; however, the system does not allow indefinite retries. After an activity exhausts its attempts, it triggers any configured onFailure alarm and will not try to run again unless you manually issue a rerun command via the CLI, API, or console button.

Q: How do I add alarms to an activity?

You can define Amazon SNS alarms to trigger on activity success, failure, or delay. Create an alarm object and reference it in the onFail,onSuccess, or onLate slots of the activity object.

Q: Can I manually rerun activities that have failed?

Yes. You can rerun a set of completed or failed activities by resetting their state to SCHEDULED. This can be done by using the Rerun button in the UI or modifying their state in the command line or API. This will immediately schedule a of re-check all activity dependencies, followed by the execution of additional activity attempts. Upon subsequent failures, the Activity will perform the original number of retry attempts.

Q: On what resources are activities run?

AWS Data Pipeline activities are run on compute resources that you own. There are two types of compute resources: AWS Data Pipeline–managed and self-managed. AWS Data Pipeline–managed resources are Amazon EMR clusters or Amazon EC2 instances that the AWS Data Pipeline service launches only when they're needed. Resources that you manage are longer running and can be any resource capable of running the AWS Data Pipeline Java-based Task Runner (on-premise hardware, a customer-managed Amazon EC2 instance, etc.).

Q: Will AWS Data Pipeline provision and terminate AWS Data Pipeline-managed compute resources for me?

Yes, compute resources will be provisioned when the first activity for a scheduled time that uses those resources is ready to run and those instances will be terminated when the final activity that uses the resources has completed successfully or failed.

Q: Can multiple compute resources be used on the same pipeline?

Yes, simply define multiple cluster objects in your definition file and associate the cluster to use for each activity via its runsOn field. This allows pipelines to combine AWS and on-premise resources, or to use a mix of instance types for their activities – for example, you may want to use a t1.micro to execute a quick script cheaply, but later on the pipeline may have an Amazon EMR job that requires the power of a cluster of larger instances.

Q: Can I execute activities on on-premise resources, or AWS resources that I manage?

Yes. To enable running activities using on-premise resources, AWS Data Pipeline supplies a Task Runner package that can be installed on your on-premise hosts. This package continuously polls the AWS Data Pipeline service for work to perform. When it's time to run a particular activity on your on-premise resources, for example, executing a DB stored procedure or a database dump, AWS Data Pipeline will issue the appropriate command to the Task Runner. In order to ensure that your pipeline activities are highly available, you can optionally assign multiple Task Runners to poll for a given job. This way, if one Task Runner becomes unavailable, the others will simply pick up its work.

Q: How do I install a Task Runner on my on-premise hosts?

You can install the Task Runner package on your on-premise hosts using the following steps:

1. Download the AWS Task Runner package.
2. Create a configuration file that includes your AWS credentials.
3. Start the Task Runner agent via the following command:
   java -jar TaskRunner-1.0.jar --config ~/credentials.json --workerGroup=[myWorkerGroup]
4. When defining activities, set the activity to run on [myWorkerGroup] in order to dispatch them to the previously installed hosts.

Back to top »

Getting Started

Q: How can I get started with AWS Data Pipeline?

To get started with AWS Data Pipeline, simply visit the AWS Management Console and go to the AWS Data Pipeline tab. From there, you can create a pipeline using a simple graphical editor.

Q: What can I do with AWS Data Pipeline?

With AWS Data Pipeline, you can schedule and manage periodic data-processing jobs. You can use this to replace simple systems which are current managed by brittle, cron-based solutions, or you can use it to build complex, multi-stage data processing jobs.

Q: Are there Sample Pipelines that I can use to try out AWS Data Pipeline?

Yes, there are sample pipelines in our documentation. Additionally, the console has several pipeline templates that you can use to get started.

Back to top »

Limits

Q: How many pipelines can I create in AWS Data Pipeline?

By default, your account can have 100 pipelines.

Q: Are there limits on what I can put inside a single pipeline?

By default, each pipeline you create can have 100 objects.

Q: Can my limits be changed?

Yes. If you would like to increase your limits, simply contact us.

Back to top »

Billing

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Back to top »

AWS Data Pipeline

Product Details

Pricing

Developer Resources

FAQs

RELATED LINKS

Documentation

Management Console

Release Notes

Discussion Forum

AWS Glue FAQs

General

Q: What is AWS Glue?

AWS Glue is a fully-managed, pay-as-you-go, extract, transform, and load (ETL) service that automates the time-consuming steps of data preparation for analytics. AWS Glue automatically discovers and profiles your data via the Glue Data Catalog, recommends and generates ETL code to transform your source data into target schemas, and runs the ETL jobs on a fully managed, scale-out Apache Spark environment to load your data into its destination. It also allows you to setup, orchestrate, and monitor complex data flows.

Q: How do I get started with AWS Glue?

To start using AWS Glue, simply sign into the AWS Management Console and navigate to “Glue” under the “Analytics” category. You can follow one of our guided tutorials that will walk you through an example use case for AWS Glue. You can also find sample ETL code in our GitHub repository under AWS Labs.

Q. What are the main components of AWS Glue?

AWS Glue consists of a Data Catalog which is a central metadata repository, an ETL engine that can automatically generate Scala or Python code, and a flexible scheduler that handles dependency resolution, job monitoring, and retries. Together, these automate much of the undifferentiated heavy lifting involved with discovering, categorizing, cleaning, enriching, and moving data, so you can spend more time analyzing your data.

Q: When should I use AWS Glue?

You should use AWS Glue to discover properties of the data you own, transform it, and prepare it for analytics. Glue can automatically discover both structured and semi-structured data stored in your data lake on Amazon S3, data warehouse in Amazon Redshift, and various databases running on AWS. It provides a unified view of your data via the Glue Data Catalog that is available for ETL, querying and reporting using services like Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum. Glue automatically generates Scala or Python code for your ETL jobs that you can further customize using tools you are already familiar with. AWS Glue is serverless, so there are no compute resources to configure and manage.

Q: What data sources does AWS Glue support?

AWS Glue natively supports data stored in Amazon Aurora, Amazon RDS for MySQL, Amazon RDS for Oracle, Amazon RDS for PostgreSQL, Amazon RDS for SQL Server, Amazon Redshift, and Amazon S3, as well as MySQL, Oracle, Microsoft SQL Server, and PostgreSQL databases in your Virtual Private Cloud (Amazon VPC) running on Amazon EC2. The metadata stored in the AWS Glue Data Catalog can be readily accessed from Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum. You can also write custom Scala or Python code and import custom libraries and Jar files into your Glue ETL jobs to access data sources not natively supported by AWS Glue. For more details on importing custom libraries, refer to our documentation.

Q: How does AWS Glue relate to AWS Lake Formation?

A: Lake Formation leverages a shared infrastructure with AWS Glue, including console controls, ETL code creation and job monitoring, a common data catalog, and a serverless architecture. While AWS Glue is still focused on these types of functions, Lake Formation encompasses all AWS Glue features AND provides additional capabilities designed to help build, secure, and manage a data lake. See the AWS Lake Formation pages for more details.

AWS Glue Data Catalog

Q: What is the AWS Glue Data Catalog?

The AWS Glue Data Catalog is a central repository to store structural and operational metadata for all your data assets. For a given data set, you can store its table definition, physical location, add business relevant attributes, as well as track how this data has changed over time.

The AWS Glue Data Catalog is Apache Hive Metastore compatible and is a drop-in replacement for the Apache Hive Metastore for Big Data applications running on Amazon EMR. For more information on setting up your EMR cluster to use AWS Glue Data Catalog as an Apache Hive Metastore, click here.

The AWS Glue Data Catalog also provides out-of-box integration with Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum. Once you add your table definitions to the Glue Data Catalog, they are available for ETL and also readily available for querying in Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum so that you can have a common view of your data between these services.

Q: How do I get my metadata into the AWS Glue Data Catalog?

AWS Glue provides a number of ways to populate metadata into the AWS Glue Data Catalog. Glue crawlers scan various data stores you own to automatically infer schemas and partition structure and populate the Glue Data Catalog with corresponding table definitions and statistics. You can also schedule crawlers to run periodically so that your metadata is always up-to-date and in-sync with the underlying data. Alternately, you can add and update table details manually by using the AWS Glue Console or by calling the API. You can also run Hive DDL statements via the Amazon Athena Console or a Hive client on an Amazon EMR cluster. Finally, if you already have a persistent Apache Hive Metastore, you can perform a bulk import of that metadata into the AWS Glue Data Catalog by using our import script.

Q: What are AWS Glue crawlers?

An AWS Glue crawler connects to a data store, progresses through a prioritized list of classifiers to extract the schema of your data and other statistics, and then populates the Glue Data Catalog with this metadata. Crawlers can run periodically to detect the availability of new data as well as changes to existing data, including table definition changes. Crawlers automatically add new tables, new partitions to existing table, and new versions of table definitions. You can customize Glue crawlers to classify your own file types.

Q: How do I import data from my existing Apache Hive Metastore to the AWS Glue Data Catalog?

You simply run an ETL job that reads from your Apache Hive Metastore, exports the data to an intermediate format in Amazon S3, and then imports that data into the AWS Glue Data Catalog.

Q: Do I need to maintain my Apache Hive Metastore if I am storing my metadata in the AWS Glue Data Catalog?

No. AWS Glue Data Catalog is Apache Hive Metastore compatible. You can point to the Glue Data Catalog endpoint and use it as an Apache Hive Metastore replacement. For more information on how to configure your cluster to use AWS Glue Data Catalog as an Apache Hive Metastore, please read our documentation here.

Q. If I am already using Amazon Athena or Amazon Redshift Spectrum and have tables in Amazon Athena's internal data catalog, how can I start using the AWS Glue Data Catalog as my common metadata repository?

Before you can start using AWS Glue Data Catalog as a common metadata repository between Amazon Athena, Amazon Redshift Spectrum, and AWS Glue, you must upgrade your Amazon Athena data catalog to AWS Glue Data Catalog. The steps required for the upgrade are detailed here.

Extract, transform, and load (ETL)

Q: What programming language can I use to write my ETL code for AWS Glue?

You can use either Scala or Python.

Q: How can I customize the ETL code generated by AWS Glue?

AWS Glue’s ETL script recommendation system generates Scala or Python code. It leverages Glue’s custom ETL library to simplify access to data sources as well as manage job execution. You can find more details about the library in our documentation. You can write ETL code using AWS Glue’s custom library or write arbitrary code in Scala or Python by using inline editing via the AWS Glue Console script editor, downloading the auto-generated code, and editing it in your own IDE. You can also start with one of the many samples hosted in our Github repository and customize that code.

Q: Can I import custom libraries as part of my ETL script?

Yes. You can import custom Python libraries and Jar files into your AWS Glue ETL job. For more details, please check our documentation here.

Q: Can I bring my own code?

Yes. You can write your own code using AWS Glue’s ETL library, or write your own Scala or Python code and upload it to a Glue ETL job. For more details, please check our documentation here.

Q: How can I develop my ETL code using my own IDE?

You can create and connect to development endpoints that offer ways to connect your notebooks and IDEs.

Q: How can I build end-to-end ETL workflow using multiple jobs in AWS Glue?

In addition to the ETL library and code generation, AWS Glue provides a robust set of orchestration features that allow you to manage dependencies between multiple jobs to build end-to-end ETL workflows. AWS Glue ETL jobs can either be triggered on a schedule or on a job completion event. Multiple jobs can be triggered in parallel or sequentially by triggering them on a job completion event. You can also trigger one or more Glue jobs from an external source such as an AWS Lambda function.

Q: How does AWS Glue monitor dependencies?

AWS Glue manages dependencies between two or more jobs or dependencies on external events using triggers. Triggers can watch one or more jobs as well as invoke one or more jobs. You can either have a scheduled trigger that invokes jobs periodically, an on-demand trigger, or a job completion trigger.

Q: How does AWS Glue handle errors?

AWS Glue monitors job event metrics and errors, and pushes all notifications to Amazon CloudWatch. With Amazon CloudWatch, you can configure a host of actions that can be triggered based on specific notifications from AWS Glue. For example, if you get an error or a success notification from Glue, you can trigger an AWS Lambda function. Glue also provides default retry behavior that will retry all failures three times before sending out an error notification.

Q: Can I run my existing ETL jobs with AWS Glue?

Yes. You can run your existing Scala or Python code on AWS Glue. Simply upload the code to Amazon S3 and create one or more jobs that use that code. You can reuse the same code across multiple jobs by pointing them to the same code location on Amazon S3.

Q: How can I use AWS Glue to ETL streaming data?

AWS Glue ETL is batch oriented, and you can schedule your ETL jobs at a minimum of 5 min intervals. While it can process micro-batches, it does not handle streaming data. If your use case requires you to ETL data while you stream it in, you can perform the first leg of your ETL using Amazon Kinesis Data Firehose or Amazon Kinesis Data Analytics, and then store data to either Amazon S3 or Amazon Redshift and trigger a Glue ETL job to pick up that dataset and continue applying additional transformations to that data.

Q: Do I have to use both AWS Glue Data Catalog and Glue ETL to use the service?

No. While we do believe that using both the AWS Glue Data Catalog and ETL provides an end-to-end ETL experience, you can use either one of them independently without using the other.

Clean and deduplicate data

Q: What kind of problems does the FindMatches ML Transform solve?

FindMatches generally solves Record Linkage and Data Deduplication problems. Deduplication is what you have to do when you are trying to identify records in a database which are conceptually “the same”, but for which you have separate records. This problem is trivial if duplicate records can be identified by a unique key (for instance if products can be uniquely identified by a UPC Code), but becomes very challenging when you have to do a “fuzzy match”.

Record linkage is basically the same problem as data deduplication under the hood, but this term usually means that you are doing a “fuzzy join” of two databases that do not share a unique key rather than deduplicating a single database. As an example, consider the problem of matching a large database of customers to a small database of known fraudsters. FindMatches can be used on both record linkage and deduplication problems.

For instance, AWS Glue's FindMatches ML Transform can help you with the following problems:

Linking patient records between hospitals so that doctors have more background information and are better able to treat patients by using FindMatches on separate databases that both contain common fields such as name, birthday, home address, phone number, etc.

Deduplicating a database of movies containing columns like “title”, “plot synopsis”, “year of release”, “run time”, and “cast”. For instance, the same movie might be variously identified as “Star Wars”, “Star Wars: A New Hope”, and “Star Wars: Episode IV—A New Hope (Special Edition)”.

Automatically group all related products together in your storefront by identifying equivalent items in an apparel product catalog where you want to define “equivalent” to mean that they are the same ignoring differences in size and color. Hence “Levi 501 Blue Jeans, size 34x34” is defined to be the same as “Levi 501 Jeans--black, Size 32x31”.

Q: How does AWS Glue deduplicate my data?

AWS Glue's FindMatches ML Transform makes it easy to find and link records that refer to the same entity but don't share a reliable identifier. Before FindMatches, developers would commonly solve data-matching problems deterministically, by writing huge numbers of hand-tuned rules. FindMatches uses machine learning algorithms behind the scenes to learn how to match records according to each developer's own business criteria. FindMatches first identifies records for the customer to label as to whether they match or do not match and then uses machine learning to create an ML Transform. Customers can then execute this Transform on their database to find matching records or they can ask FindMatches to give them additional records to label to push their ML Transform to higher levels of accuracy.

Q: What are ML Transforms?

ML Transforms provide a destination for creating and managing machine-learned transforms. Once created and trained, these ML Transforms can then be executed in standard AWS Glue scripts. Customers select a particular algorithm (for example, the FindMatches ML Transform) and input datasets and training examples, and the tuning parameters needed by that algorithm. AWS Glue uses those inputs to build an ML Transform that can be incorporated into a normal ETL Job workflow.

Q: How do ML Transforms work?

AWS Glue includes specialized ML-based dataset transformation algorithms customers can use to create their own ML Transforms. These include record de-duplication and match finding.

Customers start by navigating to the ML Transforms tab in the console (or using the ML Transforms service endpoints or accessing ML Transforms training via CLI) to create their first ML transform model. The ML Transforms tab provides a user-friendly view for management of user transforms. ML Transforms require distinct workflow requirements from other transforms, including the need for separate training, parameter tuning, and execution workflows; the need for estimating the quality metrics of generated transformations; and the need to manage and collect additional truth labels for training and active learning.

To create an ML transform via the console, customers first select the transform type (such as Record Deduplication or Record Matching) and provide the appropriate data sources previously discovered in Data Catalog. Depending on the transform, customers may then be asked to provide ground truth label data for training or additional parameters. Customers can monitor the status of their training jobs and view quality metrics for each transform. (Quality metrics are reported using a hold-out set of the customer-provided label data.)

Once satisfied with the performance, customers can promote ML Transforms models for use in production. ML Transforms can then be used during ETL workflows, both in code autogenerated by the service and in user-defined scripts submitted with other jobs, similar to pre-built transforms offered in other AWS Glue libraries.

Q: Can I see a presentation on using AWS Glue (and AWS Lake Formation) to find matches and deduplicate records?

A: Yes, the full recording of the AWS Online Tech Talk, "Fuzzy Matching and Deduplicating Data with ML Transforms for AWS Lake Formation" is available here.

AWS Product Integrations

Q: When should I use AWS Glue vs. AWS Data Pipeline?

AWS Glue provides a managed ETL service that runs on a serverless Apache Spark environment. This allows you to focus on your ETL job and not worry about configuring and managing the underlying compute resources. AWS Glue takes a data first approach and allows you to focus on the data properties and data manipulation to transform the data to a form where you can derive business insights. It provides an integrated data catalog that makes metadata available for ETL as well as querying via Amazon Athena and Amazon Redshift Spectrum.

AWS Data Pipeline provides a managed orchestration service that gives you greater flexibility in terms of the execution environment, access and control over the compute resources that run your code, as well as the code itself that does data processing. AWS Data Pipeline launches compute resources in your account allowing you direct access to the Amazon EC2 instances or Amazon EMR clusters.

Furthermore, AWS Glue ETL jobs are Scala or Python based. If your use case requires you to use an engine other than Apache Spark or if you want to run a heterogeneous set of jobs that run on a variety of engines like Hive, Pig, etc., then AWS Data Pipeline would be a better choice.

Q: When should I use AWS Glue vs. Amazon EMR?

AWS Glue works on top of the Apache Spark environment to provide a scale-out execution environment for your data transformation jobs. AWS Glue infers, evolves, and monitors your ETL jobs to greatly simplify the process of creating and maintaining jobs. Amazon EMR provides you with direct access to your Hadoop environment, affording you lower-level access and greater flexibility in using tools beyond Spark.

Q: When should I use AWS Glue vs AWS Database Migration Service?

AWS Database Migration Service (DMS) helps you migrate databases to AWS easily and securely. For use cases which require a database migration from on-premises to AWS or database replication between on-premises sources and sources on AWS, we recommend you use AWS DMS. Once your data is in AWS, you can use AWS Glue to move and transform data from your data source into another database or data warehouse, such as Amazon Redshift.

Q: When should I use AWS Glue vs AWS Batch?

AWS Batch enables you to easily and efficiently run any batch computing job on AWS regardless of the nature of the job. AWS Batch creates and manages the compute resources in your AWS account, giving you full control and visibility into the resources being used. AWS Glue is a fully-managed ETL service that provides a serverless Apache Spark environment to run your ETL jobs. For your ETL use cases, we recommend you explore using AWS Glue. For other batch oriented use cases, including some ETL use cases, AWS Batch might be a better fit.

Q: When should I use AWS Glue vs Amazon Kinesis Data Analytics?

Amazon Kinesis Data Analytics allows you to run standard SQL queries on your incoming data stream. You can specify a destination like Amazon S3 to write your results. Once your data is available in your target data source, you can kick off an AWS Glue ETL job to do further transform your data and prepare it for additional analytics and reporting.

Pricing and billing

Q: How am I charged for AWS Glue?

You will pay a simple monthly fee, above the AWS Glue Data Catalog free tier, for storing and accessing the metadata in the AWS Glue Data Catalog. Additionally, you will pay an hourly rate, billed per second, for the ETL job and crawler run, with a 10-minute minimum for each. If you choose to use a development endpoint to interactively develop your ETL code, you will pay an hourly rate, billed per second, for the time your development endpoint is provisioned, with a 10-minute minimum. For more details, please refer our pricing page.

Q: When does billing for my AWS Glue jobs begin and end?

Billing commences as soon as the job is scheduled for execution and continues until the entire job completes. With AWS Glue, you only pay for the time for which your job runs and not for the environment provisioning or shutdown time.

Security and availability

Q: How does AWS Glue keep my data secure?

We provide server side encryption for data at rest and SSL for data in motion.

Q: What are the service limits associated with AWS Glue?

Please refer our documentation to learn more about service limits.

Q: What regions is AWS Glue in?

Please refer to the AWS Region Table for details of AWS Glue service availability by region.

Q: How many DPUs (Data Processing Units) are allocated to the development endpoint?

A development endpoint is provisioned with 5 DPUs by default. You can configure a development endpoint with a minimum of 2 DPUs and a maximun of 5 DPUs.

Q: How do I scale the size and performance of my AWS Glue ETL jobs?

You can simply specify the number of DPUs (Data Processing Units) you want to allocate to your ETL job. A Glue ETL job requires a minimum of 2 DPUs. By default, AWS Glue allocates 10 DPUs to each ETL job.

Q: How do I monitor the execution of my AWS Glue jobs?

The AWS Glue provides status of each job and pushes all notifications to Amazon CloudWatch. You can setup SNS notifications via CloudWatch actions to be informed of job failures or completions.

Service Level Agreement

Q: What does the AWS Glue SLA guarantee?

Our AWS Glue SLA guarantees a Monthly Uptime Percentage of at least 99.9% for AWS Glue.

Q: How do I know if I qualify for a SLA Service Credit?

You are eligible for a SLA credit for AWS Glue under the AWS Glue SLA if more than one Availability Zone in which you are running a task, within the same region has a Monthly Uptime Percentage of less than 99.9% during any monthly billing cycle.

For full details on all of the terms and conditions of the SLA, as well as details on how to submit a claim, please see the AWS Glue SLA details page.

Visit the pricing page

Explore pricing options for AWS Glue.

Learn more »

Sign up for a free account

Instantly get access to the AWS Free Tier.

Sign up »

Start building on the console

Get started building with AWS Glue on the AWS Management Console.

Sign in »

AWS Lake Formation FAQs

General

Q: What is a data lake?

A: A data lake is a scalable central repository of large quantities and varieties of data, both structured and unstructured. Data lakes enable you to manage the full lifecycle of your data. The first step of building a data lake is ingesting and cataloging data from a variety of sources. The data is then enriched, combined, and cleaned before analysis. This makes it easy to discover and analyze the data with direct queries, visualization, and machine learning. Data lakes complement traditional data warehouses, providing more flexibility, cost-effectiveness, and scalability for ingestion, storage, transformation, and analysis of your data. The traditional challenges around the construction and maintenance of data warehouses and limitations in the types of analysis can be overcome using data lakes.

Read more about "What is a data lake?"

Q: What is AWS Lake Formation?

A: Lake Formation is an integrated data lake service that makes it easy for you to ingest, clean, catalog, transform, and secure your data and make it available for analysis and machine learning. Lake Formation gives you a central console where you can discover data sources, set up transformation jobs to move data to an Amazon S3 data lake, remove duplicates and match records, catalog data for access by analytic tools, configure data access and security policies, and audit and control access from AWS analytic and machine learning services. Lake Formation automatically manages access to the registered data in Amazon S3 via services including AWS Glue, Amazon Athena, Amazon Redshift, and (in beta) Amazon EMR Notebooks and Zeppelin notebooks with Apache Spark, to ensure compliance with your defined policies. If you've set up transformation jobs spanning AWS services, Lake Formation configures the flows, centralizes their orchestration, and lets you monitor the execution of your jobs. With Lake Formation, you can configure and manage your data lake without manually integrating multiple underlying AWS services.

Q: Why should I use Lake Formation to build my data lake?

A: Lake Formation makes it easy to build, secure, and manage your AWS data lake. Lake Formation integrates with underlying AWS security, storage, analysis, and machine learning services and automatically configures them to comply with your centrally defined access policies; and gives you a single console to monitor your jobs and data transformation and analytic workflows.

Lake Formation can manage data ingestion via AWS Glue. Data is automatically classified and relevant data definitions, schema, and metadata are stored in the central data catalog. AWS Glue also converts your data to your choice of open data formats to be stored in S3 and cleans your data to remove duplicates and link records across data sets. Once your data is in your S3 data lake, you can define access policies, including table and column level access controls, and enforce encryption for data at rest. You can then use a wide variety of AWS analytic and machine learning services to access your data lake. All access is secured, governed, and auditable.

Q: Can I see a presentation on AWS Lake Formation?

A: Yes, you can watch the full recording of the "Intro to AWS Lake Formation" session from re:Invent.

Q: What kind of problems does the FindMatches ML Transform solve?

A: FindMatches generally solves Record Linkage and Data Deduplication problems. Deduplication is what you have to do when you are trying to identify records in a database which are conceptually “the same”, but for which you have separate records. This problem is trivial if duplicate records can be identified by a unique key (for instance if products can be uniquely identified by a UPC Code), but becomes very challenging when you have to do a “fuzzy match”.

Record linkage is basically the same problem as data deduplication under the hood, but this term usually means that you are doing a “fuzzy join” of two databases that do not share a unique key rather than deduplicating a single database. As an example, consider the problem of matching a large database of customers to a small database of known fraudsters. FindMatches can be used on both record linkage and deduplication problems.

For instance, Lake Formation's FindMatches ML Transform can help you with the following problems:

- Linking patient records between hospitals so that doctors have more background information and are better able to treat patients by using FindMatches on separate databases that both contain common fields such as name, birthday, home address, phone number, etc.
- Deduplicating a database of movies containing columns like “title”, “plot synopsis”, “year of release”, “run time”, and “cast”. For instance, the same movie might be variously identified as “Star Wars”, “Star Wars: A New Hope”, and “Star Wars: Episode IV—A New Hope (Special Edition)”.
- Automatically group all related products together in your storefront by identifying equivalent items in an apparel product catalog where you want to define “equivalent” to mean that they are the same ignoring differences in size and color. Hence “Levi 501 Blue Jeans, size 34x34” is defined to be the same as “Levi 501 Jeans--black, Size 32x31”.

Q: How does Lake Formation deduplicate my data?

A: Lake Formation's FindMatches ML Transform makes it easy to find and link records that refer to the same entity but don’t share a reliable identifier. Before FindMatches, developers would commonly solve data-matching problems deterministically, by writing huge numbers of hand-tuned rules. FindMatches uses machine learning algorithms behind the scenes to learn how to match records according to each developer's own business criteria. FindMatches first identifies records for the customer to label as to whether they match or do not match and then uses machine learning to create an ML Transform. Customers can then execute this Transform on their database to find matching records or they can ask FindMatches to give them additional records to label to push their ML Transform to higher levels of accuracy.

Q: What are ML Transforms?

A: ML Transforms provide a destination for creating and managing machine-learned transforms. Once created and trained, these ML Transforms can then be executed in standard AWS Glue scripts. Customers select a particular algorithm (for example, the FindMatches ML Transform) and input datasets and training examples, and the tuning parameters needed by that algorithm. AWS Lake Formation uses those inputs to build an ML Transform that can be incorporated into a normal ETL Job workflow.

Q: How do ML Transforms work?

A: Lake Formation includes specialized ML-based dataset transformation algorithms customers can use to create their own ML Transforms. These include record de-duplication and match finding.

Customers start by navigating to the ML Transforms tab in the Lake Formation console (or using the ML Transforms service endpoints or accessing ML Transforms training via CLI) to create their first ML transform model. The ML Transforms tab provides a user-friendly view for management of user transforms. ML Transforms require distinct workflow requirements from other transforms, including the need for separate training, parameter tuning, and execution workflows; the need for estimating the quality metrics of generated transformations; and the need to manage and collect additional truth labels for training and active learning.

To create an ML transform via the console, customers first select the transform type (such as Record Deduplication or Record Matching) and provide the appropriate data sources previously discovered in Data Catalog. Depending on the transform, customers may then be asked to provide ground truth label data for training or additional parameters. Customers can monitor the status of their training jobs and view quality metrics for each transform. (Quality metrics are reported using a hold-out set of the customer-provided label data.)

Once satisfied with the performance, customers can promote ML Transforms models for use in production. ML Transforms can then be used during ETL workflows, both in code autogenerated by the service and in user-defined scripts submitted with other jobs, similar to pre-built transforms offered in AWS Glue libraries.

Q: Can I see a presentation on using AWS Lake Formation to find matches and deduplicate records?

A: Yes, the full recording of the AWS Online Tech Talk, "Fuzzy Matching and Deduplicating Data with ML Transforms for AWS Lake Formation" is available here.

Q: How does Lake Formation relate to other AWS services?

A: Lake Formation manages data access for registered data that is stored in S3, and manages query access from AWS Glue, Athena, Redshift, and (in beta) EMR Notebooks and Zeppelin notebooks for EMR with Apache Spark through a unified security model and permissions. Lake Formation can ingest data from S3, Amazon RDS databases, and AWS CloudTrail logs, understand their formats, and make data clean and queryable. Lake Formation configures the flows, centralizes their orchestration, and lets you monitor the execution of your jobs.

Read more about "Data Lakes and Analytics on AWS" including how to build a customized data lake.

Q: How does Lake Formation relate to AWS Glue?

A: Lake Formation leverages a shared infrastructure with AWS Glue, including console controls, ETL code creation and job monitoring, blueprints to create workflows for data ingest, the same data catalog, and a serverless architecture. While AWS Glue focuses on these types of functions, Lake Formation encompasses all AWS Glue features AND provides additional capabilities designed to help build, secure, and manage a data lake. See the AWS Glue features page for more details.

ETL and catalog

Q: How does Lake Formation help me discover the data I can move into my data lake?

A: Lake Formation automatically discovers all AWS data sources to which it is provided access by your AWS IAM policies. It crawls S3, RDS, and CloudTrail sources and through blueprints it identifies them to you as data that can be ingested into your data lake. No data is ever moved or made accessible to analytic services without your permission. You can also use AWS Glue to ingest data from other sources including S3 and DynamoDB.

You can also define JDBC connections to allow Lake Formation to access your AWS databases and on-premises databases including Oracle, MySQL, Postgres, SQL Server, and MariaDB.

Lake Formation ensures that all your data is described in a central data catalog, giving you one location to browse the data that you have permission to view and query. The permissions are defined in your data access policy and can be set at the table and column level.

In addition to the properties automatically populated by the crawlers, you can add additional labels including business attributes such as data sensitivity, at the table- or column-level, and add field-level comments.

Q: How does Lake Formation organize my data in a data lake?

A: You can use one of the blueprints available in Lake Formation to ingest data into your data lake. Lake Formation creates Glue workflows that crawl source tables, extract the data, and load it to S3. In S3, Lake Formation organizes the data for you, setting up partitions and data formats for optimized performance and cost. For data already in Amazon S3, you can register those buckets with Lake Formation to manage them.

Lake Formation also crawls your data lake to maintain a data catalog and provides an intuitive user interface for you to search entities (by type, classification, attribute, or free-form text.)

Q: How does Lake Formation use machine learning to clean my data?

A: Lake Formation provides jobs that run machine learning algorithms to perform de-duplication and link matching records. Creating ML Transforms is as easy as selecting your source, selecting a desired transform, and providing training data for the changes you would like performed. Once trained to your satisfaction, the ML Transforms can be run as part of your regular data movement workflows, with no machine learning expertise required.

Q: What are other ways I can ingest data to AWS for use with Lake Formation?

A: Customers can move petabytes to exabytes of data from their datacenters to AWS using physical appliances with AWS Snowball, AWS Snowball Edge, and AWS Snowmobile or connect their on-premises applications directly to AWS with AWS Storage Gateway. Customers can accelerate data transfer using a dedicated network connection between a customer's network and AWS with AWS Direct Connect or boost long distance global data transfers using Amazon's globally distributed edge locations with Amazon S3 Transfer Acceleration. Amazon Kinesis also provides a useful way to load streaming data to S3. Lake Formation Data Importers can be set up to perform ongoing ETL jobs and prepare ingested data for analysis.

Q: Can I use my existing data catalog or Hive Metastore with Lake Formation?

A: Lake Formation provides a way for you to import your existing catalog and metastore into the Data Catalog. However, Lake Formation requires your metadata to reside in the Data Catalog to ensure governed access to your data.

Security and governance

Q: How does Lake Formation protect my data?

A: Lake Formation protects your data by giving you a central location where you can configure granular data access policies that protect your data, regardless of which services are used to access it.

To centralize data access policy controls using Lake Formation, first shut down direct access to your buckets in S3 so all data access is managed by Lake Formation. Next, configure data protection and access policies using Lake Formation, which enforces those policies across all the AWS services accessing data in your lake. You can configure users and roles and define the data these roles can access, down to the table and column level.

Lake Formation currently supports Server-Side-Encryption on S3 (SSE-S3, AES-265). Lake Formation also supports private endpoints in your VPC and records all activity in AWS CloudTrail, so you have network isolation and auditability.

Q: How does Lake Formation work with AWS IAM?

A: Lake Formation integrates with IAM so authenticated users and roles can be automatically mapped to data protection policies that are stored in the Data Catalog. The IAM integration also enables you to use Microsoft Active Directory or LDAP to federate into IAM using SAML.

Enabling data access

Q: How does Lake Formation help an analyst or data scientist discover what data they can access?

A: Lake Formation ensures that all your data is described in the Data Catalog, giving you a central location to browse the data that you have permission to view and query. The permissions are defined in your data access policy and can be set at the table and column level.

Q: Can I use third party business intelligence tools with Lake Formation?

A: Yes, you can use your third-party business applications, like Tableau and Looker, to connect to your AWS data sources through services like Athena or Redshift. Access to data is managed by the underlying Data Catalog, so regardless of which application you use, you are assured that access to your data is governed and controlled.

Q: Does Lake Formation provide APIs or a CLI?

A: Yes, Lake Formation provides APIs and a CLI to integrate Lake Formation functionality into your custom applications. Java and C++ SDKs are also available to enable you to integrate your own data engines with Lake Formation.

AWS Step Functions FAQs

Overview

Q: What is AWS Step Functions?

AWS Step Functions is a fully managed service that makes it easy to coordinate the components of distributed applications and microservices using visual workflows. Building applications from individual components that each perform a discrete function lets you scale easily and change applications quickly. Step Functions is a reliable way to coordinate components and step through the functions of your application. Step Functions provides a graphical console to arrange and visualize the components of your application as a series of steps. This makes it simple to build and run multi-step applications. Step Functions automatically triggers and tracks each step, and retries when there are errors, so your application executes in order and as expected. Step Functions logs the state of each step, so when things do go wrong, you can diagnose and debug problems quickly. You can change and add steps without even writing code, so you can easily evolve your application and innovate faster.

Q: What are the benefits of designing my application using orchestration?

Breaking an application into service components (or steps) ensures that the failure of one component does not bring the whole system down, that each component scales independently, and that components may be updated without requiring the entire system to be redeployed after each change. The coordination of service components involves managing execution dependencies, scheduling, and concurrency in accordance with the logical flow of the application. In such an application, developers may use service orchestration to do this and to handle failures.

Q: What are some common AWS Step Functions use cases?

AWS Step Functions helps with any computational problem or business process that can be subdivided into a series of steps. It's also useful for creating end-to-end workflows to manage jobs with interdependencies. Common use cases include:

- Data processing: consolidate data from multiple databases into unified reports, refine and reduce large data sets into useful formats, or coordinate multi-step analytics and machine learning workflows
- DevOps and IT automation: build tools for continuous integration and continuous deployment, or create event-driven applications that automatically respond to changes in infrastructure
- E-commerce: automate mission-critical business processes, such as order fulfillment and inventory tracking
- Web applications: implement robust user registration processes and sign-on authentication

For more details, explore AWS Step Functions use cases and customer testimonials.

Q: How does AWS Step Functions work?

Using AWS Step Functions, you define state machines that describe your workflow as a series of steps, their relationships, and their inputs and outputs. State machines contain a number of states, each of which represents an individual step in a workflow diagram. States can perform work, make choices, pass parameters, initiate parallel execution, manage timeouts, or terminate your workflow with a success or failure. The visual console automatically graphs each state in the order of execution, making it easy to design multi-step applications. The console highlights the real-time status of each step and provides a detailed history of every execution. For more information, see How Step Functions Works in the AWS Step Functions Developer Guide.

Q: How does AWS Step Functions connect to my resources?

You can configure your state machines to perform work by using activity tasks and service tasks. Activity tasks let you assign a specific step in your workflow to code running somewhere else (known as an activity worker). An activity worker can be any application that can make an HTTP connection, hosted anywhere. For example, activity workers can run on an Amazon EC2 instance, on a mobile device, or on an on-premises server. The activity worker polls Step Functions for work, takes any inputs from Step Functions, performs the work using your code, and returns results. Since activity workers request work, it is easy to use workers that are deployed behind a firewall.

Service tasks let you connect a step in your workflow to a supported AWS service. Step Functions pushes requests to other services so they can perform actions for your workflow, waits for the service task to complete, and then continues to the next step.

An AWS Step Functions state machine can contain combinations of activity tasks and service tasks. AWS Step Functions applications can also combine activity workers running in a data center with service tasks that run in the cloud. The workers in the data center continue to run as usual, along with any cloud-based service tasks.

Q: How do I get started with AWS Step Functions?

There are a number of ways you can get started with AWS Step Functions:

- Explore sample projects in the Step Functions console
- Read through the AWS Step Functions Developer Guide
- Try our 10-minute Tutorials

Q: What language does AWS Step Functions use?

AWS Step Functions state machines are defined in JSON using the declarative Amazon States Language. To create an activity worker, you may use any programming language, as long as you can communicate with AWS Step Functions using web service APIs. For convenience, you may use an AWS SDK in the language of your choosing. AWS Lambda supports code written in Node.js (JavaScript), Python, Golang (Go), and C# (using the .NET Core runtime and other languages). For more information on the Lambda programming model, see the AWS Lambda Developer Guide.

Q: My workflow has some of the properties of Standard Workflows and some properties of Express Workflows. How do I get the best of both?

You can compose the two workflow types: Express Workflows can run as a child workflow of Standard Workflows. The Express Workflow is invoked from a Task state in the parent orchestration workflow and succeeds or fails as a whole from the parent's perspective. It is subject to the parent's retry policy for that Task. You can also call Express Workflows from within an Express Workflow, so long as all workflows do not exceed the duration limit of the parent. You might choose to factor your workflows this way if your use case has a combination of long-running or exactly-once, and short-lived high-rate steps.

Comparisons

Q: When should I use AWS Step Functions vs. Amazon SQS?

You should consider AWS Step Functions when you need to coordinate service components in the development of highly scalable and auditable applications. You should consider using Amazon Simple Queue Service (Amazon SQS), when you need a reliable, highly scalable, hosted queue for sending, storing, and receiving messages between services. Step Functions keeps track of all tasks and events in an application. Amazon SQS requires you to implement your own application-level tracking, especially if your application uses multiple queues. The Step Functions Console and visibility APIs provide an application-centric view that lets you search for executions, drill down into an execution's details, and administer executions. Amazon SQS requires implementing such additional functionality. Step Functions offers several features that facilitate application development, such as passing data between tasks and flexibility in distributing tasks. Amazon SQS requires you to implement some application-level functionality. While you can use Amazon SQS to build basic workflows to coordinate your distributed application, you can get this facility out-of-the-box with Step Functions, alongside other application-level capabilities.

Q: When should I use AWS Step Functions vs. Amazon Simple Workflow Service (SWF)?

You should consider using AWS Step Functions for all your new applications, since it provides a more productive and agile approach to coordinating application components using visual workflows. If you require external signals to intervene in your processes, or you would like to launch child processes that return a result to a parent, then you should consider Amazon Simple Workflow Service (Amazon SWF). With Amazon SWF, instead of writing state machines in declarative JSON, you write a decider program to separate activity steps from decision steps. This provides you complete control over your orchestration logic, but increases the complexity of developing applications. You may write decider programs in the programming language of your choice, or you may use the Flow framework to use programming constructs that structure asynchronous interactions for you.

Q: When should I use Express Workflows vs. Standard Workflows?

You should use Express Workflows for workloads with high event rates and short durations. Express Workflows support event rates of more than 100,000 per second. Express Workflows have a maximum duration of five minutes. Express Workflows guarantee “at least once” execution of each workflow step. Failed workflows must be re-run from the beginning. Express Workflows support all service integrations. Express Workflows do not support activities, Job-run (.sync), and Callback patterns.

Standard Workflows on AWS Step Functions are more suitable for long-running, durable, and auditable workflows where repeating workflow steps is expensive (e.g., restarting a long-running media transcode) or harmful (e.g., charging a credit card twice). Example workloads include training and deploying machine learning models, report generation, billing, credit card processing, and ordering and fulfillment processes. Standard Workflows guarantee exactly once execution of each workflow step, with a maximum duration of one year. These workflows track and store detailed step-by-step information about each workflow that you may inspect during and after the workflow execution. Standard Workflows support all service integrations, activities, and design patterns.

Integration

Q: How does AWS Step Functions connect and coordinate other AWS services?

Workflows that you create with AWS Step Functions can connect and coordinate other AWS services using service tasks. For example, you can:

- Invoke an AWS Lambda function
- Run an Amazon Elastic Container Service or AWS Fargate task
- Get an existing item from an Amazon DynamoDB table or put a new item into a DynamoDB table
- Submit an AWS Batch job and wait for it to complete
- Publish a message to an Amazon SNS topic
- Send a message to an Amazon SQS queue
- Start an AWS Glue job run
- Create an Amazon SageMaker job to train a machine learning model or batch transform a data set

To learn more about using Step Functions to connect to other AWS services, see the Step Functions Developer Guide. You can also create tasks in your state machines that run applications, see the FAQ in the Overview section, *How does AWS Step Functions connect to my resources?*

Q: How does AWS Step Functions work with Amazon API Gateway?

You can associate your Step Functions APIs with Amazon API Gateway so that these APIs invoke your state machines when an HTTPS request is sent to an API method that you define. You can use an Amazon API Gateway API to start Step Functions state machines that coordinate the components of a distributed backend application, and integrate human activity tasks into the steps of your application such as approval requests and responses. You can also make serverless asynchronous calls to the APIs of services that your application uses. For more information, try our tutorial, Creating a Step Functions API Using API Gateway.

Q: How does logging and monitoring work for AWS Step Functions?

AWS Step Functions sends metrics to Amazon CloudWatch and AWS CloudTrail for application monitoring. Amazon CloudWatch collects and track metrics, sets alarms, and automatically reacts to changes in AWS Step Functions. AWS CloudTrail captures all API calls for Step Functions as events, including calls from the Step Functions console and from code calls to the Step Functions APIs. Step Functions also supports Amazon CloudWatch Events managed rules for each integrated service in your workflow, and will create and manage CloudWatch Events rules in your AWS account as needed. For more information, see Monitoring and Logging in the AWS Step Functions Developer Guide.

Q: What happens if my Express Workflow fails due to exhausted retries or an unmanaged exception?

By default, Express Workflows report all outcomes to CloudWatch Logs including workflow input, output, and completed steps. You may select different levels of logging to only log errors, and you can choose to not log input and output. Workflows that exhaust retries or have an unmanaged exception should be re-run from the start.

Security

Q: Can I access Step Functions from resources behind my Amazon VPC without connecting to the internet?

Step Functions also supports VPC Endpoints (VPCE) using AWS PrivateLink. You can access Step Functions from VPC-enabled AWS Lambda functions and other AWS services without traversing the public internet. For more information, refer the Amazon Virtual Private Cloud Endpoints for AWS Step Functions in the AWS Step Functions Developer Guide.

Compliance

Q: What are the compliance standards supported by Step Functions?

AWS Step Functions conforms to HIPAA, FedRAMP, SOC, GDPR, and other common compliance standards. See AWS Cloud Security site to get a detailed list of supported compliance standards.

Amazon EventBridge FAQs

Overview

Q: What is Amazon EventBridge?

Amazon EventBridge is a service that provides real-time access to changes in data in AWS services, your own applications and Software-as-a-Service (SaaS) applications without writing code. To get started, you can choose an event source on the Amazon EventBridge console, and select a target from a number of AWS services including AWS Lambda, Amazon SNS, and Amazon Kinesis Data Firehose. Amazon EventBridge will automatically deliver the events in near real-time.

Q: How can I get started using Amazon EventBridge?

Log in to your AWS account, navigate to the Amazon EventBridge console, and choose an event source from a list of partner SaaS applications and AWS services. If you are using a partner application ensure that you have configured your SaaS account to emit events, and accept it in the offered event sources section of the Amazon EventBridge console. Amazon EventBridge will automatically create an event bus for you to which events will be routed. Alternatively, you can use the AWS SDK to instrument your application to start emitting events to your event bus. Optionally configure a filtering rule and attach a target for your events, for example, this can be a Lambda function. Amazon EventBridge will be automatically ingest, filter and send the events to the configured target in a secure and highly available way.

Q: Can I publish my own events to Amazon EventBridge?

Yes. Customers can generate custom application-level events and publish them to Amazon EventBridge via the service’s APIs. Customers can also set up scheduled events that are generated on a periodic basis, and can process these events in any of the Amazon EventBridge supported targets.

Q: What is the format of an event?

Events use a specific JSON structure. Every event has the same top-level envelope fields, such as the source of the event, timestamp, and region. This is followed by a detail field which is the body of the event. For example, when an Amazon EC2 auto scaling group creates a new Amazon EC2 instance, it emits an event with source: “aws.autoscaling” and detail: "EC2 instance created successfully".

Q: How do I filter which events are delivered to a target?

You can filter events with rules. A rule matches incoming events for a given event bus and routes them to targets for processing. A single rule can route to multiple targets, all of which are processed in parallel. Rules allow different application components to look for and process the events that are of interest to them. A rule can customize an event before it is sent to the target, by passing only certain parts or by overwriting it with a constant. For the example given in the previous question, you can create an event rule that matches on source: “aws.autoscaling” and detail: "EC2 instance created successfully" to be notified any time an auto scaling group successfully creates an Amazon EC2 instance.

Q: How do I secure access to Amazon EventBridge?

Amazon EventBridge integrates with AWS Identity and Access Management (IAM) so that you can specify which actions a user in your AWS Account can perform. For example, you could create an IAM policy that gives only certain users in your organization permission to create event buses or attach event targets.

Q: How does Amazon EventBridge relate to CloudWatch Events?

Amazon EventBridge builds upon and extends CloudWatch Events. It uses the same service API and endpoint, and the same underlying service infrastructure. For existing CloudWatch Events customers, nothing changes - you can continue to use the same API, CloudFormation templates, and console. We heard from customers that CloudWatch Events is the ideal service for building event-driven architectures, and so we built new features that would enable our customers to connect data from their own apps and third-party SaaS apps. Rather than keeping this beneath the CloudWatch service, we have released this functionality with a new name, Amazon EventBridge, to signify the expansion beyond the monitoring use case that CloudWatch Events was developed for.

Q: I currently use Amazon CloudWatch Events and I WANT to try the features of Amazon EventBridge. Do I need to move my Amazon CloudWatch Events rules and permissions to Amazon EventBridge?

No. Existing Amazon CloudWatch Events users can access their existing default bus, rules, and events in the new Amazon EventBridge console and API or in the Amazon CloudWatch Events console and API.

Q: I'm already using Amazon CloudWatch Events and I don't need the features of Amazon EventBridge. What will change for me?

Nothing. Amazon EventBridge uses the same Amazon CloudWatch Events API so all of your existing CloudWatch Events API usage will remain the same.

Q: Are you going to deprecate Amazon CloudWatch Events one day?

No, we are not going to deprecate the API or the service itself. Amazon EventBridge is using the same API, and has added additional features. Over time, the Amazon CloudWatch Events name will be replaced with Amazon EventBridge.

Q: Which AWS services are integrated as event sources for Amazon EventBridge?

There are over 90 AWS services available as event sources for EventBridge including AWS Lambda, Amazon Kinesis, and AWS Fargate. For a full list of AWS service integrations please see the EventBridge documentation.

Q: Which AWS services are integrated as event targets for Amazon EventBridge?

There are over 15 AWS services available as event targets for EventBridge including AWS Lambda, Amazon SQS, Amazon SNS, Amazon Kinesis Streams, and Amazon Kinesis Firehose. For a full list of AWS service integrations please see the EventBridge documentation.

Limits and performance

Q: What are the service limits?

See “Service Limits” page here.

Q: What is the latency I can expect between sending and receiving an event?

Typical latency is about half a second. Note that this can vary.

Q: Does Amazon EventBridge support resource tagging?

Yes, you can tag rules. You can’t tag event buses or event sources.

Q: What throughput can I expect from Amazon EventBridge?

Event bus throughput limits are given in the “Service Limits” page here. If you require higher throughput please request a service limit increase through the AWS Support Center by choosing Create Case and then choosing Service Limit Increase.

Q: Does EventBridge have a Service Level Agreement?
Yes. AWS will use commercially reasonable efforts to make EventBridge available with a Monthly Uptime Percentage for each AWS region, during any monthly billing cycle, of at least 99.99%. For details, please review the full EventBridge Service Level Agreement.

Now in Preview Schema Registry

Q: What is a schema?

A schema represents the structure of an event, and commonly includes information such as the title and format of each piece of data included in the event. For example, a schema might include fields such as name and phone number, and the fact that the name is a text string, and the phone number is an integer. The schema can also include information on patterns, such as a requirement that the phone number be 10 digits in length. The schema of an event is important because it shows what information is contained in the event, and allows you to write code based on that data.

Q: What is a schema registry?

A schema registry stores a searchable collection of schema so any developer in your organization can easily access schema generated by the application, rather than looking through documentation or finding the schema’s author for this information. You can add a schema to the registry manually, or automate this process by turning on the EventBridge schema discovery feature.

Q: What is the schema discovery feature?

Schema discovery automates the processes of finding schemas and adding them to your registry. When schema discovery is enabled for an EventBridge event bus, the schema of each event sent to the event bus is automatically added to the registry. If the schema of an event changes, schema discovery will automatically create a new version of the schema in the registry. Once a schema is added to the registry, you can generate a code binding for the schema, either in the EventBridge console or directly in your IDE, which allows you to represent the event as an strongly-typed object in your code, and take advantage of IDE features such as validation and auto-complete.

Q: Can I discover schemas from events delivered across other accounts?

During preview, schema discovery is only enabled for events originating within the same account as the discoverer on the default, custom and partner event buses.

Q: How much does the schema registry cost?

There is no cost to use the schema registry, however there is a cost per ingested event when you turn on schema discovery. Schema discovery has a free tier of 5M ingested events per month, which should cover most development usage. There is a fee of $0.10 per million ingested events for additional usage outside of the free tier. For more info on pricing, please see the EventBridge pricing page.

Q: How does the schema registry reduce the amount of code I need to write?

First, you can use schema discovery to automatically identify schema for any events sent to your EventBridge event bus, and storing them in the registry, saving you from having to manually manage your event schema. Second, when you write applications that handle events on your bus, you can generate and download code bindings for this schema so you can use strongly-typed objects directly in your code. This saves overhead for deserialization, validation, and guesswork for your event handler.

Q: Why should I use the schema registry?

With the schema registry, EventBridge gives you a way to develop event-driven applications significantly faster, allowing you to focus on your application code. Previously, you needed to find available events and their structure, and write code to interpret and translate events into a format understandable by your code. Now, with the schema registry, you can automatically find the events available from any supported event source, including AWS services, third-party, and custom applications, and detect their schema.

Q: Which IDEs does the schema registry support?

The schema registry is available via the AWS Toolkit for Jetbrains (Intellij, PyCharm, Webstorm, Rider) and VS Code, as well as in the EventBridge console and APIs. Learn more about using the EventBridge schema registry within your IDE.

Q: Can I use schema with the Serverless Application Model (SAM)?

Yes, the latest version of the SAM CLI includes an interactive mode that allows you to create new serverless applications on EventBridge for any schema as an event type. Simply choose the “EventBridge Starter App” template, and the schema of your event, and SAM will automatically generate an application with a Lambda Function invoked by EventBridge, with handling code of the event .

This means that you can treat an event trigger like a normal object in your code, and use features such as validation and auto-complete in your IDE.

The AWS Toolkit plugin for Jetbrains (Intellij, PyCharm, Webstorm, Rider) and VS Code also provide functionality to generate serverless applications from this template, with a schema as a trigger, directly from these IDEs.

Q: In which languages can I generate code from my schemas?

At preview, code generation is available in Java (8+), Python (3.6+), and Typescript (3.0+).

Q: In which regions is the schema registry preview available?

The schema registry preview is available in US East (Ohio), US West (Oregon), US East (Northern Virginia) Asia Pacific (Tokyo) Region, and Europe (Ireland) Regions.

Cost and billing

Q: What does event bridge cost?

Please see Pricing here.

Q: Will I be charged for events sent by a partner to an event source that does not have an event bus attached?

No.

Architecture and design

Q: Can I have a target that sends events to another account?

Yes. These are called cross-account events, and you can have a target that is either the default event bus or any other event bus in another account.

Q: Can I use AWS CloudFormation with Amazon EventBridge?

AWS CloudFormation is supported for Rules and EventBusPolicy resources. Event bus and event source resources are not yet supported, but will be in the future.

Q: When should I use Amazon EventBridge and when should I use Amazon SNS?

Both Amazon EventBridge and Amazon SNS can be used to develop event-driven applications, and your choice will depend on your specific needs. Amazon EventBridge is recommended when you want to build an application that reacts to events from SaaS applications and/or AWS services. Amazon EventBridge is the only event-based service that integrates directly with third-party SaaS partners. Amazon EventBridge also automatically ingests events from over 90 AWS services without requiring developers to create any resources in their account. Further, Amazon EventBridge uses a defined JSON-based structure for events, and allows you to create rules that are applied across the entire event body to select events to forward to a target. Amazon EventBridge currently supports over 15 AWS services as targets, including AWS Lambda, Amazon SQS, Amazon SNS, and Amazon Kinesis Streams and Firehose, among others. At launch, Amazon EventBridge is has limited throughput (see Service Limits) which can be increased upon request, and typical latency of around half a second.

Amazon SNS is recommended when you want to build an application that reacts to high throughput or low latency messages published by other applications or microservices (as Amazon SNS provides nearly unlimited throughput), or for applications that need very high fan-out (thousands or millions of endpoints). Messages are unstructured and can be in any format. Amazon SNS supports forwarding messages to 6 different types of targets, including AWS Lambda, Amazon SQS, HTTP/S endpoints, SMS, Mobile Push, and email. Amazon SNS typical latency is under 30 msec. A wide range of AWS services send SNS messages by configuring the service to do so (more than 30, including Amazon EC2, Amazon S3, and Amazon RDS).

Integrations

Q: Why would I integrate my SaaS application with Amazon EventBridge?

Amazon EventBridge makes it easy for SaaS vendors to integrate their service into their customers' event-driven architectures built on AWS. Amazon EventBridge makes your product directly accessible to millions of AWS developers, unlocking new use cases. It offers a fully auditable, secure, and scalable pathway to send events without the SaaS vendor managing any eventing infrastructure.

Q: My SaaS company would be a great event source. How do I get on-boarded?

SaaS vendors interested in becoming an Amazon EventBridge partner, should follow self-service instructions at the Amazon EventBridge integrations page to begin publishing events to Amazon EventBridge.

Q: How much effort will be required for a SaaS Vendor to integrate with Amazon EventBridge?

SaaS vendors that already support a webhook or other push-based integration mode can expect to perform less than 5 days of development to integrate with Amazon EventBridge.

Q: Which SaaS Integrations are supported?

For a full list of supported integrations please see here.

Amazon MQ FAQs

Q: What is Amazon MQ?

Amazon MQ is a managed message broker service for Apache ActiveMQ that makes it easy to set up and operate message brokers in the cloud. Amazon MQ manages the administration and maintenance of ActiveMQ, a popular open source message broker. Amazon MQ supports durability-optimized brokers backed by Amazon Elastic File System (Amazon EFS) to support high availability and message durability, and throughput-optimized brokers backed by Amazon Elastic Block Store (EBS) to support high-volume applications that require low latency and high throughput. With Amazon MQ, you get direct access to the ActiveMQ console and industry standard APIs and protocols for messaging, including JMS, NMS, AMQP, STOMP, MQTT, and WebSocket. You can easily move from any message broker that uses these standards to Amazon MQ because you don't have to rewrite any messaging code in your applications.

Q: Who should use Amazon MQ?

Amazon MQ is suitable for enterprise IT pros, developers, and architects who are managing a message broker themselves–whether on-premises or in the cloud–and want to move to a fully managed cloud service without rewriting the messaging code in their applications.

Q: What does Amazon MQ manage on my behalf?

Amazon MQ manages the work involved in setting up a message broker, from provisioning the infrastructure capacity you request–including broker instances and storage–to installing the broker software. Once your broker is up and running, Amazon manages ongoing software upgrades, security updates, and fault detection and recovery. Amazon MQ stores messages redundantly across multiple Availability Zones (AZs) for message durability. With active/standby brokers, Amazon MQ automatically fails over to a standby instance in the event of a failure so you can continue sending and receiving messages.

Q: What is an Amazon MQ network of brokers?

Amazon MQ uses the “network of brokers” feature that is part of Apache ActiveMQ. A network of brokers consists of multiple brokers connected together. Brokers in the network share information about the clients and destinations each broker hosts. The brokers use this information to route messages through the network. With Amazon MQ, the brokers in the network can either be active-standby brokers (each active broker in the network has a standby node, with shared storage, that will take over if the active node fails), or single-instance brokers (if the node fails, it will be unavailable until it is restarted). Each broker in the network maintains its own unique message store which is replicated across multiple AZs within a region. The nodes in the network forward messages to each other, so messages are stored by a single broker at any given time.

You should use network of brokers if you require high availability with fast reconnection in the case of broker failure, or if you need the ability to scale horizontally.

Q: When would I use Amazon MQ vs. managing ActiveMQ on Amazon EC2 myself?

The choice depends on how closely you want to manage your message broker and underlying infrastructure. Amazon MQ provides a managed message broker service that takes care of operating ActiveMQ, including broker set up, monitoring, maintenance, and provisioning the underlying infrastructure for high availability and durability. You may want to consider Amazon MQ when you want to offload operational overhead and associated costs. If you want greater control in order to customize features and configurations or to use custom ActiveMQ plugins, you may want to consider installing and running ActiveMQ on Amazon EC2 directly.

Q: How do I migrate if I'm using a different message broker instead of ActiveMQ?

Amazon MQ provides compatibility with the most common messaging APIs, such as Java Message Service (JMS) and .NET Message Service (NMS), and protocols, including AMQP, STOMP, MQTT, and WebSocket. This makes it easy to switch from any standards-based message broker to Amazon MQ without rewriting the messaging code in your applications. In most cases, you can simply update the endpoints of your Amazon MQ broker to connect to your existing applications, and start sending messages.

Q: How does Amazon MQ work with other AWS services?

Any application that runs on an AWS compute service, such as Amazon EC2, Amazon ECS, or AWS Lambda, can use Amazon MQ. Amazon MQ is also integrated with the following AWS services:

- Amazon CloudWatch - monitor metrics and generate alarms
- Amazon CloudWatch Logs - publish logs from your Amazon MQ brokers to Amazon CloudWatch Logs
- AWS CloudTrail - log, continously monitor, and retain Amazon MQ API calls
- AWS CloudFormation - automate the process of creating, updating, and deleting message brokers
- AWS Identity and Access Management (IAM) - authentication and authorization of the service API
- AWS Key Management Service (KMS) - create and control the keys used to encrypt your data

Q: What kind of messaging durability does Amazon MQ provide?

Amazon MQ provides durability-optimized brokers backed by Amazon Elastic File System (Amazon EFS) to support high availability and message durability. When the ActiveMQ broker is used in persistent mode, each message is redundantly stored across multiple Availability Zones (AZs). With durability-optimized brokers, the message store can be accessed concurrently from all AZs in the region where it is located, which means that the message broker can fail over from one AZ to another AZ in the region without message loss. Amazon MQ is designed for 99.999999999% (eleven 9's) message durability.

Q: How can I get started with Amazon MQ?

Amazon MQ makes it easy to setup and operate message brokers in the cloud. With Amazon MQ, you can use the AWS Management Console, CLI, or API calls to launch a production-ready message broker in minutes. In most cases, you can simply update the endpoints of your Amazon MQ broker to connect to your existing applications and start sending messages.

Try a short tutorial, Create a Connected Message Broker, to get started today.

Q: How am I charged for Amazon MQ?

With Amazon MQ, you pay only for what you use. You are charged for broker instance and storage usage, and standard data transfer fees. It's easy to get started with Amazon MQ with our free tier for one year. See Amazon MQ pricing for details.

Q: Which versions of ActiveMQ does Amazon MQ support?

Amazon MQ provides support for ActiveMQ versions 5.15.0, 5.15.6, 5.15.8, 5.15.9, and 5.15.10.

Q: Does Amazon MQ meet compliance standards?

Yes. Amazon MQ is HIPAA eligible, and meets standards for PCI, SOC, and ISO compliance.

Amazon MQ is HIPAA eligible, which means you can use it to store and transmit messages between healthcare systems, including messages containing protected health information (PHI). Amazon MQ is PCI DSS compliant, which means you can use it to process, store, or transmit payment information. Amazon MQ is also ISO 9001, 27001, 27017, and 27018 certified. These certifications are among the most recognized global security standards attesting to quality and information security management in the cloud, and the protection of personally identifiable information. Amazon MQ is SOC 1, 2, and 3 compliant, allowing you to get deep insight into the security processes and controls that protect customer data.

For a complete list of AWS services and compliance programs, please see AWS Services in Scope by Compliance Program.

Q: When should I use Amazon MQ vs. Amazon SQS and SNS?

Amazon MQ, Amazon SQS, and Amazon SNS are messaging services that are suitable for anyone from startups to enterprises. If you're using messaging with existing applications, and want to move your messaging to the cloud quickly and easily, we recommend you consider Amazon MQ. It supports industry-standard APIs and protocols so you can switch from any standards-based message broker to Amazon MQ without rewriting the messaging code in your applications. If you are building brand new applications in the cloud, we recommend you consider Amazon SQS and Amazon SNS. Amazon SQS and SNS are lightweight, fully managed message queue and topic services that scale almost infinitely and provide simple, easy-to-use APIs. You can use Amazon SQS and SNS to decouple and scale microservices, distributed systems, and serverless applications, and improve reliability.

Q: When should I use Amazon MQ vs. AWS IoT Message Broker?

You can use Amazon MQ when you want to offload operational overhead and associated costs with an open source messaging application such as ActiveMQ or any commercial message brokers. You can use Amazon MQ when you are migrating from commercial brokers or open source brokers such as ActiveMQ to reduce broker maintenance, licensing costs and improve broker stability. Amazon MQ is also suitable for Application Integration use cases where you are developing new cloud based applications using micro-services that communicate with complex messaging patterns and require low-latency, high availability and message durability. Amazon MQ supports industry standard APIs such as JMS and NMS, and protocols for messaging, including AMQP, STOMP, MQTT, and WebSocket.

You can use AWS IoT Message Broker when your use case involves IoT devices' telemetry, device management, device security and IoT Analysis. AWS IoT Message Broker is suitable for IoT industry customers connecting large device fleets and collecting telemetry data to send it to native AWS services. AWS IoT Message broker supports industry standard lightweight protocols such as MQTT, HTTP and MQTT over WebSocket.

Q: How do I use my own custom keys to encrypt the data in Amazon MQ?

Amazon MQ supports the AWS Key Management Service (AWS KMS) to create and manage keys for at-rest encryption of your data in Amazon MQ. When you create a broker, you can select the KMS key used to encrypt your data from the following three options: a KMS key in the Amazon MQ service account, a KMS key in your account that Amazon MQ creates and manages, or a KMS key in your account that you create and manage. In addition to encryption at rest, all data transferred between Amazon MQ and client applications is securely transmitted using TLS/SSL.

Q: How can I monitor my broker instances, queues, and topics?

Amazon MQ and Amazon CloudWatch are integrated so you can view and analyze metrics for your broker instances, as well as your queues and topics. You can view and analyze metrics from the Amazon MQ console, the CloudWatch console, the command line, or programmatically. Metrics are automatically collected and pushed to CloudWatch every minute.

Q: Does Amazon MQ have a Service Level Agreement?

Yes. AWS will use commercially reasonable efforts to make Active/Standby Brokers available with a Monthly Uptime Percentage of at least 99.9% during any monthly billing cycle (the "Service Commitment"). In the event Amazon MQ does not meet the Monthly Uptime Percentage commitment, you will be eligible to receive a Service Credit. For details, please review the full Amazon MQ Service Level Agreement.

Q: What type of storage is available with Amazon MQ?

Amazon MQ supports two types of broker storage – durability optimized using Amazon Elastic File System (Amazon EFS) and throughput optimized using Amazon Elastic Block Store (EBS). To take advantage of high durability and replication across multiple Availability Zones, use durability optimized brokers backed by Amazon EFS. To take advantage of high throughput for your high volume applications, use throughput optimized brokers backed by EBS. Throughput optimized message brokers reduce the number of brokers required, and cost of operating, high-volume applications using Amazon MQ.

Amazon SNS FAQs

Overview

Q: What is Amazon Simple Notification Service (Amazon SNS)?

Amazon Simple Notification Service (Amazon SNS) is a web service that makes it easy to set up, operate, and send notifications from the cloud. It provides developers with a highly scalable, flexible, and cost-effective capability to publish messages from an application and immediately deliver them to subscribers or other applications. It is designed to make web-scale computing easier for developers. Amazon SNS follows the "publish-subscribe" (pub-sub) messaging paradigm, with notifications being delivered to clients using a "push" mechanism that eliminates the need to periodically check or "poll" for new information and updates. With simple APIs requiring minimal up-front development effort, no maintenance or management overhead and pay-as-you-go pricing, Amazon SNS gives developers an easy mechanism to incorporate a powerful notification system with their applications.

Q: How can I get started using Amazon SNS?

You can create an Amazon SNS topic and publish messages in a few steps by completing our 10-minute tutorial, Send Fanout Notifications.

For additional detail, see the Amazon SNS Developer Guide, and sample code in the Resource Center.

Q: What are the benefits of using Amazon SNS?

Amazon SNS offers several benefits making it a versatile option for building and integrating loosely-coupled, distributed applications:

- Instantaneous, push-based delivery (no polling)
- Simple APIs and easy integration with applications
- Flexible message delivery over multiple transport protocols
- Inexpensive, pay-as-you-go model with no up-front costs
- Web-based AWS Management Console offers the simplicity of a point-and-click interface

Q: What are some example uses for Amazon SNS notifications?

The Amazon SNS service can support a wide variety of needs including event notification, monitoring applications, workflow systems, time-sensitive information updates, mobile applications, and any other application that generates or consumes notifications. For example, Amazon SNS can be used in workflow systems to relay events among distributed computer applications, move data between data stores or update records in business systems. Event updates and notifications concerning validation, approval, inventory changes and shipment status are immediately delivered to relevant system components as well as end-users. A common pattern is to use SNS to publish messages to Amazon SQS message queues to reliably send messages to one or many system components asynchronously. Another example use for Amazon SNS is to relay time-critical events to mobile applications and devices. Since Amazon SNS is both highly reliable and scalable, it provides significant advantages to developers who build applications that rely on real-time events.

Q: How does Amazon SNS work?

It is very easy to get started with Amazon SNS. Developers must first create a “topic” which is an “access point” – identifying a specific subject or event type – for publishing messages and allowing clients to subscribe for notifications. Once a topic is created, the topic owner can set policies for it such as limiting who can publish messages or subscribe to notifications, or specifying which notification protocols will be supported (i.e. HTTP/HTTPS, email, SMS). Subscribers are clients interested in receiving notifications from topics of interest; they can subscribe to a topic or be subscribed by the topic owner. Subscribers specify the protocol and end-point (URL, email address, etc.) for notifications to be delivered. When publishers have information or updates to notify their subscribers about, they can publish a message to the topic – which immediately triggers Amazon SNS to deliver the message to all applicable subscribers.

Q: How is Amazon SNS different from Amazon SQS?

Amazon Simple Queue Service (SQS) and Amazon SNS are both messaging services within AWS, which provide different benefits for developers. Amazon SNS allows applications to send time-critical messages to multiple subscribers through a "push" mechanism, eliminating the need to periodically check or "poll" for updates. Amazon SQS is a message queue service used by distributed applications to exchange messages through a polling model, and can be used to decouple sending and receiving components. Amazon SQS provides flexibility for distributed components of applications to send and receive messages without requiring each component to be concurrently available.

A common pattern is to use SNS to publish messages to Amazon SQS queues to reliably send messages to one or many system components asynchronously.

Q: How is Amazon SNS different from Amazon MQ?

Amazon MQ, Amazon SQS, and Amazon SNS are messaging services that are suitable for anyone from startups to enterprises. If you're using messaging with existing applications, and want to move your messaging to the cloud quickly and easily, we recommend you consider Amazon MQ. It supports industry-standard APIs and protocols so you can switch from any standards-based message broker to Amazon MQ without rewriting the messaging code in your applications. If you are building brand new applications in the cloud, we recommend you consider Amazon SQS and Amazon SNS. Amazon SQS and SNS are lightweight, fully managed message queue and topic services that scale almost infinitely and provide simple, easy-to-use APIs. You can use Amazon SQS and SNS to decouple and scale microservices, distributed systems, and serverless applications, and improve reliability.

Q: How can I get started using Amazon SNS?

To sign up for Amazon SNS, click the "Sign up for Amazon SNS" button on the Amazon SNS detail page. You must have an Amazon Web Services account to access this service; if you do not already have one, you will be prompted to create one when you begin the Amazon SNS sign-up process. After signing up, please refer to the Amazon SNS documentation and Getting Started Guide to begin using Amazon SNS. Using the AWS Management Console, you can easily create topics, add subscribers, send notifications, and edit topic policies – all from your browser.

Q: Is Amazon SNS supported in the AWS Management Console?

Amazon SNS is supported in the AWS Management Console which provides a point-and-click, web-based interface to access and manage Amazon SNS. Using the AWS Management Console, you can create topics, add subscribers, and send notifications – all from your browser. In addition, the AWS Management Console makes it easy to publish messages to your endpoint of choice (HTTP, SQS, Lambda, mobile push, email, or SMS) and edit topic policies to control publisher and subscriber access.

Q: What are the Amazon SNS service access points in each region?

Please refer to the AWS Regions and Endpoints section of the AWS documentation for the latest list of all Amazon SNS service access points.

Q: Can I get a history of SNS API calls made on my account for security analysis and operational troubleshooting purposes?

Yes. SNS supports AWS CloudTrail, a web service that records AWS API calls for your account and delivers log files to you. With CloudTrail, you can obtain a history of such information as the identity of the API caller, the time of the API call, the source IP address of the API caller, the request parameters, and the response elements returned by SNS.

SNS currently supports CloudTrail auditing for authenticated calls only. CloudTrail Audit logs for unauthenticated ConfirmSubscription and Unsubscribe calls are not available at this time. For more information, see the CloudTrail section of the SNS Developer Guide.

To receive a history of SNS API calls made on your account, simply turn on AWS CloudTrail in the AWS Management Console. To learn more about AWS CloudTrail, click here.

Billing

Q: How much does Amazon SNS cost?

With Amazon SNS, there is no minimum fee and you pay only for what you use. Users pay $0.50 per 1 million Amazon SNS Requests, $0.06 per 100,000 notification deliveries over HTTP, and $2.00 per 100,000 notification deliveries over email. For SMS messaging, users can send 100 free notification deliveries, and for subsequent messages charges vary by destination country.

Amazon SNS also includes a Free Tier, where users can get started with Amazon SNS for free. Each month, Amazon SNS customers incur no charges for the first 1 million Amazon SNS requests, no charges for the first 100,000 notifications over HTTP, no charges for the first 100 notifications over SMS, and no charges for the first 1,000 notifications over email.

Please refer to the Amazon SNS Features page for additional details on pricing and data transfer costs.

Q: How will I be charged and billed for my use of Amazon SNS?

There are no set-up fees to begin using the service. At the end of the month, your credit card will automatically be charged for that month's usage. You can view your charges for the current billing period at any time on the Amazon Web Services web site by logging into your Amazon Web Services account and clicking "Account Activity" under "Your Web Services Account".

Q: When does billing of my Amazon SNS use begin and end?

Your Amazon SNS billing cycle begins on the first day of each month and ends on the last day of each month. Your monthly charges will be totalled at the end of each month.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Features and functionality

Q: What is the format of an Amazon SNS topic?

Topic names are limited to 256 characters. Alphanumeric characters plus hyphens (-) and underscores (_) are allowed. Topic names must be unique within an AWS account. After you delete a topic, you can reuse the topic name. When a topic is created, Amazon SNS will assign a unique ARN (Amazon Resource Name) to the topic, which will include the service name (SNS), region, AWS ID of the user and the topic name. The ARN will be returned as part of the API call to create the topic. Whenever a publisher or subscriber needs to perform any action on the topic, they should reference the unique topic ARN.

The following is the ARN for a topic named "mytopic" created by a user with the AWS account ID "123456789012" and hosted in the US East region:

arn:aws:sns:us-east-1:1234567890123456:mytopic Note: Users should NOT attempt to build the topic ARN from its separate components – they should always use the name returned from the API call to create the topic.

Q: What are the available operations for Amazon SNS and who can perform these operations?

Amazon SNS provides a set of simple APIs to enable event notifications for topic owners, subscribers and publishers.

Owner operations:

- CreateTopic – Create a new topic.
- DeleteTopic – Delete a previously created topic.
- ListTopics – List of topics owned by a particular user (AWS ID).
- ListSubscriptionsByTopic – List of subscriptions for a particular topic
- SetTopicAttributes – Set/modify topic attributes, including setting and modifying publisher/subscriber permissions, transports supported, etc.

- GetTopicAttributes – Get/view existing attributes of a topic
- AddPermission – Grant access to selected users for the specified actions
- RemovePermission – Remove permissions for selected users for the specified actions

Subscriber operations:

- Subscribe – Register a new subscription on a particular topic, which will generate a confirmation message from Amazon SNS
- ConfirmSubscription – Respond to a subscription confirmation message, confirming the subscription request to receive notifications from the subscribed topic
- UnSubscribe – Cancel a previously registered subscription
- ListSubscriptions – List subscriptions owned by a particular user (AWS ID)

Publisher operations:

- Publish: Publish a new message to the topic.

Q: Why are there two different APIs to list subscriptions?

The two APIs to list subscriptions perform different functions and return different results:

- The ListSubscriptionsByTopic API allows a topic owner to see the list of all subscribers actively registered to a topic.
- The ListSubscriptions API allows a user to get a list of all their active subscriptions (to one or more topics).

Q: What are the different delivery formats/transports for receiving notifications?

In order for customers to have broad flexibility of delivery mechanisms, Amazon SNS supports notifications over multiple transport protocols. Customers can select one the following transports as part of the subscription requests:

- “HTTP”, “HTTPS” – Subscribers specify a URL as part of the subscription registration; notifications will be delivered through an HTTP POST to the specified URL.

- ”Email”, “Email-JSON” – Messages are sent to registered addresses as email. Email-JSON sends notifications as a JSON object, while Email sends text-based email.
- “SQS” – Users can specify an SQS standard queue as the endpoint; Amazon SNS will enqueue a notification message to the specified queue (which subscribers can then process using SQS APIs such as ReceiveMessage, DeleteMessage, etc.). Note that FIFO queues are not currently supported.
- “SMS” – Messages are sent to registered phone numbers as SMS text messages.

Q: Can topic owners control the transports that are allowed on topics they create/own?

Topic owners can configure specific transports on their topics by setting the appropriate permissions through access control policies.

Q: How does an owner set Access Control policies?

Please refer to the Amazon SNS Getting Started Guide for an overview of setting access control policies.

Q: Can a single topic support subscriptions over multiple protocols/transports?

Subscribers to an Amazon SNS topic can receive notifications on any transport supported by the topic. A topic can support subscriptions and notification deliveries over multiple transports.

Q: Can subscribers selectively receive only a subset of messages published to a topic?

Yes, you can use message filtering on Amazon Simple Notification Service (SNS) to build simpler and more streamlined pub/sub architectures. Message filtering enables Amazon SNS topic subscribers to selectively receive only a subset of the messages they are interested in, as opposed to receiving all messages published to a topic. To monitor the usage of SNS subscription filter policies, use Amazon

CloudWatch metrics, which are automatically collected for you. You can also use the AWS::SNS::Subscription resource type in AWS CloudFormation templates to quickly deploy solutions that use SNS message filtering. For more details, try our 10-minute tutorial, Filter Messages Published to Topics, or see the Filter Messages with Amazon SNS section in our documentation."

Q: Can Amazon SNS be used with other AWS services?

Amazon SNS can be used with other AWS services such as Amazon SQS, Amazon EC2 and Amazon S3. Here is an example of how an order processing workflow system uses Amazon SNS with Amazon EC2, SQS, and SimpleDB. In this workflow system, messages are sent between application components whenever a transaction occurs or an order advances through the order processing pipeline. When a customer initially places an order, the transaction is first recorded in Amazon SimpleDB and an application running on Amazon EC2 forwards the order request to a payment processor which debits the customer's credit card or bank account. Once approved, an order confirmation message is published to an Amazon SNS topic. In this case, the topic has various subscribers over Email/HTTP – merchant, customer and supply chain partners – and notifications sent by Amazon SNS for that topic can instantly update all of them that payment processing was successful. Notifications can also be used to orchestrate an order processing system running on EC2, where notifications sent over HTTP can trigger real-time processing in related components such as an inventory system or a shipping service. By integrating Amazon SNS with Amazon SQS, all notifications delivered are also persisted in an Amazon SQS queue where they are processed by an auditing application at a future time.

Q: Is Amazon SNS available in all regions where AWS services are available?

Please refer to the AWS Regions and Endpoints section of the AWS documentation for the most up to date information on Amazon SNS availability.

Q: How soon can customers recreate topics with previously used topic names?

Topic names should typically be available for reuse approximately 30-60 seconds after the previous topic with the same name has been deleted. The exact time will depend on the number of subscriptions which were active on the topic – topics with a few subscribers will be available instantly for reuse, topics with larger subscriber lists may take longer.

Transports

Q: How would a user subscribe for notifications to be delivered over email?

To receive email notifications for a particular topic, a subscriber should specify “Email” or “Email-JSON” as the protocol and provide a valid email address as the end-point. This can be done using the AWS Management Console or by calling the Amazon SNS API directly. Amazon SNS will then send an email with a confirmation link to the specified email address, and require the user monitoring the email address to explicitly opt-in for receiving email notifications from that particular topic. Once the user confirms the subscription by clicking the provided link, all messages published to the topic will be delivered to that email address.

Q: Why does Amazon SNS provide two different transports to receive notifications over email?

The two email transports are provided for two distinct types of customers/end-users. “Email-JSON” sends notifications as a JSON object, and is meant for applications to programmatically process emails. The ”Email” transport is meant for end-users/consumers and notifications are regular, text-based messages which are easily readable.

Q: Can a user change the Subject and Display name for notifications sent over Email/Email-JSON?

Amazon SNS allows users to specify the Subject field for emails as a parameter passed in to the Publish API call and can be different for every message published. The Display name for topics can be set using the SetTopicAttributes API – this name applies to all emails sent from this topic.

Q: Do subscribers need to specifically configure their email settings to receive notifications from Amazon SNS?

In most cases, users should be able to receive subscription confirmations and notifications from Amazon SNS without doing anything specific. However, there could be cases where the email provider's default settings or other user-specific configurations mistakenly redirect the emails to the junk/spam folder. To ensure that users see confirmation messages and notifications sent from Amazon SNS, users can add "no-reply@sns.amazonaws.com" to their contact lists and check their junk/spam folders for messages from Amazon SNS.

Q: In the case of passing in an SQS queue as an endpoint, will users need to create the queue prior to subscribing? What permissions will the queue require?

Using the SQS console, users should create the SQS queue prior to subscribing it to a Topic. Select this queue on the console, and from the 'Queue Actions' in the menu bar, select 'Subscribe Queue to SNS Topic' from the drop-down list. In the subscribe dialog box, select the topic from the 'Choose a Topic' drop-down list, and click the 'Subscribe' button. For complete step-by-step instructions, please refer to the Amazon SNS documentation.

Q: Are Amazon SQS FIFO queues compatible with Amazon Simple Notification Service (SNS)?

Amazon SNS does not currently support forwarding messages to Amazon SQS FIFO queues. You can use SNS to forward messages to standard queues.

Q: How would a developer setup an Amazon SQS queue to receive Amazon SNS notifications?

To have Amazon SNS deliver notifications to an SQS queue, a developer should subscribe to a topic specifying "SQS" as the transport and a valid SQS standard queue as the end-point. In order to allow the SQS queue to receive notifications from Amazon SNS, the SQS queue owner must subscribe the SQS queue to the Topic for Amazon SNS to successfully deliver messages to the queue.

If the user owns both the Amazon SNS topic being subscribed to and the SQS queue receiving the notifications, nothing further is required. Any message published to the topic will automatically be delivered to the specified SQS queue. If the user owning the SQS queue is not the owner of the topic, Amazon SNS will require an explicit confirmation to the subscription request.

Please refer to the Amazon SNS documentation for further details on subscribing an SQS queue to a topic and setting access control policies for SQS queues.

Q: How can I fanout identical messages to multiple SQS queues?

Create an SNS topic first using SNS. Then create and subscribe multiple SQS standard queues to the SNS topic. Now whenever a message is sent to the SNS topic, the message will be fanned out to the SQS queues, i.e. SNS will deliver the message to all the SQS queues that are subscribed to the topic.

Q: What is the format of structured notification messages sent by Amazon SNS?

The notification message sent by Amazon SNS for deliveries over HTTP, HTTPS, Email-JSON and SQS transport protocols will consist of a simple JSON object, which will include the following information:

- MessageId: A Universally Unique Identifier, unique for each notification published.
- Timestamp: The time (in GMT) at which the notification was published.
- TopicArn: The topic to which this message was published
- Type: The type of the delivery message, set to “Notification” for notification deliveries.
- UnsubscribeURL: A link to unsubscribe the end-point from this topic, and prevent receiving any further notifications.
- Message: The payload (body) of the message, as received from the publisher.
- Subject: The Subject field – if one was included as an optional parameter to the publish API call along with the message.
- Signature: Base64-encoded “SHA1withRSA” signature of the Message, MessageId, Subject (if present), Type, Timestamp, and Topic values.
- SignatureVersion: Version of the Amazon SNS signature used.

Notification messages sent over the “Email” transport only contain the payload (message body) as received from the publisher.

Q: How would a user subscribe for notifications to be delivered over SMS?

Please refer to the 'SMS Related Question' section below.

Security

Q: How can users secure the messages sent to my topics?

All API calls made to Amazon SNS are validated for the user’s AWS Id and the signature. In addition, we recommend that users secure their data over the wire by connecting to our secure SSL end-points.

Q: Who can create a topic?

Topics can only be created by users with valid AWS IDs who have signed up for Amazon SNS. The easiest way to create a topic is to use the AWS Management Console. It can also be created through the CreateTopic API. The AWS Management Console is available at: http://aws.amazon.com/console

Q: Can multiple users publish to a single topic?

A topic owner can set explicit permissions to allow more than one user (with a valid AWS ID) to publish to a topic. By default, only topic owners have permissions to publish to a topic.

Q: How can the owner grant/revoke publish or subscribe permissions on a topic?

The AddPermission and RemovePermission APIs provide a simple interface for developers to add and remove permissions for a topic. However, for conditional access and more advanced use cases, users should use access control policies to manage permissions. The easiest way to manage permissions is to use the AWS Management Console. The AWS Management Console is available at: http://aws.amazon.com/console

Q: How does a topic owner give access to subscribers? Do subscribers have to have valid AWS IDs?

Amazon SNS makes it easy for users with and without AWS IDs to receive notifications. The owner of the topic can grant/restrict access to subscribers by setting appropriate permissions for the topic using Access Control policies. Users can receive notifications from Amazon SNS in two ways:

- Users with AWS IDs: Subscribers with valid AWS IDs (please refer to this link for details on obtaining AWS IDs) can subscribe to any topic directly – as long as the topic owner has granted them permissions to do so. The AWS IDs will be validated as part of the subscription registration.
- Other users: Topic owners can subscribe and register end-points on behalf of users without AWS IDs.

In both cases, the owner of the subscription endpoint needs to explicitly opt-in and confirm the subscription by replying to confirmation message sent by Amazon SNS.

Q: How will Amazon SNS authenticate API calls?

All API calls made to Amazon SNS will validate authenticity by requiring that requests be signed with the secret key of the AWS ID account and verifying the signature included in the requests.

Q: How does Amazon SNS validate a subscription request to ensure that notifications will not be sent to users as spam?

As part of the subscription registration, Amazon SNS will ensure that notifications are only sent to valid, registered subscribers/end-points. To prevent spam and ensure that a subscriber end-point is really interested in receiving notifications from a particular topic, Amazon SNS requires an explicit opt-in from subscribers using a 2-part handshake:

i. When a user first calls the Subscribe API and subscribes an end-point, Amazon SNS will send a confirmation message to the specified end-point.

ii. On receiving the confirmation message at the end-point, the subscriber should confirm the subscription request by sending a valid response. Only then will Amazon SNS consider the subscription request to be valid. If there is no response to the challenge, Amazon SNS will not send any notifications to that end-point. The exact mechanism of confirming the subscription varies by the transport protocol selected:

- For HTTP/HTTPS notifications, Amazon SNS will first POST the confirmation message (containing a token) to the specified URL. The application monitoring the URL will have to call the ConfirmSubscription API with the token included token.
- For Email and Email-JSON notifications, Amazon SNS will send an email to the specified address containing an embedded link. The user will need to click on the embedded link to confirm the subscription request.
- For SQS notifications, Amazon SNS will enqueue a challenge message containing a token to the specified queue. The application monitoring the queue will have to call the ConfirmSubscription API with the token.

Note: The explicit "opt-in" steps described above are not required for the specific case where you subscribe your Amazon SQS queue to your Amazon SNS topic – and both are "owned" by the same AWS account.

Q: How long will subscription requests remain pending, while waiting to be confirmed?

Token included in the confirmation message sent to end-points on a subscription request are valid for 3 days.

Q: Who can change permissions on a topic?

Only the owner of the topic can change permissions for that topic.

Q: How can users verify that notification messages are sent from Amazon SNS?

To ensure the authenticity of the notifications, Amazon SNS will sign all notification deliveries using a cryptographically secure, asymmetric mechanism (private-public key pair based on certificates). Amazon SNS will publish its certificate to a well-known location (e.g. http://sns.us-east-1.amazonaws.com/SimpleNotificationService.pem for the US East region) and sign messages with the private key of that certificate. Developers/applications can obtain the certificate and validate the signature in the notifications with the certificate's public key, to ensure that the notification was indeed sent out by Amazon SNS. For further details on certificate locations, please refer to the Amazon SNS details page.

Q: Do publishers have to sign messages as well?

Amazon SNS requires publishers with AWS IDs to validate their messages by signing messages with their secret AWS key; the signature is then validated by Amazon SNS.

Q: Can a publisher/subscriber use SSL to secure messages?

Yes, both publishers and subscribers can use SSL to help secure the channel to send and receive messages. Publishers can connect to Amazon SNS over HTTPS and publish messages over the SSL channel. Subscribers should register an SSL-enabled end-point as part of the subscription registration, and notifications will be delivered over a SSL channel to that end-point.

Q: What permissions does a subscriber need to allow Amazon SNS to send notifications to a registered endpoint?

The owner of the end-point receiving the notifications has to grant permissions for Amazon SNS to send messages to that end-point.

Q: How can subscriptions be unsubscribed?

Subscribers can be unsubscribed either by the topic owner, the subscription owner or others – depending on the mechanism used for confirming the subscription request.

- A subscription that was confirmed with the AuthenticateOnUnsubscribe flag set to True in the call to the ConfirmSubscription API call can only be unsubscribed by a topic owner or the subscription owner.

- If the subscription was confirmed anonymously without the AuthenticateOnUnsubscribe flag set to True, then it can be anonymously unsubscribed.

In all cases except when unsubscribed by the subscription owner, a final cancellation message will be sent to the end-point, allowing the endpoint owner to easily re-subscribe to the topic (if the Unsubscribe request was unintended or in error). For further details on the ConfirmSubscription API, please refer to the Amazon SNS documentation.

Compliance

Q: Is Amazon SNS HIPAA eligible?

Yes, the AWS HIPAA compliance program includes Amazon SNS as a HIPAA eligible Service. If you have an executed Business Associate Agreement (BAA) with AWS, you can now use Amazon SNS to build HIPAA-compliant applications. If you don't have a BAA or have other questions about using AWS for your HIPAA-compliant applications, contact us for more information. Please note that Amazon SNS mobile push notification and SMS functionalities are outside the scope of the Service’s HIPAA eligibility and thus not suitable for transmitting Protected Health Information (PHI).

To learn more, see the following resources:

- AWS HIPAA Compliance page
- AWS Cloud Computing in Healthcare page

To see the current list of compliance programs that Amazon SNS is in scope for, see AWS Services in Scope by Compliance Program.

Q: What else is Amazon SNS compliant with?

Please see AWS Services in Scope by Compliance Program for the latest information about SNS and other AWS services.

Reliability

Q: How durable is my data once published to Amazon SNS?

SNS provides durable storage of all messages that it receives. Upon receiving a publish request, SNS stores multiple copies (to disk) of the message across multiple Availability Zones before acknowledging receipt of the request to the sender. Each AWS Region has multiple, isolated locations known as Availability Zones. Although rare, should a failure occur in one zone, the operation of SNS and the durability of your messages continue without disruption.

Q: Will a notification contain more than one message?

No, all notification messages will contain a single published message.

Q: How many times will a subscriber receive each message?

Although most of the time each message will be delivered to your application exactly once, the distributed nature of Amazon SNS and transient network conditions could result in occasional, duplicate messages at the subscriber end. Developers should design their applications such that processing a message more than once does not create any errors or inconsistencies.

Q: Will messages be delivered to me in the exact order they were published?

The Amazon SNS service will attempt to deliver messages from the publisher in the order they were published into the topic. However, network issues could potentially result in out-of-order messages at the subscriber end.

Q: Can a message be deleted after being published?

No, once a message has been successfully published to a topic, it cannot be recalled.

Q: Does Amazon SNS guarantee that messages are delivered to the subscribed endpoint?

Yes, as long as the subscribed endpoint is accessible. A message delivery fails when Amazon SNS can't access a subscribed endpoint, due to either a client-side or a server-side error. A client-side error happens when the subscribed endpoint has been deleted by the endpoint owner, or when its access permissions have changed in a way that prevents Amazon SNS from delivering messages to this endpoint. A server-side error happens when the service that powers the subscribed endpoint is unavailable, such as Amazon SQS or AWS Lambda. When Amazon SNS receives a client-side error, or continues to receive a server-side error for a message beyond the number of retries specified by the corresponding retry policy, Amazon SNS discards the message — unless a dead-letter queue is attached to the subscription. For more information, see Message Delivery Retries. and Amazon SNS Dead-Letter Queues.

Q: What happens to Amazon SNS messages if the subscribing endpoint is not available?

If a message cannot be successfully delivered on the first attempt, Amazon SNS executes a 4-phase retry policy: 1) retries with no delay in between attempts, 2) retries with minimum delay between attempts, 3) retries according to a back-off model, and 4) retries with maximum delay between attempts. When the message delivery retry policy is exhausted, Amazon SNS can move the message to a dead-letter queue (DLQ). For more information, see Message Delivery Retries and Amazon SNS Dead-Letter Queues.

Worldwide SMS

Q: What features are part of the new Worldwide SMS capability?

You can use Amazon SNS to deliver SMS (text) messages to 200+ countries, and you do not require recipients to explicitly opt in as before. You must obtain prior permission from recipients to send SMS messages to their phone numbers, where required by local law and regulations. Additionally, you can now mark your SMS messages as Transactional to optimize for reliable delivery, or you can mark it as Promotional to optimize for cost savings. Furthermore, you can set account and message-level spend quotas to avoid inadvertent overruns.

Q: When should I mark an SMS message as Transactional?

SMS messages that are of high priority to your business should be marked as Transactional. This ensures that messages such as those that contain one-time passwords (OTP) or PINs get delivered over routes with the highest delivery reliability. These routes tend to be more expensive than Promotional messaging routes in countries other than the US. You should never mark marketing messages as Transactional, because this violates the local regulatory policies in certain countries, and your account may be marked for abuse and suspended.

Q: When should I mark an SMS message as Promotional?

SMS messages that carry marketing messaging should be marked Promotional. Amazon SNS ensures that such messages are sent over routes that have a reasonable delivery reliability but are substantially cheaper than the most reliable routes. This also allows Amazon SNS to handle and deliver your messages in compliance with on local laws and regulation

Q: What are account-level and message-level spend quotas and how do they work?

Spend quotas can be specified for an AWS account and for individual messages, and the quotas apply only to the cost of sending SMS messages.

The default spend quota per account (if not specified) is 1.00 USD per month. If you want to raise the quota, submit an SNS Quota Increase case. For New quota value, enter your desired monthly spend quota. In the Use Case Description field, explain that you are requesting an SMS monthly spend quota increase.

Amazon SNS sends SMS messages that you publish while the total cost incurred for your SMS traffic is below your spend quota for that calendar month. Once the spend quota is exceeded, Amazon SNS stops delivering messages until you either increase the spend quota or a new calendar month begins. Similarly, you can also specify a spend quota for an individual message, and Amazon SNS will send the message only if the cost is below the quota. Amazon SNS will not send your SMS messages if the account-level spend quota is exceeded, regardless of whether the message-level spend quota is exceeded.

Q: Is two-way SMS supported?

Amazon SNS does not currently support two-way SMS capabilities, except for opt out where required by local regulations.

Q: Do I need to subscribe phone numbers to an SNS Topic before sending an SMS message to it?

You no longer need to subscribe a phone number to an Amazon SNS topic before you publish messages to it. Now, you can directly publish messages to a phone number using the Amazon SNS console or the Publish request in the Amazon SNS API.

Q: Does AWS offer short codes for purchase?

Yes. You can reserve a dedicated short code that is assigned to your account and available exclusively to you.

To reserve a short code, create a case using the AWS Support Center. For more information, see Reserving a Dedicated Short Code for SMS Messaging in the *Amazon SNS Developer Guide*.

For pricing information, see Worldwide SMS Pricing.

Q: Does AWS offer long codes for purchase?

No. AWS does not currently offer long codes for purchase.

Q: Will SMS notifications come from a specific number of short codes or long codes?

Amazon SNS uses a pool of long codes or short codes to send SMS notifications. While there is a possibility that SMS notifications come from multiple numbers, Amazon SNS ensures that the messages sent from an AWS account to a specific phone number always come from the same long code or short code. This is called "Sticky Sender ID".

You can reserve a dedicated short code to ensure that all SMS messages that you send through Amazon SNS have a persistent short code. By reserving a short code, you make it easier for your audience to recognize that your organization is the source of your messages. For more information, see

Reserving a Dedicated Short Code for SMS Messaging in the *Amazon SNS Developer Guide*.

Q: Which countries does Amazon SNS support for Worldwide SMS?

Amazon SNS supports more than 200 countries, and we keep growing our reach. Please refer to the SMS Supported Country List for a comprehensive list of supported calling countries. For SMS message sending to China, please Contact Us.

Q: Which AWS regions support Worldwide SMS?

1) US-East-1 (Virginia), 2) US-West-2 (Oregon), 3) EU-West-1 (Dublin), 4) Asia Pacific (Tokyo), 5) Asia Pacific (Singapore), and 6) Asia Pacific (Sydney).

Q: Do the AWS phone numbers change?

Yes. Amazon SNS uses a pool of long codes or short codes to send SMS notifications. So while there is a possibility that SMS notifications come from multiple numbers, Amazon SNS ensures that the messages sent from an AWS account to a specific phone number, always come from the same long code or short code. This is called "Sticky Sender ID".

Q: Why do some devices on the same carrier receive messages from different phone numbers?

Amazon SNS uses a pool of long codes or short codes to send SMS notifications. So while there is a possibility that SMS notifications come from multiple numbers, Amazon SNS ensures that the messages sent from an AWS account to a specific phone number always come from the same long code or short code. This is called "Sticky Sender ID".

Q: What is the phone number format for sending messages to other countries?

AWS strongly encourages E.164 number formatting for all phone numbers both in the ‘to’ and ‘from’ (when applicable) fields. Please refer to the SMS Supported Country List for a comprehensive list of supported countries.

Q: Does Amazon SNS determine if a phone number is a mobile, landline, or VoIP number?

No. Currently, Amazon SNS does not detect whether a phone number is mobile, landline, or VoIP.

Q: Is time-based or scheduled delivery supported for SMS messages?

No. Amazon SNS does not currently support time-based or scheduled delivery.

Q: How do I track the delivery status of my SMS messages?

By enabling the Delivery Status feature in Amazon SNS, you can get information on the following for each message: MessageID, Time Sent, Destination Phone Number, Disposition, Disposition Reason (if applicable), Price, and Dwell Time.

Q: Do you support MMS?

No. Currently Amazon SNS does not support MMS messages.

Q: What is the cost of receiving SMS messages from Amazon SNS?

Costs for receiving SMS messages depend on the Data and Messaging of the recipient's wireless / mobile carrier plans.

Q: How do recipients opt out from receiving SMS messages from AWS?

Recipients can use their devices to opt out by replying to the message with any of the following:

- ARRET (French)
- CANCEL
- END
- OPT-OUT
- OPTOUT
- QUIT
- REMOVE

- STOP
- TD
- UNSUBSCRIBE

To opt out, the recipient must reply to the same long code or short code that Amazon SNS used to deliver the message. After opting out, the recipient will no longer receive SMS messages delivered from your AWS account unless you opt in the phone number.

Q: How do I know if a recipient device has ‘opted out’ of Global SMS?

The SNS console displays the list of opted out numbers for your account. Additionally, the Amazon SNS API provides the ListPhoneNumbersOptedOut request for listing opted out phone numbers.

Q: If a user opts out, will that number be unsubscribed automatically from the SNS Topic?

No. Opt-outs do not unsubscribe a number from an Amazon SNS topic, but rather disable the subscription. This means if you opt-in a phone number you do not need to re-subscribe the phone number to the topic.

Q: How do I confirm the end user received the SMS message?

You can use our Delivery Status feature to get information on the final disposition of your SMS message. For more information on the feature and how to use it, please refer to our documentation.

Q: Does Amazon SNS provide delivery receipts for SMS messages?

Our Delivery Status feature provides information based on delivery receipts received from the destination carrier. For more information on the Delivery Status feature and how to use it, please refer to our documentation.

Q: Does SMS support delivery to VoIP services like Google Voice or Hangouts?

Yes. Amazon SNS does support delivery to VoIP services that can receive SMS messages.

SMS pricing

Q: How much do you charge for sending SMS messages?

The price you pay for sending SMS messages varies based on the recipient's country or region, and may also vary based on the recipient's mobile carrier. You can find the latest rates on the SMS Pricing page.

Q: Why does the price for sending SMS messages to the same destination country and carrier keep changing?

The costs associated with sending SMS messages to different countries and regions—and even to different carriers within those countries and regions—can change frequently and with little or no notice. Carrier policies, technological changes, and even geopolitical issues can cause the prices for sending SMS messages to change.

We strive to be transparent by publishing the current SMS sending rates on the SMS Pricing Page.

Q: Am I charged if my SMS messages aren't delivered?

You may be charged for failed deliveries if the destination carrier reports that you attempted to send a message to an invalid phone number. Phone numbers can be invalid for several reasons, such as when the phone number doesn't exist, the recipient's account doesn't have sufficient credit, or the destination number is a landline number.

Q: Does the length of a message impact the price I pay?

Yes. A single SMS message can contain a maximum of 140 bytes of information. If a message contains more than 140 bytes, Amazon SNS automatically splits it into multiple messages. When Amazon SNS splits a long message into several smaller messages, you pay for each individual message.

The maximum number of characters in a single message depends on the way the characters are encoded. A message that includes characters encoded using GSM-7 (also known as GSM 03.38) encoding can include 160 characters. A message that uses ASCII encoding can contain up to 140 characters. A message that uses UCS-2 encoding can contain up to 70 characters. When you use Amazon SNS to send an SMS message, it automatically chooses the most compact encoding that supports all of the characters in that message.

For more information about sending SMS messages, see Sending an SMS Message in the *Amazon Simple Notification Service Developer Guide*.

Q: Is there an AWS Free Tier allowance for sending SMS messages?

Yes—the first 100 SMS messages you send to destinations in the United States each month are included in the AWS Free Tier. This allowance doesn't expire. Note that a "message" in this sense refers to a single transmission that contains 140 bytes of information or less, as explained in the answer to the previous question.

If you send more than 100 messages to destinations in the United States, or if you send messages to destinations outside the United States, each message you send is charged based on our current pricing rates. You can find the latest rates on the SMS Pricing page.

Quotas and restrictions

Q: Are there quotas for the number of topics or number of subscribers per topic?

By default, SNS offers 10 million subscriptions per topic, and 100,000 topics per account. To request a higher quota, please contact Support.

Q: How much and what kind of data can go in a message?

With the exception of SMS messages, Amazon SNS messages can contain up to 256 KB of text data, including XML, JSON and unformatted text.

The following Unicode characters are accepted:

> #x9 | #xA | #xD | [#x20 to #xD7FF] | [#xE000 to #xFFFD] | [#x10000 to #x10FFFF]
>
> (according to http://www.w3.org/TR/REC-xml/#charsets).

Each 64KB chunk of published data is billed as 1 request. For example, a single API call with a 256KB payload will be billed as four requests.

SMS messages

Each SMS message can contain up to 140 bytes, and the character limit depends on the encoding scheme. For example, an SMS message can contain:

- 160 GSM characters
- 140 ASCII characters
- 70 UCS-2 characters

If you publish a message that exceeds the size limit, Amazon SNS sends it as multiple messages, each fitting within the size limit. Messages are not cut off in the middle of a word but on whole-word boundaries. The total size limit for a single SMS publish action is 1600 bytes.

Q: How many message filters can be applied to a topic?

By default, 200 filter policies per account per region can be applied to a topic. Please contact us if more is required.

Q: Are there TCP ports that should be used for cross-region communication between SNS and EC2?

Yes, cross-region communication between SNS and EC2 on ports other than 80/443/4080/8443 is not guaranteed to work and should be avoided.

Raw message delivery

Q: What is raw message delivery?

You can opt-in to get your messages delivered in raw form, i.e. exactly as you published them. By default, messages are delivered encoded in JSON that provides metadata about the message and topic. Raw message delivery can be enabled by setting the “RawMessageDelivery” property on the subscriptions. This property can be set by using the AWS Management Console, or by using the API SetSubscriptionAttributes.

Q: What is the default behavior if the raw message delivery property on the subscription is not set?

By default, if this property is not set, messages will be delivered in JSON format, which is the current behavior. This ensures existing applications will continue to operate as expected.

Q: Which types of endpoints support raw message delivery?

Raw message delivery support is supported with SQS and HTTP(S) endpoints. Deliveries to Lambda, email, and SMS endpoints will behave the same independent of the “RawMessageDelivery” property.

Q: How will raw messages be delivered to HTTP endpoints?

When raw-formatted messages are delivered to HTTP/s endpoints, the message body will be included in the body of the HTTP POST.

Mobile push notifications

Q: What is SNS Mobile Push?

SNS Mobile Push lets you use Simple Notification Service (SNS) to deliver push notifications to Apple, Google, Fire OS, and Windows devices, as well as Android devices in China with Baidu Cloud Push. With push notifications, an installed mobile application can notify its users immediately by popping a notification about an event, without opening the application. For example, if you install a sports app and enable push notifications, the app can send you the latest score of your favorite team even if the app isn't running. The notification appears on your device, and when you acknowledge it, the app launches to display more information. Users' experiences are similar to receiving an SMS, but with enhanced functionality and at a fraction of the cost.

Q: How do I get started sending push notifications?

Push notifications can only be sent to devices that have your app installed, and whose users have opted in to receive them. SNS Mobile Push does not require explicit opt-in for sending push notifications, but iOS, Android and Kindle Fire operating systems do require it. In order to send push notifications with SNS, you must also register your app and each installed device with SNS. For more information, see Using Amazon SNS Mobile Push Notifications.

Q: Which push notifications platforms are supported?

Currently, the following push notifications platforms are supported:

- Amazon Device Messaging (ADM)
- Apple Push Notification Service (APNS)
- Firebase Cloud Messaging (FCM)
- Windows Push Notification Service (WNS) for Windows 8+ and Windows Phone 8.1+
- Microsoft Push Notification Service (MPNS) for Windows Phone 7+
- Baidu Cloud Push for Android devices in China

Q: How many push notifications can I send with the SNS Free Tier?

The SNS free tier includes 1 million publishes, plus 1 million mobile push deliveries. So you can send 1 million free push notifications every month. Notifications to all mobile push endpoints are all counted together toward your 1 million free mobile push deliveries.

Q: Does enabling push notifications require any special confirmations with SNS Mobile Push?

No, they do not. End-users opt-in to receive push notifications when they first run an app, whether or not SNS delivers the push notifications.

Q: Do I have to modify my client app to use SNS Mobile Push?

SNS does not require you to modify your client app. Baidu Cloud Push requires Baidu-specific components to be added to your client code in order to work properly, whether or not you choose to use SNS.

Q: How do SNS topics work with Mobile Push?

SNS topics can have subscribers from any supported push notifications platform, as well as any other endpoint type such as SMS or email. When you publish a notification to a topic, SNS will send identical copies of that message to each endpoint subscribed to the topic. If you use platform-specific payloads to define the exact payload sent to each push platform, the publish will fail if it exceeds the maximum payload size imposed by the relevant push notifications platform.

Q: What payload size is supported for various target platforms?

SNS will support maximum payload size that is supported by the underlying native platform. Customers can use a JSON object to send platform specific messages. See Using SNS Mobile Push API for additional details.

Q: How do platform-specific payloads work?

When you publish to a topic and want to have customized messages sent to endpoints for the different push notification platforms then you need to select “Use different message body for different protocols” option on the Publish dialog box and then update the messages. You can use platform-specific payloads to specify the exact API string that is relayed to each push notifications service. For example, you can use platform-specific payloads to manipulate the badge count of your iOS application via APNS. For more information, see Using Amazon SNS Mobile Push Notifications.

Q: Can one token subscribe to multiple topics?

Yes. Each token can be subscribed to an unlimited number of SNS topics.

Q: What is direct addressing? How does it work?

Direct addressing allows you to deliver notifications directly to a single endpoint, rather than sending identical messages to all subscribers of a topic. This is useful if you want to deliver precisely targeted messages to each recipient. When you register device tokens with SNS, SNS creates an endpoint that corresponds to the token. You can publish to the token endpoint just as you would publish to a topic. You can direct publish either the text of your notification, or a platform-specific payload that takes advantage of platform-specific features such as updating the badge count of your app. Direct addressing is currently only available for push notifications endpoints.

Q: Does SNS support direct addressing for SMS or Email?

At this time, direct addressing is only supported for mobile push endpoints (APNS, FCM, ADM, WNS, MPNS, Baidu) and SMS. Email messaging requires the use of topics.

Q: How does SNS Mobile Push handle token feedback from notification services?

Push notification services such as APNS and FCM provide feedback on tokens which may have expired or may have been replaced by new tokens. If either APNS or FCM reports that a particular token has either expired or is invalid, SNS automatically "disables" the application endpoint associated with the token, and notifies you of this change via an event. FCM specifically, at times not only indicates that a token is invalid, but also provides the new token associated with the application endpoint in its response to SNS. When this happens, SNS automatically updates the associated endpoint with the new token value, leaving the endpoint enabled, and then notifies you of this change via an event.

Q: I use Google Cloud Messaging (GCM) for SNS mobile notifications. What happens when GCM is deprecated?

GCM device tokens are completely interchangeable with the newer Firebase Cloud Messaging (FCM) device tokens. If you have existing GCM tokens, you’ll still be able to use them to send notifications. This statement is also true for

GCM tokens that you generate in the future. For more information please visit The End of Google Cloud Messaging, and What it Means for Your Apps blog.

Q: Can I migrate existing apps to SNS Mobile Push?

Yes. You can perform a bulk upload of existing device tokens to Amazon SNS, either via the console interface or API. You would also register your app with SNS by uploading your credentials for the relevant push notifications services, and configure your proxy or app to register future new tokens with SNS.

Q: Can I monitor my push notifications through Amazon CloudWatch?

Yes. SNS publishes Cloudwatch metrics for number of messages published, number of successful notifications, number of failed notifications, number of notifications filtered out, and size of data published. Metrics are available on per application basis. You can access Cloudwatch metrics via AWS Management Console or CloudWatch APIs.

Q: What types of Windows Push Notifications does Amazon SNS support?

SNS supports all types of push notifications types offered by Microsoft WNS and MPNS, including toast, tile, badge and raw notifications. Use the TYPE message attribute to specify which notification type you wish to use. When you use default payloads to send the same message to all mobile platforms, SNS will select toast notifications by default for Windows platforms. It is required to specify a notification type for Windows platforms when you use platform-specific payloads.

Q: Does SNS support Windows raw push notifications?

Yes. You must encode the notification payload as text to send raw notifications via SNS.

Q: What is Baidu Cloud Push?

Baidu Cloud Push is a third-party alternative push notifications relay service for Android devices. You can use Baidu Cloud Push to reach Android customers in China, no matter what Android app store those customers choose to use for downloading your app. For more information about Baidu Cloud Push, visit: http://developer.baidu.com/cloud/push.

Q: Can I publish Baidu notifications from all public AWS regions?

Yes, SNS supports Baidu push notifications from all public AWS regions.

Q: Can I use Baidu notifications to any Android app store?

Yes, Baidu push notifications work for apps installed via any Android app store.

Q: What are message attributes?

Message attributes allow you to provide structured metadata items (such as timestamps, geospatial data, signatures, and identifiers) about the message. Message attributes are optional and separate from, but sent along with, the message body. This information can be used by the receiver of the message to help decide how to handle the message without having to first process the message body.

You can use SNS message attributes in conjunction with SQS and mobile push endpoints. To learn more about message attributes, please see the SNS Getting Started Guide.

Q: What message attributes are supported in SNS?

SNS supports different message attributes for each endpoint type, depending on what the endpoint types each support themselves.

- **For SQS endpoints**, you can specify up to 10 name-type-value triples per message. Types supported include: String, Binary and Number (including integers, floating point, and doubles).
- **For mobile push endpoints**, you can take advantage of specific message attributes that each mobile platform supports (such as notification type).

Q: What is Time to Live (TTL)?

Some messages that you can send with SNS are relevant or valuable only for a limited period of time. Amazon SNS now allows you to set a TTL (Time to Live) value for each message. When the TTL expires for a given message that was not delivered and read by an end user, the message is deleted. TTL is specified in seconds and is relative to the time Publish call is made.

Q: How do I specify a TTL for my messages?

You can specify a TTL using the console or via API. TTL can be specified at publish time for a message, using the message attribute below. There is a different attribute for each platform. An attribute specified for a platform is applicable only for notification deliveries to that platform.

Q: What is the default TTL?

SNS uses a default Time to Live (TTL) of 4 weeks for all mobile platforms.

Q: Do TTL message attributes override TTLs specified in a message payload?

Yes. Google FCM and Amazon ADM allow you to specify a TTL within the message payload. If you specify TTL within the message payload and also within a message attribute, SNS will follow the message attribute.

Q: What happens if I specify TTL=0?

Some platforms treat TTL = 0 as a special case and attempt to deliver the message immediately, else let it expire. If you specify TTL = 0, SNS will relay your message to the appropriate service with TTL = 0 in order to take advantage of this special case.

Q: What SNS endpoints support TTL?

You can use TTL with the following mobile push endpoints: APNS, APNS_Sandbox, FCM, ADM, Baidu, and WNS. Microsoft MPNS does not currently support TTL. TTL is also not supported for SQS, HTTP, email or SMS endpoints.

Q: What does the Delivery Status feature of Amazon SNS do?

The Delivery Status feature lets you collect information on success rates, failure rates and dwell times of your push notifications for the supported mobile notification platforms. The currently supported platforms include Apple (APNS), Google (FCM), Windows (WNS and MPNS), Amazon (ADM), and Baidu. The status information is captured in the Amazon CloudWatch log groups created by Amazon SNS on your behalf. Additionally, you can create actionable metrics in Amazon CloudWatch and trigger alarms based on the patterns you are interested in.

Q: Is the Delivery Status feature in Amazon SNS available only for mobile push notifications? Do you plan to support this feature for other endpoint types?

Currently the Delivery Status feature is available for mobile push notifications and SMS. We will evaluate extending this to other endpoint types based on feedback from customers.

Q: How do I activate the Delivery Status feature?

You can activate the Delivery Status feature from the Amazon SNS console. From your Application, choose the Delivery Status option in the Application Actions drop-down menu. For details, please read our documentation.

Q: Can I activate the Delivery Status feature from the Amazon SNS APIs?

Yes, you can activate this feature from Amazon SNS APIs by adding the relevant application-level attributes. Our documentation goes over the application-level attributes that you need to add and the specific API calls that need to be made to enable this feature.

Q: How much does the Delivery Status feature cost?

There is currently no additional Amazon SNS charge for using the Delivery Status feature. However, depending upon your usage, you may incur charges for using CloudWatch since this feature creates Amazon CloudWatch log groups. Read our pricing page for more information about CloudWatch pricing and free tier.

Q: Why can you only choose a sampling percentage for successful delivery attempts and not sample failed delivery attempts?

Based on feedback we received from customers, we found that most developers are interested in knowing all the delivery attempt failures for their applications – and prefer to only store sample successful deliveries rather than logging all of them.

Q: How can I set alarms based on failure metrics or dwell time metrics?

After activating the Delivery Status feature, you need to define a Log Metrics Filter in Amazon CloudWatch Logs for the log group that gets created by Amazon SNS on your behalf. This metrics filter can be defined to extract information that you are interested in, such as failure rate and dwell time. Once a Metric Filter is defined, you can create it and assign it to a Metric. This metric can then be used to set alarms or send notifications based on thresholds you define. For more information, take a look at our documentation or blog.

SNS support for AWS Lambda

Q: What does support for AWS Lambda endpoints in Amazon SNS mean?

You can invoke your AWS Lambda functions by publishing messages to Amazon SNS topics that have AWS Lambda functions subscribed to them. Because Amazon SNS supports message fan-out, publishing a single message can invoke different AWS Lambda functions or invoke Lambda functions in addition to delivering notifications to supported Amazon SNS destinations such as mobile push, HTTP endpoints, SQS, email and SMS.

Q: What is AWS Lambda?

AWS Lambda is a compute service that runs your code in response to events and automatically manages the compute resources for you, making it easy to build applications that respond quickly to new information. More information on AWS Lambda and how to create AWS Lambda functions can be found here.

Q: What can I do with AWS Lambda functions and Amazon SNS?

By subscribing AWS Lambda functions to Amazon SNS topics, you can perform custom message handling. You can invoke an AWS Lambda function to provide custom message delivery handling by first publishing a message to an AWS Lambda function, have your Lambda function modify a message (e.g. localize language) and then filter and route those messages to other topics and endpoints. Apps and services that already send Amazon SNS notifications, such as Amazon CloudWatch, can now immediately take advantage of AWS Lambda without having to provision or manage infrastructure for custom message handling. You can also use delivery to an AWS Lambda function as a way to publish to other AWS services such as Amazon Kinesis or Amazon S3. You can subscribe an AWS Lambda function to the Amazon SNS topic, and then have the Lambda function in turn write to another service.

Q: How do I activate AWS Lambda endpoint support in Amazon SNS?

You need to first create an AWS Lambda function via your AWS account and the AWS Lambda console, and then subscribe that AWS Lambda function to a topic using the Amazon SNS console or the Amazon SNS APIs. Once that is complete, any messages that you publish to the Amazon SNS topics which have Lambda functions subscribed to them will be delivered to the appropriate Lambda functions in addition to any other destinations subscribed to that topic.

Q: What does delivery of a message from Amazon SNS to an AWS Lambda function do?

A message delivery from Amazon SNS to an AWS Lambda function creates an instance of the AWS Lambda function and invokes it with your message as an input. For more information on message formats, please refer to the Amazon SNS documentation and the AWS Lambda documentation.

Q: How much does this feature cost?

Publishing a message with Amazon SNS costs $0.50 per million requests. Aside from charges incurred in using AWS services, there are no additional fees for delivering a message to an AWS Lambda function. Amazon SNS has a Free Tier of 1 million requests per month. For more information, please refer to Amazon

SNS pricing. AWS Lambda function costs are based on the number of requests for your functions and the time your code executes. The AWS Lambda Free-Tier includes 1M requests per month and 400,000 GB-seconds of compute time per month. For more information, please refer to AWS Lambda pricing.

Q: Can I subscribe AWS Lambda functions created by someone else to Amazon SNS topics that I own?

We currently do not allow an AWS account owner to subscribe an AWS Lambda function that belongs to another account. You can subscribe your own AWS Lambda functions to your own Amazon SNS topics or subscribe your AWS Lambda functions to an Amazon SNS topic that was created by another account so long as the topic policy for that SNS topic allows it.

Q: Is there a quota to the number of AWS Lambda functions that I can subscribe to an Amazon SNS topic?

Amazon SNS treats AWS Lambda functions like any other destination. By default, SNS offers 10 million subscriptions per topic. To request a higher quota, please contact us.

Q: What data can I pass to my AWS Lambda function?

When an AWS Lambda function is invoked as a result of an Amazon SNS message delivery, the AWS Lambda function receives data such as the Message ID, the topic ARN, the message payload and message attributes via an SNS Event. For more information on the event structure passed to the AWS Lambda function please read our blog.

Q: Can I track delivery status for message delivery attempts to AWS Lambda functions?

To track the success or failure status of message deliveries, you need to activate the Delivery Status feature of Amazon SNS. For more information about how to activate this feature please read our blog.

Q: What regions is AWS Lambda available in?

See AWS Regions and Endpoints for a complete list.

Q: Do my AWS Lambda functions need to be in the same region as my Amazon SNS usage?

You can subscribe your AWS Lambda functions to an Amazon SNS topic in any region.

Q: Are there any data transfer costs for invoking AWS Lambda functions?

Data transfer costs are applicable to message deliveries to AWS Lambda functions. Please refer to our pricing for more information.

Q: Are there any quotas to the concurrency of AWS Lambda functions?

AWS Lambda currently supports 1000 concurrent executions per AWS account per region. If your Amazon SNS message deliveries to AWS Lambda contribute to crossing these concurrency quotas, your Amazon SNS message deliveries will be throttled. If AWS Lambda throttles an Amazon SNS message, Amazon SNS will retry the delivery attempts. For more information about AWS Lambda concurrency quotas, please refer to AWS Lambda documentation.

Q: Can Amazon SNS use the same AWS Lambda functions that I use with other services (e.g. Amazon S3)?

You can use the same AWS Lambda functions that you use with other services as long as the same function can parse the event formats from Amazon SNS in addition to the event format of the other services. For the SNS event format please read our blog.

VoIP iOS and Mac OS notifications

Q: What are VoIP Push Notifications for iOS?

In iOS 8 and later, voice-over-IP (VoIP) apps can register for VoIP remote notifications such that iOS can launch or wake the app, as appropriate, when an incoming VoIP call arrives for the user. The procedure to register for VoIP notifications is similar to registering for regular push notifications on iOS. For more information, please refer to our documentation.

Q: Can I use VoIP Push Notifications and other Push Notifications in the same iOS app?

Yes, you can have an iOS application that is registered to receive both types of push notifications. However, you will need to obtain the VoIP push notification certificate from Apple in addition to the regular push notification certificate and create a new Platform Application in Amazon SNS and choose Apple VoIP Push as the platform type. For more information, please refer to our documentation.

Q: What are Mac OS push notifications?

You can now send push notifications to Mac OS desktops that run Mac OS X Lion (10.7) or later using Amazon SNS. For more information, please refer to our documentation.

Amazon SQS FAQs

Overview

Q: What are the benefits of Amazon SQS over homegrown or packaged message queuing systems?

Amazon SQS provides several advantages over building your own software for managing message queues or using commercial or open-source message queuing systems that require significant up-front time for development and configuration.

These alternatives require ongoing hardware maintenance and system administration resources. The complexity of configuring and managing these systems is compounded by the need for redundant storage of messages that ensures messages are not lost if hardware fails.

In contrast, Amazon SQS requires no administrative overhead and little configuration. Amazon SQS works on a massive scale, processing billions of messages per day. You can scale the amount of traffic you send to Amazon SQS up or down without any configuration. Amazon SQS also provides extremely high message durability, giving you and your stakeholders added confidence.

Q: How is Amazon SQS different from Amazon SNS?

Amazon SNS allows applications to send time-critical messages to multiple subscribers through a “push” mechanism, eliminating the need to periodically check or “poll” for updates. Amazon SQS is a message queue service used by distributed applications to exchange messages through a polling model, and can be used to decouple sending and receiving components.

Q: How is Amazon SQS different from Amazon MQ?

If you're using messaging with existing applications, and want to move your messaging to the cloud quickly and easily, we recommend you consider Amazon MQ. It supports industry-standard APIs and protocols so you can switch from any standards-based message broker to Amazon MQ without rewriting the messaging code in your applications. If you are building brand new applications in the cloud, we recommend you consider Amazon SQS and Amazon SNS. Amazon SQS and SNS are lightweight, fully managed message queue and topic services that scale almost infinitely and provide simple, easy-to-use APIs.

Q: Does Amazon SQS provide message ordering?

Yes. FIFO (first-in-first-out) queues preserve the exact order in which messages are sent and received. If you use a FIFO queue, you don't have to place sequencing information in your messages. For more information, see FIFO Queue Logic in the *Amazon SQS Developer Guide*.

Standard queues provide a loose-FIFO capability that attempts to preserve the order of messages. However, because standard queues are designed to be massively scalable using a highly distributed architecture, receiving messages in the exact order they are sent is not guaranteed.

Q: Does Amazon SQS guarantee delivery of messages?

Standard queues provide at-least-once delivery, which means that each message is delivered at least once.

FIFO queues provide exactly-once processing, which means that each message is delivered once and remains available until a consumer processes it and deletes it. Duplicates are not introduced into the queue.

Q: How is Amazon SQS different from Amazon Kinesis Streams?

Amazon SQS offers a reliable, highly-scalable hosted queue for storing messages as they travel between applications or microservices. It moves data between distributed application components and helps you decouple these components. Amazon SQS provides common middleware constructs such as dead-letter queues and poison-pill management. It also provides a generic web services API and can be accessed by any programming language that the AWS SDK supports. Amazon SQS supports both standard and FIFO queues.

Amazon Kinesis Streams allows real-time processing of streaming big data and the ability to read and replay records to multiple Amazon Kinesis Applications. The Amazon Kinesis Client Library (KCL) delivers all records for a given partition key to the same record processor, making it easier to build multiple applications that read from the same Amazon Kinesis stream (for example, to perform counting, aggregation, and filtering).

For more information, see the Amazon Kinesis Documentation.

Q: Does Amazon use Amazon SQS for its own applications?

Yes. Developers at Amazon use Amazon SQS for a variety of applications that process large numbers of messages every day. Key business processes in both Amazon.com and Amazon Web Services use Amazon SQS.

Billing

Q: How much does Amazon SQS cost?

You pay only for what you use, and there is no minimum fee.

The cost of Amazon SQS is calculated per request, plus data transfer charges for data transferred out of Amazon SQS (unless data is transferred to Amazon EC2 instances or to AWS Lambda functions within the same region). For detailed pricing breakdowns per queue type and region, see Amazon SQS Pricing.

Q: What can I do with the Amazon SQS Free Tier?

The Amazon SQS Free Tier provides you with 1 million requests per month at no charge.

Many small-scale applications are able to operate entirely within the limits of the Free Tier. However, data transfer charges might still apply. For more information, see Amazon SQS Pricing.

The Free Tier is a monthly offer. Free usage does not accumulate across months.

Q: Will I be charged for all Amazon SQS requests?

Yes, for any requests beyond the free tier. All Amazon SQS requests are chargeable, and they are billed at the same rate.

Q: Do Amazon SQS batch operations cost more than other requests?

No. Batch operations (SendMessageBatch, DeleteMessageBatch, and ChangeMessageVisibilityBatch) all cost the same as other Amazon SQS requests. By grouping messages into batches, you can reduce your Amazon SQS costs.

Q: How will I be charged and billed for my use of Amazon SQS?

There are no initial fees to begin using Amazon SQS. At the end of the month, your credit card will be automatically charged for the month's usage.

You can view your charges for the current billing period at any time on the AWS website:

1. Log into your AWS account.
2. Under **Your Web Services Account**, select **Account Activity**.

Q: How can I track and manage the costs associated with my Amazon SQS queues?

You can tag and track your queues for resource and cost management using cost allocation tags. A tag is a metadata label comprised of a key-value pair. For example, you can tag your queues by cost center and then categorize and track your costs based on these cost centers.

For more information, see Tagging Your Amazon SQS Queues in the Amazon SQS Developer Guide. For more information on cost allocation tagging of AWS resources, see Using Cost Allocation Tags in the AWS Billing and Cost Management User Guide.

Q: Do your prices include taxes?

Except as noted otherwise, our prices don't include any applicable taxes and duties such as VAT or applicable sales tax.

For customers with a Japanese billing address, the use of AWS in any region is subject to Japanese Consumption Tax. For more information, see the Amazon Web Services Consumption Tax FAQ.

Features, functionality, and interfaces

Q: Can I use Amazon SQS with other AWS services?

Yes. You can make your applications more flexible and scalable by using Amazon SQS with compute services such as Amazon EC2, Amazon EC2 Container Service (Amazon ECS), and AWS Lambda, as well as with storage and database services such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.

Q: How can I interact with Amazon SQS?

You can access Amazon SQS using the AWS Management Console, which helps you create Amazon SQS queues and send messages easily.

Amazon SQS also provides a web services API. It is also integrated with the AWS SDKs, allowing you to work in the programming language of your choice.

Q: What API actions are available for Amazon SQS?

For information on message queue operations, see the Amazon SQS API Reference.

Q: Who can perform operations on a message queue?

Only an AWS account owner (or an AWS account that the account owner has delegated rights to) can perform operations on an Amazon SQS message queue.

Q: Can I use Java Message Service (JMS) with Amazon SQS?

Yes. You can take advantage of the scale, low cost, and high availability of Amazon SQS without the worry and high overhead of running your own JMS cluster.

Amazon provides the Amazon SQS Java Messaging Library that implements the JMS 1.1 specification and uses Amazon SQS as the JMS provider. For more information, see Using JMS with Amazon SQS in the *Amazon SQS Developer Guide*.

Q: How does Amazon SQS identify messages?

All messages have a global unique ID that Amazon SQS returns when the message is delivered to the message queue. The ID isn’t required to perform any further actions on the message, but it is useful for tracking the receipt of a particular message in the message queue.

When you receive a message from the message queue, the response includes a receipt handle that you must provide when deleting the message.

For more information, see Queue and Message Identifiers in the Amazon SQS Developer Guide.

Q: How does Amazon SQS handle messages that can't be processed?

In Amazon SQS, you can use the API or the console to configure dead letter queues, which are queues that receive messages from other source queues.

If you make a queue into a dead letter queue, it receives messages after a maximum number of processing attempts cannot be completed. You can use dead letter queues to isolate messages that can't be processed for later analysis.

For more information, see "Can I use a dead letter queue with FIFO queues?" on this page and Using Amazon SQS Dead Letter Queues in the Amazon SQS Developer Guide.

Q: What is a visibility timeout?

The visibility timeout is a period of time during which Amazon SQS prevents other consuming components from receiving and processing a message. For more information, see Visibility Timeout in the *Amazon SQS Developer Guide*.

Q: Does Amazon SQS support message metadata?

Yes. An Amazon SQS message can contain up to 10 metadata attributes. You can use message attributes to separate the body of a message from the metadata that describes it. This helps process and store information with greater speed and efficiency because your applications don't have to inspect an entire message before understanding how to process it.

Amazon SQS message attributes take the form of name-type-value triples. The supported types include string, binary, and number (including integer, floating-point, and double). For more information, see Using Amazon SQS Message Attributes in the *Amazon SQS Developer Guide*.

Q: How can I determine the time-in-queue value?

To determine the time-in-queue value, you can request the SentTimestamp attribute when receiving a message. Subtracting that value from the current time results in the time-in-queue value.

Q: What is the typical latency for Amazon SQS?

Typical latencies for SendMessage, ReceiveMessage, and DeleteMessage API requests are in the tens or low hundreds of milliseconds.

Q: For anonymous access, what is the value of the SenderId attribute for a message?

When the AWS account ID is not available (for example, when an anonymous user sends a message), Amazon SQS provides the IP address.

Q: What is Amazon SQS long polling?

Amazon SQS long polling is a way to retrieve messages from your Amazon SQS queues. While the regular short polling returns immediately, even if the message queue being polled is empty, long polling doesn’t return a response until a message arrives in the message queue, or the long poll times out.

Long polling makes it inexpensive to retrieve messages from your Amazon SQS queue as soon as the messages are available. Using long polling might reduce the cost of using SQS, because you can reduce the number of empty receives. For more information, see Amazon SQS Long Polling in the *Amazon SQS Developer Guide*.

Q: Is there an additional charge for using Amazon SQS long polling?

No. Long-polling ReceiveMessage calls are billed exactly the same as short-polling ReceiveMessage calls.

Q: When should I use Amazon SQS long polling, and when should I use Amazon SQS short polling?

In almost all cases, Amazon SQS long polling is preferable to short polling. Long-polling requests let your queue consumers receive messages as soon as they arrive in your queue while reducing the number of empty ReceiveMessageResponse instances returned.

Amazon SQS long polling results in higher performance at reduced cost in the majority of use cases. However, if your application expects an immediate response from a ReceiveMessage call, you might not be able to take advantage of long polling without some modifications to your application.

For example, if your application uses a single thread to poll multiple queues, switching from short polling to long polling will probably not work, because the single thread will wait for the long-poll timeout on any empty queues, delaying the processing of any queues that might contain messages.

In such an application, it is a good practice to use a single thread to process only one queue, allowing the application to take advantage of the benefits that Amazon SQS long polling provides.

Q: What value should I use for my long-poll timeout?

In general, you should use maximum 20 seconds for a long-poll timeout. Because higher long-poll timeout values reduce the number of empty

ReceiveMessageResponse instances returned, try to set your long-poll timeout as high as possible.

If the 20-second maximum doesn't work for your application (see the example in the previous question), set a shorter long-poll timeout, as low as 1 second.

All AWS SDKs work with 20-second long polls by default. If you don't use an AWS SDK to access Amazon SQS, or if you configured your AWS SDK to specifically have a shorter timeout, you might need to modify your Amazon SQS client to allow longer requests or to use a shorter long-poll timeout.

Q: What is the AmazonSQSBufferedAsyncClient for Java?

The AmazonSQSBufferedAsyncClient for Java provides an implementation of the AmazonSQSAsyncClient interface and adds several important features:

- Automatic batching of multiple SendMessage, DeleteMessage, or ChangeMessageVisibility requests without any required changes to the application
- Prefetching of messages into a local buffer that allows your application to immediately process messages from Amazon SQS without waiting for the messages to be retrieved

Working together, automatic batching and prefetching increase the throughput and reduce the latency of your application while reducing your costs by making fewer Amazon SQS requests. For more information, see Client-Side Buffering and Request Batching in the *Amazon SQS Developer Guide*.

Q: Where can I download the AmazonSQSBufferedAsyncClient for Java?

You can download the AmazonSQSBufferedAsyncClient as part of the AWS SDK for Java.

Q: Do I have to rewrite my application to use the AmazonSQSBufferedAsyncClient for Java?

No. The AmazonSQSBufferedAsyncClient for Java is implemented as a drop-in replacement for the existing AmazonSQSAsyncClient.

If you update your application to use the latest AWS SDK and change your client to use the AmazonSQSBufferedAsyncClient for Java instead of the AmazonSQSAsyncClient, your application will receive the added benefits of automatic batching and prefetching.

Q: How can I subscribe Amazon SQS message queues to receive notifications from Amazon SNS topics?

1. In the Amazon SQS console, select an Amazon SQS standard queue.
2. Under Queue Actions, select Subscribe Queue to SNS Topic from the drop-down list.
3. In the dialog box, select the topic from the Choose a Topic drop-down list, and click Subscribe.

For more information, see Subscribing a Queue to an Amazon SNS Topic in the *Amazon SQS Developer Guide*.

Q: Can I delete all messages in a message queue without deleting the message queue itself?

Yes. You can delete all messages in an Amazon SQS message queue using the PurgeQueue action.

When you purge a message queue, all the messages previously sent to the message queue are deleted. Because your message queue and its attributes remain, there is no need to reconfigure the message queue; you can continue using it.

To delete only specific messages, use the DeleteMessage or DeleteMessageBatch actions.

For more information, see this Tutorial: Purging Messages from an Amazon SQS Queue.

FIFO queues

Q: What regions are FIFO queues available in?

FIFO queues are currently available in the following regions:

- US West (Oregon, N. California)
- US East (Ohio, N. Virginia)
- GovCloud (US-East)
- GovCloud (US-West)
- EU (Ireland, Frankfurt, London, Paris, Stockholm)
- Asia Pacific (Sydney, Tokyo, Mumbai, Seoul, Singapore)
- Canada (Central)
- South America (Sao Paulo)
- China (Ningxia), operated by NWCD
- China (Beijing), operated by SINNET
- Asia Pacific (Hong Kong)

Q: How many copies of a message will I receive?

FIFO queues are designed to never introduce duplicate messages. However, your message producer might introduce duplicates in certain scenarios: for example, if the producer sends a message, does not receive a response, and then resends the same message. Amazon SQS APIs provide deduplication functionality that prevents your message producer from sending duplicates. Any duplicates introduced by the message producer are removed within a 5-minute deduplication interval.

For standard queues, you might occasionally receive a duplicate copy of a message (at-least-once delivery). If you use a standard queue, you must design your applications to be idempotent (that is, they must not be affected adversely when processing the same message more than once).

For more information, see Exactly-Once Processing in the Amazon SQS Developer Guide.

Q: Are the Amazon SQS queues I used previously changing to FIFO queues?

No. Amazon SQS *standard* queues (the new name for existing queues) remain unchanged, and you can still create standard queues. These queues continue to provide the highest scalability and throughput; however, you will not get ordering guarantees and duplicates might occur.

Standard queues are appropriate for many scenarios, such as work distribution with multiple idempotent consumers.

Q: Can I convert my existing standard queue to a FIFO queue?

No. You must choose the queue type when you create it. However, it is possible to move to a FIFO queue. For more information, see Moving From a Standard Queue to a FIFO Queue in the *Amazon SQS Developer Guide*.

Q: Are Amazon SQS FIFO queues backwards-compatible?

To take advantage of FIFO queue functionality, you must use the latest AWS SDK.

FIFO queues use the same API actions as standard queues, and the mechanics for receiving and deleting messages and changing the visibility timeout are the same. However, when sending messages, you must specify a message group ID. For more information, see FIFO Queue Logic in the *Amazon SQS Developer Guide*.

Q: With which AWS or external services are Amazon SQS FIFO queues compatible?

Some AWS or external services that send notifications to Amazon SQS might not be compatible with FIFO queues, despite allowing you to set a FIFO queue as a target.

The following features of AWS services aren't currently compatible with FIFO queues:

- Auto Scaling Lifecycle Hooks
- AWS IoT Rule Actions

- AWS Lambda Dead Letter Queues

For information about compatibility of other services with FIFO queues, see your service documentation.

Q: Are Amazon SQS FIFO queues compatible with the Amazon SQS Buffered Asynchronous Client, the Amazon SQS Extended Client Library for Java, or the Amazon SQS Java Message Service (JMS) Client?

FIFO queues aren't currently compatible with the Amazon SQS Buffered Asynchronous Client.

FIFO queues are compatible with the Amazon SQS Extended Client Library for Java and the Amazon SQS Java Message Service (JMS) client.

Q: Which AWS CloudWatch metrics do Amazon SQS FIFO queues support?

FIFO queues support all metrics that standard queues support. For FIFO queues, all approximate metrics return accurate counts. For example, the following AWS CloudWatch metrics are supported:

- ApproximateNumberOfMessagesDelayed - The number of messages in the queue that are delayed and not available for reading immediately.
- ApproximateNumberOfMessagesVisible - The number of messages available for retrieval from the queue.
- ApproximateNumberOfMessagesNotVisible - The number of messages that are in flight (sent to a client but have not yet been deleted or have not yet reached the end of their visibility window).

Q: What are message groups?

Messages are grouped into distinct, ordered "bundles" within a FIFO queue. For each message group ID, all messages are sent and received in strict order. However, messages with different message group ID values might be sent and received out of order. You must associate a message group ID with a message. If you don't provide a message group ID, the action fails.

If multiple hosts (or different threads on the same host) send messages with the same message group ID are sent to a FIFO queue, Amazon SQS delivers the messages in the order in which they arrive for processing. To ensure that Amazon SQS preserves the order in which messages are sent and received, ensure that multiple senders send each message with a unique message group ID.

For more information, see FIFO Queue Logic in the *Amazon SQS Developer Guide*.

Q: Do Amazon SQS FIFO queues support multiple producers?

Yes. One or more producers can send messages to a FIFO queue. Messages are stored in the order that they were successfully received by Amazon SQS.

If multiple producers send messages in parallel, without waiting for the success response from SendMessage or SendMessageBatch actions, the order between producers might not be preserved. The response of SendMessage or SendMessageBatch actions contains the final ordering sequence that FIFO queues use to place messages in the queue, so your multiple-parallel-producer code can determine the final order of messages in the queue.

Q: Do Amazon SQS FIFO queues support multiple consumers?

By design, Amazon SQS FIFO queues don't serve messages from the same message group to more than one consumer at a time. However, if your FIFO queue has multiple message groups, you can take advantage of parallel consumers, allowing Amazon SQS to serve messages from different message groups to different consumers.

Q: Can I use a dead letter queue with FIFO queues?

Yes. However, you must use a FIFO dead letter queue with a FIFO queue. (Similarly, you can use only a standard dead letter queue with a standard queue.)

Q: What is the throughput limit for an Amazon SQS FIFO queue?

By default, FIFO queues support up to 3,000 messages per second with batching, or up to 300 messages per second (300 send, receive, or delete operations per second) without batching. If you require a higher throughput, submit a support ticket to request a review of your FIFO queue requirements.

Q: Are there any limits specific to FIFO queue attributes?

The name of a FIFO queue must end with the .fifo suffix. The suffix counts towards the 80-character queue name limit. To determine whether a queue is FIFO, you can check whether the queue name ends with the suffix.

Security and reliability

Q: How reliable is the storage of my data in Amazon SQS?

Amazon SQS stores all message queues and messages within a single, highly-available AWS region with multiple redundant Availability Zones (AZs), so that no single computer, network, or AZ failure can make messages inaccessible. For more information, see Regions and Availability Zones in the *Amazon Relational Database Service User Guide*.

Q: How can I secure the messages in my message queues?

Authentication mechanisms ensure that messages stored in Amazon SQS message queues are secured against unauthorized access. You can control who can send messages to a message queue and who can receive messages from a message queue. For additional security, you can build your application to encrypt messages before they are placed in a message queue.

Amazon SQS has its own resource-based permissions system that uses policies written in the same language as AWS Identity and Access Management (IAM) policies: for example, you can use variables, just like in IAM policies. For more information, see Amazon SQS Policy Examples in the *Amazon SQS Developer Guid*e.

Amazon SQS supports the HTTP over SSL (HTTPS) and Transport Layer Security (TLS) protocols. Most clients can automatically negotiate to use newer versions of TLS without any code or configuration change. Amazon SQS supports versions 1.0, 1.1, and 1.2 of the Transport Layer Security (TLS) protocol in all regions.

Q: Why are there separate ReceiveMessage and DeleteMessage operations?

When Amazon SQS returns a message to you, the message stays in the message queue whether or not you actually receive the message. You're responsible for deleting the message and the deletion request acknowledges that you’re done processing the message.

If you don’t delete the message, Amazon SQS will deliver it again on when it receives another receive request. For more information, see Visibility Timeout in the *Amazon SQS Developer Guide*.

Q: Can a deleted message be received again?

No. FIFO queues never introduce duplicate messages.

For standard queues, under rare circumstances, you might receive a previously-deleted message a second time.

Q: What happens if I issue a DeleteMessage request on a previously-deleted message?

When you issue a DeleteMessage request on a previously-deleted message, Amazon SQS returns a *success* response.

Server-Side encryption (SSE)

Q: What are the benefits of server-side encryption (SSE) for Amazon SQS?

Server-side encryption (SSE) lets you transmit sensitive data in encrypted queues. SSE protects the contents of messages in Amazon SQS queues using keys managed in the AWS Key Management Service (AWS KMS). SSE encrypts messages as soon as Amazon SQS receives them. The messages are stored in encrypted form and Amazon SQS decrypts messages only when they are sent to an authorized consumer.

For more information, see Protecting Data Using Server-Side Encryption (SSE) and AWS KMS in the Amazon SQS Developer Guide

Q: Can I use SNS, Cloud Watch Events and S3 Events with encrypted queues?

Yes. To do this you need to enable compatibility between AWS services (eg. Amazon CloudWatch Events, Amazon S3, and Amazon SNS), and Queues with SSE. For detailed instructions see the Compatibility section of the SQS Developer Guide.

Q: What regions are queues with SSE available in?

Server-side encryption (SSE) for Amazon SQS is available in the following regions: Asia Pacific (Mumbai, Osaka, Seoul, Singapore, Sydney, Tokyo, Ningxia, and Beijing), Canada (Central), EU (Frankfurt, Ireland, London, and Paris), South America (Sao Paulo), US West (N. California, Oregon), US East (N. Virginia, Ohio) and GovCloud (US).

Q: How do I enable SSE for a new or existing Amazon SQS queue?

To enable SSE for a new or existing queue using the Amazon SQS API, specify the customer master key (CMK) ID: the alias, alias ARN, key ID, or key ARN of the an AWS-managed CMK or a custom CMK by setting the KmsMasterKeyId attribute of the CreateQueue or SetQueueAttributes action.

For detailed instructions, see Creating an Amazon SQS Queue with Server-Side Encryption and Configuring Server-Side Encryption (SSE) for an Existing Amazon SQS Queue in the *Amazon SQS Developer Guide*.

Q: What Amazon SQS queue types can use SSE?

Both standard and FIFO queues support SSE.

Q: What permissions do I need to use SSE with Amazon SQS?

Before you can use SSE, you must configure AWS KMS key policies to allow encryption of queues and encryption and decryption of messages.

To enable SSE for a queue, you can use the AWS-managed customer master key (CMK) for Amazon SQS or a custom CMK. For more information, see Customer Master Keys in the *AWS KMS Developer Guide*.

To send messages to an encrypted queue, the producer must have the kms:GenerateDataKey and kms:Decrypt permissions for the CMK.

To receive messages from an encrypted queue, the consumer must have the kms:Decrypt permission for any CMK that is used to encrypt the messages in the specified queue. If the queue acts as a dead letter queue, the consumer must also have the kms:Decrypt permission for any CMK that is used to encrypt the messages in the source queue.

For more information, see What Permissions Do I Need to Use SSE? in the *Amazon SQS Developer Guide*.

Q: Are there any charges for using SSE with Amazon SQS?

There are no additional Amazon SQS charges. However, there are charges for calls from Amazon SQS to AWS KMS. For more information, see AWS Key Management Service Pricing.

The charges for using AWS KMS depend on the data key reuse period configured for your queues. For more information, see How Do I Estimate My AWS KMS Usage Costs? in the *Amazon SQS Developer Guide*.

Q: What does SSE for Amazon SQS encrypt and how is it encrypted?

SSE encrypts the body of a message in an Amazon SQS queue.

SSE doesn't encrypt the following components:

- Queue metadata (queue name and attributes)
- Message metadata (message ID, timestamp, and attributes)
- Per-queue metrics

Amazon SQS generates data keys based on the AWS-managed customer master key (CMK) for Amazon SQS or a custom CMK to provide envelope encryption and decryption of messages for a configurable time period (from 1 minute to 24 hours).

For more information, see What Does SSE for Amazon SQS Encrypt? in the *Amazon SQS Developer Guide.*

Q: What algorithm does SSE for Amazon SQS use to encrypt messages?

SSE uses the AES-GCM 256 algorithm.

Q: Does SSE limit the transactions per second (TPS) or number of queues that can be created with Amazon SQS?

SSE doesn't limit the throughput (TPS) of Amazon SQS. The number of SSE queues that you can create is limited by the following:

- The data key reuse period (1 minute to 24 hours).
- The AWS KMS per-account limit (100 TPS by default).
- The number of IAM users or accounts that access queues.
- The existence of a large backlog (a larger backlog requires more AWS KMS calls).

For example, let's assume the following limits:

- You set your data key reuse period to 5 minutes (300 seconds).
- Your KMS account has a default AWS KMS TPS limit of 100 TPS.
- You use an Amazon SQS queue without a backlog and with 1 IAM user for
- SendMessage or ReceiveMessage actions to all queues.

In this case, you can calculate the theoretical maximum of Amazon SQS queues with SSE as follows:

300 seconds × 100 TPS / 1 IAM user = 30,000 queues

Q: How can I estimate my AWS KMS usage costs?

To predict costs and better understand your AWS bill, you might want to know how often Amazon SQS uses your CMK.

> **Note**: Although the following formula can give you a very good idea of expected costs, actual costs might be higher because of the distributed nature of Amazon SQS.

To calculate the number of API requests per queue (R), use the following formula:

> **R = B / D * (2 * P + C)**

B is the billing period (in seconds)

D is the data key reuse period (in seconds)

P is the number of producing principals that send to the Amazon SQS queue.

C is the number of consuming principals that receive from the Amazon SQS queue.

> **Important**: In general, producing principals incur double the cost of consuming principals. For more information, see How Does the Data Key Reuse Period Work? in the *Amazon SQS Developer Guide*.
>
> If the producer and consumer have different IAM users, the cost increases.

For more information, see How Do I Estimate My AWS KMS Usage Costs? in the *Amazon SQS Developer Guide*

Compliance

Q: Is Amazon SQS PCI DSS certified?

Yes. Amazon SQS is PCI DSS Level 1 certified. For more information, see PCI Compliance.

Q: Is Amazon SQS HIPAA-eligible?

Yes, AWS has expanded its HIPAA compliance program to include Amazon SQS as a HIPAA Eligible Service. If you have an executed Business Associate Agreement (BAA) with AWS, you can use Amazon SQS to build your HIPAA-compliant applications, store messages in transit, and transmit messages—including messages containing protected health information (PHI).

If you already have an executed BAA with AWS, you can start using Amazon SQS right away. If you don't have a BAA or have other questions about using AWS for your HIPAA-compliant applications, contact us for more information.

Note: If you prefer not to transfer PHI through Amazon SQS (or if you have messages larger than 256 KB), you can alternatively send Amazon SQS message payloads through Amazon S3 using the Amazon SQS Extended Client Library for Java (Amazon S3 is a HIPAA Eligible Service, excluding the use of Amazon S3 Transfer Acceleration). For more information, see Using the Amazon SQS Extended Client Library for Java in the *Amazon SQS Developer Guide*.

Limits and restrictions

Q: How long can I keep my messages in Amazon SQS message queues?

Longer message retention provides greater flexibility to allow for longer intervals between message production and consumption.

You can configure the Amazon SQS message retention period to a value from 1 minute to 14 days. The default is 4 days. Once the message retention limit is reached, your messages are automatically deleted.

Q: How do I configure Amazon SQS to support longer message retention?

To configure the message retention period, set the MessageRetentionPeriod attribute using the console or using the Distributiveness method. Use this attribute to specify the number of seconds a message will be retained in Amazon SQS.

You can use the MessageRetentionPeriod attribute to set the message retention period from 60 seconds (1 minute) to 1,209,600 seconds (14 days). For more information on working with this message attribute, see the *Amazon SQS API Reference*.

Q: How do I configure the maximum message size for Amazon SQS?

To configure the maximum message size, use the console or the SetQueueAttributes method to set the MaximumMessageSize attribute. This attribute specifies the limit on bytes that an Amazon SQS message can contain. Set this limit to a value between 1,024 bytes (1 KB), and 262,144 bytes (256 KB). For more information, see Using Amazon SQS Message Attributes in the *Amazon SQS Developer Guide*.

To send messages larger than 256 KB, use the Amazon SQS Extended Client Library for Java. This library lets you send an Amazon SQS message that contains a reference to a message payload in Amazon S3 that can be as large as 2 GB.

Q: What kind of data can I include in a message?

Amazon SQS messages can contain up to 256 KB of text data, including XML, JSON and unformatted text. The following Unicode characters are accepted:

#x9 | #xA | #xD | [#x20 to #xD7FF] | [#xE000 to #xFFFD] | [#x10000 to #x10FFFF]

For more information, see the XML 1.0 Specification.

Q: How large can Amazon SQS message queues be?

A single Amazon SQS message queue can contain an unlimited number of messages. However, there is a 120,000 limit for the number of inflight messages for a standard queue and 20,000 for a FIFO queue. Messages are inflight after they have been received from the queue by a consuming component, but have not yet been deleted from the queue.

Q: How many message queues can I create?

You can create any number of message queues.

Q: Is there a size limit on the name of Amazon SQS message queues?

Queue names are limited to 80 characters.

Q: Are there restrictions on the names of Amazon SQS message queues?

You can use alphanumeric characters, hyphens (-), and underscores (_).

Q: Can I reuse a message queue name?

A message queue's name must be unique within an AWS account and region. You can reuse a message queue's name after you delete the message queue.

Queue sharing

Q: How do I share a message queue?

You can associate an access policy statement (and specify the permissions granted) with the message queue to be shared. Amazon SQS provides APIs for creating and managing access policy statements:

- AddPermission
- RemovePermission
- SetQueueAttributes
- GetQueueAttributes

For more information, see the *Amazon SQS API Reference*.

Q: Who pays for shared queue access?

The message queue owner pays for shared message queue access.

Q: How do I identify another AWS user I want to share a message queue with?

The Amazon SQS API uses the AWS account number to identify AWS users.

Q: What do I need to provide to an AWS user I want to share a message queue with?

To share a message queue with an AWS user, provide the full URL from the message queue you want to share. The CreateQueue and ListQueues operations return this URL in their responses.

Q: Does Amazon SQS support anonymous access?

Yes. You can configure an access policy that allows anonymous users to access a message queue.

Q: When should I use the permissions API?

The permissions API provides an interface for sharing access to a message queue to developers. However, this API cannot allow conditional access or more advanced use cases.

Q: When should I use the SetQueueAttributes operation with JSON objects?

The SetQueueAttributes operation supports the full access policy language. For example, you can use the policy language to restrict access to a message queue by IP address and time of day. For more information, see Amazon SQS Policy Examples in the *Amazon SQS Developer Guide*.

Service access and regions

Q: What regions is Amazon SQS available in?

For service region availability, see the AWS Global Infrastructure Region Table.

Q: Can I share messages between queues in different regions?

No. Each Amazon SQS message queue is independent within each region.

Q: Is there a pricing difference between regions?

Amazon SQS pricing is the same for all regions, except China (Beijing). For more information, see Amazon SQS Pricing.

Q: What is the pricing structure between various regions?

You can transfer data between Amazon SQS and Amazon EC2 or AWS Lambda free of charge within a single region.

When you transfer data between Amazon SQS and Amazon EC2 or AWS Lambda in different regions, you will be charged the normal data transfer rate. For more information, see Amazon SQS Pricing.

AWS AppSync FAQs

Q. What is AWS AppSync?

AWS AppSync is a new service that enables developers to manage and synchronize mobile app data in real time across devices and users, but still allows the data to be accessed and altered when the mobile device is in an offline state.

The service further allows developers to optimize the user experience by selecting which data is automatically synchronized to each user's device when changes are made, minimizing storage and bandwidth requirements, with a query language called GraphQL.

Using these capabilities, developers can, in minutes, build real time collaborative experiences spanning browsers, mobile apps, Alexa skills, and IoT devices that remain usable when network connectivity is lost.

Q. What types of apps can I build using the features released today?

AWS AppSync can be used to build mobile apps that would benefit from being able to synchronize user and app data across devices, continue functioning when disconnected, and offer real-time collaboration experiences. There are applications across all verticals. Examples include:

- Gaming apps with real-time scoreboards
- News feeds and financial data
- Customer service dashboards
- Shared wallet, travel or itinerary tracking with offline usage
- Social Media with content feeds and search/discovery/messaging
- Dating apps with likes, messaging and geo/proximity awareness

- Field service apps that need to allow for querying and CRUD operations, even when disconnected
- Document collaboration
- 3D collaboration such as shared whiteboards
- AR/VR with multiple actors (doctors in surgery with observers, teachers and students)
- Multi-device (e.g., Alexa, mobile, web, IoT) and multi-modal applications (e.g., task list) that need to work offline yet reflect the same eventually consistent state
- Chat apps, including presence indicators and conversation history

Q. What application developer languages are supported in AWS AppSync?

AWS AppSync SDKs support iOS, Android, and JavaScript. The JavaScript support spans web frameworks such as React and Angular as well as technologies such as React Native and Ionic. You can also use open source clients to connect to the AppSync GraphQL endpoint for using other platform such as generic HTTP libraries or even a simple CURL commands.

Q. What is GraphQL ?

GraphQL is a data language to enable client apps to fetch, change and subscribe to data from servers. In a GraphQL query, the client specifies how the data is to be structured when it is returned by the server. This makes it possible for the client to query only for the data it needs, in the format that it needs it in.

Q. What is a GraphQL Schema?

A GraphQL schema is a definition of what data capabilities are available for the client application to operate on. For example, a schema might say what queries are available or how an app can subscribe to data without needing to know about the underlying data source. Schemas are defined by a type system, which an application's data model can leverage.

Q. Do I need to know GraphQL to get started?

No, AWS AppSync can automatically setup your entire API, schema, and connect data sources with a simple UI builder that allows you to type in your data model in seconds. You can then immediately begin using the endpoint in a client application. The console also provides many sample schema and data sources for fully functioning applications.

Q. Can I use AWS AppSync with my existing AWS resources?

Yes. With AWS AppSync you can use existing tables, functions, and domains from Amazon DynamoDB, AWS Lambda and Amazon Elasticsearch Service with a GraphQL schema. AWS AppSync allows you to create data sources using existing AWS resources and configure the interactions using Mapping Templates.

Q. What is a Mapping Template?

GraphQL requests execute as "resolvers" and need to be converted into the appropriate message format for the different AWS Services that AWS AppSync integrates. For example, a GraphQL query on a field will need to be converted into a unique format for Amazon DynamoDB, AWS Lambda, and Amazon Elasticsearch Service respectively. AWS AppSync provides Mapping Templates for this, which are written in Apache Velocity Template Language (VTL) allowing you to provide custom logic to meet your needs. AWS AppSync also provides built-in templates for the different services and utility functions for enhanced usability.

Q. How is data secured with AWS AppSync?

Application data is stored at rest in your AWS account and not in the AWS AppSync service. You can protect access to this data from applications by using security controls with AWS AppSync including AWS Identity and Access Management (IAM), as well as Amazon Cognito User Pools. Additionally, user context can be passed through for authenticated requests so that you can perform fine-grained access control logic against your resources with Mapping Templates in AWS AppSync.

Q. Can I make my data real-time with AWS AppSync?

Yes. Subscriptions are supported with AWS AppSync against any of the data sources, so that when a mutation occurs, the results can be passed down to clients subscribing to the event stream immediately using either MQTT over WebSockets or pure WebSockets.

Q. How can I do complex queries with AWS AppSync?

The data sources available to AWS AppSync allow you to take full advantage of capabilities provided by Amazon DynamoDB, Amazon Elasticsearch Service, and AWS Lambda when using GraphQL. Features such as indexing and conditional checks, along with Mapping Templates, return comprehensive results from DynamoDB. Use cases such as fuzzy searches, geo searches and more that Amazon Elasticsearch Service offers are available to your application. Finally, Lambda can be used for serial or batched requests to return data from other sources such as Amazon Aurora.

Q. What AWS Regions are available for AWS AppSync?

AWS AppSync is available in US East (N. Virginia), US West (Oregon), US East (Ohio), EU (Ireland), EU (Frankfurt), Asia Pacific (Tokyo), Asia Pacific (Sydney), Asia Pacific (Mumbai) and Asia Pacific (Singapore).

Q. Can I import existing Amazon DynamoDB tables?

AWS AppSync can automatically generate a GraphQL schema from an existing DynamoDB table, including the inference of your table’s key schema and indexes. Once the import is complete GraphQL queries, mutations, and subscriptions can be used with zero coding. AppSync will also “auto-map” non-key attributes from your GraphQL types to DynamoDB attributes.

Q. Can AWS AppSync create a database for me?

Customers can create a GraphQL schema, either by hand or using the console, and AWS AppSync can automatically provision Amazon DynamoDB tables and appropriate indexes for you. Additionally, it will connect the data sources to "GraphQL resolvers" allowing you to just focus on your application code and data structures.

Q. What clients can I use to connect my application to my AppSync API?

You can use any HTTP or GraphQL client to connect to a GraphQL API on AppSync. We do recommend using the Amplify clients which are optimized to connect to the AppSync backend. There are some options depending on your application's use case:

- For DynamoDB data sources, use the DataStore category in the Amplify client. It provides the best developer experience and built-in conflict detection and resolution.
- For non-DynamoDB data sources in scenarios where you have no offline requirements, use the API (GraphQL) category in the Amplify client.
- For non-DynamoDB data sources in scenarios where you have offline requirements, use the AppSync SDK.

Amazon Sumerian FAQs

General

Q: What is Amazon Sumerian?
Amazon Sumerian is a managed service that lets you create and run 3D, Augmented Reality (AR) and Virtual Reality (VR) applications. You can build immersive and interactive scenes that run on AR and VR, mobile devices, and your web browser. Whether you are non-technical, a web or mobile developer, or have years of 3D development experience, getting started with Sumerian is easy. You can design scenes directly from your browser and, because Sumerian is a web-based application, you can quickly add connections in your scenes to existing AWS services.

Q: What technologies is Amazon Sumerian based on?
Amazon Sumerian is based on WebGL 2. Sumerian falls back to WebGL 1 where WebGL 2 is not available.

Q: How do I access Amazon Sumerian?
Amazon Sumerian is generally available. To get started, click here.

Q: What browers does Amazon Sumerian support?

For a list of supported browsers, please see the Amazon Sumerian Release Notes.

Q: What is WebVR?
WebVR is an open specification that makes it possible to connect a VR headset to your computer and experience VR through your web browser. We chose WebVR because it makes it easy to get into VR experiences no matter what device you have.

To view an Amazon Sumerian scene in WebVR, you will need a headset and compatible browser.

Q: What is WebGL?
WebGL (Web Graphics Library) is a JavaScript API used for rendering interactive 2D and 3D graphics in a browser.

Q: Can I use Amazon Sumerian on my mobile devices?

Yes. You can view scenes that you published from Amazon Sumerian's editor on a mobile device. Currently, Sumerian does not support editing scenes on mobile devices.

Q: What are Sumerian Hosts?
Sumerian Hosts are AI driven animated characters that can be added to your scene. Amazon Sumerian helps you design and assemble scenes to create rich 3D environments. Within those environments, you can use Sumerian Hosts to create and add animated 3D characters to your scenes. Hosts can guide your users through a scene by narrating scripts or answering questions. You can customize the Host’s characteristics such as gender, appearance, clothing, voice, and language. Using Sumerian’s integration with Amazon Lex and Amazon Polly, you can enable engaging spoken interactions between Hosts and your users. Polly lets you input text which your Host can speak in lifelike voices across a variety of languages. Using Lex’s automatic speech recognition and natural language understanding capabilities, you can build the conversational interface for your Host that allows it to understand and respond to users’ speech.

Learn more about adding a Host to your Amazon Sumerian scene by following our tutorials.

Q: Does Amazon Sumerian support Virtual Reality (VR)?
Yes. Because VR is a rapidly evolving technology, we've built Amazon Sumerian's VR support on top of WebVR, which allows us to quickly add support for new head-mounted displaysHMDs. Sumerian currently supports Oculus Go, Oculus Rift, HTC Vive, and HTC Vive Pro.

Q: Does Amazon Sumerian support Augmented Reality (AR)?
Yes. Amazon Sumerian supports AR compatible mobile devices. To get started developing for iOS or Android with Sumerian, you can follow the tutorials found here.

Q: What file types does Amazon Sumerian support?
Amazon Sumerian supports FBX 2017 and OBJ.

Q: Does Sumerian integrate with other AWS services?
Yes. We have integrated a base set of AWS services directly into the Sumerian interface. For example, you can choose an Amazon Polly persona and attach a script to a Sumerian host to make the Host speak.

Sometimes you may want to fetch real-time data from other solutions built on AWS and embed the data into your scene. With Amazon Sumerian you can configure your scene with an Amazon Cognito pool and use the AWS SDK to fetch data from the AWS services you are already using and add it to your scene.

To start integrating cloud services into your scenes, follow the IoT Thing Shadow and Script Actions tutorial.

Q: How is Sumerian different than Amazon Lumberyard?
Amazon Lumberyard and Amazon Sumerian are complimentary services that help customers design, build, and deploy 3D content. Sumerian and Lumberyard do share similarities in feature sets but serve different customer types. Sumerian uses technologies like JavaScript, HTML, WebGL, and WebVR to enable web and mobile developers to deploy experiences to target platforms through web browsers.

	Amazon Sumerian	Amazon Lumberyard
Technologies	JavaScript, HTML, WebGL, WebVR	C++, Lua, DirectX, OpenGL
Platform Examples	Web Browsers (e.g. Chrome and Firefox), Mobile Browsers (e.g. iOS, Android), Oculus Go, Oculus Rift, HTC Vive, HTC Vive Pro, Google Daydream, Lenovo Mirage	PC, Mac, iPhone, Xbox, Playstation
Experience Types	Focused, supports latest web browser technology	Big, complex, and/or high visual fidelity
Distribution Examples	Any browser and hybrid experiences for iOS ARKit and Android ARCore	Steam, Amazon.com, Google Play

Pricing

Q: Is there a free trial for Amazon Sumerian?
Yes. You can try AWS Sumerian for free in your first year of using the service. The Sumerian free tier lets you create a 50MB published scene that receives 100 views per month for free in the first year.

Q: How much does Amazon Sumerian cost?
Please see our pricing page for the latest information.

Support

Q: What kind of support is available for Amazon Sumerian?
All Amazon Sumerian customers have access to documentation, tutorials, and prebuilt scene templates and assets. Additional support for Sumerian is available via AWS Premium Support plans. Support is also provided via the Sumerian Slack channel.

Q: How do I submit feedback or suggestions?
Please email us your feedback. Suggestions can also be shared via the slack channel.

AWS Cost Management FAQs

Tools to help you to access, organize, understand, control, and optimize your AWS costs and usage

General

Q: Who should use the AWS Cost Management products?

We have yet to meet a customer who does not consider cost management a priority. AWS Cost Management tools are used by IT professionals, financial analysts, resource managers, and developers across all industries to access detailed information related to their AWS costs and usage, analyze their cost drivers and usage trends, and take action on their insights.

Getting Started

Q: How do I get started with the AWS Cost Management tools?

The quickest way to get started with the AWS Cost Management tools is to access the Billing Dashboard. From there, you can access a number of products that can help you to better understand, analyze, and control your AWS costs, including, but not limited to, AWS Cost Explorer, AWS Budgets, and the AWS Cost & Usage Report.

AWS Cost Explorer

Q: What are the benefits of using AWS Cost Explorer?

AWS Cost Explorer lets you explore your AWS costs and usage at both a high level and at a detailed level of analysis, and empowering you to dive deeper using a number of filtering dimensions (e.g., AWS Service, Region, Linked Account, etc.) AWS Cost Explorer also gives you access to a set of default reports to help you get started, while also allowing you to create custom reports from scratch.

For more information about the breadth of AWS Cost Explorer features, please click here or refer to the Managing your Usage and Costs user guide.

Q: What kinds of default reports are available?

AWS Cost Explorer provides a set of default reports to help you get familiar with the available filtering dimensions and types analyses that can be done using AWS Cost Explorer. These reports include a breakdown of your top 5 cost-accruing AWS services, and an analysis of your overall Amazon EC2 usage, an analysis of the total costs of your linked accounts, and the Reserved Instance Utilization and Coverage reports.

To access these default reports, please access Cost Explorer.

For more information about the RI Utilization report and the RI Coverage reports, please reference the Reserved Instance Reporting FAQ section.

Q: Can I create and save custom AWS Cost Explorer reports?

Yes. You can currently save up to 50 custom AWS Cost Explorer reports.

Q: What can I do with the AWS Cost Explorer API?

The AWS Cost Explorer API is the low-latency, ad-hoc query service that powers AWS Cost Explorer, and is accessible via a command-line interface and supported AWS SDKs. Using the AWS Cost Explorer API, you can build custom, interactive cost management applications without having to set up and maintain any additional infrastructure.

The AWS Cost Explorer API incurs a charge of $.01 per request. Please note that if your result set is paginated each page counts as a separate request.

To learn more about the AWS Cost Explorer API, please reference the technical documentation.

Q: When should I use AWS Compute Optimizer and when should I use AWS Cost Explorer?

You should use AWS Cost Explorer if you want to identify under-utilized EC2 instances that may be downsized on an instance by instance basis within the same instance family, and you want to understand the potential impact on your AWS bill by taking into account your RIs and Savings Plans. Cost Explorer offers recommendations for all commercial regions (outside of China) and supports the A, T, M, C, R, X, Z, I, D, H instance families.

You should use AWS Compute Optimizer if you want to look at instance type recommendations beyond downsizing within an instance family. You can use AWS Compute Optimizer to get downsizing recommendations within or across instance families, upsizing recommendations to remove performance bottlenecks, and recommendations for EC2 instances that are parts of an Auto Scaling group. AWS Compute Optimizer provides you additional capabilities to enhance recommendation quality and the user experience, such as using machine learning to identify workload types and automatically choose workload-specific recommendation methodology for them. You should also use AWS Compute Optimizer if you want to understand the performance risks and how your workload would perform on various EC2 instance options to evaluate the price-performance trade-off for your workloads. AWS Compute Optimizer is available in US East (N. Virginia), US East (Ohio), US West (Oregon), EU (Ireland), and South America (Sao Paulo), and supports the M, C, R, T and X instance families.

AWS Cost & Usage Report

Q: What is the AWS Cost & Usage Report?

The AWS Cost & Usage Report is your one-stop shop for accessing the most detailed information available about your AWS costs and usage. The AWS Cost & Usage Report can be generated at an hourly and/or daily level of granularity.

For more information about the exact information available in the AWS Cost & Usage Report, please reference the Cost & Usage Report Data Dictionary. You can enable the AWS Cost & Usage Report from here.

Q: How can I get started using the AWS Cost & Usage Report?

You can enable the AWS Cost & Usage Report from the Cost & Usage Reports page in the Billing Console. Please note that in order to receive the AWS Cost & Usage Report, you will need to create and configure an S3 bucket.

Q: How frequently is the AWS Cost & Usage Report updated?

The AWS Cost & Usage Report is updated at least once per day. An updated version of the report is delivered to your S3 bucket each time an update is completed.

Q: What else can I do with the AWS Cost & Usage Report?

You can configure your Cost & Usage Reports to integrate with Amazon Athena. Once Amazon Athena integration has been enabled for your Cost & Usage Report, your data will be delivered in compressed Apache Parquet files to an Amazon S3 bucket of your choice. From there, you can use an out-of-the-box AWS CloudFormation template to perform a one-time configuration of an AWS Glue crawler. This will ensure that your latest cost and usage information is always available to Amazon Athena – with no additional work required to prepare your data for analysis.

The AWS Cost & Usage Report can also be automatically uploaded into Amazon Redshift and/or Amazon QuickSight. In order for this to work, ensure that you select the option for receiving an Amazon Redshift and/or Amazon QuickSight manifest file when setting your report preferences.

Reserved Instance (RI) Reporting

Q: How can I use AWS Cost Management tools to better understand the costs and usage associated with my Reserved Instances (RIs)?

There are three main ways to gain insight into the costs and usage associated with your RIs: the default RI reports in Cost Explorer, the reservation-related data in the Cost & Usage Report, and AWS Cost Explorer's RI purchase recommendations.

Q: What are some of the insights you can glean using the RI reports in Cost Explorer?

AWS Cost Explorer provides two reports out-of-the-box--the RI Utilization and RI Coverage reports--to help you understand how you are using your RIs. The RI Utilization report visualizes the degree to which you are using your existing resources and helps you identify opportunities to improve your RI cost efficiencies. The RI Coverage report allows you to discover how much of your overall instance usage is covered by RIs, so that you can make informed decisions about when to purchase or modify an RI to ensure maximum coverage.

Q: What kind of RI-related information can you gain from the Cost & Usage Report?

The Cost & Usage Report gives you access to a wealth of RI-related information, including the ARN of the Reserved Instance that received the RI discount, the total reserved units in a reservation, and pricing information. This can help you trace your RI discounts, understand how well you are using your RIs, and analyze your savings compared to the On-Demand instance usage prices.

You can learn more about the Cost & Usage Report here. You can enable the Cost & Usage Report here.

AWS Budgets

Q: What is AWS Budgets and how does it work?

Using AWS Budgets, you can set a budget that alerts you when you exceed (or are forecasted to exceed) your budgeted cost or usage amount. You can also set alerts based on your RI Utilization and Coverage using AWS Budgets.

Learn more about AWS Budgets here, or reference the Monitoring your Usage and Costs user guide.

Q: What kinds of dimensions can be used to create a budget?

AWS Budgets gives you access to a number of filtering dimensions (i.e., AWS Service, Availability Zone, and Linked Account), and allows you to create budgets that are tracked on a monthly, quarterly, or yearly cadence.

Q: How many budgets can I create?

You can create up to 20,000 budgets. If you would like to increase your limit, please reach out to Customer Support.

Q: How many alerts and subscribers can I add for each budget?

For each budget, you are allowed to create up to five alerts. Each alert can be sent to 10 email subscribers and/or be published to an SNS topic.

Q: Is there a cost associated with using AWS Budgets?

You can create 2 budgets for free. Any additional active budgets accrue a cost of $.02 per budget per day.

Ready to get started?

Control your AWS costs

Learn how to control your AWS costs using the AWS Free Tier and AWS Budgets.

Learn more »

Sign up for a free account

Instantly get access to the AWS Free Tier and start experimenting with Amazon S3.

Sign up »

Start building in the console

Get started building with Amazon S3 in the AWS Console.

Get started »

Savings Plans FAQ

What is Savings Plans?

Savings Plans is a flexible pricing model that offer low prices on EC2, Lambda, and Fargate usage, in exchange for a commitment to a consistent amount of usage (measured in $/hour) for a 1 or 3 year term. When you sign up for Savings Plans, you will be charged the discounted Savings Plans price for your usage up to your commitment. For example, if you commit to $10 of compute usage an hour, you will get the Savings Plans prices on that usage up to $10 and any usage beyond the commitment will be charged On Demand rates.

What types of Savings Plans does AWS offer to reduce my bill?

AWS offers two types of Savings Plans:

1. **Compute Savings Plans** provide the most flexibility and help to reduce your costs by up to 66%. These plans automatically apply to EC2 instance usage regardless of instance family, size, AZ, region, OS or tenancy, and also apply to Fargate and Lambda usage. For example, with Compute Savings Plans, you can change from C4 to M5 instances, shift a workload from EU (Ireland) to EU (London), or move a workload from EC2 to Fargate or Lambda at any time and automatically continue to pay the Savings Plans price.
2. **EC2 Instance Savings Plans** provide the lowest prices, offering savings up to 72% in exchange for commitment to usage of individual instance families in a region (e.g. M5 usage in N. Virginia). This automatically reduces your cost on the selected instance family in that region regardless of AZ, size, OS or tenancy. EC2 Instance Savings Plans give you the flexibility to change your usage between instances within a family in that region. For example, you can move from c5.xlarge running Windows to c5.2xlarge running Linux and automatically benefit from the Savings Plans prices.

How do I get started with Savings Plans?

You can get started with Savings Plans from AWS Cost Explorer in the management console or by using the API/CLI. You can easily make a commitment to Savings Plans by using the recommendations provided in AWS Cost Explorer, to realize the biggest savings. The recommended hourly commitment is based on your historical On Demand usage and your choice of plan type, term length, and payment option. Once you sign up for a Savings Plan, your compute usage will automatically be charged at the discounted Savings Plans prices and any usage beyond your commitment will be charged at regular On Demand rates.

What payment options are available to pay for Savings Plans?

Savings Plans is available in 3 different payment options. The No Upfront option does not require any upfront payment, and your commitment will be charged purely on a monthly basis. The Partial Upfront option offers lower prices on Savings Plans. With this option you be charged at least half of your commitment upfront and the remaining will be charged on a monthly basis. With the All Upfront option, you will receive the lowest prices and your entire commitment will be charged in one payment.

Where do I see the discounted prices offered by Savings Plans?

The prices offered by Savings Plans are available on the pricing page here and can also be obtained using APIs/CLI. After you sign up, you can view the prices offered by your active Savings Plans in the AWS Management console or by using the API/CLI.

How do Savings Plans work with AWS Organizations/Consolidated Billing?

Savings Plans can be purchased in any account within an AWS Organization/Consolidated Billing family. By default, the benefit provided by Savings Plans is applicable to usage across all accounts within an AWS Organization/consolidated billing family. However you can also choose to restrict the benefit of Savings Plans to only the account that purchased them

How do I find out how much my Savings Plans have helped me save?

AWS Cost Explorer will show you details on the savings realized with your Savings Plans. You can also use the Savings Plans performance reports in AWS Cost Explorer to understand how well you are utilizing your Savings Plans.

Can I have multiple Savings Plans active at the same time?

Yes. Your AWS bill will be generated by taking into account all active Savings Plans.

How do Savings Plans compare to EC2 RIs?

Savings Plans offers significant savings over On Demand, just like EC2 RIs, but automatically reduce your bills on compute usage across any AWS region, even as usage changes. This provides you the flexibility to use the compute option that best suits your needs and continue to save money, all without having to perform exchanges or modifications.

Compute Savings Plans, which provides savings up to 66% (just like Convertible RIs), automatically reduce your cost on any EC2 instance usage regardless of region, instance family, size, OS, tenancy and even on AWS Fargate or AWS Lambda. EC2 Instance Savings Plans, which provide savings up to 72% (just like Standard RIs), automatically save you money on any instance usage within a given EC2 instance family in a chosen region (e.g. M5 in N. Virginia) regardless of size, OS or tenancy.

Do Savings Plans provide capacity reservations for EC2 instances?

No, Savings Plans does not provide a capacity reservation. You can however reserve capacity with On Demand Capacity Reservations and pay lower prices on them with Savings Plans.

Can I continue to purchase EC2 RIs?

Yes. You can continue purchasing RIs to maintain compatibility with your existing cost management processes, and your RIs will work along-side Savings Plans to reduce your overall bill. However as your RIs expire we encourage you to sign up for Savings Plans as they offer the same savings as RIs, but with additional flexibility.

Amazon Managed Blockchain FAQs

General

Q: What is Amazon Managed Blockchain?

A: Amazon Managed Blockchain is a fully managed service that makes it easy to create and manage scalable blockchain networks using the popular open source frameworks Hyperledger Fabric and Ethereum*. Managed Blockchain allows you to set up and manage a scalable blockchain network with just a few clicks. Managed Blockchain eliminates the overhead required to create the network, and automatically scales to meet the demands of thousands of applications running millions of transactions. Once your network is up and running, Managed Blockchain makes it easy to manage and maintain your blockchain network. It manages your certificates, lets you easily invite new members to join the network, and tracks operational metrics such as usage of compute, memory, and storage resources. In addition, Managed Blockchain can replicate an immutable copy of your blockchain network activity into Amazon Quantum Ledger Database (QLDB), a fully managed ledger database. This allows you to easily analyze the network activity outside the network and gain insights into trends.

*Hyperledger Fabric available today. Ethereum coming soon.

Q: What can I do with Amazon Managed Blockchain?

A: With Amazon Managed Blockchain, you can easily create blockchain networks across multiple AWS accounts with the open source frameworks, Hyperledger Fabric and Ethereum. These blockchain frameworks enable network members to securely transact and share data on a distributed and immutable ledger. Additionally, you can configure voting rules for your network so members can democratically govern it (i.e., voting on who to invite to join). To gain insights into activity on the network, you can replicate blockchain network activity data to Amazon QLDB for secure storage and off-chain analytics.

Q: How do I get started with Amazon Managed Blockchain?

A: To get started with Amazon Managed Blockchain, go to the AWS Management Console and click on Amazon Managed Blockchain. Click on "Create blockchain network". Follow the network creation wizard to create your first network and member. Then, invite other AWS accounts to join the network or create more members in your account to simulate a multi-member network.

For steps on how to get started with building your first network, please visit the Getting Started Guide in the Amazon Managed Blockchain documentation.

Q: How do you access Amazon Managed Blockchain?

A: You can access Amazon Managed Blockchain from the AWS Management Console, AWS Command Line Interface (CLI), or AWS Software Development Kit (SDK). To interact with the Hyperledger Fabric components provisioned and managed by Amazon Managed Blockchain, such as the certificate authority, ordering service, and peer nodes, you can use the open source Hyperledger Fabric CLI and SDK. Amazon Managed Blockchain provides endpoints to access these services, and you create a VPC PrivateLink endpoint for your network to access these endpoints. Please use a compatible version of the Hyperledger Fabric CLI and SDK with the version of Hyperledger Fabric specified in your network.

Q: What is a blockchain network?

A: Blockchain is a technology that makes it possible to build applications where multiple parties can record transactions without the need for a trusted, central authority to ensure that transactions are verified and secure.

Blockchain enables this by establishing a peer-to-peer network (a blockchain network) where each participant in the network has access to a shared ledger where the transactions are recorded. These transactions are by design, immutable and independently verifiable.

Q: What is the difference between Amazon Managed Blockchain and Amazon Quantum Ledger Database (QLDB)?

A: QLDB is a ledger database purpose-built for customers who need to maintain a complete and verifiable history of data changes in an application that they own and manage in a centralized way. Amazon QLDB is not a blockchain technology. Instead, blockchain technologies focus enabling multiple parties to transact and share data securely in a decentralized way; without a trusted, central authority. Every member in a network has an independently verifiable copy of an immutable ledger, and members can create and endorse transactions in the network. Amazon Managed Blockchain is a fully managed blockchain service that enables multiple parties to transact and share data directly and securely without the need for a central, trusted authority.

Q: What open source blockchain frameworks does Amazon Managed Blockchain support?

A: Amazon Managed Blockchain supports the open source Hyperledger Fabric and Ethereum frameworks. Hyperledger Fabric is available today, while Ethereum coming soon.

Q: What version of Hyperledger Fabric is supported?

A: Please visit the Amazon Managed Blockchain documentation to find the current supported versions of Hyperledger Fabric.

Q: What region is the Amazon Managed Blockchain currently available in?

A: Please visit the AWS Region Table to see the regions where you can use Amazon Managed Blockchain.

Q: What is the difference between the Amazon Managed Blockchain Starter Edition and Standard Edition network types?

A: Amazon Managed Blockchain offers two different network types: Starter Edition and Standard Edition. Each type is aimed for a particular set of use cases, and has a different hourly membership rate.

The Amazon Managed Blockchain Starter Edition network is designed for test networks and small production networks. It has several different attributes than the Standard Edition: You can have a maximum of 5 members per network and 2 peer nodes per member. Available peer node types are bc.t3.small and bc.t3.medium. The ordering service provisioned in a Starter Edition network has lower transaction throughput and availability than that in a Standard Edition network.

The Amazon Managed Blockchain Standard Edition network is designed for production networks. It has several different attributes than the Starter Edition: You can have a maximum of 14 members per network and 3 peer nodes per member. The bc.t3, bc.m5, and bc.c5 instance families are available instance types for peer nodes. The ordering service provisioned in a Standard Edition network has higher transaction throughput and availability than that in a Starter Edition network.

Building a network

Q: How do I invite other AWS accounts to join the blockchain network?

A: You can create a proposal to invite another AWS account to the blockchain network, and the current members in that network vote on the proposal. If the proposal becomes approved based on the voting rules of the network, then the other AWS account will receive an invitation to join the network.

Q: Does the account that creates an Amazon Managed Blockchain network own that resource?

A: An Amazon Managed Blockchain network is a decentralized resource where multiple AWS accounts have an equal ownership stake depending on the voting rules specified at the network’s creation. With the Approval Threshold Policy type, though an initial AWS account creates the network, governance can be distributed among multiple members after they join the network. If the initial member of the network leaves, that network will still be active among the remaining members.

Q: How do I delete an Amazon Managed Blockchain network?

A: An Amazon Managed Blockchain network is deleted once the last member in the network deletes their membership. If you have created a multi-member blockchain network in your AWS account, the network will be deleted once you delete all of the members. If you are in a blockchain network with memberships that you do not own, the network will only be deleted when the last member deletes their membership. If you delete your member and there are other members still in the network, that network will not be terminated. When Amazon Managed Blockchain is generally available, there will be configurable options to terminate a network if the founding member leaves.

Q: How do I create a VPC PrivateLink endpoint to access Hyperledger Fabric resources provisioned for the network?

A: Amazon Managed Blockchain provides endpoints to interact with your Hyperledger Fabric resources, specifically the Hyperledger Fabric certificate authority, ordering service, and peer nodes. To access these endpoints, you need to create a VPC PrivateLink endpoint in the VPC from which you would like to access the network. You can create a VPC PrivateLink endpoint from the VPC console, AWS CLI, or AWS SDK. When creating your endpoint, use the VPC Endpoint Service Name provided in the Amazon Managed Blockchain network details. If you have created multiple members in a single AWS account, you only need to create on VPC PrivateLink endpoint and not one for each member. Please note that you are billed separately for VPC PrivateLink endpoints you create and use. Please visit the Amazon Managed Blockchain documentation for more information on creating VPC PrivateLink endpoints for your network.

Q: How do I use the open source Hyperledger Fabric CLI or SDK on a client node to interact with my Amazon Managed Blockchain resources?

A: To interact with the Hyperledger Fabric certificate authority (CA), peer nodes, and ordering service created for your network, you can use the open source Hyperledger Fabric CLI or SDK and configure them with the respective endpoint information provided. Amazon Managed Blockchain exposes the endpoints for these components using a VPC PrivateLink endpoint that you create in your VPC. The Amazon EC2 instance or other resource running the Hyperledger Fabric CLI or SDK must have a route to reach this VPC PrivateLink endpoint. For instructions on how to configure these clients, please visit the Amazon Managed Blockchain documentation.

Q: What are the components of Hyperledger Fabric?

A: An Amazon Managed Blockchain for Hyperledger Fabric creates and manages the required components on your behalf that are needed to run a network. A Hyperledger Fabric network includes the ordering service, certificate authority, and peer components.

To interact with these components, you use an open source Hyperledger Fabric CLI or SDK from a client host that you create and manage. For more information about Hyperledger Fabric, please visit the Amazon Managed Blockchain documentation.

Q: How do I create a channel in my Hyperledger Fabric network?

A: Hyperledger Fabric channel is a private “subnet” of communication between two or more specific network members, for the purpose of conducting private and confidential transactions. Each transaction on the blockchain network is executed on a channel, where each party must be authenticated and authorized to transact on that channel.

To create a new channel in your Amazon Managed Blockchain network, you use the open source Hyperledger Fabric CLI or SDK with the endpoints exposed on your Hyperledger Fabric resources. You call configuration system chaincode, which creates a genesis block for the channel ledger, which stores configuration information about the channel policies, members, and anchor peer nodes for the channel. Please visit the Amazon Managed Blockchain documentation to learn more about creating a Hyperledger Fabric channel.

Q: How do I deploy chaincode applications to Hyperledger Fabric network?

A: Chaincode is a program that typically handles business logic agreed to by members of the network and is sometimes called a “smart contract.” To install and instantiate chaincode on the blockchain network, you use the open source Hyperledger Fabric CLI or SDK with the endpoints exposed on your Hyperledger Fabric resources. Additionally, only admin users in your membership can do these operations. To learn more about using chaincode with Hyperleder Fabric, please visit the Amazon Managed Blockchain documentation.

Q: How do I replicate my Amazon Managed Blockchain network activity to Amazon QLDB for secure storage and off-chain analytics?

A: Replicating blockchain network activity to Amazon QLDB will be available soon.

Security & availability

Q: How do I control access to my Amazon Managed Blockchain network?

A: Hyperledger Fabric uses certificates to identify users in each membership and determine their permissions on the network. You can create and manage these users using the Hyperledger Fabric certificate authority.

Q: How do I access the endpoints on the Hyperledger Fabric components managed by Amazon Managed Blockchain?

A: To access the endpoints on the Hyperledger Fabric components managed by Amazon Managed Blockchain, such as the Hyperledger Fabric certificate authority, ordering service, and peer nodes, you need to create a VPC PrivateLink endpoint in the VPC from which you would like to access the network. You can create a VPC PrivateLink endpoint from the VPC console, Amazon Managed Blockchain console, AWS CLI, or AWS SDK. When creating your endpoint, use the VPC Endpoint Service Name provided in the Amazon Managed Blockchain network details. If you have created multiple members in a single AWS account, you only need to create on VPC PrivateLink endpoint and not one for each member. Your client will also be able to interact with peer nodes from other members in the network to receive endorsements for proposed transactions.

Please note that you are billed separately for VPC PrivateLink endpoints you create and use. Please visit the Amazon Managed Blockchain documentation for more information on creating VPC PrivateLink endpoints for your network.

Q: Can I create multiple peer nodes to increase the availability of my blockchain components?

A: In the Amazon Managed Blockchain Standard Edition, you can create up to 3 blockchain peer nodes in each membership across Amazon EC2 availability zones for high availability. In the Starter Edition, you can create 2 peer nodes per membership.

Q: What permissions does the admin user when creating my network member?

A: The admin user you configure when creating your network member serves as the initial user in your Hyperledger Fabric membership. You can use the username and password to enroll this user with your Hyperledger Fabric certificate authority and create additional users in your membership. The admin user can also create channels on the network, and install and instantiate chaincode applications.

Pricing & billing

Q: How is Amazon Managed Blockchain priced?

A: There is no up-front commitment with Amazon Managed Blockchain. For Hyperledger Fabric on Amazon Managed Blockchain, you simply pay an hourly charge (billed per second) for your network membership, peer nodes, and peer node storage, and you pay for the amount of data you write to the network. Amazon Managed Blockchain offers two editions, the Standard Edition and the Starter Edition, and each edition has a different membership hourly rate. Additionally, you pay standard data transfer rates. To interact with your Amazon

Managed Blockchain resources, you will need a VPC PrivateLink endpiont that is billed separately.

When you are finished with an Amazon Managed Blockchain network, you can easily leave the network or terminate it and stop paying. You only pay for the resources you use. Please visit the Amazon Managed Blockchain pricing page for more information.

Q: Is there a different price for the Amazon Managed Blockchain Starter Edition and Standard Edition?

A: Yes, there is a different hourly membership rate for the Amazon Managed Blockchain Starter Edition and Standard Edition. Each edition is designed for a particular set of use cases. Please visit the Amazon Managed Blockchain pricing page for more information.

Amazon Quantum Ledger Database (QLDB) FAQs

General

Q: What is Amazon Quantum Ledger Database?

Amazon Quantum Ledger Database (QLDB) is a purpose-built ledger database that provides a complete and cryptographically verifiable history of all changes made to your application data.

Q: How is a ledger database different from other databases?

Traditional databases allow you to overwrite or delete data, so developers use techniques such as audit tables and audit trails to help track data lineage. While these approaches can work, they require custom development, can be difficult to scale, and put the onus on the application developer to ensure all the right data is being recorded. Data in Amazon QLDB is written to an append-only journal, providing the developer with full data lineage. Moreover, data in Amazon QLDB's journal is immutable and verifiable, meaning you can trust the data in your ledger.

Q: What data should I store in a ledger database?

Amazon QLDB's features make it a natural fit for system-of-record applications – those for which data integrity, completeness, and verifiability are critical. For example, in the supply chain and logistics space, an application built on Amazon QLDB would have the entire history of changes, such as movement between carriers and across borders, available for query and analysis. In finance, system-of-record applications track critical data, such as credit and debit transactions. Instead of building complex record keeping functionality within their application, banks can use QLDB to easily store a permanent and complete record of all financial transactions.

Q: Is Amazon Quantum Ledger Database a distributed ledger or blockchain service?

Amazon QLDB is not a blockchain or distributed ledger technology. Blockchain and distributed ledger technologies focus on solving the problem of decentralized applications involving multiple parties where there can be no single entity that owns the application, and the parties do not necessarily trust each other fully. On the other hand, QLDB is a ledger database purpose-built for customers who need to maintain a complete and verifiable history of data changes in an application that they own. Amazon QLDB offers history, immutability and verifiability combined with the familiarity, scalability and ease of use of a fully managed AWS database. If your application requires decentralization and involves multiple, untrusted parties, a blockchain solution may be appropriate. If your application requires a complete and verifiable history of all application data changes, but does not involve multiple, untrusted parties, Amazon QLDB is a great fit. If you have a use case for distributed ledgers or blockchain, please see Amazon Managed Blockchain.

Q: What kind of functionality does Amazon QLDB support?

In addition to providing a complete and verifiable history of application data changes, Amazon QLDB supports transactions with ACID semantics, a flexible document data model, and a familiar SQL-like API. QLDB is also fully managed and automatically scales to meet the needs of your application with no provisioning required.

Q: How do I connect to Amazon QLDB from my application?

In order to connect to Amazon QLDB and transact with the data in the ledger, you need to use the AWS-provided QLDB driver. Follow the steps in this link to download the driver and configure a connection.

Q: How do I try Amazon QLDB?

Getting started with Amazon QLDB is easy as there are no servers to manage or capacity to provision. You can create a new ledger in minutes using the AWS Management Console, AWS Command Line Interface (CLI), an AWS CloudFormation template, or by making calls to the QLDB API.

Performance

Q: What type of performance can I expect from Amazon QLDB?

Amazon QLDB can execute 2 – 3X as many transactions than ledgers in common blockchain frameworks. Blockchain frameworks are decentralized so to execute a transaction, they require a majority of members of the network to reach consensus on the validity of the transaction. On the other hand, QLDB has a centralized design, allowing its transactions to execute without the need for multi-party consensus.

Querying

Q: What is PartiQL? How does Amazon QLDB support it?

Amazon QLDB allows you to access and manipulate your data using PartiQL, which is a new open standard query language that supports SQL-compatible access to QLDB's document-oriented data model that includes semi-structured and nested data while remaining independent of any particular data source. To learn more about PartiQL read here.

Pricing

Q. How much does Amazon QLDB cost?

For Amazon QLDB pricing, please refer to our pricing page.

Q. In which AWS regions is Amazon QLDB available?

Amazon QLDB is available today in US East (Ohio), US East (N. Virginia), US West (Oregon), Asia Pacific (Tokyo), and EU (Ireland), with additional regions coming soon.

Scalability

Q. How does Amazon QLDB scale?

With Amazon QLDB, you don’t have to worry about provisioning capacity or configuring read and write limits. You create a ledger, define your tables, and QLDB automatically scales to support the demands of your application.

Q. What are the limits associated with an Amazon QLDB?

You can see the limits associated with Amazon QLDB on the AWS Service Limits page.

Backup and Restore

Q. Can I take a snapshot or backup of my ledger?

Amazon QLDB does not support a backup and restore feature as of now. At present, an export to S3 functionality is available. Using this functionality you can export the contents of your QLDB journal to S3.

Q. Can I restore my ledger to a particular point in time?

Amazon QLDB does not support a point-in-time restore feature as of now.

Availability, Durability, and Replication

Q. Is Amazon QLDB durable?

Amazon QLDB's ledger is deployed across multiple AZs with multiple copies per AZ. We maintain redundancy within the region and ensure full recovery from availability zone failures. A write is acknowledged only after being written to a durable storage in multiple AZs, and hence, QLDB is strongly durable.

Q. How does high availability work in Amazon QLDB?

Amazon QLDB is a highly available service. By default, multiple copies of your QLDB ledger are replicated across availability zones in a region. So, in the case of a zone failure you can still continue to operate QLDB.

Q. Does Amazon QLDB have cross-region replication?

Amazon QLDB does not support cross-region replication as of now. QLDB's export to S3 feature enables customers to export the contents of the QLDB journal to a S3 bucket. The S3 buckets can be configured for cross-region replication.

Security

Q: Can I use Amazon QLDB in Amazon Virtual Private Cloud (Amazon VPC)?

Amazon QLDB is integrated with AWS Private Link. Customers can create a VPC endpoint, which enables them to privately connect a VPC to supported AWS services powered by PrivateLink without requiring an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection.

Q: How does authentication work with Amazon QLDB?

Amazon QLDB uses the same authentication mechanism as other AWS services. The mechanism requires a request signature to be attached to the HTTP requests (header or query string). The signature is computed using other requests fields and the AWS credentials (Access Key ID and Secret Access Key).

Q. How does encryption work in Amazon QLDB?

By default, all data in transit and at rest is encrypted. Today, Amazon QLDB does not support customer managed CMKs (Customer Master Keys). Amazon QLDB uses AWS-owned keys to encrypt customer data.

Alexa for Business FAQs

General

What is Alexa for Business?

Alexa for Business allows organizations of all sizes to introduce Alexa to their workplace. With Alexa for Business, you can use the Alexa you know as an intelligent assistant to stay organized and focus on the work that matters. Alexa helps workers be more productive as they move throughout their day at – home and at their desks as enrolled users with personal devices, and in meeting rooms, copy rooms or other shared spaces with shared devices. Alexa for Business includes the tools and controls that administrators need to deploy and manage shared Alexa devices, skills, and users at scale.

How can I get started with Alexa for Business?

To start using Alexa for Business, you need one or more Alexa devices and an AWS account. Simply sign into the console and navigate to "Alexa for Business" under "Business Productivity".

What are some example uses for Alexa in an organization?

With Alexa for Business, you can deploy Alexa devices:

At your desk: Alexa lets you be more productive throughout your day and stay focused on important tasks. Alexa can help you manage your calendar with Alexa Smart Scheduling Assistant, keep track of your to-do list, and set reminders. Alexa can automatically dial into your conference calls and make phone calls for you. Alexa can help quickly find information for you, like the latest sales data, or the inventory levels in your warehouse.

In your meeting room: Alexa for Business simplifies meeting room experiences for your employees. You can control conferencing systems, check room availability, and book rooms with just your voice. For example, say "Alexa, join my meeting" and Alexa finds the upcoming meeting from the calendar, turns on the display, and connects you to the meeting. Alexa for Business integrates with popular video conferencing devices, room control systems, meeting room scheduling solutions, and calendar systems. You can also use Alexa for Business in your meeting rooms to offer your employees a natural interface to report service and equipment issues, answer frequently asked questions, and provide a company news briefing by building private skills or using Blueprints.

Around your workplace: Alexa helps your workplace run more efficiently. By building your own custom Alexa skills, you can easily voice-enable your workplace, and let Alexa help with common everyday tasks. Using your custom Alexa skills, Alexa can provide directions, find an open meeting room, order new supplies, report building problems, or notify IT of an equipment issue. Alexa can also provide important information, like inventory levels, and help with on-the-job training.

What is the difference between shared devices and enrolled users using personal devices?

Shared devices are the Alexa-enabled devices that you deploy to shared spaces in your workplace, like meeting rooms, lobbies, or breakout rooms. Shared devices are not linked to any specific user, and anyone with physical access to a shared device can use it. Shared devices are managed and configured directly through the Alexa for Business console, where you can assign them to locations, manage settings and enable groups of skills. To simplify setting up shared devices, you can either import existing setup devices or use the Device Setup Tool provided by Alexa for Business.

Personal devices are Alexa devices that are registered to a user’s personal Alexa account. In order to use a personal device with Alexa for Business, the user’s personal Alexa account needs to be enrolled into the organization’s Alexa for Business account. Enrolling in Alexa for Business gives users access to certain Alexa for Business features including the ability to automatically dial into conference calls, discover and enable private skills, and access to their Microsoft Exchange calendar (provided this has been configured by the Alexa for Business account administrator).

What devices can I use with Alexa for Business?

You can use the following devices with Alexa for Business:

Shared devices:

- Echo (1st and 2nd generation)
- Echo Dot (2nd and 3rd generation)
- Echo Plus (1st and 2nd generation)
- Polycom Trio 8500 and 8800

Personal devices:

- Echo Show
- Echo Spot
- Echo Plus
- Echo
- Echo Dot
- Any other Alexa-enabled device

To get started, you can purchase any of these devices individually, or purchase the Alexa for Business Starter Kit which includes a bundle of seven Echo devices and instructional materials that will help you get the most from Alexa for Business around your workplace.

Where is Alexa for Business available today?

Alexa for Business is currently available in the US East (N. Virginia) and supports Alexa devices running anywhere in the US. Access or use of Alexa for Business or its features may be restricted or limited in countries where Alexa for Business is not currently offered.

How does Alexa for Business work with Alexa Skills Kit?

Using the Alexa Skills Kit, you can build your own skills. With Alexa for Business you can make these skills available to your shared devices and enrolled users without having to publish them to the Alexa Skills store. Alexa for Business also provides skills developers an API to build context aware skills for use on shared devices. Alexa for Business supports any skill in the Alexa Skills store.

Does Alexa for Business provide a public API?

Yes, public APIs are available for creating and managing users, rooms, room profiles, skill groups, and devices. APIs are available via the AWS CLI and SDK; you can learn more about the APIs in the documentation.

Does the Alexa for Business API log actions in AWS CloudTrail?

Yes. All Alexa for Business actions performed via the AWS CLI and SDK will be included in your CloudTrail audit logs.

How is Alexa for Business different from Amazon Lex?

Alexa for Business is intended to enable organizations to take advantage of Amazon’s voice enabled assistant, Alexa. Alexa for Business provides Alexa capabilities that make workers more productive, while working alongside all of the other capabilities that Alexa has today like music, smart home controls, shopping, and thousands of third party skills.

Amazon Lex is intended to help build custom conversational interfaces and chat bots for use cases like call centers or application based bots. Bots built with Lex can be highly customized and exist separately from Alexa but they do not take advantage of Alexa's built in capabilities or third party skills. Both Alexa for Business and Amazon Lex use Amazon's deep learning capabilities that provide Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU).

Shared devices

What is a shared device?

Alexa for Business lets you use shared Alexa devices in common areas around your workplace. Shared devices can be used by anyone, and they are not associated with a personal Alexa account, and no one can use their personal skills with these devices.

Where can I deploy shared devices?

Shared devices can be deployed to any common area in your organization such as meeting rooms, lobbies, kitchens, break rooms, and copy rooms. Interactions with shared devices are not linked to a personal Alexa account, and provide your organization with Alexa's built-in capabilities and third party skills you choose to enable.

How do I set up shared devices?

To set up shared devices, Alexa for Business provides the Device Setup Tool. Run the Device Setup Tool to connect all your Alexa devices (Amazon Echo, Echo Dot, or Echo Plus) to your corporate Wi-Fi network (either WPA2 Personal or WPA2 Enterprise) and register them with your Alexa for Business account. You can also import select Echo devices registered to Amazon.com or Amazon Business accounts directly from the Alexa for Business management console. Once your devices are set up, you can assign them to your rooms. Please refer to the Alexa for Business Administrator Guide for more information about using the Device Setup Tool.

What are "rooms"?

Rooms are the shared spaces where you put your Alexa devices, such as meeting rooms, lobbies, copy rooms and breakout rooms.

What is a "room profile"?

A room profile contains the settings for your Alexa device including wake word, address, time zone, and units of measurement. A room profile simplifies the process of creating and managing rooms. For example, you can create a room profile that contains the Alexa settings that apply to all rooms in the same building.

What is a network profile?

A network profile contains the network settings for your shared Echo device including the SSID, network security type, network credentials and description. A network profile simplifies the process of creating, managing and assigning network configurations to your echo devices managed by Alexa for Business. Additionally, network profiles enable you to rotate your wireless passwords and enterprise certificates on associated Echo devices at scale without any physical interaction.

How do I use "skill groups"?

Skill groups are collections of Alexa skills you can use to enable skills on the devices in your rooms. For example, you can define a skill group with all the skills users will need in your meeting rooms. When you assign an Alexa device to a room, Alexa for Business automatically enables the skills in the skill groups assigned to the room. You can add skills to your skill groups at any time, and Alexa for Business will automatically enable them on all the Alexa devices in rooms that have been assigned that skill group.

What can users ask Alexa from a shared device?

Users can ask Alexa any of the same things they can ask on personal devices such as “Alexa, what time is it?” or “Alexa, what is the capital of Washington state?” If allowed by the administrator, users can also make outbound calls by asking “Alexa, call 206-555-1212” Users can also access any skill which you have enabled on the shared device, including private skills.

Does Alexa on shared devices provide personal responses to users?

No. Interactions with Alexa on a shared device are not linked to any personal Alexa account. Users cannot make phone calls to personal contacts, access their personal calendar, or interact with any personal skills linked to their personal Alexa account.

Do shared devices support shopping?

No, shopping is not available on shared devices.

Do shared devices support timers, alarms, and lists?

Yes, users can use a shared device to set a timer or alarm or add items to a list. However, Alexa for Business provides a capability that allows you to reset a shared device from the console, or using the AWS SDK. You can reset a device to clear timers, alarms, shopping list, to-do list, the history of Bluetooth connections, and set the volume level back to 50%.

Alexa in Meeting Rooms

How can I use Alexa in my meeting rooms?

Your employees can use Alexa for Business in the meeting room to control conferencing equipment, check room availability, and book the room. You can improve meeting room utilization in your office, by having Alexa for Business release rooms when they are booked but unused, or reminding them to end meetings when their reservation is ending. Alexa for Business also surfaces room utilization metrics to you so you can learn about how your meeting space is used and optimize it.

You can also use Alexa to offer your employees a voice interface to report service and equipment issues, answer frequently asked questions, and provide a company news briefing by building private skills or using Blueprints.

Which meeting room equipment works with Alexa for Business?

Alexa for Business can control most popular video conferencing and in-room systems including Polycom Group Series, Cisco TelePresence systems, Cisco Webex Room Kit, Crestron 3-Series Control Systems, and Zoom Rooms. Alexa for Business is also built-in to Polycom Trio 8500 and 8800.

In addition, the Alexa for Business conference device APIs allow you to build skills so that Alexa can work with additional equipment or perform specific tasks in your meeting rooms. To learn how to enable your conferencing equipment, please see our documentation.

How does Alexa for Business know what meeting to join?

You can connect Alexa for Business to Google G-Suite, Office365, or Microsoft Exchange calendars. Alexa for Business utilizes this calendar integration to look up the dial-in information of the scheduled meeting. Alexa for Business can look up meeting dial-in information from the most used conferencing providers including Amazon Chime, Cisco WebEx, Fuze, Google Meet, Zoom, BlueJeans, and Skype for Business. If there is no scheduled meeting, or Alexa cannot determine the dial-in information, users will be prompted for the meeting ID and optional PIN for the default conferencing provider that you specified in the Alexa for Business console.

What can users ask Alexa in meeting rooms?

Users can say *“Alexa, join my meeting”* to start their meeting or *“Alexa, end the meeting”* to end the meeting. Users can check availability by asking *“Alexa, is this room free?”*. User can also make reservations for the current or future time by saying *“Alexa, book this room for half an hour”* or *“Alexa, book this room at 2”*, and find out who owns the room booking if it's not free by asking *“Alexa, who booked this room?”*. Users can also make calls, as well as access any Alexa skills which have been enabled for the Alexa devices in the meeting room.

How do I get started with Room Booking?

The Room Booking feature is automatically enabled for every Alexa for Business customer once they link their calendar provider. The feature is supported by all three calendaring systems currently integrated with Alexa for Business: Microsoft Exchange, Microsoft Office 365, and Google G-Suite, and requires read and write permissions. Without write permissions, Alexa cannot create events on the rooms' resource calendars.

Can I delete events from a room’s calendar with Alexa?

No, users cannot delete a meeting from a room's calendar with Alexa.

How can users make phone calls using Alexa?

There are two ways users can make phone calls using Alexa. First, they can ask Alexa to call a contact from the address book set up by their administrator. For example, a user can say “Alexa, call IT”. Second, they can ask Alexa to call a specific phone number by speaking the numbers during the request. For example, a user can say "Alexa, call 212 555 1212"

Can I create address books to simplify calling from my shared devices?

Yes, you can create address books in the Alexa for Business console by clicking on the **Create Address Book** link in the **Calls** tab. Address books can contain frequently used contacts, such as the IT helpdesk, facilities, or the building reception. When an address book is associated with a shared Echo device, users can initiate a call from the device to a contact in the address book by speaking the contact name. For example, a user trying to reach the IT helpdesk could say “Alexa, call IT”.

Can I create different address books for different shared devices?

Yes, you can create multiple address books and have a unique list of contacts in each of them. This lets you use different numbers for the same contact when used in different contexts. For example, you might have a unique phone number for the IT helpdesk in each building; creating a unique address book for each building makes it possible for users reach the right IT helpdesk.

How do I enable outbound calling for my shared devices?

Outbound calling is by default enabled for shared devices and you can start making calls straight away. You can disable outbound calling by changing the setting in your room profiles in the Alexa for Business console.

Note: Alexa does not currently support inbound calls from mobile or landline phones

What phone number shows up when making calls via Echo devices?

When making calls via Echo device, the phone number shows up as an unknown number. For third-party conferencing devices that have Alexa built-in such as Polycom Trio, the phone number associated with the device will show up as caller ID.

How much does it cost to make calls from Echo devices ?

Making outbound PSTN calls from your Echo devices is free of charge.

Can I make international calls from Echo devices?

No, users are only allowed to call most local and toll-free US numbers. International calls, premium rate numbers, N-1-1 numbers, abbreviated dial codes, and dialing-by-letters are not supported.

What room utilization metrics can Alexa for Business surface to me?

Once you have enabled room utilization metrics in a room profile, Alexa for Business provides:

- Total meetings - how many meetings were reserved in each room.
- Total meeting minutes - how many minutes each room was reserved.

If you also have intelligent room release enabled, you will have access to the below metrics:

- Attended meetings: How many meetings were checked into
- Released meetings: How many meetings were released due to a lack of check in

- Released minutes: How many minutes were freed up due to intelligent room release
- Attendance rate: When there was a meeting on the calendar, how often did an employee show up and check in?
- Recovered meetings: A recovered meeting occurs when a room is reserved over a previously released time slot.

How does intelligent room release work?

In the Alexa for Business console, you can configure the amount of time an employee has to check into a room reservation before the room is released. Employees will be automatically checked in if they start their meetings with Alexa (ex: “Alexa, join the meeting”), or are in a call on a device with Alexa-built in. They can also check in by saying “Alexa, check in.”

If no check in has occurred within the selected time, Alexa will remind the room to check in. If check in has not occured within the next minute, Alexa will release the reservation, and notify the room that it has been released. Releasing the reservation means that Alexa will truncate the meeting invite so that the end time is updated to the release time.

Alexa will also measure how often meeting room are released - surfacing important metrics to you like attendance rate, released minutes, and recovered meetings.

Enrolled users and personal devices

What is an enrolled user?

Enrolled users are users that have linked their personal Alexa account with your Alexa for Business account. This allows them to use their personal Alexa devices for work, at their desks or in their homes. You can invite users to join your Alexa for Business account, who can then join using the Alexa for Business enrollment portal. Once they’ve enrolled, you can enable calendar access and make your private skills available to them. Enrolled users can access these on any of the devices in their personal Alexa account. Enrolled users can also use shared devices in your organization, but they can only access the skills available on those devices.

How do I invite a user to join my Alexa for Business organization?

You can send your users an invitation to join your organization via the Alexa for Business console. You can use the console to customize the content of the invitation e-mail that users will receive. The e-mail contains an enrollment URL where your users can login with the Amazon account they use to manage their Alexa devices. Once this is completed, users have access to the Alexa for Business resources you have enabled for them, including their Microsoft Exchange calendar, and your private skills. Users will also be able to auto-dial into conference calls from their Alexa devices, based on the default conferencing provider you configured for your organization in the Alexa for Business console.

What if a user does not already have an Amazon account, or doesn't use Alexa?

Your users do not need to be existing users of Alexa to use Alexa for Business. They can use their existing Amazon.com account to enroll with Alexa for Business or create a new Amazon.com account if they do not already have one. Once enrolled, users who are new to Alexa can install the Alexa mobile app for Android or iOS to customize Alexa’s settings for their personal devices.

Should my users use a different Amazon account from the one they use at home?

Users may choose to use any Amazon account they wish to enroll in your Alexa for Business organization. We recommend they use the same account that they use at home so that they can access Alexa for Business capabilities whether they are at home, on the go, or at the office.

What Alexa devices are supported for enrolled users?

Enrolled users can use any type of Alexa-enabled device. However, some features, such as dialing into conference calls, are only available from compatible Amazon Echo devices (Echo Dot, Echo, Echo Plus, Echo Show).

What can users do with Alexa after enrolling with your Alexa for Business account?

Users can continue to ask Alexa the same things they asked before they enrolled with your organization. With Alexa for Business, enrolled users get access to additional skills and features, such as asking Alexa to join their scheduled meetings. Users will also be able to access any private skills you choose to make available to your organization.

As an administrator, what access and control do I have to my enrolled users’ personal Alexa accounts?

You do not have any access or control over your users’ personal Alexa accounts, including Alexa capabilities and skills. You cannot see or delete utterances from personal devices used by enrolled users. As an administrator, you can enable your private skills for your users to access, and you can require that users use voice profiles to access their calendars.

Can a user link their Alexa account with multiple organizations?

Yes. Users can link their personal Alexa account to more than one Alexa for Business organization.

Can I help users self-enroll so that I don’t need to send them an invitation e-mail?

Yes. You can set up a self enrollment process within your organization and use the Alexa for Business SDK to automatically trigger an invitation e-mail to be sent to users. You can also choose to publish an internal LDAP connected web portal that authenticates users and verifies access before generating an invitation e-mail.

Can I remove users from my Alexa for Business account?

Yes. You can remove users from your Alexa for Business account using the Alexa for Business console. Removing a user will revoke access to all Alexa for Business features and your private skills.

Do I need my company’s IT department to do anything to enable Alexa for Business Work Updates?

As long as your company uses Office 365 or GSuite, no IT work is needed to set up Alexa for Business Work Updates. Please note, however, that your company may block third party apps and you might not be able to use Work Updates. If this happens, reach out to your IT department for clarifications or authorization.

Calendar integration

What business calendar systems are supported?

You can link your Alexa account to calendars in Google G-Suite, Microsoft Office365, and Microsoft Exchange 2013 (or later).

What can users do after they’ve linked their calendar to their personal Alexa account?

After the calendar is linked, users can ask Alexa to add new events, schedule meetings, cancel or delete events, and review upcoming events on your calendar.

How does a user link their personal Alexa account to their Microsoft Exchange 2013 (or later) account?

After a user is successfully enrolled with Alexa for Business, they can link their Microsoft Exchange account. To link a Microsoft Exchange account to Alexa, open the Alexa app, select Settings, and then select Calendar. Choose Microsoft Exchange and select Link account.

How can users restrict their calendar from being accessed?

Once a calendar is linked to Alexa, users can create a voice profile and restrict the calendar to only their voice.

How can I manage my calendar with Alexa?

Alexa helps you find one-on-one meeting time with your contacts. It offers suggested times that you and your contact are available to meet. Alexa also makes sure that you don't double-book your meeting by looking at all calendars you have linked with Alexa. Now, you are less likely to book a meeting with a coworker at the same time as your daughter's soccer game or your dentist appointment.

Because of busy schedules, we are frequently moving meetings to accommodate a higher priority work or personal event. Now you can let Alexa Smart Scheduling Assistant help you move those meetings, and send updates to all participants.

The Alexa Smart Scheduling Assistant expands Alexa's calendar management capabilities that also includes browsing the calendar, creating new events and canceling appointments.

How can a new Alexa user get started with the Alexa Smart Scheduling Assistant?

1. Link a Google Gmail, Google G Suite, Microsoft Exchange 2013 (or later), or Microsoft Office 365 calendar with Alexa; For more information about how to link your calendar, see Connect Your Calendar to Alexa.
2. Add work or personal contacts to your Alexa app. For more information about how to add your contacts, see Add and Edit Your Contacts to the Alexa App.
3. Have access to your contacts’ calendar availability information.
4. After the calendars and contacts are set up, you can ask Alexa to manage your meetings. For example, you can say ‘Alexa, schedule a meeting with John’ or 'Alexa, move my 2PM meeting.'

How can an existing Alexa user get started with the Alexa Smart Scheduling Assistant?

If you have a linked Microsoft Office 365 calendar you may need to relink your calendar. Go to your Alexa Companion App | Settings | Calendar | Microsoft and click **Unlink this Microsoft account**. Then click on **Link this Microsoft account** and follow the prompts as defined in Calendar linking flow.

If you have a linked Google G-Suite or Microsoft Exchange 2013 (or later) calendar, you can follow the steps below.

1. Add work or personal contacts to your Alexa app. For more information about how to add your contacts, see Add and Edit Your Contacts to the Alexa App.
2. Have access to your contacts' calendar availability information.
3. After the calendars and contacts are set up, you can ask Alexa to manage your meetings. For example, you can say 'Alexa, schedule a meeting with John' or 'Alexa, move my 2PM meeting.'

What are the new permissions requested from Microsoft Office 365?

Alexa Smart Scheduling Assistant requires the following permissions

1. **Calendars.Read.Shared** - Read user and shared calendars - Allows the app to read events in all calendars that the user can access, including delegate and shared calendars.
2. **Calendars.ReadWrite.Shared** - Read and write user and shared calendars - Allows the app to create, read, update and delete events in all calendars the user has permissions to access. This includes delegate and shared calendars.
3. **People.Read** - Read users' relevant people lists - Allows the app to read a scored list of people relevant to the signed-in user. The list can include local contacts, contacts from user's social networks and company directory, and people from recent communications (such as email and Skype).

The permissions help Alexa Smart Scheduling Assistant look up free/busy information for the organizer and participants, as well as determine the right contact that the user is trying to meet with.

Can I choose to decline the new permissions?

If you are new customer, you will be required to provide the permissions when linking your calendar for the first time. If you an existing customer who wants Alexa to get availability information when scheduling meetings, you will have to unlink and relink your calendar. If you choose not to relink, you will not be able to get the additional functionality.

What calendars are supported by the Alexa Smart Scheduling Assistant?

The Alexa Smart Scheduling Assistant supports Google G-Suite, Microsoft Office 365, and Microsoft Exchange 2013 (or later) calendars. It is available to Alexa for Business and Alexa consumers in the US.

How does Alexa find open spots on calendars when I use Alexa to schedule a 1:1 meeting with a contact?

Alexa's calendar works like most other calendar applications, such as the one on your mobile phone. When you link your calendar to Alexa using your calendar credentials, Alexa is able to read, create, edit, and cancel your calendar appointments. Alexa can also find open times on the calendar of the person with whom you're trying to schedule a meeting, so long as they have shared their calendar with you. When you ask Alexa to schedule a meeting with a colleague, Alexa compares your calendar with your colleague's to find and suggest mutually available free times.

How can I control which contacts can see my calendar through Alexa?

You can control calendar access by setting permissions in your calendar application. Alexa needs permission to access the free/busy information in calendars, such as Microsoft Office 365, Microsoft Exchange, Google G Suite or Google Calendar. If you have restricted access to your calendar, those restrictions will also apply to anyone using Alexa to schedule a meeting with you.

Different calendars have different default permissions. With Google Calendars, users must explicitly share their calendar with their contacts - calendars are not shared by default. Please see Share your calendar with someone to learn more. With G Suite Calendars, users can see calendar availability for contacts if your administrator has enabled it. Please see Set calendar visibility and sharing options to learn more. With Microsoft Office 365 and Microsoft Exchange 2013 (or later), users can see calendar availability for any contacts in their organization by default. The ability to hide or show your availability is controlled by your organization's IT administrator.

Can Alexa access calendar free/busy information for contacts outside my organization?

Alexa respects the sharing controls enabled by your IT administrator. Alexa will only see free/busy information if your IT administrator has granted sharing permissions with specific external organizations. By default, Microsoft Office 365, Microsoft Exchange, Google Calendar and Google G Suite do not allow sharing outside the organization.

How does Alexa schedule meetings if I have multiple calendars linked with my account?

Alexa allows only one default calendar to be used for creating meetings and then sending the email invite. You can see the default calendar in the calendar settings of your Alexa mobile app. When looking for your availability, Alexa will read the events across all your linked calendars and then attempt to find your contact's availability through your linked calendar that has access to the contact's availability information.

For example, let's say you have a personal Google Calendar and a work Microsoft Exchange 2016 calendar linked to Alexa, and you have set Microsoft Exchange 2016 as your default calendar for creating meetings in the Alexa mobile app. When you ask Alexa to schedule a meeting with a work contact, whose availability you can see because they are in your organization, Alexa will look for your availability on both your Google Calendar and Microsoft Exchange 2016 calendars. Alexa will then attempt to look for your work contact's availability using both Google Calendar and Microsoft Exchange 2016. Because your work colleague does not have a Google Calendar shared with you, but their availability is shared via your organization's Microsoft Exchange 2016 settings, Alexa will use your contact's Microsoft Exchange 2016 calendar to suggest meeting times that work for both of you. When you pick a time slot for the meeting and complete the other prompts, Alexa will send an email invite to your contact through Microsoft Exchange 2016 because you setup Exchange as your default calendar for creating and sending meeting requests.

Private skills

How do I build a private skill for my organization?

You build private skills much like how you build a public skill - by using the Alexa Skills Kit. When your skill is ready you can mark the skill as private, submit the skill, and then distribute it to your Alexa for Business account. Please refer to the Alexa Skills Kit for more information.

Does my private skills need to pass certification before I can distribute it to my AWS account?

No. Private skills are not subject to certification. As a result, you should only enable private skills that you developed or are from trusted developers.

How do I make a private skill available on my shared devices?

You make a private skill available to your shared devices by adding the skill to a skill group and then adding the skill group to your rooms. Alexa for Business automatically enables the skills for the Alexa devices assigned to your rooms.

How do I make a private skill available to my users?

Once you have published a private skill, you can navigate to the Skills section on the Alexa for Business Console. Find the skill in the Private Skills tab and check the box under the column “Available for users”. This will enable the skill for all users in the organization.

How can my users access the private skill?

Your users can view and manage private skills from the Alexa app on their phone by going to the menu, selecting Skills, and then selecting Your Skills (at the top of the screen.

Privacy and data security

How do Amazon Echo devices recognize the wake word?

Amazon Echo devices use on-device keyword spotting to detect the wake word. When these devices detect the wake word, they stream audio to the cloud, including a fraction of a second of audio before the wake word.

Can I turn off the microphone on Echo devices?

Yes, you can turn off the microphone by pushing the microphone on/off button on the top of your device. When the microphone on/off button turns red (on the Echo Show there is a red LED), the microphone is off. The device will not respond to the wake word until you reactivate the microphone by pushing the microphone on/off button again. An organization cannot turn on a device’s microphones via the Alexa for Business Console if the device’s microphones have been turned off.

How do I know when an Echo device is streaming my voice to the Cloud?

When an Echo device detects the wake word the light ring around the top of your device turns blue, to indicate that the device is streaming audio to the Cloud (for Echo Show and Echo Spot, you will see a blue bar or ring on the screen). When you use the wake word, the audio stream includes a fraction of a second of audio before the wake word. The audio stream closes once your question or request has been processed.

For personal devices you can enable a ’start of request sound,’ a short audible tone that plays after the wake word is recognized to indicate that the device is streaming audio. You can also enable an ’end of request sound’ that will play a short audible tone at the end of your request, to indicate that the connection has closed and the device is no longer streaming audio. This is available within the Sounds settings in the Alexa App (Settings > [Your Device Name] > Sounds).

What can an organization tell their users about the user’s information when using a corporate skill on an enrolled account or using a device managed by the organization?

You can tell them that the organization has no access to the information it receives about how they use a personal device, outside of when they interact with corporate skills. The organization may receive engagement metrics (device and skill usage metrics) for shared devices. In either case, the organization has no access to any voice recordings.

Voice recordings from shared devices being managed by Alexa for Business can be deleted from the Alexa for Business management console or by voice. If a user has enrolled their personal account, they can view and delete individual voice recordings associated with their account using the Alexa companion app, or all recordings by visiting Manage Your Content and Devices.

More Alexa and information on Alexa can be found here: Alexa Device FAQs.

When an organization manages shared devices using Alexa for Business, what information does that organization have access to?

The organization can see and control which skills are enabled on a shared device, the room where it’s assigned, and the settings applied to the device.

When an organization manages shared devices using Alexa for Business, does the organization have access to voice recordings made by users of the shared device?

No, unlike with a personal Alexa-enabled device where a user can review their voice recordings in the Alexa companion app, Alexa for Business organizations cannot access any voice recordings or text transcripts of what a user said. In addition, the organization doesn’t see Alexa’s responses to users’ queries.

What data do skill developers for Alexa for Business have access to?

Skill developers receive the information about their skill and its usage that is made available to skill developers in the Alexa Skills Kit developer portal. They also have access certain information about shared devices via the Alexa for Business API.

What controls do organizations have over personal accounts that they let enroll and join their Alexa for Business account?

Organizations can control which of their users can enroll and join their personal account to the organization’s Alexa for Business account. In addition, they can require a user create a voice profile to access corporate resources like calendars.

What information does an organization receive about its users' Amazon accounts when users enroll their personal account with the organization's Alexa for Business account?

The organization does not have any access to the user’s personal Amazon account. The organization does not receive the name or email that the personal account uses. As with shared devices, the organization has no access to the voice recordings on a personal device, including deleting voice recordings.

Are voice inputs processed by Alexa for Business stored, and how are they used by Alexa for Business?

Alexa for Business may store and use voice inputs processed by the service solely to provide and maintain the service and to improve and develop the quality of Alexa for Business and other Amazon machine learning and artificial intelligence services. Use of your content is necessary for continuous improvement of your Alexa for Business customer experience, including the development and training of related technologies. We do not use any personally identifiable information that may be contained in your content to target products, services, or marketing to you or your end users. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you.

How can voice recordings be deleted?

An individual using a shared device can delete their voice recordings by saying either “Alexa, forget what I just said” or “Alexa, forget what I said today.”

The organization can also delete voice recordings for shared devices they manage in one of two ways: via the Alexa for Business console or via an programmatic API call. The organization does not have any access to these voice recordings, other than the ability to delete them. Personal device users can can view and delete specific voice recordings associated with their accounts by going to History in Settings in the Alexa app, drilling down for a specific entry, and then tapping on the delete button. Or, personal device users can delete all voice recordings associated with their accounts for each of their Alexa-enabled products by selecting the applicable product at Manage Your Content and Devices.

Deleting voice recordings may degrade your Alexa for Business experience.

Who has access to my content that is processed and stored by Alexa for Business?

Only authorized employees will have access to your content that is processed by Alexa for Business. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you.

Alexa for Business provides access to non-AWS services provided by third parties and other Amazon entities (for example, skills and features that provide movie showtimes and traffic information). If you or your users use those non-AWS services, we may exchange information related to those requests with the parties providing the applicable services and that information is subject to the privacy and security practices of those parties.

Do I still own my content that is processed and stored by Alexa for Business?

You always retain ownership of your content and we will only use your content with your consent.

Is the content processed by Alexa for Business moved outside the AWS region where I am using Alexa for Business?

Any content processed by Alexa for Business is encrypted and stored at rest in the AWS region where you are using Alexa for Business. Some portion of content processed by Alexa for Business may be stored in another AWS region solely in connection with the continuous improvement and development of your Alexa for Business customer experience and other Amazon machine learning and artificial intelligence services. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you.

Can customers manage how Amazon uses voice recordings for machine learning?

Voice recordings are used to train our speech recognition and natural language understanding systems using machine learning. By default, a very small percentage of these recordings are manually reviewed in order to improve this process. Customers can now designate, at a room profile level, if voice recordings from shared devices they manage will be manually reviewed and used to improve machine learning algorithms. Using a new preference, Data Use Policy, located in the Room Profile, customers can either allow (default) or disallow manual reviews of voice recordings which are used to improve Amazon’s services.

Support and billing

What support is provided for Alexa for Business?

Depending on your AWS support contract, Alexa for Business is supported under Developer Support, Business Support and Enterprise Support plans.

How much does Alexa for Business cost?

Please see Alexa for Business Pricing for the latest information.

Does Alexa for Business offer a Free Tier?

Currently there is no Free Tier for Alexa for Business.

Amazon Chime FAQs

General

Q: What is Amazon Chime?

Q: What is the Amazon Chime web app?

Q: Which operating systems does the Amazon Chime app support?

Q: What is Amazon Chime Voice Connector?

Q: What is Amazon Chime Business Calling?

Q: What is the Amazon Chime Meetings App for Slack?

Getting Started

Q: Do I need a credit card to access the free trial?

Q: How do I download the Amazon Chime app?

Q: How do I create an Amazon Chime account?

Q: Why am I being asked to login with Amazon?

Q: Can I use my existing Amazon.com account with Amazon Chime?

Q: What is Login with Amazon?

Q: I am an end user. Where can I learn more about using Amazon Chime?

Q: I am an IT administrator. Where can I learn more about how to perform administration tasks for Amazon Chime?

Q: How do I get started with Amazon Chime call me?

Q: How do I get started with Amazon Chime Business Calling?

Q: Do I need to download the Amazon Chime client or app to use Amazon Chime with Slack?

Q: Will my scheduled meetings call me in Slack?

Application Features

Q: Can I chat with people outside my organization?

Q: Will my meeting attendees be required to download anything?

Q: How many people can participate in an online meeting or video conference?

Q: Does Amazon Chime support in-room video conferencing systems?

Q: How do I join an Amazon Chime meeting using a SIP or H.323 compatible video conferencing system?

Q: Can I join a meeting using a phone?

Q: What countries does Amazon Chime call me work in?

Q: What's the difference between a voice (VoIP) or video call and a meeting?

Q: How do I initiate a voice (VoIP) or video call?

Q: What is the maximum number of attendees I can have on a meeting?

Q: What is Alexa for Business?

Q: Can I start Amazon Chime meetings in my conference rooms using Alexa?

Q: Can I join Amazon Chime meetings from my desk using Alexa?

Q: Can I call a telephone number with Amazon Chime Business Calling?

Q: Can I text a telephone number with Amazon Chime Business Calling?

Q: Can a telephone call become a meeting?

Q: Do I have to have Amazon Chime Basic or Pro to use Amazon Chime Business Calling?

Administration

Q: How do I log in to the Amazon Chime console to access my account?

Q: Can I add or remove Amazon Chime users in my Amazon Chime organization?

Q: In which AWS region is Amazon Chime service hosted?

Q: In which AWS regions does Amazon Chime host meetings?

Q: Can an organization select which AWS regions can host its Amazon Chime meetings?

Q: Can a user select the AWS region for a specific Amazon Chime meeting?

Q: Can I assign different permissions to different users in my Amazon Chime Team or Enterprise account?

Q: What does "claim your domain" mean?

Q: How do I claim my domain?

Q: Can I use my corporate directory to manage my Amazon Chime users?

Q: How do I configure Amazon Chime to use another identity provider (such as Microsoft Active Directory or Okta SSO)?

Q: Is there an additional cost when using Microsoft Active Directory or Okta SSO to manage my Amazon Chime users?

Q: Does Amazon Chime provide a history of changes made to my account through the Amazon Chime console?

Q: What is required to set up phone numbers for users?

Q: How do I set up and access Call Detail Records (CDRs) for Amazon Chime Business Calling?

Q: How can I perform bulk provisioning of phone numbers and users?

Q: Are all features in Amazon Chime Business Calling available when I assign a phone number to a user?

Q: How do I add the Amazon Chime Meetings App for Slack?

Q: How can I upgrade the users in my Slack workspace to Amazon Chime Pro?

Billing and Permissions

Q: Can I schedule or host instant meetings of three or more people if I only have Basic permissions?

Q: Is Amazon Chime call me available to trial users?

Q: Does Amazon Chime charge a usage fee for all meeting participants?

Q: What's the difference between a 2-person voice or video call and a meeting?

Q: Is Amazon Chime Business Calling available to trial users?

Q: Will I be charged for phone numbers for Amazon Chime Business Calling if I have not assigned them to a user?

Q: What countries can I make calls to with Amazon Chime Business Calling?

Q: What countries does Amazon Chime Business Calling support for telephone numbers?

Q: Does Amazon Chime Business Calling support toll free numbers?

Q: Will I be charged separately for S3 usage associated to Call Detail Record storage?

Pricing

Q: How does the free trial for Amazon Chime work?

Q: Do I need a credit card to access the free trial?

Q: How do I purchase Amazon Chime?

Q: How does billing work for Amazon Chime?

Q: What does Amazon Chime cost?

Q: Is there an additional charge when attendees join a meeting using a standard phone line?

Q: Is there an additional cost when using Microsoft Active Directory or Okta SSO to manage my Amazon Chime users?

Q: Are there any fixed costs for Amazon Chime Business Calling?

Q: Does the Amazon Chime global expansion affect pricing?

Developers

Q: Does Amazon Chime have APIs available to integrate Chime with my internal systems and processes?

Q: Can I use incoming webhooks to send messages to Amazon Chime chat rooms?

Q: Does Amazon Chime support chat bots?

Reporting

Q: Does Amazon Chime provide administrators with usage reports?

Q: Does Amazon Chime Business Calling provide Call Detail Records?

Q: What information is included in the Amazon Chime usage reports?

Q: How do I access these reports?

Check out the Amazon Chime features page

Learn more about Amazon Chime features

Learn more »

Contact us

Contact us and let us provide you with a personalized demo.

Sign up »

Download Amazon Chime

Try Amazon Chime Pro features free for 30 days!

Download »

Amazon WorkDocs features

Secure Storage and Sharing

Store content in Amazon WorkDocs: You can store virtually any type of file on Amazon WorkDocs. Each individual user account on Amazon WorkDocs includes 1 TB of storage capacity by default. Administrators can set storage limits for individual users and also purchase additional storage for users on a pay-as-you-go basis. There is no limit on the amount of additional data and number of documents you can store.

Unlimited versioning: You can easily keep track of prior versions in Amazon WorkDocs with unlimited versioning. A new version of a file is created every time you save it. With Amazon WorkDocs, all feedback is associated with a specific file version, so you can easily refer back to comments in earlier iterations.

Search: Searching made easy with Amazon WorkDocs Smart Search: Amazon WorkDocs Smart Search speeds up your content searches so you can spend more time creating, editing, and sharing files with colleagues. It lets you search across document content, comments, and labels in addition to file and folder names. You can further refine your search by content location, last updated time and date range, and specific file types. Results will display as a sorted list, with folders listed before files.

WorkDocs drive offline content and offline search: With WorkDocs Drive offline capabilities, you can access content even when network connectivity is limited. In addition you can also find the content in both online and offline modes with Smart Search. WorkDocs Smart Search lets you query across content, comments, and document labels in addition to searching for files by name.

Share a link: You can share files with anyone using a shareable link, invite any internal user to contribute, and request feedback. When creating a link, you decide if the link is public to anyone, or only other users. You can set an expiration date and password on the link.

Invite others – internally and externally: With Amazon WorkDocs, you can invite others to view, contribute to, or co-own your files by entering user names, group names, and email addresses. You can also request specific feedback with a personal message and set a deadline. To share content externally, you can invite an external user to create guest account the first time they log in to Amazon WorkDocs.

1-click file sharing: You can also share content with a single click directly from Windows File Explorer on a Windows PCs using Amazon WorkDocs Companion. To use this feature, the Amazon WorkDocs Companion application should be installed on your Microsoft Windows PC. You can upload files to your Amazon WorkDocs site and share them with a link automatically copied to your clipboard.

File control: Amazon WorkDocs lets you control who can access, comment, and download or print your files. You can lock files to make changes and ensure that edits are not overwritten by other contributors, eliminating the need to coordinate changes. You can also disable feedback when you have completed a file.

Encryption: With Amazon WorkDocs, your content is encrypted in transit and at rest to ensure the security of your data and help you meet regulatory and compliance requirements. Amazon WorkDocs is built on AWS, where security is our number one priority.

Tasks management: With Amazon WorkDocs Tasks, you can review action items that are pending, assigned, and resolved in one place. Types of tasks you can use include sending content out for review, requesting access to a file, transferring ownership of a file, and requesting action from your site administrator. With Tasks, you and your teams can collaborate more efficiently by highlighting action items for the day.

WorkDocs Drive

Amazon WorkDocs Drive is a desktop application that combines the ease of working in Windows File Explorer or Mac Finder with the scale of Amazon WorkDocs. With Amazon WorkDocs Drive, all of your files are available on-demand from your device without consuming valuable disk space on your PC or Mac. You can use Amazon WorkDocs Drive as your primary user drive, and you don’t need to use network shares to store your content.

All changes you make to content that is accessed through Amazon WorkDocs Drive are automatically synced to your Amazon WorkDocs site, and also available to access from any other devices via Amazon WorkDocs applications.

Amazon WorkDocs Drive is available for Microsoft Windows PCs, Amazon WorkSpaces, and macOS version 10.11 and later.

Collaboration

Commenting: With Amazon WorkDocs, you can add private comments, format comment text, resolve comments, and respond to comments in a threaded conversation. When providing feedback in Amazon WorkDocs, you can add overall comments or comment on specific sections of a file.

Notifications: You can notify someone of your comment by tagging them with @username. You can also disable email notifications for files.

Request feedback: When sharing a document, you can assign colleagues a task which will then appear in the "Tasks" tab in the top right of the Amazon WorkDocs user interface.

Editing

Editing in Microsoft Office: You can edit documents directly in Microsoft Office with Amazon WorkDocs Companion. Amazon WorkDocs Companion is an app that lets you edit Microsoft Office, .pdf, and .txt files from your browser. You can save your changes as a new version on your Amazon WorkDocs site. With Amazon WorkDocs Companion, you no longer need to manually download, save, and upload files when accessing Amazon WorkDocs from your browser.

Open with Microsoft Office Online: Amazon WorkDocs also lets you open and edit Microsoft Office files directly from your browser using Open with Office Online. You can review feedback, edit, and make changes to Microsoft Office files without switching applications.

Hancom ThinkFree Office Online: Amazon WorkDocs lets you create Microsoft Office files in real time from your browser with collaborative editing powered by Hancom Thinkfree Office Online. You can create new documents, spreadsheets, and presentations without leaving your browser.

Activity Feed

Amazon WorkDocs Activity Feed helps you easily keep track of engagements with your content by other users on your Amazon WorkDocs site. You can search actions by file, folder, or user name in real-time. You can also use multiple options to filter your results. With Amazon WorkDocs, administrators can use the Activity Feed to track site-wide actions by file, folder, or user name in real-time. You can also use multiple options to filter your results.

Web and Mobile Access

You can access your Amazon WorkDocs site from any browser with a connection to the Internet. You can also access Amazon WorkDocs via mobile applications on iOS, Android, and Fire Tablet to collaborate from any device at any time

WorkDocs Approval Workflow

You can use approval workflow to adhere to business processes in your organization. You can create an approval workflow to route documents and other files stored in WorkDocs to one or more users for their approval. Approval workflow enables users to build workflows to track and manage their document approval processes in an automated manner.

For Administrators

Compliance: Amazon WorkDocs is HIPAA eligible, PCI DSS compliant, and aligns with ISO compliance requirements. Amazon WorkDocs helps you meet your regulatory and compliance requirements for collaboration and file management. With Amazon WorkDocs, you can store and collaborate on files that contain sensitive financial and medical data. To help you demonstrate your commitment to information security, Amazon WorkDocs also has ISO 9001, 27001, 27107, and 27018 certifications.

Active Directory Integration: Amazon WorkDocs lets you use your Active Directory to manage your users. If you use Active Directory, you can create user groups, enable multi-factor authentication (MFA), and configure single sign-on (SSO) for your Amazon WorkDocs site. Your users can also log in with their existing credentials when you use Active Directory with Amazon WorkDocs.

With Amazon WorkDocs, you can integrate with Active Directory in two ways – either by establishing a secure trust relationship between your on-premises Active Directory and your AWS Directory Service for Microsoft Active Directory (Enterprise Edition) domain controller, or by using the AWS Directory Service Active Directory Connector.

Data Residency: You can specify in which AWS Region to store your content to help meet data residency requirements. Your users can access your Amazon WorkDocs site from anywhere in the world regardless of which AWS region you choose. Refer to AWS Regions to see where Amazon WorkDocs is currently available.

Data Retention: Amazon WorkDocs lets you specify a site-wide data retention policy for your users' files and folders. With a data retention policy, you can recover user-deleted files and folders during the retention policy period. Retention policies are applied to all files and folders associated with an Amazon WorkDocs site, and the retention policy can be adjusted from the default of 60 days to any value from 0 to 365 days.

IP Filtering Allow Lists: IP address-based allow lists can be added in your Amazon WorkDocs Admin Console. You can set the IP address ranges from which you wish to provide access. When a user tries to connect to your Amazon WorkDocs site from a browser, Amazon WorkDocs drive, mobile device, sync, or companion app, the IP address from which the request originated is evaluated against your allow list. If it is not on the allow list, access will be denied. If you do not filter user access by IP address with an allow list, access will be open to all IP addresses.

Migration Tool: Using the Amazon WorkDocs migration service application, you can configure migration tasks, and target WorkDocs account and site to migrate data to. You can schedule the migration task to execute during a specific period as a one-time data transfer operation or have regular migrations so as to minimize downtime for your users. The migration service application provides up-to-date information and status on migration jobs including detailed reports once migration has successfully completed.

Drive Letter Selection: With Amazon WorkDocs Drive, you can select any custom drive letter based on your organization’s configurations, standards, or preferences. You can also mass deploy a specific drive letter for your organization. This feature offers your organization not only the flexibility but also the control you need to select a specific drive letter for your virtual WorkDocs Drive.

For Developers

The Amazon WorkDocs SDK helps you build content collaboration and management capabilities into your solutions and applications by providing full administrator and user level access to Amazon WorkDocs site resources. Using the Amazon WorkDocs SDK, you can also integrate the activity feed with your analytics solutions to create real-time monitoring of your Amazon WorkDocs users and files.

Extensible API: The Amazon WorkDocs SDK includes an extensible API that provides admin and user level actions for: user administration, permission management, sharing, commenting, metadata, labeling, and activity tracking.

Part of the AWS SDK: The Amazon WorkDocs SDK is part of the AWS SDK so you can easily take advantage of the power of AWS for security, monitoring, business logic, analytics, storage, artificial intelligence, and app development.

API Monitoring and Notifications: If you want to monitor API actions, Amazon WorkDocs is integrated with AWS CloudTrail and Amazon SNS. AWS CloudTrail logs API calls and Amazon SNS notifies you when new log files are delivered. You can use AWS CloudWatch to automatically monitor and alert you when certain criteria are met in your CloudTrail logs.

Check out Amazon WorkDocs pricing

Information on Amazon WorkDocs pricing

Learn more »

Sign up for a free account

Instantly get access to the AWS Free Tier.

Sign up »

Sign up for an account

Get started with Amazon WorkDocs

Sign in »

Amazon WorkMail FAQs

General

Q: What is Amazon WorkMail?

Amazon WorkMail is a secure, managed business email and calendar service with support for existing desktop and mobile clients. Amazon WorkMail gives users the ability to seamlessly access their email, contacts, and calendars using Microsoft Outlook, their web browser, or their native iOS and Android email applications. You can integrate Amazon WorkMail with your existing corporate directory and control both the keys that encrypt your data and the location in which your data is stored.

Q: How can I get started using Amazon WorkMail?

To get started with Amazon WorkMail, you will need an AWS account. You can use this account to sign into the AWS Management Console and create an organization, add your domains, and also create users, groups, or resources. Please refer to the Amazon WorkMail documentation for more information on getting started.

Q: What clients can I use to access Amazon WorkMail?

You can access Amazon WorkMail from Microsoft Outlook clients on Windows and Mac OS X, and on mobile devices that support the Microsoft Exchange ActiveSync protocol including iPhone, iPad, Kindle Fire, Fire Phone, Android, Windows Phone, and BlackBerry 10. Additionally, you can use the Apple Mail application on Mac OS X or the Amazon WorkMail web application to securely access Amazon WorkMail using your web browser.

Q: Does Amazon WorkMail support accessibility capabilities?

Yes, you can use screen readers and keyboard shortcuts with the Amazon WorkMail web application for easier accessibility; you can learn more about these capabilities on the Working with Accessibility Features documentation page here. In addition, the accessibility capabilities offered in supported desktop and mobile clients (see below for a list) can also be used with Amazon WorkMail.

Q: What is the mailbox storage limit in Amazon WorkMail?

Amazon WorkMail offers a mailbox storage limit of 50 GB per user.

Q: What is the maximum size of email that I can send from Amazon WorkMail?

The maximum size of outgoing and incoming email in Amazon WorkMail is 25 MB.

Q: Can I share my calendar with other users in my organization?

Yes. Amazon WorkMail offers the ability to share your calendar with your co-workers.

Q: Does Amazon WorkMail provide resource booking?

Yes. Amazon WorkMail provides the option to create resource mailboxes such as conference rooms, projectors, and other equipment. The resource mailboxes will allow users to reserve the room or equipment by including the resource in meeting invites.

Q: Does Amazon WorkMail support email archiving?

Email journaling can be enabled to capture and preserve messages in your existing archiving solution.

Q: Can I set up email redirect rules on Amazon WorkMail?

Yes, you can configure email redirection rules for Amazon WorkMail mailboxes. You can setup email redirection rules on your desktop email application, such as Microsoft Outlook, or using the Amazon WorkMail web application. You will need to ensure that the Amazon Simple Email Service (Amazon SES) identity policies for your domains are up-to-date to take advantage of email redirection rules. Please visit this page for more information on how to update the Amazon SES identity policy for your domain.

Q: Are there limits on the number of organizations and users I can create when using Amazon WorkMail?

No, there are no limits on the number of organizations and users you can create.

Q: Are there limits on the number of messages I can send per user?

There are limits only on sending external messages. For example, the number of messages sent to recipients outside your organization. Each user in your organization can send messages to a maximum of 10,000 external recipients per day, and the total external recipients for an AWS account is limited to 100,000 per day. New Amazon WorkMail accounts may start with limits that are lower than the limits described here; please see AWS Service Limits for more information.

Amazon WorkMail is a business e-mail service and not intended to be used for bulk e-mail services. For bulk e-mail services, please see Amazon Simple Email Service.

Q: Are there limits associated with the use of the Amazon WorkMail SMTP gateway?

Yes. To learn more about SMTP limits, please see AWS Service Limits.

Q: Are there limits on the number of messages each user can receive?

There are no limits on the number of messages each user can receive. However, we may queue or reject messages (and send a bounce to the sender) if there is a large volume of incoming email in a short period of time. Please see AWS Service Limits for more information.

Q: Do meeting requests count when evaluating usage against message limits?

All messages that are sent to another user are considered when evaluating these limits. These include e-mails, meeting requests, meeting responses, task requests, as well as all messages that are forwarded or redirected automatically as a result of a rule.

Q: Does Amazon WorkMail support public folders?

No, WorkMail does not offer public folders.

Web Application

Q: What features does the Amazon WorkMail web application provide?

The Amazon WorkMail web application provides users anywhere with access to email, calendar, contacts, and tasks. Users can also access shared calendars, access the global address book, manage their out-of-office replies, and book resources.

Q: Which browsers does the Amazon WorkMail web application work on?

The Amazon WorkMail web application supports the following browsers: Firefox, Chrome, Safari and Edge. For more information, please see Log On to the Amazon WorkMail Web Application.

Q: In which languages is the Amazon WorkMail web application available?

The Amazon WorkMail web application is currently available in English, French, and Russian.

Mobility

Q: Can I use Amazon WorkMail on my mobile device?

Yes. Amazon WorkMail is compatible with most major mobile devices supporting the Microsoft Exchange ActiveSync protocol, including iPad, iPhone, Kindle Fire, Fire Phone, Android, Windows Phone, and BlackBerry 10.

Q: What mobile device policies does Amazon WorkMail support?

Amazon WorkMail gives you the ability to require a PIN or password on your devices, configure the password strength, require a device lock after a number of failed login attempts, require a screen lock for idle timeouts, and require device and storage card encryption.

Q: Does Amazon WorkMail offer the ability to remotely wipe mobile devices?

Yes. Amazon WorkMail offers a remote wipe feature. A remote wipe can be performed by the IT administrator using the AWS Management Console.

Desktop Clients

Q: Can I use Amazon WorkMail with Microsoft Outlook on Microsoft Windows?

Yes. Amazon WorkMail offers native support for Microsoft Outlook 2007, 2010, 2013, and 2016 on Microsoft Windows.

Q: Do I need any additional software to connect Microsoft Outlook to Amazon WorkMail?

No. Amazon WorkMail offers native support for the most recent versions of Microsoft Outlook and does not require any additional software to connect Microsoft Outlook.

Q: Can I use Amazon WorkMail with Microsoft Outlook on Mac OS X?

Yes. Amazon WorkMail offers native support for Microsoft Outlook 2011 and Microsoft Outlook 2016 on Mac OS X.

Q: Can I use Amazon WorkMail with other clients on Mac OS X?

Yes. Amazon WorkMail offers native support for the Apple Mail and Calendar applications on Mac OS X (10.6 and above).

Q: Does the Amazon WorkMail user subscription include a license for Microsoft Outlook?

Amazon WorkMail does not include a license for Microsoft Outlook. To use Microsoft Outlook with Amazon WorkMail, you must have a valid license from Microsoft.

Q: Does Amazon WorkMail support the Click-to-run version of Microsoft Outlook 2010, 2013, and 2016?

Yes. Amazon WorkMail supports the Click-to-run versions of Microsoft Outlook 2010, 2013, and 2016.

Q: Can I access my Amazon WorkMail mailbox with my existing POP3 or IMAP client applications?

You can access your Amazon WorkMail mailbox with client applications that support the IMAP protocol. Amazon WorkMail currently does not offer support for POP3 email access.

Q: When using an IMAP client application, can I access all items in my Amazon WorkMail mailbox?

The IMAP protocol provides access to email, but not to calendar items, contacts, notes, or tasks.

Q: When using an IMAP client application, will I be able to see all my email folders?

Yes, any folder which contains email will be visible and accessible using an IMAP client application.

Q: How do I send email when using an IMAP email client application?

You can send email by configuring your IMAP email client to use the Amazon WorkMail SMTP gateway. Amazon WorkMail SMTP addresses can be found at AWS Regions and Endpoints.

Q. What is the Amazon WorkMail SMTP Gateway?

The Simple Mail Transfer Protocol (SMTP) gateway is an Amazon WorkMail service which allows you to submit email messages for delivery to both internal and external recipients. To learn more, please see Connect your Client IMAP Application.

Q. What email client applications can I use to send email using the Amazon WorkMail SMTP gateway?

You can use the Amazon WorkMail SMTP gateway to send email using any email client that supports the SMTP protocol. This includes popular email clients like Microsoft Outlook, Apple Mail or Mozilla Thunderbird.

Setup and Maintenance

Q: Do I need to set up a directory to use Amazon WorkMail?

Each user you add to your Amazon WorkMail organization needs to exist in a directory, but you do not have to provision a directory yourself. You can integrate your existing Microsoft Active Directory with Amazon WorkMail using AWS Directory Service AD Connector or run AWS Directory Service for Microsoft Active Directory Enterprise Edition ("Microsoft AD") so you don’t have to manage users in two places and users can continue to use their existing Microsoft Active Directory credentials. Alternatively, you can have Amazon WorkMail create and manage a Simple AD directory for you and have users in that directory created when you add them to your Amazon WorkMail organization.

Q: How can I integrate with an existing Microsoft Active Directory?

You can integrate with an existing Microsoft Active Directory by setting up an AWS Directory Service AD Connector or Microsoft AD and enabling Amazon WorkMail for this directory. After you've configured this integration, you can choose which users you would like to enable for Amazon WorkMail from a list of users in your existing directory, and users can log in to Amazon WorkMail using their existing Active Directory credentials.

Q: Can I use my existing domain name with Amazon WorkMail?

Yes. You can add your existing domain name to Amazon WorkMail using the AWS Management Console. Before the domain name can be used, you must verify the ownership of the domain name. You can verify the ownership by adding a DNS record to your DNS server.

Q: Can I assign multiple email addresses to a user account?

Yes. You can assign multiple email addresses to a user account using the AWS Management Console.

Q: Can I create distribution groups to deliver email to multiple users?

Yes. You can create new distribution group or enable an existing group from your Microsoft Active Directory using the AWS Management Console. These distribution groups are available in the Global Address Book. Users can also create personal distribution groups using Microsoft Outlook or the Amazon WorkMail web application.

Q: What happens if a user forgets their password to access Amazon WorkMail?

If Amazon WorkMail is integrated with an existing Active Directory domain, then the user would follow the existing lost password process for your existing domain, such as contacting an internal helpdesk. If the account is integrated with a Simple AD directory and a user forgets their password, then the account's IT administrator can reset the password from the AWS Management Console.

Q: How does an IT administrator remove a user's access to Amazon WorkMail?

The account's IT administrator can remove a user's access to Amazon WorkMail using the AWS Management Console.

Q: Does Amazon WorkMail provide a management API?

No. Amazon WorkMail does not currently provide a management API.

Administrative SDK

Q: Does Amazon WorkMail offer an SDK?

Yes. Amazon WorkMail provides an administrative SDK so you can natively integrate WorkMail with your existing services. The SDK enables programmatic user, email group, and meeting room or equipment resource management through API calls. This means your existing IT service management tools, workflows, and third party applications can automate WorkMail migration and management. To learn more, please visit our our API reference.

Additionally, Amazon WorkMail offers an SDK for accessing the content of email messages that are in the process of being sent from or delivered to your organization. You can use this SDK in combination with Email Flow Rules to trigger Lambdas on mail sending or delivery and access the full message for analysis or integration with other systems. To learn more, see the API Reference.

Journaling

Q: How can I start using email journaling?

Email journaling can be setup from the Amazon WorkMail Management Console under Organization Settings. You can enable email journaling, specify the email address to which journaled emails are sent, and specify the email address to which reports are sent.

Q: Can I apply email journaling to a specific set of actions or users?

No. Today email journaling is a global setting that is applied to all inbound and outbound email, and all users.

Q: Does email journaling apply to recipients in the blind carbon copy (BCC) field?

Yes. Email sent using BCC recipients is recorded using email journaling.

Q: Will journaling reports show email recipients in the BCC field?

For outbound email, journaling reports will contain the details of recipients in the BCC field. For inbound email, the journaling report will only contain of details of recipients in the BCC field if those recipients are in your Amazon WorkMail organization.

Q: Will emails marked as spam be journaled?

Yes, they will.

Q: Will emails marked as containing viruses be journaled?

No. Emails that contain viruses will be dropped and will not be journaled.

Q: What actions will be taken in case of delivery failures to the journaling destination mailbox?

Amazon WorkMail will continue to try to deliver the journaled messages to the journaling destination mailbox for 12 hours. In case of continuous failure, the failure reports will be delivered to the address you specify in the Amazon WorkMail Management Console.

Q: What do journaling failed delivery reports contain?

Whenever journaled email fails to be delivered to the primary journaling address, a report is sent to the failed delivery report email address you specify in the Amazon WorkMail Management Console. This report contains information about each journaled message that failed to be delivered, but does not show the contents of the original message.

Q: What is the email address from which journaled emails are sent?

Journaled emails are be sent from amazonjournaling@<alias>.awsapps.com where <alias> is your Amazon WorkMail organization name.

Q: Is there an additional cost to using email journaling?

No, there is no additional cost to using email journaling.

Q: Which SMTP headers will identify a journaled message by the journaling agent?

“X-WM-Journal-Report” will be used as the header to identify journaled messages. This header will be signed so that it cannot be mimicked.

Q: Do journaling messages count against the sending limits?

No, journaling messages are always sent as long as the user is allowed to send a message. They are not counted against that user’s sending limit. On receiving a message, the journaling message is always sent as long as it can be delivered to a user.

Migration & Interoperability with Microsoft Exchange Server

Q: How can I migrate mailboxes from my existing email solution to Amazon WorkMail?

You can migrate your existing mailboxes to Amazon WorkMail using solutions from a preferred Amazon WorkMail migration provider. To see a list of providers, please visit this webpage. If you’re migrating from Microsoft Exchange Server 2013 or 2010, you can set up interoperability to minimize disruption for your end users.

Q: Does Amazon WorkMail support interoperability with Microsoft Exchange Server?

Yes, Amazon WorkMail supports interoperability with Microsoft Exchange Server 2013 and 2010. You can learn about how to set up interoperability here.

Q: What interoperability capabilities does Amazon WorkMail support?

Interoperability allows you to use the same corporate domain for all mailboxes on both Microsoft Exchange and Amazon WorkMail. Your users can seamlessly schedule meetings with bi-directional sharing of calendar free-busy information between the two environments, and access user and resource information through a unified global address book.

Q: Which versions of Microsoft Exchange Server are supported with Amazon WorkMail interoperability?

Amazon WorkMail offers interoperability support with Microsoft Exchange Server 2013 and 2010.

Q: Are there additional charges to use interoperability features?

No. Interoperability features are included in Amazon WorkMail per mailbox pricing.

Q: Can users access Amazon WorkMail using their existing Microsoft Active Directory credentials?

Yes, users can connect to Amazon WorkMail using their existing Microsoft Active Directory credentials.

Q: Will mailboxes on Amazon WorkMail use the same domain as mailboxes on my Microsoft Exchange server?

Yes. To make this possible, you need to enable email routing between Microsoft Exchange and Amazon WorkMail so that mailboxes on both environments use the same corporate domain. To set up email routing, you can follow the steps outlined here.

Q: Which email platform handles incoming email traffic when interoperability is established?

Your on-premises Microsoft Exchange Server handles and processes all incoming email. If you’re using interoperability for migration, you can switch your MX record to point to Amazon WorkMail when your migration is complete.

Q: Can I restrict access to my Microsoft Exchange Server to just my VPC?

No, you can’t restrict access to the Exchange Server to your VPC. As of now, the EWS endpoint of your on-premises Microsoft Exchange environment needs to be publicly available.

Q: Does Amazon WorkMail support bi-directional sharing of calendar free-busy information with Microsoft Exchange?

Yes, interoperability provides you bi-directional sharing of calendar free-busy information between your Amazon WorkMail and Microsoft Exchange environments. Please follow the steps here.

Q: How does Amazon WorkMail interact with my on-premises Microsoft Exchange Server to perform bi-directional calendar free-busy lookups?

You will need to configure availability settings on Amazon WorkMail and Microsoft Exchange to share calendar free-busy information. Amazon WorkMail uses the EWS URL for your Microsoft Exchange server to perform free-busy lookups. Amazon WorkMail uses an Exchange service account to login to Exchange and read free-busy data of the users in the Microsoft Exchange organization.

For free-busy lookups of Amazon WorkMail users from your Microsoft Exchange Server, Exchange performs an Autodiscover request and connects to the Amazon WorkMail EWS endpoint using an Amazon WorkMail service account.
You can find more information on this here.

Q: Do I need to set up federation on my on-premises Microsoft Exchange server?

No, for interoperability support with Amazon WorkMail, you don’t need to set up federation on your Microsoft Exchange server.

Q: Can I also view subject and location in the free-busy details when interoperability is enabled?

Yes, to view subject and location information, the service account user needs to have access to this information.

Q: Can an Amazon WorkMail user manage the shared calendar or shared folder of a user on Microsoft Exchange (and vice versa).

No, for calendar delegation or accessing shared folders, both users need to be on the same email platform. We recommend migrating users who use calendar and mailbox delegation in the same batch.

Q: How does Amazon WorkMail interact with my on-premises Microsoft Exchange Server to create a unified global address book?

Once interoperability support is enabled, Amazon WorkMail performs a synchronization of the address book with your on-premises Active Directory every four hours, using AD Connector. All Microsoft Exchange users, groups, and resources are automatically added to your Amazon WorkMail address book.

Q: Will all Microsoft Exchange Server objects synchronize to the Amazon WorkMail global address book?

Amazon WorkMail will synchronize users, groups, resources, and contacts that reside in Microsoft Exchange Server. Amazon WorkMail will not synchronize dynamic groups or address lists. When your Microsoft Exchange global address book contains these objects, they won't be available in Amazon WorkMail.

Q: Will Amazon WorkMail still synchronize with my Active Directory when interoperability support isn’t enabled?

Yes, Amazon WorkMail will still synchronize with your Active Directory when interoperability support is disabled. In this scenario only changes to Amazon WorkMail users and groups are synchronized.

Q: Does the Microsoft Outlook offline address book also contain all my Microsoft Exchange users, and groups, and resources?

Yes, the Microsoft Outlook offline address book will contain both Amazon WorkMail Microsoft Exchange users, groups, and resources.

Q: Can my distribution groups contain both Amazon WorkMail and Microsoft Exchange users as members?

Yes, you can have both Amazon WorkMail and Microsoft Exchange users as members of distribution groups.

Q: Can I still create new resource in Amazon WorkMail when interoperability support is enabled?

No. To create new resources in Amazon WorkMail, you first need to disable interoperability support. Once your new resources have been created, you can then turn interoperability support back on. This is done to ensure resources are synchronized back to your Microsoft Exchange Server.

Email Flow Rules

Q: What are email flow rules?

Amazon WorkMail allows you to use email flow rules to filter or route email traffic for your Amazon WorkMail organizations. On inbound emails, this can help you reduce email from unwanted senders, route suspicious mail to junk folders, and trigger AWS Lambda functions. On outbound, you can block sending to certain domains, route mail through custom SMTP endpoints, or trigger Lambda functions. Email flow rules can be applied based on specific email addresses, or entire email domains.

Q: What types of email flow rules can I create?

For inbound mail, mail flow rules can be created to filter email based on specific email addresses, or entire email domains. Examples include:

- Reject all incoming mail from example.com and its subdomains, generating a bounce message to the sender.
- Reject all incoming mail from example.com, except when from myemail@example.com.
- Reject all incoming mail from user@example.com.
- Bypass the spam check for all incoming email from example.com, delivering the messages instead to users' inbox.
- Deliver all messages from example.com to users' junk folders.
- Trigger AWS Lambda function that you define on receiving mail.

For outbound mail, mail flow rules can be created to filter email or route to an SMTP endpoint or AWS Lambda function. Examples include:

- Reject all outgoing mail from example.com to example2.com, generating an non-deliverable report (NDR) to the sender.
- Reject all outgoing mail to example.com silently.
- Route all mail from example.com that is going to a domain other than example.com through a custom SMTP endpoint that you define.
- Trigger AWS Lambda function that you define on all outgoing mail.

Q:What types of email data are passed to the Lambda function?

The Lambda function will receive the message id, sender, recipient, and subject of an email.

Q: Can I retrieve more information about an email message from within my Lambda?

Yes, you can retrieve the full content of the email message using WorkMail's SDKs. See the Admin Guide for more information.

Q: What format does the email content come in when retrieving it from my Lambda?

The WorkMail SDK will return the raw MIME content of the message that is being processed. You can use common MIME-processing libraries, such as JavaMail for Java or email.parser for Python to convert this to a structured format for easier parsing.

Q: Can I stop or re-route an email using a Lambda function?

No, Lambda functions are triggered asynchronously, and therefore cannot be used to affect the mail flow.

Q: How can I start using email flow rules?

Rules can be set up from the Amazon WorkMail management console by navigating to Organization Settings. You can create, modify, and delete flow rules under the Email Flow Rules tab.

Q: Can I perform filtering based on IP address or range?

IP based filtering is already supported by Amazon Simple Email Service. Please see Creating IP Address Filters for Amazon SES Email Receiving to learn more about IP-based filtering.

Q: What happens if email containing a virus is received from a source specified to bypass spam checks?

Amazon WorkMail scans all incoming and outgoing email for spam, malware, and viruses. All email containing viruses is dropped and not delivered, regardless of the configured flow rules.

Q: What happens if email flow rules overlap?

If you have email for which multiple email flow rules match, the action of the most specific rule will be applied. For example, a rule for a specific email address will take precedence over a rule for an entire domain. If multiple rules have the same specificity, the most restrictive action will be applied (for example, Drop will take precedence over Bounce). Please see Managing Email Flows for more information.

Q: Are there limits on the number of rules I can create?

Yes. To learn more about limits related to email flow rules, please see AWS Service Limits.

Q: How long does a rule need to take effect?

Rules take effect immediately after creation.

Q: Is there any additional charge for defining email flow rules?

No, there is no additional charge for using email flow rules.

Security

Q: How is data transmitted to Amazon WorkMail?

All data in transit is encrypted using industry-standard SSL. Our web application, and mobile and desktop clients transmit data to Amazon WorkMail using SSL.

Q: Can I choose the AWS region where my data is stored?

Yes. You choose the AWS region where your organization’s data is stored. Please refer to the Regional Products and Services page for details of Amazon WorkMail availability by region.

Q. How do I decide which AWS region to use?

There are several factors to consider, based on your needs, including whether using a specific AWS region enables you to meet regulatory and compliance requirements. We generally recommend that you set up your Amazon WorkMail organization in the region nearest to where most of your users are located, to reduce data access latencies.

Q: How is Amazon WorkMail protected from malware/viruses?

Amazon WorkMail scans all incoming and outgoing email for spam, malware, and viruses to help protect customers from malicious email.

Q: Does Amazon WorkMail offer support for mobile device policies, to protect data stored on mobile devices?

Yes. Amazon WorkMail gives you the ability to require a PIN or password on your users' devices, configure the password strength, require a device lock after a number of failed login attempts, require a screen lock for idle timeouts, and require device and storage card encryption.

Q: How can I manage my encryption key used for the data encryption in Amazon WorkMail?

Amazon WorkMail is integrated with Amazon Key Management Service for the encryption of your data. Key management can be performed from the Amazon IAM console. For more information about AWS Key Management Service, please see Amazon AWS Key Management developer guide.

Q: What data is encrypted with my encryption keys?

All email content, attachments, and metadata for a mailbox is encrypted using the customer-managed keys of that user's organization.

Q: Is my email encrypted when using the IMAP protocol to access my Amazon WorkMail mailbox?

Yes. All email communication is encrypted in transit by the secure connections made between the client and the server, and all email stored in Amazon WorkMail is encrypted at rest.

Q: Does Amazon WorkMail support S/MIME for signing and encrypting email?

Yes. Amazon WorkMail supports S/MIME signing and encryption in the Microsoft Outlook client and certain mobile devices like Apple iPhone and iPad. The Amazon WorkMail web application currently does not support S/MIME signing and encryption.

Q. What compliance certifications does Amazon WorkMail support?

Amazon Web Services has achieved the ISO 27001, ISO 27017 and ISO 27018 certifications. Amazon WorkMail regions in US East (N.Virginia), US West (Oregon) and EU (Ireland) are within the scope of the certifications. You can learn more about these certifications on the AWS Cloud Compliance section of the website.
You can also request a copy of the Service Organization Controls (SOC) report available from AWS Compliance to learn more about the security controls AWS uses to protect your data.

Q: How does AWS use my Amazon WorkMail email content?

You own your content in Amazon WorkMail, and you retain full ownership and control of your Amazon WorkMail email. We will not view, use, or move the contents of your Amazon WorkMail account unless authorized by you.

Integration with AWS Services

Amazon WorkDocs Integration

Q: How does Amazon WorkMail integrate with Amazon WorkDocs?

Amazon WorkDocs integration offers users the ability to distribute large documents easily from the Amazon WorkMail web application, keep control of sensitive documents distributed by email, and securely save email attachments in Amazon WorkDocs.

Q: How can I start using the Amazon WorkDocs integration?

To use the integration with Amazon WorkDocs, your organization first needs to be activated for Amazon WorkDocs. You can activate Amazon WorkDocs for your organization in the AWS Management Console. After this is done, you can enable Amazon WorkDocs for your users using the Amazon WorkDocs admin panel. After your users are enabled for Amazon WorkDocs, they can start using the Amazon WorkDocs integration in the Amazon WorkMail web application.
If your organization and users are already using Amazon WorkDocs, your users can start using the integration right after they are enabled for Amazon WorkMail.

Q: Can I use Amazon WorkMail without using Amazon WorkDocs?

Yes, however you will not be able to use the Amazon WorkDocs integration in the Amazon WorkMail web application.

Amazon Simple Email Service Integration

Q: How does Amazon WorkMail integrate with Amazon Simple Email Service?

Amazon WorkMail uses Amazon Simple Email Service to send all outgoing email. The test mail domain and your production domains are available for management in the Amazon Simple Email Service console.

Q: Will I be charged for outgoing email sent from Amazon WorkMail?

No. You won’t be charged for outgoing email sent from Amazon WorkMail.

Q: Do I need to increase Amazon SES sending limits to use Amazon WorkMail?

No. This is not needed to use with Amazon WorkMail. The SES limits only apply when you are using Amazon SES using the Amazon SES API for sending bulk email from your AWS account.

AWS CloudTrail Integration

Q: Does Amazon WorkMail integrate with AWS CloudTrail?

Yes. CloudTrail captures API calls from the WorkMail console or from WorkMail or WorkMailMessageFlow API operations. Using the information collected by CloudTrail, you can track requests made to WorkMail, the source IP address from which the requests were made, who made the requests, when they were made, and so on. To learn more about CloudTrail, including how to configure and enable it, see the AWS CloudTrail User Guide. To learn more about logging WorkMail API calls, see Logging Amazon WorkMail API Calls with AWS CloudTrail.

Q: Will I be charged for using AWS CloudTrail with Amazon WorkMail?

There is no additional WorkMail charge to use WorkMail with CloudTrail. There may be charges associated with delivering events using CloudTrail. For details, please see the CloudTrail Pricing.

Amazon CloudWatch Integration

Q: Does WorkMail offer email metrics?

Yes, WorkMail logs metrics for emails sent, received, and bounced free of charge in CloudWatch metrics

Q: Does WorkMail offer message tracking?

Yes, WorkMail offers the option to enable WorkMail Monitoring in CloudWatch logs. When activating logging, you can define the CloudWatch log group to log into, as well as the log retention period. WorkMail will then log detailed information for messages received and sent, when rules are applied, when message journaling is initiated, and for bounce messages.

Q: What data is logged in WorkMail Monitoring?

If logging is activated, WorkMail logs envelope data such as sender and recipients. Message bodies are not logged.

Q: How can I run queries on messages?

CloudWatch offers insights which allows for fast and easy querying on CloudWatch logs.

Pricing

Q: How will my business be charged for use of Amazon WorkMail?

There are no upfront fees or commitments to begin using Amazon WorkMail. At the end of the month, you are billed for that month's usage. You can view estimated charges for the current billing period by logging into the AWS Management Console and clicking on "Account Activity." You can get started with a free trial of Amazon WorkMail and activate up to 25 user accounts at no charge for the first 30 days. You can use the WorkMail console to get started today.

You are charged for the number of user accounts per month. The number of users billed in a month is based on the average number of active user accounts throughout the month. For every user account, your business is charged a monthly subscription fee. If a user account is deactivated during the month, the monthly subscription fee for that account will be prorated based on the number of active days. For more information on how pricing works, see the pricing page.

Q: Is there a free trial for Amazon WorkMail?

Yes. You can activate up to 25 users at no charge for the first 30 days after you sign up for Amazon WorkMail. After this period ends, you are charged for all active users unless you remove them or deregister your Amazon WorkMail account.

Q: Will I be charged for creating or using resources (such as meeting rooms)?

No. Creating or using of resources within Amazon WorkMail is available free of charge.

Q: Is there an additional charge for using IMAP client applications?

No. IMAP access is included in the Amazon WorkMail mailbox pricing.

Amazon EC2 FAQs

General

Overview | EC2 On-Demand Instance limits | Changes to EC2 SMTP endpoint policy | Service level agreement (SLA)

Overview

Q: What is Amazon Elastic Compute Cloud (Amazon EC2)?

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.

Q: What can I do with Amazon EC2?

Just as Amazon Simple Storage Service (Amazon S3) enables storage in the cloud, Amazon EC2 enables “compute” in the cloud. Amazon EC2’s simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon’s proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use.

Q: How can I get started with Amazon EC2?

To sign up for Amazon EC2, click the “Sign up for This Web Service” button on the Amazon EC2 detail page. You must have an Amazon Web Services account to access this service; if you do not already have one, you will be prompted to create one when you begin the Amazon EC2 sign-up process. After signing up, please refer to the Amazon EC2 documentation, which includes our Getting Started Guide.

Q: Why am I asked to verify my phone number when signing up for Amazon EC2?

Amazon EC2 registration requires you to have a valid phone number and email address on file with AWS in case we ever need to contact you. Verifying your phone number takes only a couple of minutes and involves receiving a phone call during the registration process and entering a PIN number using the phone key pad.

Q: What can developers now do that they could not before?

Until now, small developers did not have the capital to acquire massive compute resources and ensure they had the capacity they needed to handle unexpected spikes in load. Amazon EC2 enables any developer to leverage Amazon's own benefits of massive scale with no up-front investment or performance compromises. Developers are now free to innovate knowing that no matter how successful their businesses become, it will be inexpensive and simple to ensure they have the compute capacity they need to meet their business requirements.

The "Elastic" nature of the service allows developers to instantly scale to meet spikes in traffic or demand. When computing requirements unexpectedly change (up or down), Amazon EC2 can instantly respond, meaning that developers have the ability to control how many resources are in use at any given point in time. In contrast, traditional hosting services generally provide a fixed number of resources for a fixed amount of time, meaning that users have a limited ability to easily respond when their usage is rapidly changing, unpredictable, or is known to experience large peaks at various intervals.

Q: How do I run systems in the Amazon EC2 environment?

Once you have set up your account and select or create your AMIs, you are ready to boot your instance. You can start your AMI on any number of On-Demand instances by using the RunInstances API call. You simply need to indicate how many instances you wish to launch. If you wish to run more than 20 On-Demand instances, complete the Amazon EC2 instance request form.

If Amazon EC2 is able to fulfill your request, RunInstances will return success, and we will start launching your instances. You can check on the status of your instances using the DescribeInstances API call. You can also programmatically terminate any number of your instances using the TerminateInstances API call.

If you have a running instance using an Amazon EBS boot partition, you can also use the StopInstances API call to release the compute resources but preserve the data on the boot partition. You can use the StartInstances API when you are ready to restart the associated instance with the Amazon EBS boot partition.

In addition, you have the option to use Spot Instances to reduce your computing costs when you have flexibility in when your applications can run. Read more about Spot Instances for a more detailed explanation on how Spot Instances work.

If you prefer, you can also perform all these actions from the AWS Management Console or through the command line using our command line tools, which have been implemented with this web service API.

Q: What is the difference between using the local instance store and Amazon Elastic Block Store (Amazon EBS) for the root device?

When you launch your Amazon EC2 instances you have the ability to store your root device data on Amazon EBS or the local instance store. By using Amazon EBS, data on the root device will persist independently from the lifetime of the instance. This enables you to stop and restart the instance at a subsequent time, which is similar to shutting down your laptop and restarting it when you need it again.

Alternatively, the local instance store only persists during the life of the instance. This is an inexpensive way to launch instances where data is not stored to the root device. For example, some customers use this option to run large web sites where each instance is a clone to handle web traffic.

Q: How quickly will systems be running?

It typically takes less than 10 minutes from the issue of the RunInstances call to the point where all requested instances begin their boot sequences. This time depends on a number of factors including: the size of your AMI, the number of instances you are launching, and how recently you have launched that AMI. Images launched for the first time may take slightly longer to boot.

Q: How do I load and store my systems with Amazon EC2?

Amazon EC2 allows you to set up and configure everything about your instances from your operating system up to your applications. An Amazon Machine Image (AMI) is simply a packaged-up environment that includes all the necessary bits to set up and boot your instance. Your AMIs are your unit of deployment. You might have just one AMI or you might compose your system out of several building block AMIs (e.g., webservers, appservers, and databases). Amazon EC2 provides a number of tools to make creating an AMI easy. Once you create a custom AMI, you will need to bundle it. If you are bundling an image with a root device backed by Amazon EBS, you can simply use the bundle command in the AWS Management Console. If you are bundling an image with a boot partition on the instance store, then you will need to use the AMI Tools to upload it to Amazon S3. Amazon EC2 uses Amazon EBS and Amazon S3 to provide reliable, scalable storage of your AMIs so that we can boot them when you ask us to do so.

Or, if you want, you don’t have to set up your own AMI from scratch. You can choose from a number of globally available AMIs that provide useful instances. For example, if you just want a simple Linux server, you can choose one of the standard Linux distribution AMIs.

Q: How do I access my systems?

The RunInstances call that initiates execution of your application stack will return a set of DNS names, one for each system that is being booted. This name can be used to access the system exactly as you would if it were in your own data center. You own that machine while your operating system stack is executing on it.

Q: Is Amazon EC2 used in conjunction with Amazon S3?

Yes, Amazon EC2 is used jointly with Amazon S3 for instances with root devices backed by local instance storage. By using Amazon S3, developers have access to the same highly scalable, reliable, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites. In order to execute systems in the Amazon EC2 environment, developers use the tools provided to load their AMIs into Amazon S3 and to move them between Amazon S3 and Amazon EC2. See How do I load and store my systems with Amazon EC2? for more information about AMIs.

We expect developers to find the combination of Amazon EC2 and Amazon S3 to be very useful. Amazon EC2 provides cheap, scalable compute in the cloud while Amazon S3 allows users to store their data reliably.

Q: How many instances can I run in Amazon EC2?

You are limited to running On-Demand Instances per your vCPU-based On-Demand Instance limit, purchasing 20 Reserved Instances, and requesting Spot Instances per your dynamic Spot limit per region. New AWS accounts may start with limits that are lower than the limits described here.

If you need more instances, complete the Amazon EC2 limit increase request form with your use case, and your limit increase will be considered. Limit increases are tied to the region they were requested for.

Q: Are there any limitations in sending email from Amazon EC2 instances?

Yes. In order to maintain the quality of Amazon EC2 addresses for sending email, we enforce default limits on the amount of email that can be sent from EC2 accounts. If you wish to send larger amounts of email from EC2, you can apply to have these limits removed from your account by filling out this form.

Q: How quickly can I scale my capacity both up and down?

Amazon EC2 provides a truly elastic computing environment. Amazon EC2 enables you to increase or decrease capacity within minutes, not hours or days. You can commission one, hundreds or even thousands of server instances simultaneously. When you need more instances, you simply call RunInstances, and Amazon EC2 will typically set up your new instances in a matter of minutes. Of course, because this is all controlled with web service APIs, your application can automatically scale itself up and down depending on its needs.

Q: What operating system environments are supported?

Amazon EC2 currently supports a variety of operating systems including: Amazon Linux, Ubuntu, Windows Server, Red Hat Enterprise Linux, SUSE Linux Enterprise Server, Fedora, Debian, CentOS, Gentoo Linux, Oracle Linux, and FreeBSD. We are looking for ways to expand it to other platforms.

Q: Does Amazon EC2 use ECC memory?

In our experience, ECC memory is necessary for server infrastructure, and all the hardware underlying Amazon EC2 uses ECC memory.

Q: How is this service different than a plain hosting service?

Traditional hosting services generally provide a pre-configured resource for a fixed amount of time and at a predetermined cost. Amazon EC2 differs fundamentally in the flexibility, control and significant cost savings it offers developers, allowing them to treat Amazon EC2 as their own personal data center with the benefit of Amazon.com's robust infrastructure.

When computing requirements unexpectedly change (up or down), Amazon EC2 can instantly respond, meaning that developers have the ability to control how many resources are in use at any given point in time. In contrast, traditional hosting services generally provide a fixed number of resources for a fixed amount of time, meaning that users have a limited ability to easily respond when their usage is rapidly changing, unpredictable, or is known to experience large peaks at various intervals.

Secondly, many hosting services don't provide full control over the compute resources being provided. Using Amazon EC2, developers can choose not only to initiate or shut down instances at any time, they can completely customize the configuration of their instances to suit their needs – and change it at any time. Most hosting services cater more towards groups of users with similar system requirements, and so offer limited ability to change these.

Finally, with Amazon EC2 developers enjoy the benefit of paying only for their actual resource consumption – and at very low rates. Most hosting services require users to pay a fixed, up-front fee irrespective of their actual computing power used, and so users risk overbuying resources to compensate for the inability to quickly scale up resources within a short time frame.

EC2 On-Demand Instance limits

Q: What is changing?

Amazon EC2 is transitioning On-Demand Instance limits from the current instance count-based limits to the new vCPU-based limits to simplify the limit management experience for AWS customers. Usage toward the vCPU-based limit is measured in terms of number of vCPUs (virtual central processing units) for the Amazon EC2 Instance Types to launch any combination of instance types that meet your application needs.

Q: What are vCPU-based limits?

You are limited to running one or more On-Demand Instances in an AWS account, and Amazon EC2 measures usage towards each limit based on the total number of vCPUs (virtual central processing unit) that are assigned to the running On-Demand instances in your AWS account. The following table shows the number of vCPUs for each instance size. The vCPU mapping for some instance types may differ; see Amazon EC2 Instance Types for details.

Instance Size	vCPUs
nano	1
micro	1
small	1
medium	1
large	2
xlarge	4
2xlarge	8
3xlarge	12

4xlarge	16
8xlarge	32
9xlarge	36
10xlarge	40
12xlarge	48
16xlarge	64
18xlarge	72
24xlarge	96
32xlarge	128

Q: How many On-Demand instances can I run in Amazon EC2?

There are five vCPU-based instance limits, each defines the amount of capacity you can use of a given instance family. All usage of instances in a given family, regardless of generation, size, or configuration variant (e.g. disk, processor type), will accrue towards the family’s total vCPU limit, listed in the table below. New AWS accounts may start with limits that are lower than the limits described here.

On-Demand Instance Limit Name	Default vCPU Limit
Running On-Demand Standard (A, C, D, H, I, M, R, T, Z) instances	1152 vCPUs
Running On-Demand F instances	128 vCPUs
Running On-Demand G instances	128 vCPUs
Running On-Demand Inf instances	128 vCPUs
Running On-Demand P instances	128 vCPUs
Running On-Demand X instances	128 vCPUs

Q: Are these On-Demand Instance vCPU-based limits regional?

Yes, the On-Demand Instance limits for an AWS account are set on a per-region basis.

Q: Will these limits change over time?

Yes, limits can change over time. Amazon EC2 is constantly monitoring your usage within each region and your limits are raised automatically based on your use of EC2.

Q: How can I request a limit increase?

Even though EC2 automatically increases your On-Demand Instance limits based on your usage, if needed you can request a limit increases from the Limits Page on Amazon EC2 console, the Amazon EC2 service page on the Service Quotas console, or the Service Quotas API/CLI.

Q: How can I calculate my new vCPU limit?

You can find the vCPU mapping for each of the Amazon EC2 Instance Types or use the simplified vCPU Calculator to compute the total vCPU limit requirements for your AWS account.

Q: Do vCPU limits apply when purchasing Reserved Instances or requesting Spot Instances?

No, the vCPU-based limits only apply to running On-Demand instances.

Q: How can I view my current On-Demand Instance limits?

You can find your current On-Demand Instance limits on the EC2 Service Limits page in the Amazon EC2 console, or from the Service Quotas console and APIs.

Q: Will this affect running instances?

No, opting into vCPU-based limits will not affect any running instances.

Q: Can I still launch the same number of instances?

Yes, the vCPU-based instance limits allow you to launch at least the same number of instances as count-based instance limits.

Q: Will I be able to view instance usage against these limits?

With the Amazon CloudWatch metrics integration, you can view EC2 usage against limits in the Service Quotas console. Service Quotas also enables customers to use CloudWatch for configuring alarms to warn customers of approaching limits. In addition, you can continue to track and inspect your instance usage in Trusted Advisor and Limit Monitor.

Q: Will I still be able to use the DescribeAccountAttributes API?

With the vCPU limits, we no longer have total instance limits governing the usage. Hence the DescribeAccountAttributes API will no longer return the max-instances value. Instead you can now use the Service Quotas APIs to retrieve information about EC2 limits. You can find more information about the Service Quotas APIs in the AWS documentation.

Q: Will the vCPU limits have an impact on my monthly bill?

No. EC2 usage is still calculated either by the hour or the second, depending on which AMI you're running and the instance type and size you’ve launched.

Q: Will vCPU limits be available in all Regions?

vCPU-based instance limits are available in all commercial AWS Regions except the AWS China (Beijing and Ningxia) Regions and they are available in AWS GovCloud (US).

Changes to EC2 SMTP endpoint policy

Q. What is changing?

Starting Jan-27 2020, Amazon Elastic Compute Cloud (EC2) will begin rolling out a change to restrict email traffic over port 25 by default to protect customers and other recipients from spam and email abuse. Port 25 is typically used as the default SMTP port to send emails. AWS accounts that have requested and had Port 25 throttles removed in the past will not be impacted by this change.

Q. I have a valid use-case for sending emails to port 25 from EC2. How can I have these port 25 restrictions removed?

If you have a valid use-case for sending emails to port 25 (SMTP) from EC2, please submit a Request to Remove Email Sending Limitations to have these restrictions lifted. You can alternately send emails using a different port, or leverage an existing authenticated email relay service such as Amazon Simple Email Service (SES).

Service level agreement (SLA)

Q. What does your Amazon EC2 Service Level Agreement guarantee?

Our SLA guarantees a Monthly Uptime Percentage of at least 99.99% for Amazon EC2 and Amazon EBS within a Region.

Q. How do I know if I qualify for a SLA Service Credit?

You are eligible for a SLA credit for either Amazon EC2 or Amazon EBS (whichever was Unavailable, or both if both were Unavailable) if the Region that you are operating in has an Monthly Uptime Percentage of less than 99.95% during any monthly billing cycle. For full details on all of the terms and conditions of the SLA, as well as details on how to submit a claim, please see http://aws.amazon.com/ec2/sla/

Instance types

Accelerated Computing instances | Compute Optimized instances | General Purpose instances | High Memory instances | Memory Optimized instances | Previous Generation instances | Storage Optimized instances

Accelerated Computing instances

Q: What are Accelerated Computing instances?

Accelerated Computing instance family is a family of instances which use hardware accelerators, or co-processors, to perform some functions, such as floating-point number calculation and graphics processing, more efficiently than is possible in software running on CPUs. Amazon EC2 provides three types of Accelerated Computing instances – GPU compute instances for general-purpose computing, GPU graphics instances for graphics intensive applications, and FPGA programmable hardware compute instances for advanced scientific workloads.

Q. When should I use GPU Graphics and Compute instances?

GPU instances work best for applications with massive parallelism such as workloads using thousands of threads. Graphics processing is an example with huge computational requirements, where each of the tasks is relatively small, the set of operations performed form a pipeline, and the throughput of this pipeline is more important than the latency of the individual operations. To be able build applications that exploit this level of parallelism, one needs GPU device specific knowledge by understanding how to program against various graphics APIs (DirectX, OpenGL) or GPU compute programming models (CUDA, OpenCL).

Q: How are P3 instances different from G3 instances?

P3 instances are the next-generation of EC2 general-purpose GPU computing instances, powered by up to 8 of the latest-generation NVIDIA Tesla V100 GPUs. These new instances significantly improve performance and scalability, and add many new features, including new Streaming Multiprocessor (SM) architecture for machine learning (ML)/deep learning (DL) performance optimization, second-generation NVIDIA NVLink high-speed GPU interconnect, and highly tuned HBM2 memory for higher-efficiency.

G3 instances use NVIDIA Tesla M60 GPUs and provide a high-performance platform for graphics applications using DirectX or OpenGL. NVIDIA Tesla M60 GPUs support NVIDIA GRID Virtual Workstation features, and H.265 (HEVC) hardware encoding. Each M60 GPU in G3 instances supports 4 monitors with resolutions up to 4096x2160, and is licensed to use NVIDIA GRID Virtual Workstation for one Concurrent Connected User. Example applications of G3 instances include 3D visualizations, graphics-intensive remote workstation, 3D rendering, application streaming, video encoding, and other server-side graphics workloads.

Q: What are the benefits of NVIDIA Volta GV100 GPUs?

The new NVIDIA Tesla V100 accelerator incorporates the powerful new Volta GV100 GPU. GV100 not only builds upon the advances of its predecessor, the Pascal GP100 GPU, it significantly improves performance and scalability, and adds many new features that improve programmability. These advances will supercharge HPC, data center, supercomputer, and deep learning systems and applications.

Q: Who will benefit from P3 instances?

P3 instances with their high computational performance will benefit users in artificial intelligence (AI), machine learning (ML), deep learning (DL) and high performance computing (HPC) applications. Users includes data scientists, data architects, data analysts, scientific researchers, ML engineers, IT managers and software developers. Key industries include transportation, energy/oil & gas, financial services (banking, insurance), healthcare, pharmaceutical, sciences, IT, retail, manufacturing, high-tech, transportation, government, academia, among many others.

Q: What are some key use cases of P3 instances?

P3 instance use GPUs to accelerate numerous deep learning systems and applications including autonomous vehicle platforms, speech, image, and text recognition systems, intelligent video analytics, molecular simulations, drug discovery, disease diagnosis, weather forecasting, big data analytics, financial modeling, robotics, factory automation, real-time language translation, online search optimizations, and personalized user recommendations, to name just a few.

Q: Why should customers use GPU-powered Amazon P3 instances for AI/ML and HPC?

GPU-based compute instances provide greater throughput and performance because they are designed for massively parallel processing using thousands of specialized cores per GPU, versus CPUs offering sequential processing with a few cores. In addition, developers have built hundreds of GPU-optimized scientific HPC applications such as quantum chemistry, molecular dynamics, meteorology, among many others. Research indicates that over 70% of the most popular HPC applications provide built-in support for GPUs.

Q: Will P3 instances support EC2 Classic networking and Amazon VPC?

P3 instances will support VPC only.

Q: How are G3 instances different from P2 instances?

G3 instances use NVIDIA Tesla M60 GPUs and provide a high-performance platform for graphics applications using DirectX or OpenGL. NVIDIA Tesla M60 GPUs support NVIDIA GRID Virtual Workstation features, and H.265 (HEVC) hardware encoding. Each M60 GPU in G3 instances supports 4 monitors with resolutions up to 4096x2160, and is licensed to use NVIDIA GRID Virtual Workstation for one Concurrent Connected User. Example applications of G3 instances include 3D visualizations, graphics-intensive remote workstation, 3D rendering, application streaming, video encoding, and other server-side graphics workloads.

P2 instances use NVIDIA Tesla K80 GPUs and are designed for general purpose GPU computing using the CUDA or OpenCL programming models. P2 instances provide customers with high bandwidth 25 Gbps networking, powerful single and double precision floating-point capabilities, and error-correcting code (ECC) memory, making them ideal for deep learning, high performance databases, computational fluid dynamics, computational finance, seismic analysis, molecular modeling, genomics, rendering, and other server-side GPU compute workloads.

Q: How are P3 instances different from P2 instances?

P3 Instances are the next-generation of EC2 general-purpose GPU computing instances, powered by up to 8 of the latest-generation NVIDIA Volta GV100 GPUs. These new instances significantly improve performance and scalability and add many new features, including new Streaming Multiprocessor (SM) architecture, optimized for machine learning (ML)/deep learning (DL) performance, second-generation NVIDIA NVLink high-speed GPU interconnect, and highly tuned HBM2 memory for higher-efficiency.

P2 instances use NVIDIA Tesla K80 GPUs and are designed for general purpose GPU computing using the CUDA or OpenCL programming models. P2 instances provide customers with high bandwidth 25 Gbps networking, powerful single and double precision floating-point capabilities, and error-correcting code (ECC) memory.

Q: What APIs and programming models are supported by GPU Graphics and Compute instances?

P3 instances support CUDA 9 and OpenCL, P2 instances support CUDA 8 and OpenCL 1.2 and G3 instances support DirectX 12, OpenGL 4.5, CUDA 8, and OpenCL 1.2.

Q: Where do I get NVIDIA drivers for P3 and G3 instances?

There are two methods by which NVIDIA drivers may be obtained. There are listings on the AWS Marketplace which offer Amazon Linux AMIs and Windows Server AMIs with the NVIDIA drivers pre-installed. You may also launch 64-bit, HVM AMIs and install the drivers yourself. You must visit the NVIDIA driver website and search for the NVIDIA Tesla V100 for P3, NVIDIA Tesla K80 for P2, and NVIDIA Tesla M60 for G3 instances.

Q: Which AMIs can I use with P3, P2 and G3 instances?

You can currently use Windows Server, SUSE Enterprise Linux, Ubuntu, and Amazon Linux AMIs on P2 and G3 instances. P3 instances only support HVM AMIs. If you want to launch AMIs with operating systems not listed here, contact AWS Customer Support with your request or reach out through EC2 Forums.

Q: Does the use of G2 and G3 instances require third-party licenses?

Aside from the NVIDIA drivers and GRID SDK, the use of G2 and G3 instances does not necessarily require any third-party licenses. However, you are responsible for determining whether your content or technology used on G2 and G3 instances requires any additional licensing. For example, if you are streaming content you may need licenses for some or all of that content. If you are using third-party technology such as operating systems, audio and/or video encoders, and decoders from Microsoft, Thomson, Fraunhofer IIS, Sisvel S.p.A., MPEG-LA, and Coding Technologies, please consult these providers to determine if a license is required. For example, if you leverage the on-board h.264 video encoder on the NVIDIA GRID GPU you should reach out to MPEG-LA for guidance, and if you use mp3 technology you should contact Thomson for guidance.

Q: Why am I not getting NVIDIA GRID features on G3 instances using the driver downloaded from NVIDIA website?

The NVIDIA Tesla M60 GPU used in G3 instances requires a special NVIDIA GRID driver to enable all advanced graphics features, and 4 monitors support with resolution up to 4096x2160. You need to use an AMI with NVIDIA GRID driver pre-installed, or download and install the NVIDIA GRID driver following the AWS documentation.

Q: Why am I unable to see the GPU when using Microsoft Remote Desktop?

When using Remote Desktop, GPUs using the WDDM driver model are replaced with a non-accelerated Remote Desktop display driver. In order to access your GPU hardware, you need to utilize a different remote access tool, such as VNC.

Q: What is Amazon EC2 F1?

Amazon EC2 F1 is a compute instance with programmable hardware you can use for application acceleration. The new F1 instance type provides a high performance, easy to access FPGA for developing and deploying custom hardware accelerations.

Q: What are FPGAs and why do I need them?

FPGAs are programmable integrated circuits that you can configure using software. By using FPGAs you can accelerate your applications up to 30x when compared with servers that use CPUs alone. And, FPGAs are reprogrammable, so you get the flexibility to update and optimize your hardware acceleration without having to redesign the hardware.

Q: How does F1 compare with traditional FPGA solutions?

F1 is an AWS instance with programmable hardware for application acceleration. With F1, you have access to FPGA hardware in a few simple clicks, reducing the time and cost of full-cycle FPGA development and scale deployment from months or years to days. While FPGA technology has been available for decades, adoption of application acceleration has struggled to be successful in both the development of accelerators and the business model of selling custom hardware for traditional enterprises, due to time and cost in development infrastructure, hardware design, and at-scale deployment. With this offering, customers avoid the undifferentiated heavy lifting associated with developing FPGAs in on-premises data centers.

Q: What is an Amazon FPGA Image (AFI)?

The design that you create to program your FPGA is called an Amazon FPGA Image (AFI). AWS provides a service to register, manage, copy, query, and delete AFIs. After an AFI is created, it can be loaded on a running F1 instance. You can load multiple AFIs to the same F1 instance, and can switch between AFIs in runtime without reboot. This lets you quickly test and run multiple hardware accelerations in rapid sequence. You can also offer to other customers on the AWS Marketplace a combination of your FPGA acceleration and an AMI with custom software or AFI drivers.

Q: How do I list my hardware acceleration on the AWS Marketplace?

You would develop your AFI and the software drivers/tools to use this AFI. You would then package these software tools/drivers into an Amazon Machine Image (AMI) in an encrypted format. AWS manages all AFIs in the encrypted format you provide to maintain the security of your code. To sell a product in the AWS Marketplace, you or your company must sign up to be an AWS Marketplace reseller, you would then submit your AMI ID and the AFI ID(s) intended to be packaged in a single product. AWS Marketplace will take care of cloning the AMI and AFI(s) to create a product, and associate a product code to these artifacts, such that any end-user subscribing to this product code would have access to this AMI and the AFI(s).

Q: What is available with F1 instances?

For developers, AWS is providing a Hardware Development Kit (HDK) to help accelerate development cycles, a FPGA Developer AMI for development in the cloud, an SDK for AMIs running the F1 instance, and a set of APIs to register, manage, copy, query, and delete AFIs. Both developers and customers have access to the AWS Marketplace where AFIs can be listed and purchased for use in application accelerations.

Q: Do I need to be a FPGA expert to use an F1 instance?

AWS customers subscribing to a F1-optimized AMI from AWS Marketplace do not need to know anything about FPGAs to take advantage of the accelerations provided by the F1 instance and the AWS Marketplace. Simply subscribe to an F1-optimized AMI from the AWS Marketplace with an acceleration that matches the workload. The AMI contains all the software necessary for using the FPGA acceleration. Customers need only write software to the specific API for that accelerator and start using the accelerator.

Q: I'm a FPGA developer, how do I get started with F1 instances?

Developers can get started on the F1 instance by creating an AWS account and downloading the AWS Hardware Development Kit (HDK). The HDK includes documentation on F1, internal FPGA interfaces, and compiler scripts for generating AFI. Developers can start writing their FPGA code to the documented interfaces included in the HDK to create their acceleration function. Developers can launch AWS instances with the FPGA Developer AMI. This AMI includes the development tools needed to compile and simulate the FPGA code. The Developer AMI is best run on the latest C5, M5, or R4 instances. Developers should have experience in the programming languages used for creating FPGA code (i.e. Verilog or VHDL) and an understanding of the operation they wish to accelerate.

Q: I'm not an FPGA developer, how do I get started with F1 instances?

Customers can get started with F1 instances by selecting an accelerator from the AWS Marketplace, provided by AWS Marketplace sellers, and launching an F1 instance with that AMI. The AMI includes all of the software and APIs for that accelerator. AWS manages programming the FPGA with the AFI for that accelerator. Customers do not need any FPGA experience or knowledge to use these accelerators. They can work completely at the software API level for that accelerator.

Q: Does AWS provide a developer kit?

Yes. The Hardware Development Kit (HDK) includes simulation tools and simulation models for developers to simulate, debug, build, and register their acceleration code. The HDK includes code samples, compile scripts, debug interfaces, and many other tools you will need to develop the FPGA code for your F1 instances. You can use the HDK either in an AWS provided AMI, or in your on-premises development environment. These models and scripts are available publically with an AWS account.

Q: Can I use the HDK in my on-premises development environment?

Yes. You can use the Hardware Development Kit HDK either in an AWS-provided AMI, or in your on-premises development environment.

Q: Can I add an FPGA to any EC2 instance type?

No. F1 instances comes in two instance sizes f1.2xlarge, f1.4xlarge, and f1.16 xlarge.

Q: How do I use the Inferentia chip in Inf1 instances?

You can start your workflow by building and training your model in one of the popular ML frameworks such as TensorFlow, PyTorch, or MXNet using GPU instances such as P3 or P3dn. Once the model is trained to your required accuracy, you can use the ML framework's API to invoke Neuron, a software development kit for Inferentia, to compile the model for execution on Inferentia chips, load it in to Inferentia's memory, and then execute inference calls. In order to get started quickly, you can use AWS Deep Learning AMIs that come pre-installed with ML frameworks and the Neuron SDK. For a fully managed experience you will be able to use Amazon SageMaker which will enable you to seamlessly deploy your trained models on Inf1 instances.

Q: When would I use Inf1 vs. C5 vs. G4 instances for inference?

Customers running machine learning models that are sensitive to inference latency and throughput can use Inf1 instances for high-performance cost-effective inference. For those ML models that are less sensitive to inference latency and throughput, customers can use EC2 C5 instances and utilize the AVX-512/VNNI instruction set. For ML models that require access to NVIDIA's CUDA, CuDNN or TensorRT libraries, we recommend using G4 instances.

Model Characteristics and Libraries Used	EC2 Inf1	EC2 C5	EC2 G4
Models that benefit from low latency and high throughput at low cost	X		
Models not sensitive to latency and throughput		X	
Models requiring NVIDIA's developer libraries			X

Q: When should I choose Elastic Inference (EI) for inference vs Amazon EC2 Inf1 instances?

There are two cases where developers would choose EI over Inf1 instances: (1) if you need different CPU and memory sizes than what Inf1 offers, then you can use EI to attach acceleration to the EC2 instance with the right mix of CPU and memory for your application (2) if your performance requirements are significantly lower than what the smallest Inf1 instance provides, then using EI could be a more cost effective choice. For example, if you only need 5 TOPS, enough for processing up to 6 concurrent video streams, then using the smallest slice of EI with a C5.large instance could be up to 50% cheaper than using the smallest size of an Inf1 instance.

Q: What ML models types and operators are supported by EC2 Inf1 instances using the Inferentia chip?

Inferentia chips support the commonly used machine learning models such as single shot detector (SSD) and ResNet for image recognition/classification and Transformer and BERT for natural language processing and translation and many others. A list of supported operators can be found on GitHub.

Q: How do I take advantage of AWS Inferentia's NeuronCore Pipeline capability to lower latency?

Inf1 instances with multiple Inferentia chips, such as Inf1.6xlarge or Inf1.24xlarge, support a fast chip-to-chip interconnect. Using the Neuron Processing Pipeline capability, you can split your model and load it to local cache memory across multiple chips. The Neuron compiler uses ahead-of-time (AOT) compilation technique to analyze the input model and compile it to fit across the on-chip memory of single or multiple Inferentia chips. Doing so enables the Neuron Cores to have high-speed access to models and not require access to off-chip memory, keeping latency bounded while increasing the overall inference throughput.

Q: What is the difference between AWS Neuron and Amazon SageMaker Neo?

AWS Neuron is a specialized SDK for AWS Inferentia chips that optimizes the machine learning inference performance of Inferentia chips. It consists of a compiler, run-time, and profiling tools for AWS Inferentia and is required to run inference workloads on EC2 Inf1 instances. On the other hand, Amazon SageMaker Neo is a hardware agnostic service that consists of a compiler and run-time that enables developers to train machine learning models once, and run them on many different hardware platforms.

Compute Optimized instances

Q. When should I use Compute Optimized instances?

Compute Optimized instances are designed for applications that benefit from high compute power. These applications include compute-intensive applications like high-performance web servers, high-performance computing (HPC), scientific modelling, distributed analytics and machine learning inference.

Q. Can I launch C4 instances as Amazon EBS-optimized instances?

Each C4 instance type is EBS-optimized by default. C4 instances 500 Mbps to 4,000 Mbps to EBS above and beyond the general-purpose network throughput provided to the instance. Since this feature is always enabled on C4 instances, launching a C4 instance explicitly as EBS-optimized will not affect the instance's behavior.

Q. How can I use the processor state control feature available on the c4.8xlarge instance?

The c4.8xlarge instance type provides the ability for an operating system to control processor C-states and P-states. This feature is currently available only on Linux instances. You may want to change C-state or P-state settings to increase processor performance consistency, reduce latency, or tune your instance for a specific workload. By default, Amazon Linux provides the highest-performance configuration that is optimal for most customer workloads; however, if your application would benefit from lower latency at the cost of higher single- or dual-core frequencies, or from lower-frequency sustained performance as opposed to bursty Turbo Boost frequencies, then you should consider experimenting with the C-state or P-state configuration options that are available to these instances. For additional information on this feature, see the Amazon EC2 User Guide section on Processor State Control.

Q. Which instances are available within Compute Optimized instances category?

C5 instances: C5 instances are the latest generation of EC2 Compute Optimized instances. C5 instances are based on Intel Xeon Platinum processors, part of the Intel Xeon Scalable (codenamed Skylake-SP) processor family, and are available in 6 sizes and offer up to 72 vCPUs and 144 GiB memory. C5 instances deliver 25% improvement in price/performance compared to C4 instances.

C4 instances: C4 instances are based on Intel Xeon E5-2666 v3 (codenamed Haswell) processors. C4 instances are available in 5 sizes and offer up to 36 vCPUs and 60 GiB memory.

Q. Should I move my workloads from C3 or C4 instances to C5 instances?

The generational improvement in CPU performance and lower price of C5 instances, which combined result in a 25% price/performance improvement relative to C4 instances, benefit a broad spectrum of workloads that currently run on C3 or C4 instances. For floating point intensive applications, Intel AVX-512 enables significant improvements in delivered TFLOPS by effectively extracting data level parallelism. Customers looking for absolute performance for graphics rendering and HPC workloads that can be accelerated with GPUs or FPGAs should also evaluate other instance families in the Amazon EC2 portfolio that include those resources to find the ideal instance for their workload.

Q. Which operating systems/AMIs are supported on C5 Instances?

EBS backed HVM AMIs with support for ENA networking and booting from NVMe-based storage can be used with C5 instances. The following AMIs are supported on C5:

- Amazon Linux 2014.03 or newer
- Ubuntu 14.04 or newer
- SUSE Linux Enterprise Server 12 or newer
- Red Hat Enterprise Linux 7.4 or newer
- CentOS 7 or newer
- Windows Server 2008 R2
- Windows Server 2012
- Windows Server 2012 R2
- Windows Server 2016
- FreeBSD 11.1-RELEASE

For optimal local NVMe-based SSD storage performance on C5d, Linux kernel version 4.9+ is recommended.

Q. What are the storage options available to C5 customers?

C5 instances use EBS volumes for storage, are EBS-optimized by default, and offer up to 9 Gbps throughput to both encrypted and unencrypted EBS volumes. C5 instances access EBS volumes via PCI attached NVM Express

(NVMe) interfaces. NVMe is an efficient and scalable storage interface commonly used for flash based SSDs such as local NVMe storage provided with I3 and I3en instances. Though the NVMe interface may provide lower latency compared to Xen paravirtualized block devices, when used to access EBS volumes the volume type, size, and provisioned IOPS (if applicable) will determine the overall latency and throughput characteristics of the volume. When NVMe is used to provide EBS volumes, they are attached and detached by PCI hotplug.

Q. What network interface is supported on C5 instances?

C5 instances use the Elastic Network Adapter (ENA) for networking and enable Enhanced Networking by default. With ENA, C5 instances can utilize up to 25 Gbps of network bandwidth.

Q. Which storage interface is supported on C5 instances?

C5 instances will support only NVMe EBS device model. EBS volumes attached to C5 instances will appear as NVMe devices. NVMe is a modern storage interface that provides latency reduction and results in increased disk I/O and throughput.

Q. How many EBS volumes can be attached to C5 instances?

C5 instances support a maximum for 27 EBS volumes for all Operating systems. The limit is shared with ENI attachments which can be found here http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-eni.html. For example: since every instance has at least 1 ENI, if you have 3 additional ENI attachments on the c4.2xlarge, you can attach 24 EBS volumes to that instance.

Q. What is the underlying hypervisor on C5 instances?

C5 instances use a new EC2 hypervisor that is based on core KVM technology.

Q: Why does the total memory reported by the operating system not match the advertised memory of the C5 instance type?

In C5, portions of the total memory for an instance are reserved from use by the Operating System including areas used by the virtual BIOS for things like ACPI tables and for devices like the virtual video RAM.

General Purpose instances

Q: What are Amazon EC2 M6g instances?

Amazon EC2 M6g instances are the next-generation of general purpose instances powered by Arm-based AWS Graviton2 Processors. M6g instances deliver up to 40% better price/performance over current generation M5 instances. They are built on the AWS Nitro System (https://aws.amazon.com/ec2/nitro/), a combination of dedicated hardware and Nitro hypervisor. M6g instances are currently available in preview and will be generally available in the coming months.

Q: What are the specifications of the new AWS Graviton2 Processors?

The AWS Graviton2 processors deliver up to 7x performance, 4x the number of compute cores, 2x larger caches, 5x faster memory, and 50% faster per core encryption performance than first generation AWS Graviton processors. Each core of the AWS Graviton2 processor is a single-threaded vCPU. These processors also offer always-on fully encrypted DRAM memory, hardware acceleration for compression workloads, dedicated engines per vCPU that double the floating point performance for workloads such as video encoding, and instructions for int8/fp16 CPU-based machine learning inference acceleration. The CPUs are built utilizing 64-bit Arm Neoverse cores and custom silicon designed by AWS on the advanced 7 nm manufacturing technology.

Q: Is memory encryption supported by AWS Graviton2 processors?

AWS Graviton2 processors support always-on 256-bit memory encryption to further enhance security. Encryption keys are securely generated within the host system, do not leave the host system, and are irrecoverably destroyed when the host is rebooted or powered down. Memory encryption does not support integration with AWS KMS system and customers cannot bring their own keys for the same.

Q: What are some of the ideal use cases for M6g instances?

M6g instances deliver significant performance and price performance benefits for a broad spectrum of general purpose workloads such as application servers, gaming servers, microservices, mid-size databases, and caching fleets. Customers deploying applications built on open source software across the M instance family will find the M6g instances an appealing option to realize the best price-performance within the instance family. Arm developers can also build their applications directly on native Arm hardware as opposed to cross-compilation or emulation.

Q: What are the various storage options available on M6g instances?

M6g instances are EBS-optimized by default and offer up to 18,000 Mbps of dedicated EBS bandwidth to both encrypted and unencrypted EBS volumes. M6g instances only support Non-Volatile Memory Express (NVMe) interface to access EBS storage volumes. Additionally, options with local NVMe instance storage will also be available through the M6gd instance types.

Q: Which network interface is supported on M6g instances?

M6g instances support ENA based Enhanced Networking. With ENA, M6g instances can deliver up to 25 Gbps of network bandwidth between instances when launched within a Placement Group.

Q: Will customers need to modify their applications and workloads to be able to run on the M6g instances?

The changes required are dependent on the application. Customers running applications built on open source software will find that the Arm ecosystem is well developed and already likely supports their applications. Most Linux distributions as well as containers (Docker, Kubernetes, Amazon ECS, Amazon EKS) support the Arm architecture. Customers will find Arm versions of commonly used software packages available for installation through the same mechanisms that they currently use. While some applications may require re-compilation, applications that are based on interpreted languages (such as Java, Node, Python, Go) not reliant on native CPU instruction sets should run with minimal to no changes.

Q: Will there be more compute choices offered with the M6 instance families?

Yes, Intel and AMD CPU powered instances will also be offered in the future as part of the M6 instance families.

Q: How can I sign up for the M6g instance preview?

Please register your interest for the preview using the application form. You can also reach out to your account representative.

Q: Is there a charge for the M6g preview access?

Yes, please refer to the pricing information on the EC2 pricing pages.

Q: What are Amazon EC2 A1 instances?

Amazon EC2 A1 instances are general purpose instances powered by the first-generation AWS Graviton Processors that are custom designed by AWS.

Q: What are the specifications of the first-generation AWS Graviton Processors?

AWS Graviton processors are custom designed by AWS utilizing Amazon's extensive expertise in building platform solutions for cloud applications running at scale. These processors are based on the 64-bit Arm instruction set and feature Arm Neoverse cores as well as custom silicon designed by AWS. The cores operate at a frequency of 2.3 GHz.

Q: When should I use A1 instances?

A1 instances deliver significant cost savings for scale-out workloads that can fit within the available memory footprint. A1 instances are ideal for scale-out applications such as web servers, containerized microservices, and data/log processing. These instances will also appeal to developers, enthusiasts, and educators across the Arm developer community.

Q: Will customers have to modify applications and workloads to be able to run on the A1 instances?

The changes required are dependent on the application. Applications based on interpreted or run-time compiled languages (e.g. Python, Java, PHP, Node.js) should run without modifications. Other applications may need to be recompiled and those that don't rely on x86 instructions will generally build with minimal to no changes.

Q: Which operating systems/AMIs are supported on A1 Instances?

The following AMIs are supported on A1 instances: Amazon Linux 2, Ubuntu 16.04.4 or newer, Red Hat Enterprise Linux (RHEL) 7.6 or newer, SUSE Linux Enterprise Server 15 or newer. Additional AMI support for Fedora, Debian, NGINX Plus are also available through community AMIs and the AWS Marketplace. . EBS backed HVM AMIs launched on A1 instances require NVMe and ENA drivers installed at instance launch.

Q: Are there specific AMI requirements to run on M6g and A1 instances?

You will need to use the “arm64” AMIs with the M6g and A1 instances. x86 AMIs are not compatible with M6g and A1 instances.

Q: When should customers use A1 instances versus the new M6g instances?

A1 instances continue to offer significant cost benefits for scale-out workloads that can run on multiple smaller cores and fit within the available memory footprint. The new M6g instances are a good fit for a broad spectrum of applications that require more compute, memory, networking resources and/or can benefit from scaling up across platform capabilities. M6g instances will deliver the best price-performance within the instance family for these applications. M6g supports up to 16xlarge instance size (A1 supports up to 4xlarge), 4GB of memory per vCPU (A1 supports 2GB memory per vCPU), and up to 25 Gbps of networking bandwidth (A1 supports up to 10 Gbps).

Q: What are the various storage options available to A1 customers?

A1 instances are EBS-optimized by default and offer up to 3,500 Mbps of dedicated EBS bandwidth to both encrypted and unencrypted EBS volumes. A1 instances only support Non-Volatile Memory Express (NVMe) interface to access EBS storage volumes. A1 instances will not support the blkfront interface.

Q: Which network interface is supported on A1 instances?

A1 instances support ENA based Enhanced Networking. With ENA, A1 instances can deliver up to 10 Gbps of network bandwidth between instances when launched within a Placement Group.

Q: Do A1 instances support the AWS Nitro System?

Yes, A1 instances are powered by the AWS Nitro System, a combination of dedicated hardware and Nitro hypervisor.

Q: Why does the total memory reported by Linux not match the advertised memory of the A1 instance type?

In A1 instances, portions of the total memory for an instance are reserved from use by the operating system including areas used by the virtual UEFI for things like ACPI tables.

Q: What are the key use cases for Amazon EC2 M5 Instances?

M5 instances offer a good choice for running development and test environments, web, mobile and gaming applications, analytics applications, and business critical applications including ERP, HR, CRM, and collaboration apps. Customers who are interested in running their data intensive workloads (e.g. HPC, or SOLR clusters) on instances with a higher memory footprint will also find M5 to be a good fit. Workloads that heavily use single and double precision floating point operations and vector processing such as video processing workloads and need higher memory can benefit substantially from the AVX-512 instructions that M5 supports.

Q: Why should customers choose EC2 M5 Instances over EC2 M4 Instances?

Compared with EC2 M4 Instances, the new EC2 M5 Instances deliver customers greater compute and storage performance, larger instance sizes for less cost, consistency and security. The biggest benefit of EC2 M5 Instances is based on its usage of the latest generation of Intel Xeon Scalable processors (aka Skylake), which deliver up to 20% improvement in price/performance compared to M4. With AVX-512 support in M5 vs. the older AVX2 in M4, customers will gain 2x higher performance in workloads requiring floating point operations. M5 instances offer up to 25 Gbps of network bandwidth and up to 10 Gbps of dedicated bandwidth to Amazon EBS. M5 instances also feature significantly higher networking and Amazon EBS performance on smaller instance sizes with EBS burst capability.

Q: How does support for Intel AVX-512 benefit EC2 M5 and M5d Instance customers?

Intel Advanced Vector Extension 512 (AVX-512) is a set of new CPU instructions available on the latest Intel Xeon Scalable processor family, that can accelerate performance for workloads and usages such as scientific simulations, financial analytics, artificial intelligence, machine learning/deep learning, 3D modeling and analysis, image and video processing, cryptography and data compression, among others. Intel AVX-512 offers exceptional processing of encryption algorithms, helping to reduce the performance overhead for cryptography, which means EC2 M5 and M5d customers can deploy more secure data and services into distributed environments without compromising performance.

Q: What are the various processor options available to M5 customers?

The M5 and M5d instance types use a 3.1 GHz Intel Xeon Platinum 8000 series processor. The M5a and M5ad instance types use a 2.5 GHz AMD EPYC 7000 series processor.

Q: What are the various storage options available to M5 customers?

The M5 and M5a instance types leverage EBS volumes for storage. The M5d and M5ad instance types support up to 3.6TB (4 x 900GB) of local NVMe storage.

Q: When should I use the different M5 instance types?

Customers should consider using the M5a and M5ad instance types if they are looking to save money on price when their workloads do not fully utilize the compute resources of their chosen instance, resulting in them paying for performance that they don’t actually need. For workloads that require the highest processor performance or high floating-point performance capabilities, including vectorized computing with AVX-512 instructions, then we suggest you use the M5 or M5d instance types.

Q: Which network interface is supported on M5 instances?

M5, M5a, M5d, and M5ad instances support only ENA based Enhanced Networking and will not support netback. With ENA, M5 and M5d instances can deliver up to 25 Gbps of network bandwidth between instances and the M5a and M5ad instance types can support up to 20Gbps of network bandwidth between instances.

Q. Which operating systems/AMIs are supported on M5 Instances?

EBS backed HVM AMIs with support for ENA networking and booting from NVMe-based storage can be used with M5 instances. The following AMIs are supported on M5, M5a, M5ad, and M5d:

- Amazon Linux 2014.03 or newer
- Ubuntu 14.04 or newer
- SUSE Linux Enterprise Server 12 or newer
- Red Hat Enterprise Linux 7.4 or newer
- CentOS 7 or newer
- Windows Server 2008 R2
- Windows Server 2012
- Windows Server 2012 R2
- Windows Server 2016
- FreeBSD 11.1-RELEASE

For optimal local NVMe-based SSD storage performance on M5d, Linux kernel version 4.9+ is recommended.

Q. What interface connects EBS storage to my M5 instances?

M5, M5a, M5ad, and M5d instances use EBS volumes for storage, are EBS-optimized by default, and offer up to 10 Gbps throughput to both encrypted and unencrypted EBS volumes. M5 instances access EBS volumes via PCI attached NVM Express (NVMe) interfaces. NVMe is an efficient and scalable storage interface commonly used for flash based SSDs such as local NVMe storage provided with I3 and I3en instances. Though the NVMe interface may provide lower latency compared to Xen paravirtualized block devices, when used to access EBS volumes the volume type, size, and provisioned IOPS (if applicable) will determine the overall latency and throughput characteristics of the volume. When NVMe is used to provide EBS volumes, they are attached and detached by PCI hotplug.

Q. How many EBS volumes can be attached to M5 instances?

M5 and M5a instances support a maximum for 27 EBS volumes for all Operating systems. The limit is shared with ENI attachments which can be found here http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-eni.html. For example: since every instance has at least 1 ENI, if you have 3 additional ENI attachments on the m5.2xlarge, you can attach 24 EBS volumes to that instance.

Q. What is the underlying hypervisor on M5 instances?

M5, M5a, M5ad, and M5d instances use a new lightweight Nitro Hypervisor that is based on core KVM technology.

Q: Why does the total memory reported by the operating system not match the advertised memory of the M5 instance type?

In M5, M5a, M5ad, and M5d, portions of the total memory for an instance are reserved from use by the operating system including areas used by the virtual BIOS for things like ACPI tables and for devices like the virtual video RAM.

Q: How are Burstable Performance Instances different?

Amazon EC2 allows you to choose between Fixed Performance Instances (e.g. C, M and R instance families) and Burstable Performance Instances (e.g. T2). Burstable Performance Instances provide a baseline level of CPU performance with the ability to burst above the baseline.

T2 instances' baseline performance and ability to burst are governed by CPU Credits. Each T2 instance receives CPU Credits continuously, the rate of which depends on the instance size. T2 instances accrue CPU Credits when they are idle, and consume CPU credits when they are active. A CPU Credit provides the performance of a full CPU core for one minute.

Model	vCPUs	CPU Credits / hour	Maximum CPU Credit Balance	Baseline CPU Performance
t2.nano	1	3	72	5% of a core
t2.micro	1	6	144	10% of a core
t2.small	1	12	288	20% of a core
t2.medium	2	24	576	40% of a core*
t2.large	2	36	864	60% of a core**
t2.xlarge	4	54	1,296	90% of a core***
t2.2xlarge	8	81	1,944	135% of a core****

** For the t2.medium, single threaded applications can use 40% of 1 core, or if needed, multithreaded applications can use 20% each of 2 cores.*

***For the t2.large, single threaded applications can use 60% of 1 core, or if needed, multithreaded applications can use 30% each of 2 cores.*

**** For the t2.xlarge, single threaded applications can use 90% of 1 core, or if needed, multithreaded applications can use 45% each of 2 cores or 22.5% of all 4 cores.*

***** For the t2.large, single threaded applications can use all of 1 core, or if needed, multithreaded applications can use 67.5% each of 2 cores or 16.875% of all 8 cores.*

Q. How do I choose the right Amazon Machine Image (AMI) for my T2 instances?

You will want to verify that the minimum memory requirements of your operating system and applications are within the memory allocated for each T2 instance size (e.g. 512 MiB for t2.nano). Operating systems with Graphical User Interfaces (GUI) that consume significant memory and CPU, for example Microsoft Windows, might need a t2.micro or larger instance size for many use cases. You can find AMIs suitable for the t2.nano instance types on AWS Marketplace. Windows customers who do not need the GUI can use the Microsoft Windows Server 2012 R2 Core AMI.

Q: When should I choose a Burstable Performance Instance, such as T2?

T2 instances provide a cost-effective platform for a broad range of general purpose production workloads. T2 Unlimited instances can sustain high CPU performance for as long as required. If your workloads consistently require CPU usage much higher than the baseline, consider a dedicated CPU instance family such as the M or C.

Q: How can I see the CPU Credit balance for each T2 instance?

You can see the CPU Credit balance for each T2 instance in EC2 per-Instance metrics in Amazon CloudWatch. T2 instances have four metrics, CPUCreditUsage, CPUCreditBalance, CPUSurplusCreditBalance and CPUSurplusCreditsCharged. CPUCreditUsage indicates the amount of CPU Credits used. CPUCreditBalance indicates the balance of CPU Credits. CPUSurplusCredit Balance indicates credits used for bursting in the absence of earned credits. CPUSurplusCreditsCharged indicates credits that are charged when average usage exceeds the baseline.

Q: What happens to CPU performance if my T2 instance is running low on credits (CPU Credit balance is near zero)?

If your T2 instance has a zero CPU Credit balance, performance will remain at baseline CPU performance. For example, the t2.micro provides baseline CPU performance of 10% of a physical CPU core. If your instance's CPU Credit balance is approaching zero, CPU performance will be lowered to baseline performance over a 15-minute interval.

Q: Does my T2 instance credit balance persist at stop / start?

No, a stopped instance does not retain its previously earned credit balance.

Q: Can T2 instances be purchased as Reserved Instances or Spot Instances?

T2 instances can be purchased as On-Demand Instances, Reserved Instances or Spot Instances.

High Memory instances

Q. What are EC2 High Memory instances?

Amazon EC2 High Memory instances offer 6 TB, 9 TB, 12 TB, 18 TB, or 24 TB of memory in a single instance. These instances are designed to run large in-memory databases, including production installations of SAP HANA, in the cloud.

EC2 High Memory instances with 6 TB, 9 TB, and 12 TB are powered by an 8-socket platform with Intel® Xeon® Platinum 8176M (Skylake) processors. EC2 High Memory instances with 18 TB and 24 TB are the first Amazon EC2 instances powered by an 8-socket platform with 2nd Generation Intel® Xeon® Scalable (Cascade Lake) processors that are optimized for mission-critical enterprise workloads. EC2 High Memory instances deliver high networking throughput and low-latency with up to 100 Gbps of aggregate network bandwidth using Amazon Elastic Network Adapter (ENA)-based Enhanced Networking. EC2 High Memory instances are EBS-Optimized by default, and support encrypted and unencrypted EBS volumes.

Q. Are High Memory instances certified by SAP to run SAP HANA workloads?

High Memory instances are certified by SAP for running Business Suite on HANA, the next-generation Business Suite S/4HANA, Data Mart Solutions on HANA, Business Warehouse on HANA, and SAP BW/4HANA in production environments. For details, see SAP's Certified and Supported SAP HANA Hardware Directory.

Q. Which instances are available within High Memory instance category?

Five High Memory instances are available. u-6tb1.metal offers 6 TB memory; u-9tb1.metal offers 9 TB memory; u-12tb1.metal offers 12 TB memory; u-18tb1.metal offers 18 TB memory; and u-24tb1.metal offers 24 TB memory. Each High Memory instance offers 448 logical processors, where each logical processor is a hyperthread on the 8-socket platform with total of 224 CPU cores.

Q. What are the storage options available with High Memory instances?

High Memory instances support Amazon EBS volumes for storage. High Memory instances are EBS-optimized by default, and offer up to 28 Gbps of storage bandwidth to both encrypted and unencrypted EBS volumes.

Q. Which storage interface is supported on High Memory instances?

High Memory instances access EBS volumes via PCI attached NVM Express (NVMe) interfaces. EBS volumes attached to High Memory instances appear as NVMe devices. NVMe is an efficient and scalable storage interface, which is commonly used for flash based SSDs and provides latency reduction and results in increased disk I/O and throughput. The EBS volumes are attached and detached by PCI hotplug.

Q. What network performance is supported on High Memory instances?

High Memory instances use the Elastic Network Adapter (ENA) for networking and enable Enhanced Networking by default. With ENA, High Memory instances can utilize up to 100 Gbps of network bandwidth.

Q. Can I run High Memory instances in my existing Amazon Virtual Private Cloud (VPC)?

You can run High Memory instances in your existing and new Amazon VPCs.

Q. What is the underlying hypervisor on High Memory instances?

High Memory instances are EC2 bare metal instances built on the AWS Nitro System, a rich collection of building blocks that offloads many of the traditional virtualization functions to dedicated hardware. These instances do not run on a hypervisor and allow the operating systems to run directly on the underlying hardware, while still providing access to the benefits of the cloud.

Q. Do High Memory instances enable CPU power management state control?

Yes. You can configure C-states and P-states on High Memory instances. You can use C-states to enable higher turbo frequencies (as much as 4.0 GHz). You can also use P-states to lower performance variability by pinning all cores at P1 or higher P states, which is similar to disabling Turbo, and running consistently at the base CPU clock speed.

Q. What purchase options are available for High Memory instances?

High Memory instances are available on EC2 Dedicated Hosts on a 3-year Reservation. After the 3-year reservation expires, you can continue using the host at an hourly rate or release it anytime.

Q. What is the lifecycle of a Dedicated Host?

Once a Dedicated Host is allocated within your account, it will be standing by for your use. You can then launch an instance with a tenancy of "host" using the RunInstances API, and can also stop/start/terminate the instance through the API. You can use the AWS Management Console to manage the Dedicated Host and the instance. The Dedicated Host will be allocated to your account for the period of 3-year reservation. After the 3-year reservation expires, you can continue using the host or release it anytime.

Q. Can I launch, stop/start, and terminate High Memory instances using AWS CLI/SDK?

You can launch, stop/start, and terminate instances on your EC2 Dedicated Hosts using AWS CLI/SDK.

Q. Which AMIs are supported with High memory instances?

EBS-backed HVM AMIs with support for ENA networking can be used with High Memory instances. The latest Amazon Linux, Red Hat Enterprise Linux, SUSE Enterprise Linux Server, and Windows Server AMIs are supported. Operating system support for SAP HANA workloads on High Memory instances include: SUSE Linux Enterprise Server 12 SP3 for SAP, Red Hat Enterprise Linux 7.4 for SAP, Red Hat Enterprise Linux 7.5 for SAP, SUSE Linux Enterprise Server 12 SP4 for SAP, SUSE Linux Enterprise Server 15 for SAP, Red Had Enterprise Linux 7.6 for SAP. Refer to SAP's Certified and Supported SAP HANA Hardware Directory for latest detail on supported operating systems.

Q. Are there standard SAP HANA reference deployment frameworks available for the High Memory instance and the AWS Cloud?

You can use the AWS Quick Start reference SAP HANA deployments to rapidly deploy all the necessary SAP HANA building blocks on High Memory instances following SAP's recommendations for high performance and reliability. AWS Quick Starts are modular and customizable, so you can layer additional functionality on top or modify them for your own implementations.

Previous Generation instances

Q: Why don't I see M1, C1, CC2 and HS1 instances on the pricing pages any more?

These have been moved to the Previous Generation Instance page.

Q: Are these Previous Generation instances still being supported?

Yes. Previous Generation instances are still fully supported.

Q: Can I still use/add more Previous Generation instances?

Yes. Previous Generation instances are still available as On-Demand, Reserved Instances, and Spot Instance, from our APIs, CLI and EC2 Management Console interface.

Q: Are my Previous Generation instances going to be deleted?

No. Your C1, C3, CC2, CR1, G2, HS1, M1, M2, M3, R3 and T1 instances are still fully functional and will not be deleted because of this change.

Q: Are Previous Generation instances being discontinued soon?

Currently, there are no plans to end of life Previous Generation instances. However, with any rapidly evolving technology the latest generation will typically provide the best performance for the price and we encourage our customers to take advantage of technological advancements.

Q: Will my Previous Generation instances I purchased as a Reserved Instance be affected or changed?

No. Your Reserved Instances will not change, and the Previous Generation instances are not going away.

Memory Optimized instances

Q. When should I use Memory-optimized instances?

Memory-optimized instances offer large memory size for memory intensive applications including in-memory applications, in-memory databases, in-memory analytics solutions, High Performance Computing (HPC), scientific computing, and other memory-intensive applications.

Q. When should I use X1 instances?

X1 instances are ideal for running in-memory databases like SAP HANA, big data processing engines like Apache Spark or Presto, and high performance computing (HPC) applications. X1 instances are certified by SAP to run production environments of the next-generation Business Suite S/4HANA, Business Suite on HANA (SoH), Business Warehouse on HANA (BW), and Data Mart Solutions on HANA on the AWS cloud.

Q. When should I use X1e instances?

X1e instances are ideal for running in-memory databases like SAP HANA, high-performance databases and other memory optimized enterprise applications. X1e instances offer twice the memory per vCPU compared to the X1 instances. The x1e.32xlarge instance is certified by SAP to run production environments of the next-generation Business Suite S/4HANA, Business Suite on HANA (SoH), Business Warehouse on HANA (BW), and Data Mart Solutions on HANA on the AWS Cloud.

Q. How do X1 and X1e instances differ?

X1e instances offer 32GB of memory per vCPU whereas X1 instances offer 16GB of memory per vCPU. X1e instance sizes enable six instance configurations starting from 4 vCPUs and 122 GiB memory up to 128 vCPUs and 3,904 GiB of memory. X1 instances enable two instance configurations, 64 vCPUs with 976 GiB memory and 128 vCPUs with 1,952 GiB memory.

Q. What are the key specifications of Intel E7 (codenamed Haswell) processors that power X1 and X1e instances?

The E7 processors have a high core count to support workloads that scale efficiently on large number of cores. The Intel E7 processors also feature high memory bandwidth and larger L3 caches to boost the performance of in-memory applications. In addition, the Intel E7 processor:

- Enables increased cryptographic performance via the latest Intel AES-NI feature.
- Supports Transactional Synchronization Extensions (TSX) to boost the performance of in-memory transactional data processing.
- Supports Advanced Vector Extensions 2 (Intel AVX2) processor instructions to expand most integer commands to 256 bits.

Q. Do X1 and X1e instances enable CPU power management state control

Yes. You can configure C-states and P-states on x1e.32xlarge, x1e.16xlarge, x1e.8xlarge, x1.32xlarge and x1.16xlarge instances. You can use C-states to enable higher turbo frequencies (as much as 3.1 GHz with one or two core turbo). You can also use P-states to lower performance variability by pinning all cores at P1 or higher P states, which is similar to disabling Turbo, and running consistently at the base CPU clock speed.

Q: What operating systems are supported on X1 and X1e instances?

X1 and X1e instances provide high number of vCPUs, which might cause launch issues in some Linux operating systems that have a lower vCPU limit. We strongly recommend that you use the latest AMIs when you launch these instances.

AMI support for SAP HANA workloads include: SUSE Linux 12, SUSE Linux 12 SP1, SLES for SAP 12 SP1, SLES for SAP 12 SP2, and RHEL 7.2 for SAP HANA.

x1e.32xlarge will also support Windows Server 2012 R2 and 2012 RTM. x1e.xlarge, x1e.2xlarge, x1e.4xlarge, x1e.8xlarge, x1e.16xlarge and x1.32xlarge will also support Windows Server 2012 R2, 2012 RTM and 2008 R2 64bit (Windows Server 2008 SP2 and older versions will not be supported) and x1.16xlarge will support Windows Server 2012 R2, 2012 RTM, 2008 R2 64bit, 2008 SP2 64bit, and 2003 R2 64bit (Windows Server 32bit versions will not be supported).

Q. What storage options are available for X1 customers?

X1 instances offer SSD based instance store, which is ideal for temporary storage of information such as logs, buffers, caches, temporary tables, temporary computational data, and other temporary content. X1 instance store provides the best I/O performance when you use a Linux kernel that supports persistent grants, an extension to the Xen block ring protocol.

X1 instances are EBS-optimized by default and offer up to 14 Gbps of dedicated bandwidth to EBS volumes. EBS offers multiple volume types to support a wide variety of workloads. For more information see the EC2 User Guide.

Q. How do I build cost-effective failover solution on X1 and X1e instances?

You can design simple and cost-effective failover solutions on X1 instances using Amazon EC2 Auto Recovery, an Amazon EC2 feature that is designed to better manage failover upon instance impairment. You can enable Auto Recovery for X1 instances by creating an AWS CloudWatch alarm. Choose the

“EC2 Status Check Failed (System)” metric and select the “Recover this instance” action. Instance recovery is subject to underlying limitations, including those reflected in the Instance Recovery Troubleshooting documentation. For more information visit Auto Recovery documentation and Creating Amazon CloudWatch Alarms respectively.

Q. Are there standard SAP HANA reference deployment frameworks available for the X1 instance and the AWS Cloud?

You can use the AWS Quick Start reference HANA deployments to rapidly deploy all the necessary HANA building blocks on X1 instances following SAP’s recommendations for high performance and reliability. AWS Quick Starts are modular and customizable, so you can layer additional functionality on top or modify them for your own implementations. For additional information on deploying HANA on AWS, please refer to SAP HANA on AWS Cloud: Quick Start Reference Deployment Guide.

Storage Optimized instances

Q. What is a Dense-storage Instance?

Dense-storage instances are designed for workloads that require high sequential read and write access to very large data sets, such as Hadoop distributed computing, massively parallel processing data warehousing, and log processing applications. The Dense-storage instances offer the best price/GB-storage and price/disk-throughput across other EC2 instances.

Q. How do Dense-storage and HDD-storage instances compare to High I/O instances?

High I/O instances (I2) are targeted at workloads that demand low latency and high random I/O in addition to moderate storage density and provide the best price/IOPS across other EC2 instance types. Dense-storage instances (D2) and HDD-storage instances (H1) are optimized for applications that require high sequential read/write access and low cost storage for very large data sets and provide the best price/GB-storage and price/disk-throughput across other EC2 instances.

Q. How much disk throughput can Dense-storage and HDD-storage instances deliver?

The largest current generation of Dense-storage instances, d2.8xlarge, can deliver up to 3.5 GBps read and 3.1 GBps write disk throughput with a 2 MiB block size. The largest H1 instances size, h1.16xlarge, can deliver up to 1.15 GBps read and write. To ensure the best disk throughput performance from your D2 instances on Linux, we recommend that you use the most recent version of the Amazon Linux AMI, or another Linux AMI with a kernel version of 3.8 or later that supports persistent grants - an extension to the Xen block ring protocol that significantly improves disk throughput and scalability.

Q. Do Dense-storage and HDD-storage instances provide any failover mechanisms or redundancy?

The primary data storage for Dense-storage instances is HDD-based instance storage. Like all instance storage, these storage volumes persist only for the life of the instance. Hence, we recommend that you build a degree of redundancy (e.g. RAID 1/5/6) or use file systems (e.g. HDFS and MapR-FS) that support redundancy and fault tolerance. You can also back up data periodically to more durable data storage solutions such as Amazon Simple Storage Service (S3) for additional data durability. Please refer to Amazon S3 for reference.

Q. How do Dense-storage and HDD-storage instances differ from Amazon EBS?

Amazon EBS offers simple, elastic, reliable (replicated), and persistent block level storage for Amazon EC2 while abstracting the details of the underlying storage media in use. Amazon EC2 instance storage provides directly attached non-persistent, high performance storage building blocks that can be used for a variety of storage applications. Dense-storage instances are specifically targeted at customers who want high sequential read/write access to large data sets on local storage, e.g. for Hadoop distributed computing and massively parallel processing data warehousing.

Q. Can I launch H1 instances as Amazon EBS-optimized instances?

Each H1 instance type is EBS-optimized by default. H1 instances offer 1,750 Mbps to 14,000 Mbps to EBS above and beyond the general-purpose network throughput provided to the instance. Since this feature is always enabled on H1 instances, launching a H1 instance explicitly as EBS-optimized will not affect the instance's behavior.

Q. Can I launch D2 instances as Amazon EBS-optimized instances?

Each D2 instance type is EBS-optimized by default. D2 instances 500 Mbps to 4,000 Mbps to EBS above and beyond the general-purpose network throughput provided to the instance. Since this feature is always enabled on D2 instances, launching a D2 instance explicitly as EBS-optimized will not affect the instance's behavior.

Q. Are HDD-storage instances offered in EC2 Classic?

The current generation of HDD-storage instances (H1 instances) can only be launched in Amazon VPC. With Amazon VPC, you can leverage a number of features that are available only on the Amazon VPC platform – such as enabling enhanced networking, assigning multiple private IP addresses to your instances, or changing your instances' security groups. For more information about the benefits of using a VPC, see Amazon EC2 and Amazon Virtual Private Cloud (Amazon VPC).

Q. Are Dense-storage instances offered in EC2 Classic?

The current generation of Dense-storage instances (D2 instances) can be launched in both EC2-Classic and Amazon VPC. However, by launching a Dense-storage instance into a VPC, you can leverage a number of features that are available only on the Amazon VPC platform – such as enabling enhanced networking, assigning multiple private IP addresses to your instances, or changing your instances' security groups. For more information about the benefits of using a VPC, see Amazon EC2 and Amazon Virtual Private Cloud (Amazon VPC). You can take steps to migrate your resources from EC2-Classic to Amazon VPC. For more information, see Migrating a Linux Instance from EC2-Classic to a VPC.

Q. What is a High I/O instance?

High I/O instances use NVMe based local instance storage to deliver very high, low latency, I/O capacity to applications, and are optimized for applications that require millions of IOPS. Like Cluster instances, High I/O instances can be clustered via cluster placement groups for low latency networking.

Q. Are all features of Amazon EC2 available for High I/O instances?

High I/O instances support all Amazon EC2 features. I3 and I3en instances offer NVMe only storage, while previous generation I2 instances allow legacy blkfront storage access. Currently you can only purchase High I/O instances as On-Demand, Reserved Instances or as Spot instances.

Q. Is there a limit on the number of High I/O instances I can use?

Currently, you can launch 2 i3.16xlarge instances by default. If you wish to run more than 2 On-Demand instances, please complete the Amazon EC2 instance request form.

Q. How many IOPS can i3.16.xlarge instances deliver?

Using HVM AMIs, High I/O I3 instances can deliver up to 3.3 million IOPS measured at 100% random reads using 4KB block size, and up to 300,000 100% random write IOPs, measured at 4KB block sizes to applications across 8 x 1.9 TB NVMe devices.

Q. What is the sequential throughput of i3 instances?

The maximum sequential throughput, measured at 128K block sizes is 16 GB/s read throughput and 6.4 GB/s write throughput.

Q. AWS has other database and Big Data offerings. When or why should I use High I/O instances?

High I/O instances are ideal for applications that require access to millions of low latency IOPS, and can leverage data stores and architectures that manage data redundancy and availability. Example applications are:

- NoSQL databases like Cassandra and MongoDB
- In-memory databases like Aerospike

- Elasticsearch and analytics workloads
- OLTP systems

Q. Do High I/O instances provide any failover mechanisms or redundancy?

Like other Amazon EC2 instance types, instance storage on I3 and I3en instances persists during the life of the instance. Customers are expected to build resilience into their applications. We recommend using databases and file systems that support redundancy and fault tolerance. Customers should back up data periodically to Amazon S3 for improved data durability.

Q. Do High I/O instances support TRIM?

The TRIM command allows the operating system to inform SSDs which blocks of data are no longer considered in use and can be wiped internally. In the absence of TRIM, future write operations to the involved blocks can slow down significantly. I3 and I3en instances support TRIM.

Q. How many IOPS can I3en.24xlarge instances deliver?

Using HVM AMIs, high I/O I3en instances can deliver up to 2 million IOPS measured at 100% random reads using 4KB block sizes, and up to 1.6 million 100% random write IOPs, measured at 4KB block sizes to applications across 8 x 7.5 TB NVMe devices.

Q. What is the sequential throughput of I3en instances?

The maximum sequential throughput, measured at 128K block sizes is 16 GB/s read throughput and 8 GB/s write throughput.

Storage

Amazon Elastic Block Store (EBS) | Amazon Elastic File System (EFS) | NVMe Instance storage

Amazon Elastic Block Store (EBS)

Q: What happens to my data when a system terminates?

The data stored on a local instance store will persist only as long as that instance is alive. However, data that is stored on an Amazon EBS volume will persist independently of the life of the instance. Therefore, we recommend that you use the local instance store for temporary data and, for data requiring a higher level of durability, we recommend using Amazon EBS volumes or backing up the data to Amazon S3. If you are using an Amazon EBS volume as a root partition, you will need to set the Delete On Terminate flag to "N" if you want your Amazon EBS volume to persist outside the life of the instance.

Q: What kind of performance can I expect from Amazon EBS volumes?

Amazon EBS provides four current generation volume types and are divided into two major categories: SSD-backed storage for transactional workloads and HDD-backed storage for throughput intensive workloads. These volume types differ in performance characteristics and price, allowing you to tailor your storage performance and cost to the needs of your applications. For more information on see the EBS product details page, and for additional information on performance, see the Amazon EC2 User Guide's EBS Performance section.

Q: What are Throughput Optimized HDD (st1) and Cold HDD (sc1) volume types?

ST1 volumes are backed by hard disk drives (HDDs) and are ideal for frequently accessed, throughput intensive workloads with large datasets and large I/O sizes, such as MapReduce, Kafka, log processing, data warehouse, and ETL workloads. These volumes deliver performance in terms of throughput, measured in MB/s, and include the ability to burst up to 250 MB/s per TB, with a baseline throughput of 40 MB/s per TB and a maximum throughput of 500 MB/s per volume. ST1 is designed to deliver the expected throughput performance 99% of the time and has enough I/O credits to support a full-volume scan at the burst rate.

SC1 volumes are backed by hard disk drives (HDDs) and provides the lowest cost per GB of all EBS volume types. It is ideal for less frequently accessed workloads with large, cold datasets. Similar to st1, sc1 provides a burst model: these volumes can burst up to 80 MB/s per TB, with a baseline throughput of 12 MB/s per TB and a maximum throughput of 250 MB/s per volume. For infrequently accessed data, sc1 provides extremely inexpensive storage. SC1 is designed to deliver the expected throughput performance 99% of the time and has enough I/O credits to support a full-volume scan at the burst rate.

To maximize the performance of st1 and sc1, we recommend using EBS-optimized EC2 instances.

Q: Which volume type should I choose?

Amazon EBS includes two major categories of storage: SSD-backed storage for transactional workloads (performance depends primarily on IOPS) and HDD-backed storage for throughput workloads (performance depends primarily on throughput, measured in MB/s). SSD-backed volumes are designed for transactional, IOPS-intensive database workloads, boot volumes, and workloads that require high IOPS. SSD-backed volumes include Provisioned IOPS SSD (io1) and General Purpose SSD (gp2). HDD-backed volumes are designed for throughput-intensive and big-data workloads, large I/O sizes, and sequential I/O patterns. HDD-backed volumes include Throughput Optimized HDD (st1) and Cold HDD (sc1). For more information on Amazon EBS see the EBS product details page.

Q: Do you support multiple instances accessing a single volume?

While you are able to attach multiple volumes to a single instance, attaching multiple instances to one volume is not supported at this time.

Q: Will I be able to access my EBS snapshots using the regular Amazon S3 APIs?

No, EBS snapshots are only available through the Amazon EC2 APIs.

Q: Do volumes need to be un-mounted in order to take a snapshot? Does the snapshot need to complete before the volume can be used again?

No, snapshots can be done in real time while the volume is attached and in use. However, snapshots only capture data that has been written to your Amazon EBS volume, which might exclude any data that has been locally cached by your application or OS. In order to ensure consistent snapshots on volumes attached to an instance, we recommend cleanly detaching the volume, issuing the snapshot command, and then reattaching the volume. For Amazon EBS volumes that serve as root devices, we recommend shutting down the machine to take a clean snapshot.

Q: Are snapshots versioned? Can I read an older snapshot to do a point-in-time recovery?

Each snapshot is given a unique identifier, and customers can create volumes based on any of their existing snapshots.

Q: What charges apply when using Amazon EBS shared snapshots?

If you share a snapshot, you won't be charged when other users make a copy of your snapshot. If you make a copy of another user's shared volume, you will be charged normal EBS rates.

Q: Can users of my Amazon EBS shared snapshots change any of my data?

Users who have permission to create volumes based on your shared snapshots will first make a copy of the snapshot into their account. Users can modify their own copies of the data, but the data on your original snapshot and any other volumes created by other users from your original snapshot will remain unmodified.

Q: How can I discover Amazon EBS snapshots that have been shared with me?

You can find snapshots that have been shared with you by selecting "Private Snapshots" from the viewing dropdown in the Snapshots section of the AWS Management Console. This section will list both snapshots you own and snapshots that have been shared with you.

Q: How can I find what Amazon EBS snapshots are shared globally?

You can find snapshots that have been shared globally by selecting "Public Snapshots" from the viewing dropdown in the Snapshots section of the AWS Management Console.

Q: Do you offer encryption on Amazon EBS volumes and snapshots?

Yes. EBS offers seamless encryption of data volumes and snapshots. EBS encryption better enables you to meet security and encryption compliance requirements.

Q: How can I find a list of Amazon Public Data Sets?

All information on Public Data Sets is available in our Public Data Sets Resource Center. You can also obtain a listing of Public Data Sets within the AWS Management Console by choosing "Amazon Snapshots" from the viewing dropdown in the Snapshots section.

Q: Where can I learn more about EBS?

You can visit the Amazon EBS FAQ page.

Amazon Elastic File System (EFS)

Q. How do I access a file system from an Amazon EC2 instance?

To access your file system, you mount the file system on an Amazon EC2 Linux-based instance using the standard Linux mount command and the file system's DNS name. Once you've mounted, you can work with the files and directories in your file system just like you would with a local file system.

Amazon EFS uses the NFSv4.1 protocol. For a step-by-step example of how to access a file system from an Amazon EC2 instance, please see the Amazon EFS Getting Started guide.

Q. What Amazon EC2 instance types and AMIs work with Amazon EFS?

Amazon EFS is compatible with all Amazon EC2 instance types and is accessible from Linux-based AMIs. You can mix and match the instance types connected to a single file system. For a step-by-step example of how to access a file system from an Amazon EC2 instance, please see the Amazon EFS Getting Started guide.

Q. How do I load data into a file system?

You can load data into an Amazon EFS file system from your Amazon EC2 instances or from your on-premises datacenter servers.

Amazon EFS file systems can be mounted on an Amazon EC2 instance, so any data that is accessible to an Amazon EC2 instance can also be read and written to Amazon EFS. To load data that is not currently stored on the Amazon cloud, you can use the same methods you use to transfer files to Amazon EC2 today, such as Secure Copy (SCP).

Amazon EFS file systems can also be mounted on an on-premises server, so any data that is accessible to an on-premises server can be read and written to Amazon EFS using standard Linux tools. For more information about accessing a file system from an on-premises server, please see the On-premises Access section of the Amazon EFS FAQ.

For more information about moving data to the Amazon cloud, please see the Cloud Data Migration page.

Q. How do I access my file system from outside my VPC?

Amazon EC2 instances within your VPC can access your file system directly, and Amazon EC2 Classic instances outside your VPC can mount a file system via ClassicLink. On-premises servers can mount your file systems via an AWS Direct Connect connection to your VPC.

Q. How many Amazon EC2 instances can connect to a file system?

Amazon EFS supports one to thousands of Amazon EC2 instances connecting to a file system concurrently.

Q: Where can I learn more about EFS?

You can visit the Amazon EFS FAQ page.

NVMe Instance storage

Q: Which instance types offer NVMe instance storage?

Today, I3en, I3, C5d, M5d, M5ad, R5d, R5ad, z1d, and F1 instances offer NVMe instance storage.

Q: Is data stored on Amazon EC2 NVMe instance storage encrypted?

Yes, all data is encrypted in an AWS Nitro hardware module prior to being written on the locally attached SSDs offered via NVMe instance storage.

Q: What encryption algorithm is used to encrypt Amazon EC2 NVMe instance storage?

Amazon EC2 NVMe instance storage is encrypted using an XTS-AES-256 block cipher.

Q: Are encryption keys unique to an instance or a particular device for NVMe instance storage?

Encryption keys are securely generated within the Nitro hardware module, and are unique to each NVMe instance storage device that is provided with an EC2 instance.

Q: What is the lifetime of encryption keys on NVMe instance storage?

All keys are irrecoverably destroyed on any de-allocation of the storage, including instance stop and instance terminate actions.

Q: Can I disable NVMe instance storage encryption?

No, NVMe instance storage encryption is always on, and cannot be disabled.

Q: Do the published IOPS performance numbers on I3 and I3en include data encryption?

Yes, the documented IOPS numbers for I3 and I3en NVMe instance storage include encryption.

Q: Does Amazon EC2 NVMe instance storage support AWS Key Management Service (KMS)?

No, disk encryption on NVMe instance storage does not support integration with AWS KMS system. Customers cannot bring in their own keys to use with NVMe instance storage.

Networking and security

Elastic Fabric Adapter (EFA) | Elastic IP | Elastic Load Balancing | Enhanced networking | Security

Elastic Fabric Adapter (EFA)

Q: Why should I use EFA?

EFA brings the scalability, flexibility, and elasticity of cloud to tightly-coupled HPC applications. With EFA, tightly-coupled HPC applications have access to lower and more consistent latency and higher throughput than traditional TCP channels, enabling them to scale better. EFA support can be enabled dynamically, on-demand on any supported EC2 instance without pre-reservation, giving you the flexibility to respond to changing business/workload priorities.

Q: What types of applications can benefit from using EFA?

High Performance Computing (HPC) applications distribute computational workloads across a cluster of instances for parallel processing. Examples of HPC applications include computational fluid dynamics (CFD), crash simulations, and weather simulations. HPC applications are generally written using the Message Passing Interface (MPI) and impose stringent requirements for inter-instance communication in terms of both latency and bandwidth. Applications using MPI and other HPC middleware which supports the libfabric communication stack can benefit from EFA.

Q: How does EFA communication work?

EFA devices provide all ENA devices functionalities plus a new OS bypass hardware interface that allows user-space applications to communicate directly with the hardware-provided reliable transport functionality. Most applications will use existing middleware, such as the Message Passing Interface (MPI), to interface with EFA. AWS has worked with a number of middleware providers to ensure support for the OS bypass functionality of EFA. Please note that communication using the OS bypass functionality is limited to instances within a single subnet of a Virtual Private Cloud (VPC).

Q: Which instance types support EFA?

EFA is currently available on the m5n.24xlarge, m5dn.24xlarge, r5n.24xlarge, r5dn.24xlarge, c5n.18xlarge, c5n.metal, p3dn.24xlarge, i3en.24xlarge, and i3en.metal instance sizes. Support for more instance types and sizes being added in the coming months.

Q: What are the differences between an EFA ENI and an ENA ENI?

An ENA ENI provides traditional IP networking features necessary to support VPC networking. An EFA ENI provides all the functionality of an ENA ENI, plus hardware support for applications to communicate directly with the EFA ENI without involving the instance kernel (OS-bypass communication) using an extended programming interface. Due to the advanced capabilities of the EFA ENI, EFA ENIs can only be attached at launch or to stopped instances.

Q: What are the pre-requisites to enabling EFA on an instance?

EFA support can be enabled either at the launch of the instance or added to a stopped instance. EFA devices cannot be attached to a running instance.

Elastic IP

Q: Why am I limited to 5 Elastic IP addresses per region?

Public (IPV4) internet addresses are a scarce resource. There is only a limited amount of public IP space available, and Amazon EC2 is committed to helping use that space efficiently.

By default, all accounts are limited to 5 Elastic IP addresses per region. If you need more the 5 Elastic IP addresses, we ask that you apply for your limit to be raised. We will ask you to think through your use case and help us understand your need for additional addresses. You can apply for more Elastic IP address here. Any increases will be specific to the region they have been requested for.

Q: Why am I charged when my Elastic IP address is not associated with a running instance?

In order to help ensure our customers are efficiently using the Elastic IP addresses, we impose a small hourly charge for each address when it is not associated to a running instance.

Q: Do I need one Elastic IP address for every instance that I have running?

No. You do not need an Elastic IP address for all your instances. By default, every instance comes with a private IP address and an internet routable public IP address. The private IP address remains associated with the network interface when the instance is stopped and restarted, and is released when the instance is terminated. The public address is associated exclusively with the instance until it is stopped, terminated or replaced with an Elastic IP address. These IP addresses should be adequate for many applications where you do not need a long lived internet routable end point. Compute clusters, web crawling, and backend services are all examples of applications that typically do not require Elastic IP addresses.

Q: How long does it take to remap an Elastic IP address?

The remap process currently takes several minutes from when you instruct us to remap the Elastic IP until it fully propagates through our system.

Q: Can I configure the reverse DNS record for my Elastic IP address?

All Elastic IP addresses come with reverse DNS, in a standard template of the form ec2-1-2-3-4.region.compute.amazonaws.com. For customers requiring custom reverse DNS settings for internet-facing applications that use IP-based mutual authentication (such as sending email from EC2 instances), you can configure the reverse DNS record of your Elastic IP address by filling out this form. Alternatively, please contact AWS Customer Support if you want AWS to delegate the management of the reverse DNS for your Elastic IPs to your authoritative DNS name servers (such as Amazon Route 53), so that you can manage your own reverse DNS PTR records to support these use-cases. Note that a corresponding forward DNS record pointing to that Elastic IP address must exist before we can create the reverse DNS record.

Elastic Load Balancing

Q: What load balancing options does the Elastic Load Balancing service offer?

Elastic Load Balancing offers two types of load balancers that both feature high availability, automatic scaling, and robust security. These include the Classic Load Balancer that routes traffic based on either application or network level information, and the Application Load Balancer that routes traffic based on advanced application level information that includes the content of the request.

Q: When should I use the Classic Load Balancer and when should I use the Application Load Balancer?

The Classic Load Balancer is ideal for simple load balancing of traffic across multiple EC2 instances, while the Application Load Balancer is ideal for applications needing advanced routing capabilities, microservices, and container-based architectures. Please visit Elastic Load Balancing for more information.

Enhanced networking

Q: What networking capabilities are included in this feature?

We currently support enhanced networking capabilities using SR-IOV (Single Root I/O Virtualization). SR-IOV is a method of device virtualization that provides higher I/O performance and lower CPU utilization compared to traditional implementations. For supported Amazon EC2 instances, this feature provides higher packet per second (PPS) performance, lower inter-instance latencies, and very low network jitter.

Q: Why should I use Enhanced Networking?

If your applications benefit from high packet-per-second performance and/or low latency networking, Enhanced Networking will provide significantly improved performance, consistence of performance and scalability.

Q: How can I enable Enhanced Networking on supported instances?

In order to enable this feature, you must launch an HVM AMI with the appropriate drivers. C5, C5d, F1, G3, H1, I3, I3en, m4.16xlarge, M5, M5a, M5ad, M5d, P2, P3, R4, R5, R5a, R5ad, R5d, T3, T3a, X1, X1e, and z1d instances use the Elastic Network Adapter (which uses the “ena” Linux driver) for Enhanced Networking. C3, C4, D2, I2, M4 (excluding m4.16xlarge), and R3 instances use Intel® 82599g Virtual Function Interface (which uses the “ixgbevf” Linux driver). Amazon Linux AMI includes both of these drivers by default. For AMIs that do not contain these drivers, you will need to download and install the appropriate drivers based on the instance types you plan to use. You can use Linux or Windows instructions to enable Enhanced Networking in AMIs that do not include the SR-IOV driver by default. Enhanced Networking is only supported in Amazon VPC.

Q: Do I need to pay an additional fee to use Enhanced Networking?

No, there is no additional fee for Enhanced Networking. To take advantage of Enhanced Networking you need to launch the appropriate AMI on a supported instance type in a VPC.

Q: Why is Enhanced Networking only supported in Amazon VPC?

Amazon VPC allows us to deliver many advanced networking features to you that are not possible in EC2-Classic. Enhanced Networking is another example of a capability enabled by Amazon VPC.

Q: Which instance types support Enhanced Networking?

Depending on your instance type, enhanced networking can be enabled using one of the following mechanisms:

Intel 82599 Virtual Function (VF) interface - The Intel 82599 Virtual Function interface supports network speeds of up to 10 Gbps for supported instance types. C3, C4, D2, I2, M4 (excluding m4.16xlarge), and R3 instances use the Intel 82599 VF interface for enhanced networking.

Elastic Network Adapter (ENA) - The Elastic Network Adapter (ENA) supports network speeds of up to 25 Gbps for supported instance types. C5, C5d, F1, G3, H1, I3, I3en, m4.16xlarge, M5, M5a, M5ad, M5d, P2, P3, R4, R5, R5a, R5ad, R5d, T3, X1, X1e, and z1d instances use the Elastic Network Adapter for enhanced networking.

Q. Which instance types offer NVMe instance storage?

High I/O instances use NVMe based local instance storage to deliver very high, low latency, I/O capacity to applications, and are optimized for applications that require millions of IOPS. Like Cluster instances, High I/O instances can be clustered via cluster placement groups for high bandwidth networking.

Security

Q: How do I prevent other people from viewing my systems?

You have complete control over the visibility of your systems. The Amazon EC2 security systems allow you to place your running instances into arbitrary groups of your choice. Using the web services interface, you can then specify which groups may communicate with which other groups, and also which IP subnets on the Internet may talk to which groups. This allows you to control access to your instances in our highly dynamic environment. Of course, you should also secure your instance as you would any other server.

Q: Can I get a history of all EC2 API calls made on my account for security analysis and operational troubleshooting purposes?

Yes. To receive a history of all EC2 API calls (including VPC and EBS) made on your account, you simply turn on CloudTrail in the AWS Management Console. For more information, visit the CloudTrail home page.

Q: Where can I find more information about security on AWS?

For more information on security on AWS please refer to our Amazon Web Services: Overview of Security Processes white paper and to our Amazon EC2 running Windows Security Guide.

Management

Amazon CloudWatch | Amazon EC2 Auto Scaling | Hibernate | VM Import/Export

Amazon CloudWatch

Q: What is the minimum time interval granularity for the data that Amazon CloudWatch receives and aggregates?

Metrics are received and aggregated at 1 minute intervals.

Q: Which operating systems does Amazon CloudWatch support?

Amazon CloudWatch receives and provides metrics for all Amazon EC2 instances and should work with any operating system currently supported by the Amazon EC2 service.

Q: Will I lose the metrics data if I disable monitoring for an Amazon EC2 instance?

You can retrieve metrics data for any Amazon EC2 instance up to 2 weeks from the time you started to monitor it. After 2 weeks, metrics data for an Amazon EC2 instance will not be available if monitoring was disabled for that Amazon EC2 instance. If you want to archive metrics beyond 2 weeks you can do so by calling mon-get-stats command from the command line and storing the results in Amazon S3 or Amazon SimpleDB.

Q: Can I access the metrics data for a terminated Amazon EC2 instance or a deleted Elastic Load Balancer?

Yes. Amazon CloudWatch stores metrics for terminated Amazon EC2 instances or deleted Elastic Load Balancers for 2 weeks.

Q: Does the Amazon CloudWatch monitoring charge change depending on which type of Amazon EC2 instance I monitor?

No, the Amazon CloudWatch monitoring charge does not vary by Amazon EC2 instance type.

Q: Why does the graphing of the same time window look different when I view in 5 minute and 1 minute periods?

If you view the same time window in a 5 minute period versus a 1 minute period, you may see that data points are displayed in different places on the graph. For the period you specify in your graph, Amazon CloudWatch will find all the available data points and calculates a single, aggregate point to represent the entire period. In the case of a 5 minute period, the single data point is placed at the beginning of the 5 minute time window. In the case of a 1 minute period, the single data point is placed at the 1 minute mark. We recommend using a 1 minute period for troubleshooting and other activities that require the most precise graphing of time periods.

Amazon EC2 Auto Scaling

Q: Can I automatically scale Amazon EC2 Auto Scaling Groups?

Yes. Amazon EC2 Auto Scaling is a fully managed service designed to launch or terminate Amazon EC2 instances automatically to help ensure you have the correct number of Amazon EC2 instances available to handle the load for your application. EC2 Auto Scaling helps you maintain application availability through fleet management for EC2 instances, which detects and replaces unhealthy instances, and by scaling your Amazon EC2 capacity up or down automatically according to conditions you define. You can use EC2 Auto Scaling to automatically increase the number of Amazon EC2 instances during demand spikes to maintain performance and decrease capacity during lulls to reduce costs.

Allocation strategies in EC2 Auto Scaling determines how Spot Instances in your fleet are fulfilled from Spot Instance pools. The capacity-optimized allocation strategy attempts to provision Spot Instances from the most available Spot Instance pools by analyzing capacity metrics. This strategy is a good choice for workloads that have a higher cost of interruption such as big data and analytics, image and media rendering, machine learning, and high performance computing. The lowest-price allocation strategy launches Spot Instances strictly based on diversification across 'N' lowest priced pools.

For more information see the Amazon EC2 Auto Scaling FAQ.

Hibernate

Q: Why should I hibernate an instance?

You can hibernate an instance to get your instance and applications up and running quickly, if they take long time to bootstrap (e.g. load memory caches). You can start instances, bring them to a desired state and hibernate them. These "pre-warmed" instances can then be resumed to reduce the time it takes for an instance to return to service. Hibernation retains memory state across Stop/Start cycles.

Q: What happens when I hibernate my instance?

When you hibernate an instance, data from your EBS root volume and any attached EBS data volumes is persisted. Additionally, contents from the instance's memory (RAM) are persisted to EBS root volume. When the instance is restarted, it returns to its previous state and reloads the RAM contents.

Q: What is the difference between hibernate and stop?

In the case of hibernate, your instance gets hibernated and the RAM data persisted. In the case of Stop, your instance gets shutdown and RAM is cleared.

In both the cases, data from your EBS root volume and any attached EBS data volumes is persisted. Your private IP address remains the same, as does your elastic IP address (if applicable). The network layer behavior will be similar to that of EC2 Stop-Start workflow. Stop and hibernate are available for Amazon EBS backed instances only. Local instance storage is not persisted.

Q: How much does it cost to hibernate an instance?

Hibernating instances are charged at standard EBS rates for storage. As with a stopped instance, you do not incur instance usage fees while an instance is hibernating.

Q: How can I hibernate an instance?

Hibernation needs to be enabled when you launch the instance. Once enabled, you can use the StopInstances API with an additional ‘Hibernate’ parameter to trigger hibernation. You can also do this through the console by selecting your instance, then clicking Actions> Instance State > Stop - Hibernate. For more information on using hibernation, refer to the user guide.

Q: How can I resume a hibernating instance?

You can resume by calling the StartInstances API as you would for a regular stopped instance. You can also do this through the console by selecting your instance, then clicking Actions > Instance State > Start

Q: Can I enable hibernation on an existing instance?

No, you cannot enable hibernation on an existing instance (running or stopped). This needs to be enabled during instance launch.

Q: How can I tell that an instance is hibernated?

You can tell that an instance is hibernated by looking at the state reason. It should be ‘Client.UserInitiatedHibernate’. This is visible on the console under “Instances - Details” view or in the DescribeInstances API response as “reason” field.

Q: What is the state of an instance when it is hibernating?

Hibernated instances are in ‘Stopped’ state.

Q: What data is saved when I hibernate an instance?

EBS volume storage (boot volume and attached data volumes) and memory (RAM) are saved. Your private IP address remains the same (for VPC), as does your elastic IP address (if applicable). The network layer behavior will be similar to that of EC2 Stop-Start workflow.

Q: Where is my data stored when I hibernate an instance?

As with the Stop feature, root device and attached device data are stored on the corresponding EBS volumes. Memory (RAM) contents are stored on the EBS root volume.

Q: Is my memory (RAM) data encrypted when it is moved to EBS?

Yes, RAM data is always encrypted when it is moved to the EBS root volume. Encryption on the EBS root volume is enforced at instance launch time. This is to ensure protection for any sensitive content that is in memory at the time of hibernation.

Q: How long can I keep my instance hibernated?

We do not support keeping an instance hibernated for more than 60 days. You need to resume the instance and go through Stop and Start (without hibernation) if you wish to keep the instance around for a longer duration. We are constantly working to keep our platform up-to-date with upgrades and security patches, some of which can conflict with the old hibernated instances. We will notify you for critical updates that require you to resume the hibernated instance to perform a shutdown or a reboot.

Q: What are the prerequisites to hibernate an instance?

To use hibernation, the root volume must be an encrypted EBS volume. The instance needs to be configured to receive the ACPID signal for hibernation (or use the Amazon published AMIs that are configured for hibernation). Additionally, your instance should have sufficient space available on your EBS root volume to write data from memory.

Q: Which instances and operating systems support hibernation?

Hibernation is currently supported across M3, M4, M5, C3, C4, C5, R3, R4, and R5 instances running Amazon Linux, Amazon Linux 2, Ubuntu and Windows. For Windows, Hibernation is supported for instances up to 16 GB of RAM. For other operating systems, Hibernation is supported for instances with less than 150 GB of RAM. To review the list of supported OS versions, refer to the user guide.

Q: Should I use specific Amazon Machine Image (AMIs) if I want to hibernate my instance?

You can use any AMI that is configured to support hibernation. You can use AWS published AMIs that support hibernation by default. Alternatively, you can create a custom image from an instance after following the hibernation pre-requisite checklist and configuring your instance appropriately.

Q: What if my EBS root volume is not large enough to store memory state (RAM) for hibernate?

To enable hibernation, space is allocated on the root volume to store the instance memory (RAM). Make sure that the root volume is large enough to store the RAM contents and accommodate your expected usage, e.g. OS, applications. If the EBS root volume does not enough space, hibernation will fail and the instance will get shutdown instead.

VM Import/Export

Q. What is VM Import/Export?

VM Import/Export enables customers to import Virtual Machine (VM) images in order to create Amazon EC2 instances. Customers can also export previously imported EC2 instances to create VMs. Customers can use VM Import/Export to leverage their previous investments in building VMs by migrating their VMs to Amazon EC2.

Q. What operating systems are supported?

VM Import/Export currently supports Windows and Linux VMs, including Windows Server 2003, Windows Server 2003 R2, Windows Server 2008, Windows Server 2012 R1, Red Hat Enterprise Linux (RHEL) 5.1-6.5 (using Cloud Access), Centos 5.1-6.5, Ubuntu 12.04, 12.10, 13.04, 13.10, and Debian 6.0.0-6.0.8, 7.0.0-7.2.0. For more details on VM Import, including supported file formats, architectures, and operating system configurations, please see the VM Import/Export section of the Amazon EC2 User Guide.

Q. What virtual machine file formats are supported?

You can import VMware ESX VMDK images, Citrix Xen VHD images, Microsoft Hyper-V VHD images and RAW images as Amazon EC2 instances. You can export EC2 instances to VMware ESX VMDK, VMware ESX OVA, Microsoft Hyper-V VHD or Citrix Xen VHD images. For a full list of support operating systems, please see What operating systems are supported?.

Q. What is VMDK?

VMDK is a file format that specifies a virtual machine hard disk encapsulated within a single file. It is typically used by virtual IT infrastructures such as those sold by VMware, Inc.

Q. How do I prepare a VMDK file for import using the VMware vSphere client?

The VMDK file can be prepared by calling File-Export-Export to OVF template in VMware vSphere Client. The resulting VMDK file is compressed to reduce the image size and is compatible with VM Import/Export. No special preparation is required if you are using the Amazon EC2 VM Import Connector vApp for VMware vCenter.

Q. What is VHD?

VHD (Virtual Hard Disk) is a file format that that specifies a virtual machine hard disk encapsulated within a single file. The VHD image format is used by virtualization platforms such as Microsoft Hyper-V and Citrix Xen.

Q. How do I prepare a VHD file for import from Citrix Xen?

Open Citrix XenCenter and select the virtual machine you want to export. Under the Tools menu, choose "Virtual Appliance Tools" and select "Export Appliance" to initiate the export task. When the export completes, you can locate the VHD image file in the destination directory you specified in the export dialog.

Q. How do I prepare a VHD file for import from Microsoft Hyper-V?

Open the Hyper-V Manager and select the virtual machine you want to export. In the Actions pane for the virtual machine, select "Export" to initiate the export task. Once the export completes, you can locate the VHD image file in the destination directory you specified in the export dialog.

Q. Are there any other requirements when importing a VM into Amazon EC2?

The virtual machine must be in a stopped state before generating the VMDK or VHD image. The VM cannot be in a paused or suspended state. We suggest that you export the virtual machine with only the boot volume attached. You can import additional disks using the ImportVolume command and attach them to the virtual machine using AttachVolume. Additionally, encrypted disks (e.g. Bit Locker) and encrypted image files are not supported. You are also responsible for ensuring that you have all necessary rights and licenses to import into AWS and run any software included in your VM image.

Q. Does the virtual machine need to be configured in any particular manner to enable import to Amazon EC2?

Ensure Remote Desktop (RDP) or Secure Shell (SSH) is enabled for remote access and verify that your host firewall (Windows firewall, iptables, or similar), if configured, allows access to RDP or SSH. Otherwise, you will not be able to access your instance after the import is complete. Please also ensure that Windows VMs are configured to use strong passwords for all users including the administrator and that Linux VMs and configured with a public key for SSH access.

Q. How do I import a virtual machine to an Amazon EC2 instance?

You can import your VM images using the Amazon EC2 API tools:

- Import the VMDK, VHD or RAW file via the ec2-import-instance API. The import instance task captures the parameters necessary to properly configure the Amazon EC2 instance properties (instance size, Availability Zone, and security groups) and uploads the disk image into Amazon S3.
- If ec2-import-instance is interrupted or terminates without completing the upload, use ec2-resume-import to resume the upload. The import task will resume where it left off.
- Use the ec2-describe-conversion-tasks command to monitor the import progress and obtain the resulting Amazon EC2 instance ID.
- Once your import task is completed, you can boot the Amazon EC2 instance by specifying its instance ID to the ec2-run-instances API.
- Finally, use the ec2-delete-disk-image command line tool to delete your disk image from Amazon S3 as it is no longer needed.

Alternatively, if you use the VMware vSphere virtualization platform, you can import your virtual machine to Amazon EC2 using a graphical user interface provided through AWS Management Portal for vCenter. Please refer to Getting Started Guide in AWS Management Portal for vCenter. AWS Management Portal for vCenter includes integrated support for VM Import. Once the portal is installed within vCenter, you can right-click on a VM and select “Migrate to EC2” to create an EC2 instance from the VM. The portal will handle exporting the VM from vCenter, uploading it to S3, and converting it into an EC2 instance for you, with no additional work required. You can also track the progress of your VM migrations within the portal.

Q. How do I export an Amazon EC2 instance back to my on-premise virtualization environment?

You can export your Amazon EC2 instance using the Amazon EC2 CLI tools:

- Export the instance using the ec2-create-instance-export-task command. The export command captures the parameters necessary (instance ID, S3 bucket to hold the exported image, name of the exported image, VMDK, OVA or VHD format) to properly export the instance to your chosen format. The exported file is saved in an S3 bucket that you previously created
- Use ec2-describe-export-tasks to monitor the export progress

- Use ec2-cancel-export-task to cancel an export task prior to completion

Q. Are there any other requirements when exporting an EC2 instance using VM Import/Export?

You can export running or stopped EC2 instances that you previously imported using VM Import/Export. If the instance is running, it will be momentarily stopped to snapshot the boot volume. EBS data volumes cannot be exported. EC2 instances with more than one network interface cannot be exported.

Q. Can I export Amazon EC2 instances that have one or more EBS data volumes attached?

Yes, but VM Import/Export will only export the boot volume of the EC2 instance.

Q. What does it cost to import a virtual machine?

You will be charged standard Amazon S3 data transfer and storage fees for uploading and storing your VM image file. Once your VM is imported, standard Amazon EC2 instance hour and EBS service fees apply. If you no longer wish to store your VM image file in S3 after the import process completes, use the ec2-delete-disk-image command line tool to delete your disk image from Amazon S3.

Q. What does it cost to export a virtual machine?

You will be charged standard Amazon S3 storage fees for storing your exported VM image file. You will also be charged standard S3 data transfer charges when you download the exported VM file to your on-premise virtualization environment. Finally, you will be charged standard EBS charges for storing a temporary snapshot of your EC2 instance. To minimize storage charges, delete the VM image file in S3 after downloading it to your virtualization environment.

Q. When I import a VM of Windows Server 2003 or 2008, who is responsible for supplying the operating system license?

When you launch an imported VM using Microsoft Windows Server 2003 or 2008, you will be charged standard instance hour rates for Amazon EC2 running the appropriate Windows Server version, which includes the right to utilize that operating system within Amazon EC2. You are responsible for ensuring that all other installed software is properly licensed.

So then, what happens to my on-premise Microsoft Windows license key when I import a VM of Windows Server 2003 or 2008? Since your on-premise Microsoft Windows license key that was associated with that VM is not used when running your imported VM as an EC2 instance, you can reuse it for another VM within your on-premise environment.

Q. Can I continue to use the AWS-provided Microsoft Windows license key after exporting an EC2 instance back to my on-premise virtualization environment?

No. After an EC2 instance has been exported, the license key utilized in the EC2 instance is no longer available. You will need to reactivate and specify a new license key for the exported VM after it is launched in your on-premise virtualization platform.

Q. When I import a VM with Red Hat Enterprise Linux (RHEL), who is responsible for supplying the operating system license?

When you import Red Hat Enterprise Linux (RHEL) VM images, you can use license portability for your RHEL instances. With license portability, you are responsible for maintaining the RHEL licenses for imported instances, which you can do using Cloud Access subscriptions for Red Hat Enterprise Linux. Please contact Red Hat to learn more about Cloud Access and to verify your eligibility.

Q. How long does it take to import a virtual machine?

The length of time to import a virtual machine depends on the size of the disk image and your network connection speed. As an example, a 10 GB Windows Server 2008 SP2 VMDK image takes approximately 2 hours to import when it’s transferred over a 10 Mbps network connection. If you have a slower network connection or a large disk to upload, your import may take significantly longer.

Q. In which Amazon EC2 regions can I use VM Import/Export?

Visit the Region Table page to see product service availability by region.

Q. How many simultaneous import or export tasks can I have?

Each account can have up to five active import tasks and five export tasks per region.

Q. Can I run imported virtual machines in Amazon Virtual Private Cloud (VPC)?

Yes, you can launch imported virtual machines within Amazon VPC.

Q. Can I use the AWS Management Console with VM Import/Export?

No. VM Import/Export commands are available via EC2 CLI and API. You can also use the AWS Management Portal for vCenter to import VMs into Amazon EC2. Once imported, the resulting instances are available for use via the AWS Management Console.

Billing and purchase options

Billing | Savings Plans | Convertible Reserved Instances | EC2 Fleet | On-Demand Capacity Reservation | Reserved Instances | Reserved Instance Marketplace | Spot Instances

Billing

Q: How will I be charged and billed for my use of Amazon EC2?

You pay only for what you use. Displayed pricing is an hourly rate but depending on which instances you choose, you pay by the hour or second (minimum of 60 seconds) for each instance type. Partial instance-hours consumed are billed based on instance usage. Data transferred between AWS services in different regions will be charged as Internet Data Transfer on both sides of the transfer. Usage for other Amazon Web Services is billed separately from Amazon EC2.

For EC2 pricing information, please visit the pricing section on the EC2 detail page.

Q: When does billing of my Amazon EC2 systems begin and end?

Billing commences when Amazon EC2 initiates the boot sequence of an AMI instance. Billing ends when the instance terminates, which could occur through a web services command, by running "shutdown -h", or through instance failure. When you stop an instance, we shut it down but don't charge hourly usage for a stopped instance, or data transfer fees, but we do charge for the storage for any Amazon EBS volumes. To learn more, visit the AWS Documentation.

Q: What defines billable EC2 instance usage?

Instance usages are billed for any time your instances are in a "running" state. If you no longer wish to be charged for your instance, you must "stop" or "terminate" the instance to avoid being billed for additional instance usage. Billing starts when an instance transitions into the running state.

Q: If I have two instances in different availability zones, how will I be charged for regional data transfer?

Each instance is charged for its data in and data out at corresponding Data Transfer rates. Therefore, if data is transferred between these two instances, it is charged at "Data Transfer Out from EC2 to Another AWS Region" for the first instance and at "Data Transfer In from Another AWS Region" for the second instance. Please refer to this page for detailed data transfer.

Q. If I have two instances in different regions, how will I be charged for data transfer?

Each instance is charged for its data in and data out at Internet Data Transfer rates. Therefore, if data is transferred between these two instances, it is charged at Internet Data Transfer Out for the first instance and at Internet Data Transfer In for the second instance.

Q: How will my monthly bill show per-second versus per-hour?

Although EC2 charges in your monthly bill will now be calculated based on a per second basis, for consistency, the monthly EC2 bill will show cumulative usage for each instance that ran in a given month in decimal hours. An example would be an instance running for 1 hour 10 minutes and 4 seconds would look like 1.1677. Read this blog for an example of the detailed billing report.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Savings Plans

What is Savings Plans?

Savings Plans is a flexible pricing model that offers low prices on EC2, Lambda and Fargate usage, in exchange for a commitment to a consistent amount of usage (measured in $/hour) for a 1 or 3 year term. When you sign up for Savings Plans, you will be charged the discounted Savings Plans price for your usage up to your commitment. For example, if you commit to $10 of compute usage an hour, you will get the Savings Plans prices on that usage up to $10 and any usage beyond the commitment will be charged On Demand rates.

What types of Savings Plans does AWS offer?

AWS offers two types of Savings Plans:

1. Compute Savings Plans provides the most flexibility and help to reduce your costs by up to 66%. These plans automatically apply to EC2 instance usage regardless of instance family, size, AZ, region, OS or tenancy, and also apply to AWS Fargate and Lambda usage. For example, with Compute Savings Plans, you can change from C4 to M5 instances, shift a workload from EU (Ireland) to EU (London), or move a workload from EC2 to Fargate or Lambda at any time and automatically continue to pay the Savings Plans price.
2. EC2 Instance Savings Plans provides the lowest prices, offering savings up to 72% in exchange for commitment to usage of individual instance families in a region (e.g. M5 usage in N. Virginia). This automatically reduces your cost

on the selected instance family in that region regardless of AZ, size, OS or tenancy. EC2 Instance Savings Plans give you the flexibility to change your usage between instances within a family in that region. For example, you can move from c5.xlarge running Windows to c5.2xlarge running Linux and automatically benefit from the Savings Plan prices.

How do Savings Plans compare to EC2 RIs?

Savings Plans offers significant savings over On Demand, just like EC2 RIs, but automatically reduce your bills on compute usage across any AWS region, even as usage changes. This provides you the flexibility to use the compute option that best suits your needs and continue to save money, all without having to perform exchanges or modifications.

Compute Savings Plans, which provides savings up to 66% (just like Convertible RIs), automatically reduce your cost on any EC2 instance usage regardless of region, instance family, size, OS, tenancy and even on AWS Fargate and Lambda. EC2 Instance Savings Plans, which provide savings up to 72% (just like Standard RIs), automatically save you money on any instance usage within a given EC2 instance family in a chosen region (e.g. M5 in N. Virginia) regardless of size, OS or tenancy.

Do Savings Plans provide capacity reservations for EC2 instances?

No, Savings Plans do not provide a capacity reservation. You can however reserve capacity with On Demand Capacity Reservations and pay lower prices on them with Savings Plans.

How do I get started with Savings Plans?

You can get started with Savings Plans from AWS Cost Explorer in the management console or by using the API/CLI. You can easily make a commitment to a Savings Plan by using the recommendations provided in AWS Cost Explorer, to realize the biggest savings. The recommended hourly commitment is based on your historical On Demand usage and your choice of plan type, term length, and payment option. Once you sign up for a Savings Plan, your compute usage will automatically be charged at the discounted

Savings Plan prices and any usage beyond your commitment will be charged at regular On Demand rates.

Can I continue to purchase EC2 RIs?

Yes. You can continue purchasing RIs to maintain compatibility with your existing cost management processes, and your RIs will work along-side Savings Plans to reduce your overall bill. However as your RIs expire we encourage you to sign up for Savings Plans as they offer the same savings as RIs, but with additional flexibility.

Convertible Reserved Instances

Q: What is a Convertible RI?

A Convertible RI is a type of Reserved Instance with attributes that can be changed during the term.

Q: When should I purchase a Convertible RI instead of a Standard RI?

The Convertible RI is useful for customers who can commit to using EC2 instances for a three-year term in exchange for a significant discount on their EC2 usage, are uncertain about their instance needs in the future, or want to benefit from changes in price.

Q: What term length options are available on Convertible RIs?

Like Standard RIs, Convertible RIs are available for purchase for a one-year or three-year term.

Q: Can I exchange my Convertible RI to benefit from a Convertible RI matching a different instance type, operating system, tenancy, or payment option?

Yes, you can select a new instance type, operating system, tenancy, or payment option when you exchange your Convertible RIs. You also have the flexibility to exchange a portion of your Convertible RI or merge the value of multiple Convertible RIs in a single exchange.

Q: Can I transfer a Convertible or Standard RI from one region to another?

No, a RI is associated with a specific region, which is fixed for the duration of the reservation's term.

Q: How do I change the configuration of a Convertible RI?

You can change the configuration of your Convertible RI using the EC2 Management Console or the GetReservedInstancesExchangeQuote API. You also have the flexibility to exchange a portion of your Convertible RI or merge the value of multiple Convertible RIs in a single exchange. Click here to learn more about exchanging Convertible RIs.

Q: Do I need to pay a fee when I exchange my Convertible RIs?

No, you do not pay a fee when you exchange your RIs. However may need to pay a one-time true-up charge that accounts for differences in pricing between the Convertible RIs that you have and the Convertible RIs that you want.

Q: How do Convertible RI exchanges work?

When you exchange one Convertible RI for another, EC2 ensures that the total value of the Convertible RIs is maintained through a conversion. So, if you are converting your RI with a total value of $1000 for another RI, you will receive a quantity of Convertible RIs with a value that’s equal to or greater than $1000. You cannot convert your Convertible RI for Convertible RI(s) of a lesser total value.

Q: Can you define total value?

The total value is the sum of all expected payments that you’d make during the term for the RI.

Q: Can you walk me through how the true-up cost is calculated for a conversion between two All Upfront Convertible RIs?

Sure, let’s say you purchased an All Upfront Convertible RI for $1000 upfront, and halfway through the term you decide to change the attributes of the RI. Since you’re halfway through the RI term, you have $500 left of prorated value remaining on the RI. The All Upfront Convertible RI that you want to convert into costs $1,200 upfront today. Since you only have half of the term left on your existing Convertible RI, there is $600 of value remaining on the desired new Convertible RI. The true-up charge that you’ll pay will be the difference in upfront value between original and desired Convertible RIs, or $100 ($600 - $500).

Q: Can you walk me through a conversion between No Upfront Convertible RIs?

Unlike conversions between Convertible RIs with an upfront value, since you’re converting between RIs without an upfront cost, there will not be a true-up charge. However, the amount you pay on an hourly basis before the exchange will need to be greater than or equal to the amount you pay on a total hourly basis after the exchange.

For example, let’s say you purchased one No Upfront Convertible RI (A) with a $0.10/hr rate, and you decide to exchange Convertible RI A for another RI (B) that costs $0.06/hr. When you convert, you will receive two RIs of B because the amount that you pay on an hourly basis must be greater than or equal to the amount you’re paying for A on an hourly basis.

Q: Can I customize the number of instances that I receive as a result of a Convertible RI exchange?

No, EC2 uses the value of the Convertible RIs you’re trading in to calculate the minimal number of Convertible RIs you’ll receive while ensuring the result of the exchange gives you Convertible RIs of equal or greater value.

Q: Are there exchange limits for Convertible RIs?

No, there are no exchange limits for Convertible RIs.

Q: Do I have the freedom to choose any instance type when I exchange my Convertible RIs?

No, you can only exchange into Convertible RIs that are currently offered by AWS.

Q: Can I upgrade the payment option associated with my Convertible RI?

Yes, you can upgrade the payment option associated with your RI. For example, you can exchange your No Upfront RIs for Partial or All Upfront RIs to benefit from better pricing. You cannot change the payment option from All Upfront to No Upfront, and cannot change from Partial Upfront to No Upfront.

Q: Do Convertible RIs allow me to benefit from price reductions when they happen?

Yes, you can exchange your RIs to benefit from lower pricing. For example, if the price of new Convertible RIs reduces by 10%, you can exchange your Convertible RIs and benefit from the 10% reduction in price.

EC2 Fleet

Q. What is Amazon EC2 Fleet?

With a single API call, EC2 Fleet lets you provision compute capacity across different instance types, Availability Zones and across On-Demand, Reserved Instances (RI) and Spot Instances purchase models to help optimize scale, performance and cost.

Q. If I currently use Amazon EC2 Spot Fleet should I migrate to Amazon EC2 Fleet?

If you are leveraging Amazon EC2 Spot Instances with Spot Fleet, you can continue to use that. Spot Fleet and EC2 Fleet offer the same functionality. There is no requirement to migrate.

Q. Can I use Reserved Instance (RI) discounts with Amazon EC2 Fleet?

Yes, Similar to other EC2 APIs or other AWS services that launches EC2 instances, if the On-Demand instance launched by EC2 Fleet matches an existing RI, that instance will receive the RI discount. For example, if you own Regional RIs for M4 instances and you have specified only M4 instances in your EC2 Fleet, RI discounts will be automatically applied to this usage of M4.

Q. Will Amazon EC2 Fleet failover to On-Demand if EC2 Spot capacity is not fully fulfilled?

No, EC2 Fleet will continue to attempt to meet your desired Spot capacity based on the number of Spot instances you requested in your Fleet launch specification.

Q. What is the pricing for Amazon EC2 Fleet?

EC2 Fleet comes at no additional charge, you only pay for the underlying resources that EC2 Fleet launches.

Q. Can you provide a real world example of how I can use Amazon EC2 Fleet?

There are a number of ways to take advantage of Amazon EC2 Fleet, such as in big data workloads, containerized application, grid processing workloads etc. In this example of a genomic sequencing workload, you can launch a grid of worker nodes with a single API call: select your favorite instances, assign weights for these instances, specify target capacity for On-Demand and Spot Instances, and build a fleet within seconds to crunch through genomic data quickly.

Q. How can I allocate resources in an Amazon EC2 Fleet?

By default, EC2 Fleet will launch the On-Demand option that is lowest price. For Spot Instances, EC2 Fleet provides three allocation strategies: capacity-optimized, lowest price and diversified. The capacity-optimized allocation strategy attempts to provision Spot Instances from the most available Spot Instance pools by analyzing capacity metrics. This strategy is a good choice for workloads that have a higher cost of interruption such as big data and analytics, image and media rendering, machine learning, and high performance computing.

The lowest price strategy allows you to provision Spot Instances in pools that provide the lowest price per unit of capacity at the time of the request. The diversified strategy allows you to provision Spot Instances across multiple Spot pools and you can maintain your fleet’s target capacity to increase application.

Q. Can I submit a multi-region Amazon EC2 Fleet request?

No, we do not support multi-region EC2 Fleet requests.

Q. Can I tag an Amazon EC2 Fleet?

Yes. You can tag a EC2 Fleet request to create business-relevant tag groupings to organize resources along technical, business, and security dimensions.

Q. Can I modify my Amazon EC2 Fleet?

Yes, you can modify the total target capacity of your EC2 Fleet when in maintain mode. You may need to cancel the request and submit a new one to change other request configuration parameters.

Q. Can I specify a different AMI for each instance type that I want to use?

Yes, simply specify the AMI you’d like to use in each launch specification you provide in your EC2 Fleet.

On-Demand Capacity Reservation

On-Demand Capacity Reservation is an EC2 offering that lets you create and manage reserved capacity on Amazon EC2. You can create a Capacity Reservation by choosing an Availability Zone and quantity (number of instances) along with other instance specifications such as instance type and tenancy. Once created, the EC2 capacity is held for you regardless of whether you run the instances or not.

Q. How much do Capacity Reservations cost?

When the Capacity Reservation is active, you will pay equivalent instance charges whether you run the instances or not. If you do not use the reservation, the charge will show up as unused reservation on your EC2 bill. When you run an instance that matches the attributes of a reservation, you just pay for the instance and nothing for the reservation. There are no upfront or additional charges.

For example, if you create a Capacity Reservation for 20 c5.2xlarge instances and you run 15 c5.2xlarge instances, you will be charged for 15 instances and 5 unused instances in the reservation (effectively charged for 20 instances).

Q: Can I get a discount for Capacity Reservation usage?

Yes. Savings Plans or Regional RI (RI scoped to a region) discounts apply to Capacity Reservations. AWS Billing automatically applies your RI the discount when the attributes of a Capacity Reservation match the attributes of a an active Savings Plan or a Regional RI. When a Capacity Reservation is used by an instance, you are only charged for the instance (with RI Savings Plan or RI discounts applied). Regional RI dDiscounts are preferentially applied to running iinstance usages before covering unused Capacity Reservations.

For example, if you have a Regional RI for 50 c5.2xlarge instances and a Capacity Reservation for 50 c5.2xlarge instances in the same region, the RI discount will apply to the unused portion of the reservation. Note that discounts will first apply to any c5 instance usage (across instances sizes and Availability Zones) within that region before applying to unused reservations.

Note: A Regional RI is an EC2 RI scoped to an AWS Region. Zonal RIs (RIs scoped to an Availability Zone within a Region) discounts do not apply to On-Demand Capacity Reservations, as Zonal RIs already come with a capacity reservation.

Q. When should I use Savings Plans, EC2 RIs, and Capacity Reservations?

Use Savings Plans or Regional RIs to reduce your bill while committing to a one- or three-year term. Savings Plans offer significant savings over On Demand, just like EC2 RIs, but automatically reduce customers’ bills on compute usage across any AWS region, even as usage changes. Use Capacity Reservations if you need the additional confidence in your ability to launch instances. Capacity Reservations can be created for any duration and can be managed independently of your Savings Plans or RIs. If you have Savings Plans or Regional RIs, they will automatically apply to matching Capacity Reservations.

This gives you the flexibility to selectively add Capacity Reservations to a portion of your instance footprint and still reduce your bill for that usage.

Q. I have a Zonal RI (RI scoped to an Availability Zone) that also provides a capacity reservation? How does this compare with a Capacity Reservation?

A Zonal RI provides both a discount and a capacity reservation in a specific Availability Zone in return for a 1-to-3 year commitment. Capacity Reservation allows you to create and manage reserved capacity independently of your RI commitment and term length.

You can use On-Demand Capacity Reservations with a Savings Plan or a Regional RI to get, at the minimum, all the benefits of a Zonal RI for no additional cost. You also get the enhanced flexibility of Savings Plan (or Regional RI) and the features of Capacity Reservation: the ability to add or subtract from the reservation at any time, view reservation utilization in real-time, and the ability to target a Capacity Reservation for specific workloads.

Re-scoping your Zonal RIs to a region immediately gives you the Availability Zone and instance size flexibility in how RI discounts are applied. You can convert your Standard Zonal RIs to a Regional RI by modifying the scope of the RI from a specific Availability Zone to a region using the EC2 management console or the ModifyReservedInstances API.

Q. I created a Capacity Reservation. How can I use it?

A Capacity Reservation is tied to a specific Availability Zone and, by default automatically utilized by running instances in that Availability Zone. When you launch new instances that match the reservation attributes, they will automatically match to the reservation.

You can also target a reservation for specific workloads/instances if you prefer. Refer to Linux or windows technical documentation to learn more about the targeting option.

Q. How many instances am I allowed to reserve?

The number of instances you are allowed to reserve is based on your account's On-Demand instance limit. You can reserve as many instances as that limit allows, minus the number of instances that are already running.

If you need a higher limit, contact your AWS sales representative or complete the Amazon EC2 instance request form with your use case and your instance increase will be considered. Limit increases are tied to the region they are requested for.

Q. Can I modify a Capacity Reservation after it has started?

Yes. You can reduce the number of instances you reserved at any time. You can also increase the number of instances (subject to availability). You can also modify the end time of your reservation. You cannot modify a Capacity Reservation that has ended or has been deleted.

Q. Can I end a Capacity Reservation after it has started?

Yes. You can end a Capacity Reservation by canceling it using the console or API/SDK, or by modifying your reservation to specify an end time that makes it expire automatically. Running instances are unaffected by changes to your Capacity Reservation including deletion or expiration of a reservation.

Q. Where can I find more information about using Capacity Reservations?

Refer to Linux or windows technical documentation to learn about creating and using a Capacity Reservation.

Q. Can I share a Capacity Reservation with another AWS Account?

Yes, you can can share Capacity Reservations with other AWS accounts or within your AWS Organization via AWS Resource Access Manager service. You can share EC2 Capacity Reservations in three easy steps: create a Resource Share using AWS Resource Access Manager, add resources (Capacity Reservations) to the Resource Share, and specify the target accounts that you wish to share the resources with.

Note that sharing of Capacity Reservation is not available to new AWS accounts or AWS accounts that have a limited billing history. New accounts that are linked to a qualified master (payer) account or through an AWS Organization are exempt from this restriction.

Q. What happens when I share a Capacity Reservation with another AWS account?

When a Capacity Reservation is shared with other accounts, those accounts can consume the reserved capacity to run their EC2 Instances. The exact behavior depends by the preferences set on the Capacity Reservation. By default, Capacity Reservations automatically match existing and new instances from other accounts that have shared access to the reservation. You can also target a Capacity Reservation for specific workloads/instances. Individual accounts can control which of their instances consume Capacity Reservations. Refer to Linux or windows technical documentation to learn more about the instance matching options.

Q. Is there an additional charge for sharing a reservation?

There is no additional charge for sharing a reservation.

Q. Who gets charged when a Capacity Reservation is shared across multiple accounts?

If multiple accounts are consuming a Capacity Reservation, each account gets charged for its own instance usage. Unused reserved capacity, if any, gets charged to the account that owns the Capacity Reservation. If there is a consolidated billing arrangement among the accounts that share a Capacity Reservation, the master account gets billed for instance usage across all the linked accounts.

Q. Can I prioritize access to Capacity Reservation among the AWS accounts that have shared access?

No. Instance spots in a Capacity Reservation are available on a first-come-first-serve basis to any account that has shared access.

Q. How can I communicate the Availability Zone (AZ) of a CR with another account, given AZ name mappings could be different across AWS accounts?

You can now use Availability Zone ID (AZ ID) instead of AZ name. Availability Zone ID is a static reference and provides a consistent way of identifying the location of a resource across all your accounts. This makes it easier for you to provision resources centrally in a single account and share them across multiple accounts.

Q. Can I stop sharing my Capacity Reservation once I have shared it?

Yes, you can stop sharing a reservation after you have shared it. When you stop sharing a CR, with specific accounts or stop sharing entirely, other account(s) lose the ability to launch new instances into the CR. Any capacity occupied by instances running from other accounts will be restored to the CR for your use (subject to availability).

Q. Where can I find more information about sharing Capacity Reservations?

Refer to Linux or windows technical documentation to learn about sharing Capacity Reservations.

Q: Can I get a discount for Capacity Reservation usage?

Yes. Savings Plans or Regional RI discounts apply to Capacity Reservations. AWS Billing automatically applies the discount when the attributes of a Capacity Reservation match the attributes of a Savings Plan or a Regional RI. When a Capacity Reservation is used by an instance, you are only charged for the instance (with Savings Plan or RI discounts applied). Discounts are preferentially applied to instance usage before covering unused Capacity Reservations.

Note: A Regional RI is an EC2 RI scoped to an AWS Region. Zonal RI (RIs scoped to an Availability Zone within a Region) discounts do not apply to On-Demand Capacity Reservations, as Zonal RIs already come with a capacity reservation.

Reserved Instances

Q: What is a Reserved Instance?

A Reserved Instance (RI) is an EC2 offering that provides you with a significant discount on EC2 usage when you commit to a one-year or three-year term.

Q: What are the differences between Standard RIs and Convertible RIs?

Standard RIs offer a significant discount on EC2 instance usage when you commit to a particular instance family. Convertible RIs offer you the option to change your instance configuration during the term, and still receive a discount on your EC2 usage. For more information on Convertible RIs, please click here.

Q: Do RIs provide a capacity reservation?

Yes, when a Standard or Convertible RI is scoped to a specific Availability Zone (AZ), instance capacity matching the exact RI configuration is reserved for your use (these are referred to as “zonal RIs”). Zonal RIs give you additional confidence in your ability to launch instances when you need them.

You can also choose to forego the capacity reservation and purchase Standard or Convertible RIs that are scoped to a region (referred to as “regional RIs”). Regional RIs automatically apply the discount to usage across Availability Zones and instance sizes in a region, making it easier for you to take advantage of the RI’s discounted rate.

Q: When should I purchase a zonal RI?

If you want to take advantage of the capacity reservation, then you should buy an RI in a specific Availability Zone.

Q: When should I purchase a regional RI?

If you do not require the capacity reservation, then you should buy a regional RI. Regional RIs provide AZ and instance size flexibility, which offers broader applicability of the RI’s discounted rate.

Q: What are Availability Zone and instance size flexibility?

Availability Zone and instance size flexibility make it easier for you to take advantage of your regional RI’s discounted rate. Availability Zone flexibility applies your RI’s discounted rate to usage in any Availability Zone in a region, while instance size flexibility applies your RI’s discounted rate to usage of any size within an instance family. Let’s say you own an m5.2xlarge Linux/Unix regional RI with default tenancy in US East (N.Virginia). Then this RI’s discounted rate can automatically apply to two m5.xlarge instances in us-east-1a or four m5.large instances in us-east-1b.

Q: What types of RIs provide instance size flexibility?

Linux/Unix regional RIs with the default tenancy provide instance size flexibility. Instance size flexibility is not available on RIs of other platforms such as Windows, Windows with SQL Standard, Windows with SQL Server Enterprise, Windows with SQL Server Web, RHEL, and SLES or G4 instances.

Q: Do I need to take any action to take advantage of Availability Zone and instance size flexibility?

Regional RIs do not require any action to take advantage of Availability Zone and instance size flexibility.

Q: I own zonal RIs how do I assign them to a region?

You can assign your Standard zonal RIs to a region by modifying the scope of the RI from a specific Availability Zone to a region from the EC2 management console or by using the ModifyReservedInstances API.

Q: How do I purchase an RI?

To get started, you can purchase an RI from the EC2 Management Console or by using the AWS CLI. Simply specify the instance type, platform, tenancy, term, payment option, and region or Availability Zone.

Q: Can I purchase an RI for a running instance?

Yes, AWS will automatically apply an RI's discounted rate to any applicable instance usage from the time of purchase. Visit the Getting Started page to learn more.

Q: Can I control which instances are billed at the discounted rate?

No. AWS automatically optimizes which instances are charged at the discounted rate to ensure you always pay the lowest amount. For information about billing, and how it applies to RIs, see Billing Benefits and Payment Options.

Q: How does instance size flexibility work?

EC2 uses the scale shown below, to compare different sizes within an instance family. In the case of instance size flexibility on RIs, this scale is used to apply the discounted rate of RIs to the normalized usage of the instance family. For example, if you have an m5.2xlarge RI that is scoped to a region, then your discounted rate could apply towards the usage of 1 m5.2xlarge or 2 m5.xlarge instances.

Click here to learn more about how instance size flexibility of RIs applies to your EC2 usage. And click here to learn about how instance size flexibility of RIs is presented in the Cost and Usage Report.

Instance Size	Normalization Factor
nano	0.25
micro	0.5
small	1
medium	2
large	4
xlarge	8
2xlarge	16
4xlarge	32
8xlarge	64
9xlarge	72
10xlarge	80
12xlarge	96
16xlarge	128

18xlarge	144
24xlarge	192
32xlarge	256

Q: Can I change my RI during its term?

Yes, you can modify the Availability Zone of the RI, change the scope of the RI from Availability Zone to region (and vice-versa), change the network platform from EC2-VPC to EC2-Classic (and vice versa) or modify instance sizes within the same instance family (on the Linux/Unix platform).

Q: Can I change the instance type of my RI during its term?

Yes, Convertible RIs offer you the option to change the instance type, operating system, tenancy or payment option of your RI during its term. Please refer to the Convertible RI section of the FAQ for additional information.

Q: What are the different payment options for RIs?

You can choose from three payment options when you purchase an RI. With the All Upfront option, you pay for the entire RI term with one upfront payment. With the Partial Upfront option, you make a low upfront payment and are then charged a discounted hourly rate for the instance for the duration of the RI term. The No Upfront option does not require any upfront payment and provides a discounted hourly rate for the duration of the term.

Q: When are RIs activated?

The billing discount and capacity reservation (if applicable) is activated once your payment has successfully been authorized. You can view the status (pending | active | retired) of your RIs on the "Reserved Instances" page of the Amazon EC2 Console.

Q: Do RIs apply to Spot instances or instances running on a Dedicated Host?

No, RIs do not apply to Spot instances or instances running on Dedicated Hosts. To lower the cost of using Dedicated Hosts, purchase Dedicated Host

Reservations.

Q: How do RIs work with Consolidated Billing?

Our system automatically optimizes which instances are charged at the discounted rate to ensure that the consolidated accounts always pay the lowest amount. If you own RIs that apply to an Availability Zone, then only the account which owns the RI will receive the capacity reservation. However, the discount will automatically apply to usage in any account across your consolidated billing family.

Q: Can I get a discount on RI purchases?

Yes, EC2 provides tiered discounts on RI purchases. These discounts are determined based on the total list value (non-discounted price) for the active RIs you have per region. Your total list value is the sum of all expected payments for an RI within the term, including both the upfront and recurring hourly payments. The tier ranges and corresponding discounts are shown alongside.

Tier Range of List Value	Discount on Upfront	Discount on Hourly
Less than $500k	0%	0%
$500k-$4M	5%	5%
$4M-$10M	10%	10%
More than $10M	Call Us	

Q: Can you help me understand how volume discounts are applied to my RI purchases?

Sure. Let's assume that you currently have $400,000 worth of active RIs in the US-east-1 region. Now, if you purchase RIs worth $150,000 in the same region, then the first $100,000 of this purchase would not receive a discount. However, the remaining $50,000 of this purchase would be discounted by 5 percent, so you would only be charged $47,500 for this portion of the purchase over the term based on your payment option.

To learn more, please visit the Understanding Reserved Instance Discount Pricing Tier portion of the Amazon EC2 User Guide.

Q: How do I calculate the list value of an RI?

Here is a sample list value calculation for three-year Partial Upfront Reserved Instances:

3yr Partial Upfront Volume Discount Value in US-East

	Upfront $	Recurring Hourly $	Recurring Hourly Value	List Value
m3.xlarge	$ 1,345	$ 0.060	$ 1,577	$ 2,922
c3.xlarge	$ 1,016	$ 0.045	$ 1,183	$ 2,199

Q: How are volume discounts calculated if I use Consolidated Billing?

If you leverage Consolidated Billing, AWS will use the aggregate total list price of active RIs across all of your consolidated accounts to determine which volume discount tier to apply. Volume discount tiers are determined at the time of purchase, so you should activate Consolidated Billing prior to purchasing RIs to ensure that you benefit from the largest possible volume discount that your consolidated accounts are eligible to receive.

Q: Do Convertible RIs qualify for Volume Discounts?

No, however the value of each Convertible RI that you purchase contributes to your volume discount tier standing.

Q: How do I determine which volume discount tier applies to me?

To determine your current volume discount tier, please consult the Understanding Reserved Instance Discount Pricing Tiers portion of the Amazon EC2 User Guide.

Q: Will the cost of my RIs change, if my future volume qualifies me for other discount tiers?

No. Volume discounts are determined at the time of purchase, therefore the cost of your RIs will continue to remain the same as you qualify for other discount tiers. Any new purchase will be discounted according to your eligible volume discount tier at the time of purchase.

Q: Do I need to take any action at the time of purchase to receive volume discounts?

No, you will automatically receive volume discounts when you use the existing PurchaseReservedInstance API or EC2 Management Console interface to purchase RIs. If you purchase more than $10M worth of RIs contact us about receiving discounts beyond those that are automatically provided.

Reserved Instance Marketplace

Q. What is the Reserved Instance Marketplace?

The Reserved Instance Marketplace is an online marketplace that provides AWS customers the flexibility to sell their Amazon Elastic Compute Cloud (Amazon EC2) Reserved Instances to other businesses and organizations. Customers can also browse the Reserved Instance Marketplace to find an even wider selection of Reserved Instance term lengths and pricing options sold by other AWS customers.

Q. When can I list a Reserved Instance on the Reserved Instance Marketplace?

You can list a Reserved Instance when:

- You've registered as a seller in the Reserved Instance Marketplace.
- You've paid for your Reserved Instance.

- You've owned the Reserved Instance for longer than 30 days.

Q. How will I register as a seller for the Reserved Instance Marketplace?

To register for the Reserved Instance Marketplace, you can enter the registration workflow by selling a Reserved Instance from the EC2 Management Console or setting up your profile from the "Account Settings" page on the AWS portal. No matter the route, you will need to complete the following steps:

1. Start by reviewing the overview of the registration process.
2. Log in to your AWS Account.
3. Enter in the bank account into which you want us to disburse funds. Once you select "Continue", we will set that bank account as the default disbursement option.
4. In the confirmation screen, choose "Continue to Console to Start Listing".

If you exceed $20,000 in sales of Reserved Instances, or plan to sell 50 or more Reserved Instances, you will need to provide tax information before you can list your Reserved Instances. Choose "Continue with Tax Interview". During the tax interview pipeline, you will be prompted to enter your company name, contact name, address, and Tax Identification Number using the TIMS workflow.

Additionally, if you plan to sell Reserved Instances worth more than $50,000 per year you will also need to file a limit increase.

Q. How will I know when I can start selling on the Reserved Instance Marketplace?

You can start selling on the Reserved Instance Marketplace after you have added a bank account through the registration pipeline. Once activation is complete, you will receive a confirmation email. However, it is important to note that you will not be able to receive disbursements until we are able to receive verification from your bank, which may take up to two weeks, depending on the bank you use.

Q. How do I list a Reserved Instance for sale?

To list a Reserved Instance, simply complete these steps in the Amazon EC2 Console:

1. Select the Reserved Instances you wish to sell, and choose "Sell Reserved Instances". If you have not completed the registration process, you will be prompted to register using the registration pipeline.
2. For each Reserved Instance type, set the number of instances you'd like to sell, and the price for the one-time fee you would like to set. Note that you can set the one-time price to different amounts depending on the amount of time remaining so that you don't have to keep adjusting your one-time price if your Reserved Instance doesn't sell quickly. By default you just need to set the current price and we will automatically decrease the one-time price by the same increment each month.
3. Once you have configured your listing, a final confirmation screen will appear. Choose "Sell Reserved Instance".

Q. Which Reserved Instances can I list for sale?

You can list any Reserved Instances that have been active for at least 30 days, and for which we have received payment. Typically, this means that you can list your reservations once they are in the **active** state. It is important to note that if you are an invoice customer, your Reserved Instance can be in the **active** state prior to AWS receiving payment. In this case, your Reserved Instance will not be listed until we have received your payment.

Q. How are listed Reserved Instances displayed to buyers?

Reserved Instances (both third-party and those offered by AWS) that have been listed on the Reserved Instance Marketplace can be viewed in the "Reserved Instances" section of the Amazon EC2 Console. You can also use the DescribeReservedInstancesListings API call.

The listed Reserved Instances are grouped based on the type, term remaining, upfront price, and hourly price. This makes it easier for buyers to find the right Reserved Instances to purchase.

Q. How much of my Reserved Instance term can I list?

You can sell a Reserved Instance for the term remaining, rounded down to the nearest month. For example, if you had 9 months and 13 days remaining, you will list it for sale as a 9-month-term Reserved Instance.

Q. Can I remove my Reserved Instance after I've listed it for sale?

Yes, you can remove your Reserved Instance listings at any point until a sale is pending (meaning a buyer has bought your Reserved Instance and confirmation of payment is pending).

Q. Which pricing dimensions can I set for the Reserved Instances I want to list?

Using the Reserved Instance Marketplace, you can set an upfront price you'd be willing to accept. You cannot set the hourly price (which will remain the same as was set on the original Reserved Instance), and you will not receive any funds collected from payments associated with the hourly prices.

Q. Can I still use my reservation while it is listed on the Reserved Instance Marketplace?

Yes, you will continue to receive the capacity and billing benefit of your reservation until it is sold. Once sold, any running instance that was being charged at the discounted rate will be charged at the On-Demand rate until and unless you purchase a new reservation, or terminate the instance.

Q. Can I resell a Reserved Instance that I purchased from the Reserved Instance Marketplace?

Yes, you can resell Reserved Instances purchased from the Reserved Instance Marketplace just like any other Reserved Instance.

Q. Are there any restrictions when selling Reserved Instances?

Yes, you must have a US bank account to sell Reserved Instances in the Reserved Instance Marketplace. Support for non-US bank accounts will be coming soon. Also, you may not sell Reserved Instances in the US GovCloud region.

Q. Can I sell Reserved Instances purchased from the public volume pricing tiers?

No, this capability is not yet available.

Q. Is there a charge for selling Reserved Instances on the Reserved Instance Marketplace?

Yes, AWS charges a service fee of 12% of the total upfront price of each Reserved Instance you sell in the Reserved Instance Marketplace.

Q. Can AWS sell subsets of my listed Reserved Instances?

Yes, AWS may potentially sell a subset of the quantity of Reserved Instances that you have listed. For example, if you list 100 Reserved instances, we may only have a buyer interested in purchasing 50 of them. We will sell those 50 instances and continue to list your remaining 50 Reserved Instances until and unless you decide not to list them any longer.

Q. How do buyers pay for Reserved Instances that they've purchased?

Payment for completed Reserved Instance sales are done via ACH wire transfers to a US bank account.

Q. When will I receive my money?

Once AWS has received funds from the customer that has bought your reservation, we will disburse funds via wire transfer to the bank account you specified when you registered for the Reserved Instance Marketplace.

Then, we will send you an email notification letting you know that we’ve wired you the funds. Typically, funds will appear in your account within 3-5 days of when your Reserved Instance was been sold.

Q. If I sell my Reserved Instance in the Reserved Instance Marketplace, will I get refunded for the Premium Support I was charged too?

No, you will not receive a pro-rated refund for the upfront portion of the AWS Premium Support Fee.

Q. Will I be notified about Reserved Instance Marketplace activities?

Yes, you will receive a single email once a day that details your Reserved Instance Marketplace activity whenever you create or cancel Reserved Instance listings, buyers purchase your listings, or AWS disburses funds to your bank account.

Q. What information is exchanged between the buyer and seller to help with the transaction tax calculation?

The buyer's city, state, zip+4, and country information will be provided to the seller via a disbursement report. This information will enable sellers to calculate any necessary transaction taxes they need to remit to the government (e.g., sales tax, value-added tax, etc.). The legal entity name of the seller will also be provided on the purchase invoice.

Q. Are there any restrictions on the customers when purchasing third-party Reserved Instances?

Yes, you cannot purchase your own listed Reserved Instances, including those in any of your linked accounts (via Consolidated Billing).

Q. Do I have to pay for Premium Support when purchasing Reserved Instances from the Reserved Instance Marketplace?

Yes, if you are a Premium Support customer, you will be charged for Premium Support when you purchase a Reserved Instance through the Reserved Instance Marketplace.

Spot Instances

Q. What is a Spot Instance?

Spot Instances are spare EC2 capacity that can save you up 90% off of On-Demand prices that AWS can interrupt with a 2-minute notification. Spot uses the same underlying EC2 instances as On-Demand and Reserved Instances, and is best suited for fault-tolerant, flexible workloads. Spot Instances provides an additional option for obtaining compute capacity and can be used along with On-Demand and Reserved Instances.

Q. How is a Spot Instance different than an On-Demand instance or Reserved Instance?

While running, Spot Instances are exactly the same as On-Demand or Reserved instances. The main differences are that Spot Instances typically offer a significant discount off the On-Demand prices, your instances can be interrupted by Amazon EC2 for capacity requirements with a 2-minute notification, and Spot prices adjust gradually based on long term supply and demand for spare EC2 capacity.

See here for more details on Spot Instances.

Q. How do I purchase and start up a Spot instance?

Spot instances can be launched using the same tools you use launch instances today, including AWS Management Console, Auto-Scaling Groups, Run Instances and Spot Fleet. In addition many AWS services support launching Spot instances such as EMR, ECS, Datapipeline, Cloudformation and Batch.

To start up a Spot Instance, you simply need to choose a Launch Template and the number of instances you would like to request.

See here for more details on how to request Spot Instances.

Q. How many Spot Instances can I request?

You can request Spot Instances up to your Spot limit for each region. Note that customers new to AWS might start with a lower limit. To learn more about Spot Instance limits, please refer to the Amazon EC2 User Guide.

If you would like a higher limit, complete the Amazon EC2 instance request form with your use case and your instance increase will be considered. Limit increases are tied to the region they were requested for.

Q. What price will I pay for a Spot Instance?

You pay the Spot price that’s in effect at the beginning of each instance-hour for your running instance. If Spot price changes after you launch the instance, the new price is charged against the instance usage for the subsequent hour.

Q. What is a Spot capacity pool?

A Spot capacity pool is a set of unused EC2 instances with the same instance type, operating system, Availability Zone, and network platform (EC2-Classic or EC2-VPC). Each spot capacity pool can have a different price based on supply and demand.

Q. What are the best practices to use Spot Instances?

We highly recommend using multiple Spot capacity pools to maximize the amount of Spot capacity available to you. EC2 provides built-in automation to find the most cost-effective capacity across multiple Spot capacity pools using EC2 Auto Scaling, EC2 Fleet or Spot Fleet. For more information, please see Spot Best Practices.

Q. How can I determine the status of my Spot request?

You can determine the status of your Spot request via Spot Request Status code and message. You can access Spot Request Status information on the Spot Instance page of the EC2 console of the AWS Management Console, API and CLI. For more information, please visit the Amazon EC2 Developer guide.

Q. Are Spot Instances available for all instance families and sizes and in all regions?

Spot Instances are available in all public AWS regions. Spot is available for nearly all EC2 instance families and sizes, including the newest compute-optimized instances, accelerated graphics, and FPGA instance types. A full list of instance types supported in each region are listed here.

Q. Which operating systems are available as Spot Instances?

Linux/UNIX, Windows Server and Red Hat Enterprise Linux (RHEL) are available. Windows Server with SQL Server is not currently available.

Q. Can I use a Spot Instance with a paid AMI for third-party software (such as IBM's software packages)?

Not at this time.

Q. Can I stop my running Spot Instances?

Yes, you can "stop" your running Spot Instances when they are not needed and keep these stopped instances for later use, instead of terminating instances or cancelling the Spot request. Stop is available for persistent Spot requests.

Q. How can I stop the Spot Instances?

You can stop your Spot Instances by calling the StopInstances API and providing Instance Ids of the Spot Instances similar to stopping your On-Demand Instances. You can also do this through the AWS Management Console by selecting your instance, then clicking Actions > Instance State > Stop.

Q. How can I start the stopped Spot Instances?

You can start the stopped Spot Instances by calling the StartInstances API and providing Instance Ids of the Spot Instances similar to starting your On-Demand Instances. You can also do this through the AWS Management Console by selecting your instance, then clicking Actions > Instance State > Start.

Note: The Spot Instances will start only if Spot capacity is still available within your maximum price. Spot evaluates capacity availability every time whenever you will start the stopped Spot instances.

Q: How can I tell whether I have stopped my Spot Instance or it has been interrupted?

You can tell that the Spot Instance has been stopped by you or interrupted by looking at the Spot Request Status code. This is visible as Spot Request Status on the Spot Requests page of the AWS Management Console or in the DescribeSpotInstanceRequests API response as "status-code" field.

If the Spot request status code is “instance-stopped-by-user”, it means that you have stopped your spot instance.

Q. How will I be charged if my Spot instance is stopped or interrupted?

If your Spot instance is terminated or stopped by Amazon EC2 in the first instance hour, you will not be charged for that usage. However, if you stop or terminate the Spot instance yourself, you will be charged to the nearest second. If the Spot instance is terminated or stopped by Amazon EC2 in any subsequent hour, you will be charged for your usage to the nearest second. If you are running on Windows or Red Hat Enterprise Linux (RHEL) and you stop or terminate the Spot instance yourself, you will be charged for an entire hour.

Q. When would my Spot Instance get interrupted?

Over the last 3 months, 92% of Spot Instance interruptions were from a customer manually terminating the instance because the application had completed its work.

In the circumstance EC2 needs to reclaim your Spot Instance it can be for two possible reasons, with the primary one being Amazon EC2 capacity requirements (e.g. On Demand or Reserved Instance usage). Secondarily, if you have chosen to set a “maximum Spot price” and the Spot price rises above this, your instance will be reclaimed with a two-minute notification. This parameter determines the maximum price you would be willing to pay for a Spot instance hour, and by default, is set at the On-Demand price. As before, you continue to pay the Spot market price, not your maximum price, at the time your instance was running, charged in per-second increments.

Q. What happens to my Spot instance when it gets interrupted?

You can choose to have your Spot instances terminated, stopped or hibernated upon interruption. Stop and hibernate options are available for persistent Spot requests and Spot Fleets with the “maintain” option enabled. By default, your instances are terminated.

Refer to Spot Hibernation to learn more about handling interruptions.

Q. What is the difference between Stop and Hibernate interruption behaviors?

In the case of Hibernate, your instance gets hibernated and the RAM data persisted. In the case of Stop, your instance gets shutdown and RAM is cleared.

In both the cases, data from your EBS root volume and any attached EBS data volumes is persisted. Your private IP address remains the same, as does your elastic IP address (if applicable). The network layer behavior will be similar to that of EC2 Stop-Start workflow. Stop and Hibernate are available for Amazon EBS backed instances only. Local instance storage is not persisted.

Q. What if my EBS root volume is not large enough to store memory state (RAM) for Hibernate?

You should have sufficient space available on your EBS root volume to write data from memory. If the EBS root volume does not enough space, hibernation will fail and the instance will get shutdown instead. Ensure that your EBS volume is large enough to persist memory data before choosing the hibernate option.

Q. What is the benefit if Spot hibernates my instance on interruption?

With hibernate, Spot instances will pause and resume around any interruptions so your workloads can pick up from exactly where they left off. You can use hibernation when your instance(s) need to retain instance state across shutdown-startup cycles, i.e. when your applications running on Spot depend on contextual, business, or session data stored in RAM.

Q. What do I need to do to enable hibernation for my Spot instances?

Refer to Spot Hibernation to learn about enabling hibernation for your Spot instances.

Q. Do I have to pay for hibernating my Spot instance?

There is no additional charge for hibernating your instance beyond the EBS storage costs and any other EC2 resources you may be using. You are not charged instance usage fees once your instance is hibernated.

Q. Can I resume a hibernated instance?

No, you will not be able to resume a hibernated instance directly. Hibernate-resume cycle is controlled by Amazon EC2. If an instance is hibernated by Spot, it will be resumed by Amazon EC2 when the capacity becomes available.

Q. Which instances and operating systems support hibernation?

Spot Hibernation is currently supported for Amazon Linux AMIs, Ubuntu and Microsoft Windows operating systems running on any instance type across C3, C4, C5, M4, M5, R3, R4 instances with memory (RAM) size less than 100 GiB.

To review the list of supported OS versions, refer to Spot Hibernation.

Q. How am I charged if Spot price changes while my instance is running?

You will pay the price per instance-hour set at the beginning of each instance-hour for the entire hour, billed to the nearest second.

Q. Where can I see my usage history for Spot instances and see how much I was billed?

The AWS Management Console makes a detailed billing report available which shows Spot instance start and termination/stop times for all instances. Customers can check the billing report against historical Spot prices via the API to verify that the Spot price they were billed is correct.

Q: Are Spot blocks (Fixed Duration Spot instances) ever interrupted?

Spot blocks are designed not to be interrupted and will run continuously for the duration you select, independent of Spot market price. In rare situations, Spot blocks may be interrupted due to AWS capacity needs. In these cases, we will provide a two-minute warning before we terminate your instance (termination notice), and you will not be charged for the affected instance(s).

Q. What is a Spot fleet?

A Spot Fleet allows you to automatically request and manage multiple Spot instances that provide the lowest price per unit of capacity for your cluster or application, like a batch processing job, a Hadoop workflow, or an HPC grid computing job. You can include the instance types that your application can use. You define a target capacity based on your application needs (in units including instances, vCPUs, memory, storage, or network throughput) and update the target capacity after the fleet is launched. Spot fleets enable you to launch and maintain the target capacity, and to automatically request resources to replace any that are disrupted or manually terminated. Learn more about Spot fleets.

Q. Is there any additional charge for making Spot Fleet requests

No, there is no additional charge for Spot Fleet requests.

Q. What limits apply to a Spot Fleet request?

Visit the Spot Fleet Limits section of the Amazon EC2 User Guide to learn about the limits that apply to your Spot Fleet request.

Q. What happens if my Spot Fleet request tries to launch Spot instances but exceeds my regional Spot request limit?

If your Spot Fleet request exceeds your regional Spot instance request limit, individual Spot instance requests will fail with a Spot request limit exceeded request status. Your Spot Fleet request’s history will show any Spot request limit errors that the Fleet request received. Visit the Monitoring Your Spot Fleet section of the Amazon EC2 User Guide to learn how to describe your Spot Fleet request's history.

Q. Are Spot fleet requests guaranteed to be fulfilled?

No. Spot fleet requests allow you to place multiple Spot Instance requests simultaneously, and are subject to the same availability and prices as a single Spot Instance request. For example, if no resources are available for the instance types listed in your Spot Fleet request, we may be unable to fulfill your request partially or in full. We recommend you to include all the possible instance types and availability zones that are suitable for your workloads in the Spot Fleet.

Q. Can I submit a multi-Availability Zone Spot Fleet request?

Yes, visit the Spot Fleet Examples section of the Amazon EC2 User Guide to learn how to submit a multi-Availability Zone Spot Fleet request.

Q. Can I submit a multi-region Spot Fleet request?

No, we do not support multi-region Fleet requests.

Q. How does Spot Fleet allocate resources across the various Spot Instance pools specified in the launch specifications?

The RequestSpotFleet API provides three allocation strategies: capacity-optimized, lowestPrice and diversified. The capacity-optimized allocation strategy attempts to provision Spot Instances from the most available Spot Instance pools by analyzing capacity metrics. This strategy is a good choice for workloads that have a higher cost of interruption such as big data and analytics, image and media rendering, machine learning, and high performance computing.

The lowestPrice strategy allows you to provision your Spot Fleet resources in instance pools that provide the lowest price per unit of capacity at the time of the request. The diversified strategy allows you to provision your Spot Fleet resources across multiple Spot Instance pools. This enables you to maintain your fleet's target capacity and increase your application's availability as Spot capacity fluctuates.

Running your application's resources across diverse Spot Instance pools also allows you to further reduce your fleet's operating costs over time. Visit the Amazon EC2 User Guide to learn more.

Q. Can I tag a Spot Fleet request?

You can request to launch Spot Instances with tags via Spot Fleet. The Fleet by itself cannot be tagged.

Q. How can I see which Spot fleet owns my Spot Instances?

You can identify the Spot Instances associated with your Spot Fleet by describing your fleet request. Fleet requests are available for 48 hours after all its Spot Instances have been terminated. See the Amazon EC2 User Guide to learn how to describe your Spot Fleet request.

Q. Can I modify my Spot Fleet request?

Yes, you can modify the target capacity of your Spot Fleet request. You may need to cancel the request and submit a new one to change other request configuration parameters.

Q. Can I specify a different AMI for each instance type that I want to use?

Yes, simply specify the AMI you'd like to use in each launch specification you provide in your Spot Fleet request.

Q. Can I use Spot Fleet with Elastic Load Balancing, Auto Scaling, or Elastic MapReduce?

You can use Auto Scaling features with Spot Fleet such as target tracking, health checks, cloudwatch metrics etc and can attach instances to your Elastic load balancers (both classic and application load balancers). Elastic MapReduce has a feature named "Instance fleets" that provides capabilities similar to Spot Fleet.

Q. Does a Spot Fleet request terminate Spot Instances when they are no longer running in the lowest priced or capacity-optimized Spot pools and relaunch them?

No, Spot Fleet requests do not automatically terminate and re-launch instances while they are running. However, if you terminate a Spot Instance, Spot Fleet will replenish it with a new Spot Instance in the new lowest priced pool or capacity-optimized pool based on your allocation strategy.

Q: Can I use stop or Hibernation interruption behaviors with Spot Fleet?

Yes, stop-start and hibernate-resume are supported with Spot Fleet with "maintain" fleet option enabled.

Platform

Amazon Time Sync Service | Availability zones | Cluster instances | Hardware information | Micro instances | Nitro Hypervisor | Optimize CPUs

Amazon Time Sync Service

Q. How do I use this service?

The service provides an NTP endpoint at a link-local IP address (169.254.169.123) accessible from any instance running in a VPC. Instructions for configuring NTP clients are available for Linux and Windows.

Q. What are the key benefits of using this service?

A consistent and accurate reference time source is crucial for many applications and services. The Amazon Time Sync Service provides a time reference that can be securely accessed from an instance without requiring VPC configuration changes and updates. It is built on Amazon's proven network infrastructure and uses redundant reference time sources to ensure high accuracy and availability.

Q. Which instance types are supported for this service?

All instances running in a VPC can access the service.

Availability zones

Q: How isolated are Availability Zones from one another?

Each Availability Zone runs on its own physically distinct, independent infrastructure, and is engineered to be highly reliable. Common points of failures like generators and cooling equipment are not shared across Availability Zones. Additionally, they are physically separate, such that even extremely uncommon disasters such as fires, tornados or flooding would only affect a single Availability Zone.

Q: Is Amazon EC2 running in more than one region?

Yes. Please refer to Regional Products and Services for more details of our product and service availability by region.

Q: How can I make sure that I am in the same Availability Zone as another developer?

We do not currently support the ability to coordinate launches into the same Availability Zone across AWS developer accounts. One Availability Zone name (for example, us-east-1a) in two AWS customer accounts may relate to different physical Availability Zones.

Q: If I transfer data between Availability Zones using public IP addresses, will I be charged twice for Regional Data Transfer (once because it's across zones, and a second time because I'm using public IP addresses)?

No. Regional Data Transfer rates apply if at least one of the following is true, but is only charged once for a given instance even if both are true:

- The other instance is in a different Availability Zone, regardless of which type of address is used.
- Public or Elastic IP addresses are used, regardless of which Availability Zone the other instance is in.

Cluster instances

Q. What is a Cluster Compute Instance?

Cluster Compute Instances combine high compute resources with a high performance networking for High Performance Compute (HPC) applications and other demanding network-bound applications. Cluster Compute Instances provide similar functionality to other Amazon EC2 instances but have been specifically engineered to provide high performance networking.

Amazon EC2 cluster placement group functionality allows users to group Cluster Compute Instances in clusters – allowing applications to get the low-latency network performance necessary for tightly-coupled node-to-node communication typical of many HPC applications. Cluster Compute Instances also provide significantly increased network throughput both within the Amazon EC2 environment and to the Internet. As a result, these instances are also well suited for customer applications that need to perform network-intensive operations.

Learn more about use of this instance type for HPC applications.

Q. What kind of network performance can I expect when I launch instances in cluster placement group?

The bandwidth an EC2 instance can utilize in a cluster placement group depends on the instance type and its networking performance specification. Inter-instance traffic within the same region can utilize 5 Gbps for single-flow and up to 25 Gbps for multi-flow traffic. When launched in a placement group, select EC2 instances can utilize up to 10 Gbps for single-flow traffic.

Q. What is a Cluster GPU Instance?

Cluster GPU Instances provide general-purpose graphics processing units (GPUs) with proportionally high CPU and increased network performance for applications benefiting from highly parallelized processing that can be accelerated by GPUs using the CUDA and OpenCL programming models. Common applications include modeling and simulation, rendering and media processing.

Cluster GPU Instances give customers with HPC workloads an option beyond Cluster Compute Instances to further customize their high performance clusters in the cloud for applications that can benefit from the parallel computing power of GPUs.

Cluster GPU Instances use the same cluster placement group functionality as Cluster Compute Instances for grouping instances into clusters – allowing applications to get the low-latency, high bandwidth network performance required for tightly-coupled node-to-node communication typical of many HPC applications.

Learn more about HPC on AWS.

Q. What is a High Memory Cluster Instance?

High Memory Cluster Instances provide customers with large amounts of memory and CPU capabilities per instance in addition to high network capabilities. These instance types are ideal for memory intensive workloads including in-memory analytics systems, graph analysis and many science and engineering applications

High Memory Cluster Instances use the same cluster placement group functionality as Cluster Compute Instances for grouping instances into clusters – allowing applications to get the low-latency, high bandwidth network performance required for tightly-coupled node-to-node communication typical of many HPC and other network intensive applications.

Q. Does use of Cluster Compute and Cluster GPU Instances differ from other Amazon EC2 instance types?

Cluster Compute and Cluster GPU Instances use differs from other Amazon EC2 instance types in two ways.

First, Cluster Compute and Cluster GPU Instances use Hardware Virtual Machine (HVM) based virtualization and run only Amazon Machine Images (AMIs) based on HVM virtualization. Paravirtual Machine (PVM) based AMIs used with other Amazon EC2 instance types cannot be used with Cluster Compute or Cluster GPU Instances.

Second, in order to fully benefit from the available low latency, full bisection bandwidth between instances, Cluster Compute and Cluster GPU Instances must be launched into a cluster placement group through the Amazon EC2 API or AWS Management Console.

Q. What is a cluster placement group?

A cluster placement group is a logical entity that enables creating a cluster of instances by launching instances as part of a group. The cluster of instances then provides low latency connectivity between instances in the group. Cluster placement groups are created through the Amazon EC2 API or AWS Management Console.

Q. Are all features of Amazon EC2 available for Cluster Compute and Cluster GPU Instances?

Currently, Amazon DevPay is not available for Cluster Compute or Cluster GPU Instances.

Q. Is there a limit on the number of Cluster Compute or Cluster GPU Instances I can use and/or the size of cluster I can create by launching Cluster Compute Instances or Cluster GPU into a cluster placement group?

There is no limit specific for Cluster Compute Instances. For Cluster GPU Instances, you can launch 2 Instances on your own. If you need more capacity, please complete the Amazon EC2 instance request form (selecting the appropriate primary instance type).

Q. Are there any ways to optimize the likelihood that I receive the full number of instances I request for my cluster via a cluster placement group?

We recommend that you launch the minimum number of instances required to participate in a cluster in a single launch. For very large clusters, you should launch multiple placement groups, e.g. two placement groups of 128 instances, and combine them to create a larger, 256 instance cluster.

Q. Can Cluster GPU and Cluster Compute Instances be launched into a single cluster placement group?

While it may be possible to launch different cluster instance types into a single placement group, at this time we only support homogenous placement groups.

Q. If an instance in a cluster placement group is stopped then started again, will it maintain its presence in the cluster placement group?

Yes. A stopped instance will be started as part of the cluster placement group it was in when it stopped. If capacity is not available for it to start within its cluster placement group, the start will fail.

Hardware information

Q: What kind of hardware will my application stack run on?

Visit Amazon EC2 Instance Type for a list of EC2 instances available by region.

Q: How does EC2 perform maintenance?

AWS regularly performs routine hardware, power and network maintenance without disrupting customer instances. To achieve this we employ a combination of tools and methods across the entire AWS Global infrastructure, such as redundant and concurrently maintainable systems, as well as live system updates and migration. For example, in these cases - Example 1, Example 2 - EC2 used live system updates to perform the required security maintenance non-disruptively for over 90% of EC2 Instances, with each maintenance completing in less than two seconds. AWS continuously invests in technology and processes to complete routine maintenance ever more safely and quickly, often with no disruption to customer instances.

Q: How do I select the right instance type?

Amazon EC2 instances are grouped into 5 families: General Purpose, Compute Optimized, Memory Optimized, Storage Optimized and Accelerated Computing instances. General Purpose Instances have memory to CPU ratios suitable for most general purpose applications and come with fixed performance (M5, M4) or burstable performance (T2); Compute Optimized instances (C5, C4) have proportionally more CPU resources than memory (RAM) and are well suited for scale out compute-intensive applications and High Performance Computing (HPC) workloads; Memory Optimized Instances (X1e, X1, R4) offer larger memory sizes for memory-intensive applications, including database and memory caching applications; Accelerating Computing instances (P3, P2, G3, F1) take advantage of the parallel processing capabilities of NVIDIA Tesla GPUs for high performance computing and machine/deep learning; GPU Graphics instances (G3) offer high-performance 3D graphics capabilities for applications using OpenGL and DirectX; F1 instances deliver Xilinx FPGA-based reconfigurable computing; Storage Optimized Instances (H1, I3, I3en, D2) that provide very high, low latency, I/O capacity using SSD-based local instance storage for I/O-intensive applications, with D2 or H1, the dense-storage and HDD-storage instances, provide local high storage density and sequential I/O performance for data warehousing, Hadoop and other data-intensive applications. When choosing instance types, you should consider the characteristics of your application with regards to resource utilization (i.e. CPU, Memory, Storage) and select the optimal instance family and instance size.

Q: What is an “EC2 Compute Unit” and why did you introduce it?

Transitioning to a utility computing model fundamentally changes how developers have been trained to think about CPU resources. Instead of purchasing or leasing a particular processor to use for several months or years, you are renting capacity by the hour. Because Amazon EC2 is built on commodity hardware, over time there may be several different types of physical hardware underlying EC2 instances. Our goal is to provide a consistent amount of CPU capacity no matter what the actual underlying hardware.

Amazon EC2 uses a variety of measures to provide each instance with a consistent and predictable amount of CPU capacity. In order to make it easy for developers to compare CPU capacity between different instance types, we have defined an Amazon EC2 Compute Unit. The amount of CPU that is allocated to a particular instance is expressed in terms of these EC2 Compute Units. We use several benchmarks and tests to manage the consistency and predictability of the performance from an EC2 Compute Unit. The EC2 Compute Unit (ECU) provides the relative measure of the integer processing power of an Amazon EC2 instance. Over time, we may add or substitute measures that go into the definition of an EC2 Compute Unit, if we find metrics that will give you a clearer picture of compute capacity.

Q: How does EC2 ensure consistent performance of instance types over time?

AWS conducts yearly performance benchmarking of Linux and Windows compute performance on EC2 instance types. Benchmarking results, a test suite that customers can use to conduct independent testing, and guidance on expected performance variance is available under NDA for M,C,R, T and z1d instances; please contact your sales representative to request them.

Q: What is the regional availability of Amazon EC2 instance types?

For a list of all instances and regional availability, visit Amazon EC2 Pricing.

Micro instances

Q. How much compute power do Micro instances provide?

Micro instances provide a small amount of consistent CPU resources and allow you to burst CPU capacity up to 2 ECUs when additional cycles are available. They are well suited for lower throughput applications and web sites that consume significant compute cycles periodically but very little CPU at other times for background processes, daemons, etc. Learn more about use of this instance type.

Q. How does a Micro instance compare in compute power to a Standard Small instance?

At steady state, Micro instances receive a fraction of the compute resources that Small instances do. Therefore, if your application has compute-intensive or steady state needs we recommend using a Small instance (or larger, depending on your needs). However, Micro instances can periodically burst up to 2 ECUs (for short periods of time). This is double the number of ECUs available from a Standard Small instance. Therefore, if you have a relatively low throughput application or web site with an occasional need to consume significant compute cycles, we recommend using Micro instances.

Q. How can I tell if an application needs more CPU resources than a Micro instance is providing?

The CloudWatch metric for CPU utilization will report 100% utilization if the instance bursts so much that it exceeds its available CPU resources during that CloudWatch monitored minute. CloudWatch reporting 100% CPU utilization is your signal that you should consider scaling – manually or via Auto Scaling – up to a larger instance type or scale out to multiple Micro instances.

Q. Are all features of Amazon EC2 available for Micro instances?

Currently Amazon DevPay is not available for Micro instances.

Nitro Hypervisor

Q. What is the Nitro Hypervisor?

The launch of C5 instances introduced a new hypervisor for Amazon EC2, the Nitro Hypervisor. As a component of the Nitro system, the Nitro Hypervisor primarily provides CPU and memory isolation for EC2 instances. VPC networking and EBS storage resources are implemented by dedicated hardware components, Nitro Cards that are part of all current generation EC2 instance families. The Nitro Hypervisor is built on core Linux Kernel-based Virtual Machine (KVM) technology, but does not include general-purpose operating system components.

Q. How does the Nitro Hypervisor benefit customers?

The Nitro Hypervisor provides consistent performance and increased compute and memory resources for EC2 virtualized instances by removing host system software components. It allows AWS to offer larger instance sizes (like c5.18xlarge) that provide practically all of the resources from the server to customers. Previously, C3 and C4 instances each eliminated software components by moving VPC and EBS functionality to hardware designed and built by AWS. This hardware enables the Nitro Hypervisor to be very small and uninvolved in data processing tasks for networking and storage.

Q. Will all EC2 instances use the Nitro Hypervisor?

Eventually all new instance types will use the Nitro Hypervisor, but in the near term, some new instance types will use Xen depending on the requirements of the platform.

Q. Will AWS continue to invest in its Xen-based hypervisor?

Yes. As AWS expands its global cloud infrastructure, EC2's use of its Xen-based hypervisor will also continue to grow. Xen will remain a core component of EC2 instances for the foreseeable future. AWS is a founding member of the Xen Project since its establishment as a Linux Foundation Collaborative Project and remains an active participant on its Advisory Board. As AWS expands its global cloud infrastructure, EC2's Xen-based hypervisor also continues to grow. Therefore EC2's investment in Xen continues to grow, not shrink

Q. How many EBS volumes and Elastic Network Interfaces (ENIs) can be attached to instances running on the Nitro Hypervisor?

Instances running on the Nitro Hypervisor support a maximum of 27 additional PCI devices for EBS volumes and VPC ENIs. Each EBS volume or VPC ENI uses a PCI device. For example, if you attach 3 additional network interfaces to an instance that uses the Nitro Hypervisor, you can attach up to 24 EBS volumes to that instance.

Q. Will the Nitro Hypervisor change the APIs used to interact with EC2 instances?

No, all the public facing APIs for interacting with EC2 instances that run using the Nitro Hypervisor will remain the same. For example, the "hypervisor" field of the DescribeInstances response, which will continue to report "xen" for all EC2 instances, even those running under the Nitro Hypervisor. This field may be removed in a future revision of the EC2 API.

Q. Which AMIs are supported on instances that use the Nitro Hypervisor?

EBS backed HVM AMIs with support for ENA networking and booting from NVMe storage can be used with instances that run under the Nitro Hypervisor. The latest Amazon Linux AMI and Windows AMIs provided by Amazon are supported, as are the latest AMI of Ubuntu, Debian, Red Hat Enterprise Linux, SUSE Enterprise Linux, CentOS, and FreeBSD.

Q. Will I notice any difference between instances using Xen hypervisor and those using the Nitro Hypervisor?

Yes. For example, instances running under the Nitro Hypervisor boot from EBS volumes using an NVMe interface. Instances running under Xen boot from an emulated IDE hard drive, and switch to the Xen paravirtualized block device drivers.

Operating systems can identify when they are running under a hypervisor. Some software assumes that EC2 instances will run under the Xen hypervisor and rely on this detection. Operating systems will detect they are running under KVM when an instance uses the Nitro Hypervisor, so the process to identify EC2 instances should be used to identify EC2 instances that run under both hypervisors.

All the features of EC2 such as Instance Metadata Service work the same way on instances running under both Xen and the Nitro Hypervisor. The majority of applications will function the same way under both Xen and the Nitro Hypervisor as long as the operating system has the needed support for ENA networking and NVMe storage.

Q. How are instance reboot and termination EC2 API requests implemented by the Nitro Hypervisor?

The Nitro Hypervisor signals the operating system running in the instance that it should shut down cleanly by industry standard ACPI methods. For Linux instances, this requires that acpid be installed and functioning correctly. If acpid is not functioning in the instance, termination events will be delayed by multiple minutes and will then execute as a hard reset or power off.

Q. How do EBS volumes behave when accessed by NVMe interfaces?

There are some important differences in how operating system NVMe drivers behave compared to Xen paravirtual (PV) block drivers.

First, the NVMe device names used by Linux based operating systems will be different than the parameters for EBS volume attachment requests and block device mapping entries such as /dev/xvda and /dev/xvdf. NVMe devices are enumerated by the operating system as /dev/nvme0n1, /dev/nvme1n1, and so on. The NVMe device names are not persistent mappings to volumes, therefore other methods like file system UUIDs or labels should be used when configuring the automatic mounting of file systems or other startup activities. When EBS volumes are accessed via the NVMe interface, the EBS volume ID is available via the controller serial number and the device name specified in EC2 API requests is provided by an NVMe vendor extension to the Identify Controller command. This enables backward compatible symbolic links to be created by a utility script. For more information see the EC2 documentation on device naming and NVMe based EBS volumes.

Second, by default the NVMe drivers included in most operating systems implement an I/O timeout. If an I/O does not complete in an implementation specific amount of time, usually tens of seconds, the driver will attempt to cancel the I/O, retry it, or return an error to the component that issued the I/O. The Xen PV block device interface does not time out I/O, which can result in processes that cannot be terminated if it is waiting for I/O. The Linux NVMe driver behavior can be modified by specifying a higher value for the nvme.io timeout kernel module parameter.

Third, the NVMe interface can transfer much larger amounts of data per I/O, and in some cases may be able to support more outstanding I/O requests, compared to the Xen PV block interface. This can cause higher I/O latency if very large I/Os or a large number of I/O requests are issued to volumes designed to support throughput workloads like EBS Throughput Optimized HDD (st1) and Cold HDD (sc1) volumes. This I/O latency is normal for throughput optimized volumes in these scenarios, but may cause I/O timeouts in NVMe drivers. The I/O timeout can be adjusted in the Linux driver by specifying a larger value for the nvme_core.io_timeout kernel module parameter.

Optimize CPUs

Q: What is Optimize CPUs?

Optimize CPUs gives you greater control of your EC2 instances on two fronts. First, you can specify a custom number of vCPUs when launching new instances to save on vCPU-based licensing costs. Second, you can disable Intel Hyper-Threading Technology (Intel HT Technology) for workloads that perform well with single-threaded CPUs, such as certain high-performance computing (HPC) applications.

Q: Why should I use Optimize CPUs feature?

You should use Optimize CPUs if:

- You are running EC2 workloads that are not compute bound and are incurring vCPU-based licensing costs. By launching instances with custom number of vCPUs you may be able to optimize your licensing spend.

- You are running workloads that will benefit from disabling hyper-threading on EC2 instances.

Q: How will the CPU optimized instances be priced?

CPU optimized instances will be priced the same as equivalent full-sized instance.

Q: How will my application performance change when using Optimize CPUs on EC2?

Your application performance change with Optimize CPUs will be largely dependent on the workloads you are running on EC2. We encourage you to benchmark your application performance with Optimize CPUs to arrive at the right number of vCPUs and optimal hyper-threading behavior for your application.

Q: Can I use Optimize CPUs on EC2 Bare Metal instance types (such as i3.metal)?

No. You can use Optimize CPUs with only virtualized EC2 instances.

Q. How can I get started with using Optimize CPUs for EC2 Instances?

For more information on how to get started with Optimize CPUs and supported instance types, please visit the Optimize CPUs documentation page here.

Workloads

Amazon EC2 running IBM | Amazon EC2 running Microsoft Windows and other third-party software

Amazon EC2 running IBM

Q. How am I billed for my use of Amazon EC2 running IBM?

You pay only for what you use and there is no minimum fee. Pricing is per instance-hour consumed for each instance type. Partial instance-hours consumed are billed as full hours. Data transfer for Amazon EC2 running IBM is billed and tiered separately from Amazon EC2. There is no Data Transfer charge between two Amazon Web Services within the same region (i.e. between Amazon EC2 US West and another AWS service in the US West). Data transferred between AWS services in different regions will be charged as Internet Data Transfer on both sides of the transfer.

For Amazon EC2 running IBM pricing information, please visit the pricing section on the Amazon EC2 running IBM detail page.

Q. Can I use Amazon DevPay with Amazon EC2 running IBM?

No, you cannot use DevPay to bundle products on top of Amazon EC2 running IBM at this time.

Amazon EC2 running Microsoft Windows and other third-party software

Q. Can I use my existing Windows Server license with EC2?

Yes you can. After you’ve imported your own Windows Server machine images using the ImportImage tool, you can launch instances from these machine images on EC2 Dedicated Hosts and effectively manage instances and report usage. Microsoft typically requires that you track usage of your licenses against physical resources such as sockets and cores and Dedicated Hosts helps you to do this. Visit the Dedicated Hosts detail page for more information on how to use your own Windows Server licenses on Amazon EC2 Dedicated Hosts.

Q. What software licenses can I bring to the Windows environment?

Specific software license terms vary from vendor to vendor. Therefore, we recommend that you check the licensing terms of your software vendor to determine if your existing licenses are authorized for use in Amazon EC2.

Amazon EC2 Auto Scaling FAQs

General

Q: What is Amazon EC2 Auto Scaling?

Amazon EC2 Auto Scaling is a fully managed service designed to launch or terminate Amazon EC2 instances automatically to help ensure you have the correct number of Amazon EC2 instances available to handle the load for your application. Amazon EC2 Auto Scaling helps you maintain application availability through fleet management for EC2 instances, which detects and replaces unhealthy instances, and by scaling your Amazon EC2 capacity up or down automatically according to conditions you define. You can use Amazon EC2 Auto Scaling to automatically increase the number of Amazon EC2 instances during demand spikes to maintain performance and decrease capacity during lulls to reduce costs.

Q. When should I use Amazon EC2 Auto Scaling vs. AWS Auto Scaling?

You should use AWS Auto Scaling to manage scaling for multiple resources across multiple services. AWS Auto Scaling lets you define dynamic scaling policies for multiple EC2 Auto Scaling groups or other resources using predefined scaling strategies. Using AWS Auto Scaling to configure scaling policies for all of the scalable resources in your application is faster than managing scaling policies for each resource via its individual service console. It's also easier, as AWS Auto Scaling includes predefined scaling strategies that simplify the setup of scaling policies. You should also use AWS Auto Scaling if you want to create predictive scaling for EC2 resources.

You should use EC2 Auto Scaling if you only need to scale Amazon EC2 Auto Scaling groups, or if you are only interested in maintaining the health of your EC2 fleet. You should also use EC2 Auto Scaling if you need to create or configure Amazon EC2 Auto Scaling groups, or if you need to set up scheduled or step scaling policies (as AWS Auto Scaling supports only target tracking scaling policies).

EC2 Auto Scaling groups must be created and configured outside of AWS Auto Scaling, such as through the EC2 console, Auto Scaling API or via CloudFormation. AWS Auto Scaling can help you configure dynamic scaling policies for your existing EC2 Auto Scaling groups.

Q: What are the benefits of using Amazon EC2 Auto Scaling?

Amazon EC2 Auto Scaling helps to maintain your Amazon EC2 instance availability. Whether you are running one Amazon EC2 instance or thousands, you can use Amazon EC2 Auto Scaling to detect impaired Amazon EC2 instances, and replace the instances without intervention. This ensures that your application has the compute capacity that you expect. You can use Amazon EC2 Auto Scaling to automatically scale your Amazon EC2 fleet by following the demand curve for your applications, reducing the need to manually provision Amazon EC2 capacity in advance. For example, you can set a condition to add new Amazon EC2 instances in increments to the ASG when the average utilization of your Amazon EC2 fleet is high; and similarly, you can set a condition to remove instances in increments when CPU utilization is low. You can also use Amazon CloudWatch to send alarms to trigger scaling activities and Elastic Load Balancing (ELB) to distribute traffic to your instances within the ASG. If you have predictable load changes, you can set a schedule through Amazon EC2 Auto Scaling to plan your scaling activities. Amazon EC2 Auto Scaling enables you to run your Amazon EC2 fleet at optimal utilization.

Q: What is fleet management and how is it different from dynamic scaling?

If your application runs on Amazon EC2 instances, then you have what's referred to as a 'fleet'. *Fleet management* refers to the functionality that automatically replaces unhealthy instances and maintains your fleet at the desired capacity. Amazon EC2 Auto Scaling fleet management ensures that your application is able to receive traffic and that the instances themselves are working properly. When Auto Scaling detects a failed health check, it can replace the instance automatically.

The *dynamic scaling* capabilities of Amazon EC2 Auto Scaling refers to the functionality that automatically increases or decreases capacity based on load or other metrics. For example, if your CPU spikes above 80% (and you have an alarm setup) Amazon EC2 Auto Scaling can add a new instance dynamically.

Q: What is target tracking?

Target tracking is a new type of scaling policy that you can use to set up dynamic scaling for your application in just a few simple steps. With target tracking, you select a load metric for your application, such as CPU utilization or request count, set the target value, and Amazon EC2 Auto Scaling adjusts the number of EC2 instances in your ASG as needed to maintain that target. It acts like a home thermostat, automatically adjusting the system to keep the environment at your desired temperature. For example, you can configure target tracking to keep CPU utilization for your fleet of web servers at 50%. From there, Amazon EC2 Auto Scaling launches or terminates EC2 instances as required to keep the average CPU utilization at 50%.

Q: What is an EC2 Auto Scaling group (ASG)?

An Amazon EC2 Auto Scaling group (ASG) contains a collection of EC2 instances that share similar characteristics and are treated as a logical grouping for the purposes of fleet management and dynamic scaling. For example, if a single application operates across multiple instances, you might want to increase the number of instances in that group to improve the performance of the application, or decrease the number of instances to reduce costs when demand is low. Amazon EC2 Auto Scaling will automaticallly adjust the number of instances in the group to maintain a fixed number of instances even if a instance becomes unhealthy, or based on criteria that you specify. You can find more information about ASG in the Amazon EC2 Auto Scaling User Guide.

Q: What happens to my Amazon EC2 instances if I delete my ASG?

If you have an EC2 Auto Scaling group (ASG) with running instances and you choose to delete the ASG, the instances will be terminated and the ASG will be deleted.

Q: How do I know when EC2 Auto Scaling is launching or terminating the EC2 instances in an EC2 Auto Scaling group?

When you use Amazon EC2 Auto Scaling to scale your applications automatically, it is useful to know when EC2 Auto Scaling is launching or terminating the EC2 instances in your EC2 Auto Scaling group. Amazon SNS coordinates and manages the delivery or sending of notifications to subscribing clients or endpoints. You can configure EC2 Auto Scaling to send an SNS notification whenever your EC2 Auto Scaling group scales. Amazon SNS can deliver notifications as HTTP or HTTPS POST, email (SMTP, either plain-text or in JSON format), or as a message posted to an Amazon SQS queue. For example, if you configure your EC2 Auto Scaling group to use the autoscaling: EC2_INSTANCE_TERMINATE notification type, and your EC2 Auto Scaling group terminates an instance, it sends an email notification. This email contains the details of the terminated instance, such as the instance ID and the reason that the instance was terminated.

For more information read Getting SNS Notifications when your EC2 Auto Scaling Group Scales.

Q: What is a launch configuration?

A launch configuration is a template that an EC2 Auto Scaling group uses to launch EC2 instances. When you create a launch configuration, you specify information for the instances such as the ID of the Amazon Machine Image (AMI), the instance type, a key pair, one or more security groups, and a block device mapping. If you've launched an EC2 instance before, you specified the same information in order to launch the instance. When you create an EC2 Auto Scaling group, you must specify a launch configuration. You can specify your launch configuration with multiple EC2 Auto Scaling groups. However, you can only specify one launch configuration for an EC2 Auto Scaling group at a time, and you can't modify a launch configuration after you've created it. Therefore, if you want to change the launch configuration for your EC2 Auto Scaling group, you must create a launch configuration and then update your EC2 Auto Scaling group with the new launch configuration. When you change the launch configuration for your EC2 Auto Scaling group, any new instances are launched using the new configuration parameters, but existing instances are not affected.

You can see the launch configurations section of the EC2 Auto Scaling User Guide for more details.

Q: How many instances can an EC2 Auto Scaling group have?

You can have as many instances in your EC2 Auto Scaling group as your EC2 quota allows.

Q: What happens if a scaling activity causes me to reach my Amazon EC2 limit of instances?

Amazon EC2 Auto Scaling cannot scale past the Amazon EC2 limit of instances that you can run. If you need more Amazon EC2 instances, complete the Amazon EC2 instance request form.

Q: Can EC2 Auto Scaling groups span multiple AWS regions?

EC2 Auto Scaling groups are regional constructs. They can span Availability Zones, but not AWS regions.

Q: How can I implement changes across multiple instances in an EC2 Auto Scaling group?

You can use AWS CodeDeploy or CloudFormation to orchestrate code changes to multiple instances in your EC2 Auto Scaling group.

Q: If I have data installed in an EC2 Auto Scaling group, and a new instance is dynamically created later, is the data copied over to the new instances?

Data is not automatically copied from existing instances to new instances. You can use lifecycle hooks to copy the data, or an Amazon RDS database including replicas.

Q: When I create an EC2 Auto Scaling group from an existing instance, does it create a new AMI (Amazon Machine Image)?

When you create an Auto Scaling group from an existing instance, it does not create a new AMI. For more information see Creating an Auto Scaling Group Using an EC2 Instance.

Q: How does Amazon EC2 Auto Scaling balance capacity?

Balancing resources across Availability Zones is a best practice for well-architected applications, as this greatly increases aggregate system availability. Amazon EC2 Auto Scaling automatically balances EC2 instances across zones when you configure multiple zones in your EC2 Auto Scaling group settings. Amazon EC2 Auto Scaling always launches new instances such that they are balanced between zones as evenly as possible across the entire fleet. What's more, Amazon EC2 Auto Scaling only launches into Availability Zones in which there is available capacity for the requested instance type.

Q: What are lifecycle hooks?

Lifecycle hooks let you take action before an instance goes into service or before it gets terminated. This can be especially useful if you are not baking your software environment into an Amazon Machine Image (AMI). For example, launch hooks can perform software configuration on an instance to ensure that it's fully prepared to handle traffic before Amazon EC2 Auto Scaling proceeds to connect it to your load balancer. One way to do this is by connecting the launch hook to an AWS Lambda function that invokes RunCommand on the instance. Terminate hooks can be useful for collecting important data from an instance before it goes away. For example, you could use a terminate hook to preserve your fleet's log files by copying them to an Amazon S3 bucket when instances go out of service.

Visit lifecycle hooks in our Amazon EC2 Auto Scaling User Guide for more information.

Q: What are the characteristics of an "unhealthy" instance?

An unhealthy instance is one where the hardware has become impaired for some reason (bad disk, etc.), or it is not passing a user-configured ELB health check. Amazon EC2 Auto Scaling performs health checks on each individual EC2 instance at regular intervals, and if the instance is connected to an Elastic Load Balancing load balancer, it can also perform ELB health checks.

Q: Can I customize a health check?

Yes, there is an API called *SetInstanceHealth* that allows you to change an instance's state to UNHEALTHY, which will then result in a termination and replacement.

Q: Can I suspend health checks (for example, to evaluate unhealthy instances)?

Yes, you can temporarily suspend Amazon EC2 Auto Scaling health checks by using the SuspendProcesses API. You can use the ResumeProcesses API to resume automatic health checks.

Q: Which health check type should I select?

If you are using Elastic Load Balancing (ELB) with your group, you should select an ELB health check. If you're not using ELB with your group, you should select the EC2 health check.

Q: Can I use Amazon EC2 Auto Scaling for health checks and to replace unhealthy instances if I'm not using Elastic Load Balancing (ELB)?

You don't have to use ELB to use Auto Scaling. You can use the EC2 health check to identify and replace unhealthy instances.

Q: Do the Elastic Load Balancing (ELB) health checks work with Application Load Balancers and Network Load Balancers? Will an instance be marked as unhealthy if any target group associated with it becomes unhealthy?

Yes, Amazon EC2 Auto Scaling works with Application Load Balancers and Network Load Balancers including their health check feature.

Q: Is there any way to use Amazon EC2 Auto Scaling to only add a volume without adding an instance?

A volume is attached to a new instance when it is added. Amazon EC2 Auto Scaling doesn't automatically add a volume when the existing one is approaching capacity. You can use the EC2 API to add a volume to an existing instance.

Q: What does the term "stateful instances" refer to?

When we refer to a stateful instance, we mean an instance that has data on it, which exists only on that instance. In general, terminating a stateful instance means that the data (or state information) on the instance is lost. You may want to consider using lifecycle hooks to copy the data off of a stateful instance before it's terminated, or enable instance protection to prevent Amazon EC2 Auto Scaling from terminating it.

Replacing Impaired Instances

Q: How does Amazon EC2 Auto Scaling replace an impaired instance?

When an impaired instance fails a health check, Amazon EC2 Auto Scaling automatically terminates it and replaces it with a new one. If you're using an Elastic Load Balancing load balancer, Amazon EC2 Auto Scaling gracefully detaches the impaired instance from the load balancer before provisioning a new one and attaching it to the load balancer. This is all done automatically, so you don't need to respond manually when an instance needs replacing.

Q: How do I control which instances Amazon EC2 Auto Scaling terminates when scaling in, and how do protect data on an instance?

With each Amazon EC2 Auto Scaling group, you control when Amazon EC2 Auto Scaling adds instances (referred to as scaling out) or remove instances (referred to as scaling in) from your group. You can scale the size of your group manually by attaching and detaching instances, or you can automate the process through the use of a scaling policy. When you have Amazon EC2 Auto Scaling automatically scale in, you must decide which instances Amazon EC2 Auto Scaling should terminate first. You can configure this through the use of a termination policy. You can also use instance protection to prevent Amazon EC2 Auto Scaling from selecting specific instances for termination when scaling in. If you have data on an instance, and you need that data to be persistent even if your instance is scaled in, then you can use a service like S3, RDS, or DynamoDB, to make sure that it is stored off the instance.

Q: How long is the turn-around time for Amazon EC2 Auto Scaling to spin up a new instance at inService state after detecting an unhealthy server?

The turnaround time is within minutes. The majority of replacements happen within less than 5 minutes, and on average it is significantly less than 5 minutes. It depends on a variety of factors, including how long it takes to boot up the AMI of your instance.

Q: If Elastic Load Balancing (ELB) determines that an instance is unhealthy, and moved offline, will the previous requests sent to the failed instance be queued and rerouted to other instances within the group?

When ELB notices that the instance is unhealthy, it will stop routing requests to it. However, prior to discovering that the instance is unhealthy, some requests to that instance will fail.

Q: If you don’t use Elastic Load Balancing (ELB) how would users be directed to the other servers in a group if there was a failure?

You can integrate with Route53 (which Amazon EC2 Auto Scaling does not currently support out of the box, but many customers use). You can also use your own reverse proxy, or for internal microservices, can use service discovery solutions.

Security

Q: How do I control access to Amazon EC2 Auto Scaling resources?

Amazon EC2 Auto Scaling integrates with AWS Identity and Access Management (IAM), a service that enables you to do the following:

- Create users and groups under your organization's AWS account
- Assign unique security credentials to each user under your AWS account
- Control each user's permissions to perform tasks using AWS resources
- Allow the users in another AWS account to share your AWS resources
- Create roles for your AWS account and define the users or services that can assume them

- Use existing identities for your enterprise to grant permissions to perform tasks using AWS resources

For example, you could create an IAM policy that grants the Managers group permission to use only the *DescribeAutoScalingGroups*, *DescribeLaunchConfigurations*, *DescribeScalingActivities*, and *DescribePolicies* API operations. Users in the Managers group could then use those operations with any Amazon EC2 Auto Scaling groups and launch configurations. With Amazon EC2 Auto Scaling resource-level permissions, you can restrict access to a particular EC2 Auto Scaling group or launch configuration.

For more information, see the Controlling Access to Your Auto Scaling Resources section of the Amazon EC2 Auto Scaling user guide.

Q: Can you define a default admin password on Windows instances with Amazon EC2 Auto Scaling?

You can use the Key Name parameter to CreateLaunchConfiguration to associate a key pair with your instance. You can then use the *GetPasswordData* API in EC2. This is also possible through the AWS Management Console.

Q: Are CloudWatch agents automatically installed on EC2 instances when you create an Amazon EC2 Auto Scaling group?

If your AMI contains a CloudWatch agent, it’s automatically installed on EC2 instances when you create an EC2 Auto Scaling group. With the stock Amazon Linux AMI, you need to install it (recommended, via yum).

Cost Optimization

Q: Can I create a single ASG to scale instances across different purchase options?

Yes. You can provision and automatically scale EC2 capacity across different EC2 instance types, Availability Zones, and On-Demand, RIs and Spot purchase options in a single Auto Scaling Group. You have the option to define the desired split between On-Demand and Spot capacity, select which instance types work for your application, and specify preference for how EC2 Auto Scaling should distribute the ASG capacity within each purchasing model.

Q: Can I use ASGs to launch and manage just Spot Instances or just On-Demand instances and RIs?

Yes. You can configure your ASG specifying all capacity to be only Spot instances or all capacity to be only On-Demand instances and RIs.

Q: Can I have a base capacity with On-Demand instances and RIs, and scale my ASG out on Spot instances?

Yes. When setting up an ASG to combine purchasing models, you can specify the base capacity of the group to be fulfilled by On-Demand instances. As the ASG scales in or scale out, EC2 Auto Scaling ensures the base capacity be fulfilled with On-Demand instances and anything beyond that be fulfilled with either only Spot instances or a specified percentage mix of On-Demand or Spot instances.

Q: Can I modify the configuration of an ASG to update the different properties pertaining to combining purchasing models and specifying multiple instance types?

Yes. Similar to other ASG parameters, customers can update an existing ASG to modify one or all parameters pertaining to combining purchasing models and specifying multiple instance types, including instance types, prioritization order for On-Demand instances, percentage split between On-Demand and Spot instances, and allocation strategy.

Q: Can I use RI discounts with On-Demand Instances in an ASG?

Yes. For example, if you have RIs for C4 instances and EC2 Auto Scaling launches a C4 you will receive your RI pricing for On-Demand Instances.

Q: Can I specify instances of different sizes (CPU cores, memory) in my Auto Scaling group?

Yes. You can specify any instance type available in a region. Additionally, you can specify an optional weight for each instance type, which defines the capacity units that each instance would contribute to your application’s performance.

Q: What if the instance types I like are not available in an Availability Zone?

If none of the specified instance types are available in an Availability Zone, Auto Scaling will retarget the launches in other Availability Zones associated with the Auto Scaling group. Auto Scaling will always prefer keeping your compute balanced across Availability Zones and retarget if all instance types are not available in an Availability Zone.

Pricing

Q: What are the costs for using Amazon EC2 Auto Scaling?

Amazon EC2 Auto Scaling fleet managment for EC2 instances carries no additional fees. The dynamic scaling capabilities of Amazon EC2 Auto Scaling are enabled by Amazon CloudWatch and also carry no additional fees. Amazon EC2 and Amazon CloudWatch service fees apply and are billed separately.

Amazon Elastic Container Registry FAQs

General

Q: What is Amazon Elastic Container Registry (ECR)?
Amazon Elastic Container Registry (ECR) is a fully-managed Docker container registry that makes it easy for developers to store, manage, and deploy Docker container images. Amazon ECR is integrated with Amazon Elastic Container Service (ECS), simplifying your development to production workflow. Amazon ECR eliminates the need to operate your own container repositories or worry about scaling the underlying infrastructure. Amazon ECR hosts your images in a highly available and scalable architecture, allowing you to reliably deploy containers for your applications. Integration with AWS Identity and Access Management (IAM) provides resource-level control of each repository.

Q: Why should I use Amazon ECR?
Amazon ECR eliminates the need to operate and scale the infrastructure required to power your container registry. Amazon ECR uses Amazon S3 for storage to make your container images highly available and accessible, allowing you to reliably deploy new containers for your applications. Amazon ECR transfers your container images over HTTPS and automatically encrypts your images at rest. You can configure policies to manage permissions for each repository and restrict access to IAM users, roles, or other AWS accounts. Amazon ECR integrates with Amazon ECS and the Docker CLI, allowing you to simplify your development and production workflows. You can easily push your container images to Amazon ECR using the Docker CLI from your development machine, and Amazon ECS can pull them directly for production deployments.

Q: What is the pricing for Amazon ECR?
With Amazon ECR, there are no upfront fees or commitments. You pay only for

the amount of data you store in your repositories and data transferred to the Internet. Please see our Pricing page for more details.

Q: Is Amazon ECR a global service?
Amazon ECR is a regional service and is designed to give you flexibility in how images are deployed. You have the ability to push/pull images to the same region where your Docker cluster runs for the best performance. You can also access Amazon ECR anywhere that Docker runs such as desktops and on-premises environments. Pulling images between regions or out to the internet will have additional latency and data transfer costs.

Q: Can Amazon ECR host public container images?
Amazon ECR currently supports private images. However, using IAM resource-based permissions, you can configure policies for each repository to allow access to IAM users, roles, or other AWS accounts.

Q: What compliance capabilities can I enable on Amazon ECR?
You can use AWS CloudTrail on Amazon ECR to provide a history of all API actions such as who pulled an image and when tags were moved between images. Administrators can also find which EC2 instances pulled which images.

Using Amazon ECR

Q: How do I get started using Amazon ECR?
The best way to get started with Amazon ECR is to use the Docker CLI to push and pull and your first image. Visit our Getting Started page for more information.

Q: Can I access Amazon ECR inside a VPC?
Yes. You can now setup AWS PrivateLink endpoints to allow your instances pull images without traversing through the public internet.

Q: What's the best way to manage my repositories and images?
Amazon ECR provides a command line interface and APIs to create, monitor, and delete repositories and set repository permissions. You can perform the same actions in the Amazon ECR Management Console, which can be accessed

via the “Repositories” section of the Amazon ECS Console. Amazon ECR also integrates with the Docker CLI allowing you to push, pull, and tag images on your development machine.

Q: Does Amazon ECR replicate images across regions?
No. Amazon ECR is designed to give you flexibility in where you store and how you deploy your images. You can create deployment pipelines that build images, push them to Amazon ECR in selected regions, and then deploy the images to your Docker cluster.

Q: Can I use Amazon ECR within local and on-premises environments?
Yes. You can access Amazon ECR anywhere that Docker runs such as desktops and on-premises environments.

Q: Does Amazon ECR provide an Amazon Linux container image?
Yes. Amazon ECR provides Amazon Linux container images, and detailed steps can be found on the forums. Customers can use these container images to run workloads in their Linux-based Docker environment. The container image has a minimal set of packages and is able to install the full set of Amazon Linux AMI packages. Similar to the Amazon Linux AMI in EC2, Amazon Linux container images will get ongoing updates from Amazon in the form of security updates, rolling releases, and package updates.

Q: Does Amazon ECR work with Amazon ECS?
Yes. Amazon ECR is integrated with Amazon ECS allowing you to easily store, run, and manage container images for applications running on Amazon ECS. All you need to do is specify the Amazon ECR repository in your Task Definition and Amazon ECS will retrieve the appropriate images for your applications.

Q: Does Amazon ECR work with AWS Elastic Beanstalk?
Yes. AWS Elastic Beanstalk supports Amazon ECR for both single and multi-container Docker environments allowing you to easily deploy container images stored in Amazon ECR with AWS Elastic Beanstalk. All you need to do is specify the Amazon ECR repository in your Dockerrun.aws.json configuration and attach the AmazonEC2ContainerRegistryReadOnly policy to your container instance role.

Q: What version of Docker Engine does Amazon ECR support?
Amazon ECR currently supports Docker Engine 1.7.0 and up.

Q: What version of the Docker Registry API does Amazon ECR support?
Amazon ECR supports the Docker Registry V2 API specification.

Q: Will Amazon ECR automatically build images from a Dockerfile?
No. However, Amazon ECR integrates with a number of popular CI/CD solutions to provide this capability. See the Amazon ECR Partners Page for more information.

Q: Does Amazon ECR support federated access?
Yes. Amazon ECR is integrated with AWS Identity and Access Management, which supports identity federation for delegated access to the AWS Management Console or AWS APIs.

Q: What version of the Docker Image Manifest specification does Amazon ECR support?
Amazon ECR supports the Docker Image Manifest V2, Schema 2 format. In order to maintain backwards compatibility with Schema 1 images, Amazon ECR will continue to accept images uploaded in the Schema 1 format. Additionally, Amazon ECR can down-translate from a Schema 2 to a Schema 1 image when pulling with an older version of Docker Engine (1.9 and below).

Q: Does Amazon ECR support the Open Container Initiative (OCI) format?
Yes. Amazon ECR is compatible with the Open Container Initiative (OCI) image specification letting you push and pull OCI images. Amazon ECR can also translate between Docker Image Manifest V2, Schema 2 images and OCI images on pull.

Security

Q: How does Amazon ECR help ensure that container images are secure?
Amazon ECR automatically encrypts images at rest using S3 server side encryption and transfers your container images over HTTPS. You can configure policies to manage permissions and control access to your images using AWS

Identity and Access Management (IAM) users and roles without having to manage credentials directly on your EC2 instances.

Q: How can I use AWS Identity and Access Management for permissions?
You can use IAM resource-based policies to control and monitor who and what (e.g., EC2 instances) can access your container images as well as how, when, and where they can access them. To get started, use the Management Console to create resource-based policies for your repositories. Alternatively, you can use sample policies and attach them to your repositories via the Amazon ECR CLI.

Q: Can I share my images across AWS accounts?
Yes. Here is an example of how to create and set a policy for cross-account image sharing.

Amazon Elastic Container Service FAQs

General

Q: What is Amazon Elastic Container Service?

Q: Why should I use Amazon ECS?

Q: What is the pricing for Amazon ECS?

Q: How is Amazon ECS different from AWS Elastic Beanstalk?

Q: How is Amazon ECS different from AWS Lambda?

Using Amazon ECS

Q: How do I get started using Amazon ECS?

Q: Does Elastic Container Service support any other container types?

Q: I want to launch containers. Why do I have to launch Tasks?

Q: Does Amazon ECS support applications and services?

Q: Does Amazon ECS support dynamic port mapping?

Q: Does Amazon ECS support batch jobs?

Q: Can I use my own scheduler with Amazon ECS?

Q: Can I use my own AMI?

Q: How can I configure my container instances to pull from Amazon Elastic Container Registry?

Q: How does AWS Fargate work with Amazon ECS?

Q: How should I choose between using AWS Fargate with Amazon ECS or just using ECS?

Security and Compliance

Q: How does Amazon ECS isolate containers belonging to different customers?

Q: Can I apply additional security configuration and isolation frameworks to my container instances?

Q: Can I operate container instances with different security settings or segregate different tasks across different environments?

Q: Does Amazon ECS support retrieving Docker images from a private or internal source?

Q: How do I configure IAM roles for ECS tasks?

Q: With which compliance programs does Amazon ECS conform?

Q: Can I use Amazon ECS for Protected Health Information (PHI) and other HIPAA regulated workloads?

Q: Can I use Amazon ECS for US Government-regulated workloads or processing sensitive Controlled Unclassified Information (CUI)?

Service Level Agreement (SLA)

Q: What does the Amazon ECS SLA guarantee?

Q: How do I know if I qualify for a SLA Service Credit?

Transition to new ARN and ID format

Q: What is changing?

Q: How does this impact me?

Q: Will this affect existing resources?

Q: Which ECS resource ARNs are changing?

Q: What will the new ARNs look like?

Q: Which ECS resource IDs are changing?

Q: What will the new IDs look like?

Q: I have now opted-in, how do I transition my resources to the new formats so that I can tag them?

Q: What is the deadline to adopt the new formats?

Q: Will new accounts be automatically opted-in?

Q: How do I opt in and out of receiving new ARNs/IDs?

Q: Can I opt-in only a specific IAM user or IAM role in my account?

Q: What will happen if I take no action?

Q: What if I prefer to continue using the old ARN/IDs for my resources created after December 31st 2019?

Q: If I opt in to new ARNs/IDs formats and then opt back out during the transition period, what will happen to resources that were created with new formats?

Q: What should I do if my systems are not working as expected before the transition period ends?

Q: What will happen if I launch resources in multiple regions during the transition period?

Amazon EKS FAQs

General

Q: What is Amazon Elastic Kubernetes Service (Amazon EKS)?

A: Amazon EKS is a managed service that makes it easy for you to run Kubernetes on AWS without needing to install and operate your own Kubernetes control plane or worker nodes.

Q: What is Kubernetes?

A: Kubernetes is open source software that allows you to deploy and manage containerized applications at scale. Kubernetes groups containers into logical groupings for management and discoverability, then launches them onto clusters of EC2 instances. Using Kubernetes you can run containerized applications including microservices, batch processing workers, and platforms as a service (PaaS) using the same toolset on premises and in the cloud.

Q: Why should I use Amazon EKS?

A: Amazon EKS provisions and scales the Kubernetes control plane, including the API servers and backend persistence layer, across multiple AWS availability zones for high availability and fault tolerance. Amazon EKS automatically detects and replaces unhealthy control plane nodes and provides patching for the control plane. You can run EKS using AWS Fargate, which is serverless compute for containers. Fargate removes the need to provision and manage servers, lets you specify and pay for resources per application, and improves security through application isolation by design. Amazon EKS is integrated with many AWS services to provide scalability and security for your applications. These services include Elastic Load Balancing for load distribution, IAM for authentication, Amazon VPC for isolation, and AWS CloudTrail for logging.

Q: How does Amazon EKS work?

A: Amazon EKS works by provisioning (starting) and managing the Kubernetes control plane and worker nodes for you. At a high level, Kubernetes consists of two major components - a cluster of 'worker nodes' that run your containers and the control plane that manages when and where containers are started on your cluster and monitors their status.

Without Amazon EKS, you have to run both the Kubernetes control plane and the cluster of worker nodes yourself. With Amazon EKS, you provision your worker nodes using a single command in the EKS console, CLI, or API, and AWS handles provisioning, scaling, and managing the Kubernetes control plane in a highly available and secure configuration. This removes a significant operational burden for running Kubernetes and allows you to focus on building applications instead of managing AWS infrastructure.

Q: Which operating systems does Amazon EKS support?

A: Amazon EKS supports Linux x86 and Windows Server operating system distributions that are compatible with Kubernetes. Amazon EKS provides optimized AMIs for Amazon Linux 2 and Windows Server 2019. EKS-optimized AMIs for other Linux distributions, such as Ubuntu, are available from their respective vendors.

Q: I have a feature request, who do I tell?

A: Please let us know what we can add or do better by opening a feature request on the AWS Container Services Public Roadmap

Integrations

Q: Does Amazon EKS work with my existing Kubernetes applications and tools?

A: Amazon EKS runs the open-source Kubernetes software, so you can use all the existing plugins and tooling from the Kubernetes community. Applications running on Amazon EKS are fully compatible with applications running on any standard Kubernetes environment, whether running in on-premises datacenters or public clouds. This means that you can easily migrate any standard Kubernetes application to Amazon EKS without any code modifications.

Q: Does Amazon EKS work with AWS Fargate?

A: Yes. You can run Kubernetes applications as serverless containers using AWS Fargate and Amazon EKS.

Kubernetes versions and updates

Q: Which Kubernetes versions does Amazon EKS support?

A: Amazon EKS supports Kubernetes version 1.12, 1.13, and 1.14 and will continue to add support for additional Kubernetes versions in the future.

Q: Can I update my Kubernetes cluster to a new version?

A: Yes. Amazon EKS performs managed, in-place cluster upgrades for both Kubernetes and Amazon EKS platform versions. This simplifies cluster operations and lets you take advantage of the latest Kubernetes features, as well as the updates to Amazon EKS configuration and security patches.

There are two types of updates that you can apply to your Amazon EKS cluster, Kubernetes version updates and Amazon EKS platform version updates. As new Kubernetes versions are released and validated for use with Amazon EKS, we will support three stable Kubernetes versions as part of the update process at any given time.

Q: What is an EKS platform version?

A: Amazon EKS platform versions represent the capabilities of the cluster control plane, such as which Kubernetes API server flags are enabled, as well as the current Kubernetes patch version. Each Kubernetes minor version has one or more associated Amazon EKS platform versions. The platform versions for different Kubernetes minor versions are independent.

When a new Kubernetes minor version is available in Amazon EKS (for example, 1.13), the initial Amazon EKS platform version for that Kubernetes minor version starts at eks.1. However, Amazon EKS releases new platform versions periodically to enable new Kubernetes control plane settings and to provide security fixes.

Q: Why would I want manual control over Kubernetes version updates?

A: New versions of Kubernetes introduce significant change to the Kubernetes API, and as a result, can result in changed application behavior. Manual control over the version of Kubernetes on your cluster allows you to test applications against new versions of Kubernetes before upgrading production clusters. Amazon EKS provides you the ability to choose when you introduce changes to your EKS cluster.

Q: How do I update my worker nodes?

A: AWS publishes EKS-optimized Amazon Machine Images (AMIs) that include the necessary worker node binaries (Docker and Kubelet). This AMI is updated regularly and includes the most up to date version of these components. You can update your EKS managed nodes to the latest versions of the EKS-optimized AMIs with a single command in the EKS console, API, or CLI.

If you are building your own custom AMIs to use for EKS worker nodes, AWS also publishes Packer scripts that document our build steps, allowing you to identify the binaries included in each version of the AMI.

Pricing and availability

Q: How much does Amazon EKS cost?

A: You pay $0.10 per hour for each Amazon EKS cluster that you create and for the AWS resources you create to run your Kubernetes worker nodes. You only pay for what you use, as you use it; there are no minimum fees and no upfront commitments. Find more information in the EKS pricing page.

Q: Where is Amazon EKS available?

A: Please visit the AWS global infrastructure region table for the most up-to-date information on Amazon EKS regional availability.

Service Level Agreement

Q: What is Amazon EKS Service Level Agreement (SLA)?

A: The Amazon EKS SLA can be found here.

AWS Batch FAQs

General information

Q: What is AWS Batch?
AWS Batch is a set of batch management capabilities that enables developers, scientists, and engineers to easily and efficiently run hundreds of thousands of batch computing jobs on AWS. AWS Batch dynamically provisions the optimal quantity and type of compute resources (e.g., CPU or memory optimized instances) based on the volume and specific resource requirements of the batch jobs submitted. With AWS Batch, there is no need to install and manage batch computing software or server clusters, allowing you to instead focus on analyzing results and solving problems. AWS Batch plans, schedules, and executes your batch computing workloads using Amazon EC2 and Spot Instances.

Q: What is Batch Computing?
Batch computing is the execution of a series of programs ("jobs") on one or more computers without manual intervention. Input parameters are pre-defined through scripts, command-line arguments, control files, or job control language. A given batch job may depend on the completion of preceding jobs, or on the availability of certain inputs, making the sequencing and scheduling of multiple jobs important, and incompatible with interactive processing.

Q: What are the benefits of batch computing?

- It can shift the time of job processing to periods when greater or less expensive capacity is available.
- It avoids idling compute resources with frequent manual intervention and supervision.
- It increases efficiency by driving higher utilization of compute resources.
- It enables the prioritization of jobs, aligning resource allocation with business objectives.

Why AWS Batch

Q: Why should I use AWS Batch?
AWS Batch handles job execution and compute resource management, allowing you to focus on developing applications or analyzing results instead of setting up and managing infrastructure. If you are considering running or moving batch workloads to AWS, you should consider using AWS Batch.

Q: What use cases is AWS Batch optimized for?
AWS Batch is optimized for batch computing and applications that scale through the execution of multiple jobs in parallel. Deep learning, genomics analysis, financial risk models, Monte Carlo simulations, animation rendering, media transcoding, image processing, and engineering simulations are all excellent examples of batch computing applications.

Features

Q: What are the key features of AWS Batch?
AWS Batch manages compute environments and job queues, allowing you to easily run thousands of jobs of any scale using Amazon EC2 and EC2 Spot. You simply define and submit your batch jobs to a queue. In response, AWS Batch chooses where to run the jobs, launching additional AWS capacity if needed. AWS Batch carefully monitors the progress of your jobs. When capacity is no longer needed, AWS Batch will remove it. AWS Batch also provides the ability to submit jobs that are part of a pipeline or workflow, enabling you to express any interdependencies that exist between them as you submit jobs.

Q: What types of batch jobs does AWS Batch support?
AWS Batch supports any job that can executed as a Docker container. Jobs specify their memory requirements and number of vCPUs.

Q: What is a Compute Resource?
An AWS Batch Compute Resource is an EC2 instance.

Q: What is a Compute Environment?
An AWS Batch Compute Environment is a collection of compute resources on

which jobs are executed. AWS Batch supports two types of Compute Environments; Managed Compute Environments which are provisioned and managed by AWS and Unmanaged Compute Environments which are managed by customers. Unmanaged Compute Environments provide a mechanism to leverage specialized resources such as Dedicated Hosts, larger storage configurations, and Amazon EFS.

Q: What is a Job Definition?
A Job Definition describes the job to be executed, parameters, environmental variables, compute requirements, and other information that is used to optimize the execution of a job. Job Definitions are defined in advance of submitting a job and can be shared with others.

Q: What is the Amazon ECS Agent and how is it used by AWS Batch?
AWS Batch uses Amazon ECS to execute containerized jobs and therefore requires the ECS Agent to be installed on compute resources within your AWS Batch Compute Environments. The ECS Agent is pre-installed in Managed Compute Environments.

Q: How does AWS Batch make it easier to use EC2 Spot?
AWS Batch Compute Environments can be comprised of EC2 Spot instances. When creating a Managed Compute Environment, simplify specify that you would like to use EC2 Spot and provide a percentage of On Demand pricing that you are willing to pay and AWS Batch will take care of the rest. Unmanaged Compute Environments can also include Spot instances that you launch, including those launched by EC2 Spot Fleet.

Pricing

Q. What is the pricing for AWS Batch?
There is no additional charge for AWS Batch. You only pay for the AWS Resources (e.g. EC2 Instances) you create to store and run your batch jobs.

GPU Scheduling

Q: Can I use accelerators with AWS Batch?
Yes, you can use Batch to specify the number and type of accelerators your jobs require as job definition input variables, alongside the current options of vCPU and memory. AWS Batch will scale up instances appropriate for your jobs based on the required accelerators and isolate the accelerators according to each job's needs, so only the appropriate containers can access them.

Q: Why should I use accelerators with AWS Batch?
By using accelerators with Batch, you can dynamically schedule and provision your jobs according to their accelerator needs, and Batch will ensure that the appropriate number of accelerators are reserved against your jobs. Batch will scale up your EC2 Accelerated Instances when you need them, and scale them down when you're done, allowing you to focus on your applications. Batch has native integration with the EC2 Spot, meaning your accelerated jobs can take advantage of up to 90% savings when using accelerated instances.

Q: What accelerators can I use with AWS Batch?
You can use GPU's on P accelerated instances currently.

Q: How do I submit jobs requiring accelerated instances to Batch?
You can specify the number and type of accelerators in the Job Definition. You specify the accelerator by describing the accelerator type (e.g., GPU – currently the only supported accelerator) and the number of that type your job requires. Your specified accelerator type must be present on one of the instance types specified in your Compute Environments. For example, if your job needs 2 GPUs, also make sure that you have specified a P-family instance in your Compute Environment.

From the API:

```
{
"containerProperties": {
"vcpus": 1,
"image": "nvidia/cuda:9.0-base",
"memory": 2048,
"resourceRequirements" : [
{
"type" : "GPU",
"value" : "1"
}
],
```

Q: Can accelerator variables in the job definition be overwritten at job submission?
Similar to vCPU and memory requirements, you can overwrite the number and type of accelerators at job submission.

Q: Can accelerated instances be used for jobs that don't need the accelerators?
With today's behavior, Batch will avoid scheduling jobs that do not require acceleration on accelerated instances when possible. This is to avoid cases where long-running jobs occupy the accelerated instance without taking advantage of the accelerator, increasing cost. In rare cases with Spot pricing and with accelerated instances as allowed types, it is possible that Batch will determine that an accelerated instance is the least expensive way to run your jobs, regardless of accelerator needs.

If you submit a job to a CE that only allows Batch to launch accelerated instances, Batch will run the jobs on those instances, regardless of their accelerator needs.

Q: How does Batch use the ECS GPU-Optimized AMI?
From now on, p-type instances will launch by default with the ECS GPU-optimized AMI. This AMI contains libraries and runtimes needed to run GPU-based applications. You can always point to a custom AMI as needed when creating a CE.

Getting Started

Q. How do I get started?
Follow the Getting Started Guide in our documentation to get started.

Q. What do I need to provision to get started?

There is no need to manually launch your own compute resources in order to get started. The AWS Batch web console will guide you through the process of creating your first Compute Environment and Job Queue so that you can submit your first job. Resources within your compute environment will scale up as additional jobs are ready to run and scale down as the number of runnable jobs decreases.

AWS Elastic Beanstalk FAQs

General

Q: What is AWS Elastic Beanstalk? >>

Q: Who should use AWS Elastic Beanstalk? >>

Q: Which languages and development stacks does AWS Elastic Beanstalk support? >>

Q: Will AWS Elastic Beanstalk support other languages? >>

Q: What can developers now do with AWS Elastic Beanstalk that they could not before? >>

Q: How is AWS Elastic Beanstalk different from existing application containers or platform-as-a-service solutions? >>

Q: What elements of my application can I control when using AWS Elastic Beanstalk? >>

Q: What are the Cloud resources powering my AWS Elastic Beanstalk application? >>

Q: What kinds of applications are supported by AWS Elastic Beanstalk? >>

Q: Which operating systems does AWS Elastic Beanstalk use? >>

Getting Started

Q: How do I sign up for AWS Elastic Beanstalk? >>

Q: Why am I asked to verify my phone number when signing up for AWS Elastic Beanstalk? >>

Q: How do I get started after I have signed up? >>

Q: Is there a sample application that I can use to check out AWS Elastic Beanstalk? >>

Databases and Storage

Q: Does AWS Elastic Beanstalk store anything in Amazon S3? >>

Q: Can I use Amazon S3 to store application data, like images? >>

Q: What database solutions can I use with AWS Elastic Beanstalk? >>

Q: How do I set up a database for use with AWS Elastic Beanstalk? >>

Q: Does this mean I need to modify the application code when moving from test to production? >>

Security

Q: How do I make my application private? >>

Q: Can I run my application inside a Virtual Private Cloud (VPC)? >>

Q: Where can I find more information about security and running applications on AWS? >>

Q: Is it possible to use Identity & Access Management (IAM) with AWS Elastic Beanstalk? >>

Q: Why should I use IAM with AWS Elastic Beanstalk? >>

Q: How do I create IAM users? >>

Q: How do I grant an IAM user access to AWS Elastic Beanstalk? >>

Q: Can I restrict access to specific AWS Elastic Beanstalk resources? >>

Q: Who gets billed for the AWS resources that an IAM user creates? >>

Q: Who has access to an AWS Elastic Beanstalk environment launched by an IAM user? >>

Q: Can an IAM user access the AWS Elastic Beanstalk console? >>

Q: Can an IAM user call the AWS Elastic Beanstalk API? >>

Q: Can an IAM user use the AWS Elastic Beanstalk command line interface? >>

Managed Platform Updates

Q: How can I keep the underlying platform of the environment running my application automatically up-to-date? >>

Q: How can I get started with managed platform updates? >>

Q: What kinds of platform version updates will managed platform updates apply? >>

Q: How does AWS Elastic Beanstalk distinguish between “major,” “minor,” and “patch” version releases? >>

Q: When and how can I perform major version updates? >>

Q: How does Elastic Beanstalk apply managed platform updates? >>

Q: Will my application be available during the maintenance windows? >>

Q: What does it cost to use managed platform updates? >>

Q: What is a maintenance window? >>

Q: How will I be notified of the availability of new platform versions? >>

Q: Where can I find details of changes between platform versions? >>

Q: What operations can I perform on the environment while a managed update is in progress? >>

Q: Which platform version will my environment be updated to if there are multiple new versions released in between maintenance windows? >>

Q: Where can I find details of all the managed platform updates that have been performed on my environment? >>

Q: How often are platform version updates released? >>

Billing

Q: How much does AWS Elastic Beanstalk cost? >>

Q: How much do the AWS resources powering my application on AWS Elastic Beanstalk cost? >>

Q: How do I check how many AWS resources have been used by my application and access my bill? >>

Support

Q: Does AWS Support cover AWS Elastic Beanstalk? >>

Q: What other support options are available? >>

AWS Fargate FAQs

Security and Compliance

Q: With which compliance programs does AWS Fargate conform?

Q: Can I use AWS Fargate for Protected Health Information (PHI) and other HIPAA regulated workloads?

Q: Can I use AWS Fargate for US Government-regulated workloads or processing sensitive Controlled Unclassified Information (CUI)?

Service Level Agreement (SLA)

Q: What does the AWS Fargate SLA guarantee?

Q: How do I know if I qualify for a SLA Service Credit?

AWS Lambda FAQs

General

Q: What is AWS Lambda?

AWS Lambda lets you run code without provisioning or managing servers. You pay only for the compute time you consume - there is no charge when your code is not running. With Lambda, you can run code for virtually any type of application or backend service - all with zero administration. Just upload your code and Lambda takes care of everything required to run and scale your code with high availability. You can set up your code to automatically trigger from other AWS services or call it directly from any web or mobile app.

Q: What is serverless computing?

Serverless computing allows you to build and run applications and services without thinking about servers. With serverless computing, your application still runs on servers, but all the server management is done by AWS. At the core of serverless computing is AWS Lambda, which lets you run your code without provisioning or managing servers.

Q: What events can trigger an AWS Lambda function?

Please see our documentation for a complete list of event sources.

Q: When should I use AWS Lambda versus Amazon EC2?

Amazon Web Services offers a set of compute services to meet a range of needs.

Amazon EC2 offers flexibility, with a wide range of instance types and the option to customize the operating system, network and security settings, and the entire software stack, allowing you to easily move existing applications to the cloud. With Amazon EC2 you are responsible for provisioning capacity, monitoring fleet health and performance, and designing for fault tolerance and scalability. AWS Elastic Beanstalk offers an easy-to-use service for deploying and scaling web applications in which you retain ownership and full control over the underlying EC2 instances. Amazon EC2 Container Service is a scalable management service that supports Docker containers and allows you to easily run distributed applications on a managed cluster of Amazon EC2 instances.

AWS Lambda makes it easy to execute code in response to events, such as changes to Amazon S3 buckets, updates to an Amazon DynamoDB table, or custom events generated by your applications or devices. With Lambda you do not have to provision your own instances; Lambda performs all the operational and administrative activities on your behalf, including capacity provisioning, monitoring fleet health, applying security patches to the underlying compute resources, deploying your code, running a web service front end, and monitoring and logging your code. AWS Lambda provides easy scaling and high availability to your code without additional effort on your part.

Q: What kind of code can run on AWS Lambda?

AWS Lambda offers an easy way to accomplish many activities in the cloud. For example, you can use AWS Lambda to build mobile back-ends that retrieve and transform data from Amazon DynamoDB, handlers that compress or transform objects as they are uploaded to Amazon S3, auditing and reporting of API calls made to any Amazon Web Service, and server-less processing of streaming data using Amazon Kinesis.

Q: What languages does AWS Lambda support?

AWS Lambda natively supports Java, Go, PowerShell, Node.js, C#, Python, and Ruby code, and provides a Runtime API which allows you to use any additional programming languages to author your functions. Please read our documentation on using Node.js, Python, Java, Ruby, C#, Go and PowerShell.

Q: Can I access the infrastructure that AWS Lambda runs on?

No. AWS Lambda operates the compute infrastructure on your behalf, allowing it to perform health checks, apply security patches, and do other routine maintenance.

Q: How does AWS Lambda isolate my code?

Each AWS Lambda function runs in its own isolated environment, with its own resources and file system view. AWS Lambda uses the same techniques as Amazon EC2 to provide security and separation at the infrastructure and execution levels.

Q: How does AWS Lambda secure my code?

AWS Lambda stores code in Amazon S3 and encrypts it at rest. AWS Lambda performs additional integrity checks while your code is in use.

Q: What AWS regions are available for AWS Lambda?

Please refer to the AWS Global Infrastructure Region Table.

AWS Lambda functions

Q: What is an AWS Lambda function?

The code you run on AWS Lambda is uploaded as a "Lambda function". Each function has associated configuration information, such as its name, description, entry point, and resource requirements. The code must be written in a "stateless" style i.e. it should assume there is no affinity to the underlying compute infrastructure. Local file system access, child processes, and similar artifacts may not extend beyond the lifetime of the request, and any persistent state should be stored in Amazon S3, Amazon DynamoDB, or another Internet-available storage service. Lambda functions can include libraries, even native ones.

Q: Will AWS Lambda reuse function instances?

To improve performance, AWS Lambda may choose to retain an instance of your function and reuse it to serve a subsequent request, rather than creating a new copy. To learn more about how Lambda reuses function instances, visit our documentation. Your code should not assume that this will always happen.

Q: What if I need scratch space on disk for my AWS Lambda function?

Each Lambda function receives 500MB of non-persistent disk space in its own /tmp directory.

Q: Why must AWS Lambda functions be stateless?

Keeping functions stateless enables AWS Lambda to rapidly launch as many copies of the function as needed to scale to the rate of incoming events. While AWS Lambda's programming model is stateless, your code can access stateful data by calling other web services, such as Amazon S3 or Amazon DynamoDB.

Q: Can I use threads and processes in my AWS Lambda function code?

Yes. AWS Lambda allows you to use normal language and operating system features, such as creating additional threads and processes. Resources allocated to the Lambda function, including memory, execution time, disk, and network use, must be shared among all the threads/processes it uses. You can launch processes using any language supported by Amazon Linux.

Q: What restrictions apply to AWS Lambda function code?

Lambda attempts to impose as few restrictions as possible on normal language and operating system activities, but there are a few activities that are disabled: Inbound network connections are blocked by AWS Lambda, and for outbound connections only TCP/IP and UDP/IP sockets are supported, and ptrace (debugging) system calls are blocked. TCP port 25 traffic is also blocked as an anti-spam measure.

Q: How do I create an AWS Lambda function using the Lambda console?

If you are using Node.js or Python, you can author the code for your function using code editor in the AWS Lambda console which lets you author and test your functions, and view the results of function executions in a robust, IDE-like environment. Go to the console to get started.

You can also package the code (and any dependent libraries) as a ZIP and upload it using the AWS Lambda console from your local environment or specify an Amazon S3 location where the ZIP file is located. Uploads must be no larger than 50MB (compressed). You can use the AWS Eclipse plugin to author and deploy Lambda functions in Java. You can use the Visual Studio plugin to author and deploy Lambda functions in C#, and Node.js.

Q: How do I create an AWS Lambda function using the Lambda CLI?

You can package the code (and any dependent libraries) as a ZIP and upload it using the AWS CLI from your local environment, or specify an Amazon S3 location where the ZIP file is located. Uploads must be no larger than 50MB (compressed). Visit the Lambda Getting Started guide to get started.

Q: Does AWS Lambda support environment variables?

Yes. You can easily create and modify environment variables from the AWS Lambda Console, CLI or SDKs. To learn more about environment variables, see the documentation.

Q: Can I store sensitive information in environment variables?

For sensitive information, such as database passwords, we recommend you use client-side encryption using AWS Key Management Service and store the resulting values as ciphertext in your environment variable. You will need to include logic in your AWS Lambda function code to decrypt these values.

Q: How can I manage my AWS Lambda functions?

You can easily list, delete, update, and monitor your Lambda functions using the dashboard in the AWS Lambda console. You can also use the AWS CLI and AWS SDK to manage your Lambda functions. Visit the Lambda Developer Guide to learn more.

Q: Can I share code across functions?

Yes, you can package any code (frameworks, SDKs, libraries, and more) as a Lambda Layer and manage and share them easily across multiple functions.

Q: How do I monitor an AWS Lambda function?

AWS Lambda automatically monitors Lambda functions on your behalf, reporting real-time metrics through Amazon CloudWatch, including total requests, account-level and function-level concurrency usage, latency, error rates, and throttled requests. You can view statistics for each of your Lambda functions via the Amazon CloudWatch console or through the AWS Lambda console. You can also call third-party monitoring APIs in your Lambda function.

Visit Troubleshooting CloudWatch metrics to learn more. Standard charges for AWS Lambda apply to use Lambda's built-in metrics.

Q: How do I troubleshoot failures in an AWS Lambda function?

AWS Lambda automatically integrates with Amazon CloudWatch logs, creating a log group for each Lambda function and providing basic application lifecycle event log entries, including logging the resources consumed for each use of that function. You can easily insert additional logging statements into your code. You can also call third-party logging APIs in your Lambda function. Visit Troubleshooting Lambda functions to learn more. Amazon CloudWatch Logs rates will apply.

Q: How do I scale an AWS Lambda function?

You do not have to scale your Lambda functions – AWS Lambda scales them automatically on your behalf. Every time an event notification is received for your function, AWS Lambda quickly locates free capacity within its compute fleet and runs your code. Since your code is stateless, AWS Lambda can start as many copies of your function as needed without lengthy deployment and configuration delays. There are no fundamental limits to scaling a function. AWS Lambda will dynamically allocate capacity to match the rate of incoming events.

Q: How are compute resources assigned to an AWS Lambda function?

In the AWS Lambda resource model, you choose the amount of memory you want for your function, and are allocated proportional CPU power and other resources. For example, choosing 256MB of memory allocates approximately twice as much CPU power to your Lambda function as requesting 128MB of memory and half as much CPU power as choosing 512MB of memory. To learn more, see our Function Configuration documentation.

You can set your memory in 64MB increments from 128MB to 3GB.

Q: How long can an AWS Lambda function execute?

AWS Lambda functions can be configured to run up to 15 minutes per execution. You can set the timeout to any value between 1 second and 15 minutes.

Q: How will I be charged for using AWS Lambda functions?

AWS Lambda is priced on a pay per use basis. Please see the AWS Lambda pricing page for details.

Q: Can I save money on AWS Lambda with a Compute Savings Plan?

Yes. In addition to saving money on Amazon EC2 and AWS Fargate, you can also use Compute Savings Plans to save money on AWS Lambda. Compute Savings Plans offer up to 17% discount on Duration, Provisioned Concurrency, and Duration (Provisioned Concurrency). Compute Savings Plans do not offer a discount on Requests in your Lambda bill. However, your Compute Savings Plans commitment can apply to Requests at regular rates.

Q: Does AWS Lambda support versioning?

Yes. By default, each AWS Lambda function has a single, current version of the code. Clients of your Lambda function can call a specific version or get the latest implementation. Please read out documentation on versioning Lambda functions.

Q: How long after uploading my code will my AWS Lambda function be ready to call?

Deployment times may vary with the size of your code, but AWS Lambda functions are typically ready to call within seconds of upload.

Q: Can I use my own version of a supported library?

Yes. you can include your own copy of a library (including the AWS SDK) in order to use a different version than the default one provided by AWS Lambda.

Using AWS Lambda to process AWS events

Q: What is an event source?

An event source is an AWS service or developer-created application that produces events that trigger an AWS Lambda function to run. Some services publish these events to Lambda by invoking the cloud function directly (for example, Amazon S3). Lambda can also poll resources in other services that do not publish events to Lambda. For example, Lambda can pull records from an Amazon Kinesis stream or an Amazon SQS queue and execute a Lambda function for each fetched message.

Many other services, such as AWS CloudTrail, can act as event sources simply by logging to Amazon S3 and using S3 bucket notifications to trigger AWS Lambda functions.

Q: What event sources can be used with AWS Lambda?

Please see our documentation for a complete list of event sources.

Q: How are events represented in AWS Lambda?

Events are passed to a Lambda function as an event input parameter. For event sources where events arrive in batches, such as Amazon SQS, Amazon Kinesis, and Amazon DynamoDB Streams, the event parameter may contain multiple events in a single call, based on the batch size you request.To learn more about Amazon S3 event notifications visit Configuring Notifications for Amazon S3 Events. To learn more about Amazon DynamoDB Streams visit the DynamoDB Stream Developers Guide. To learn more about invoking Lambda functions using Amazon SNS, visit the Amazon SNS Developers Guide. For more information on Amazon Cognito events, visit Amazon Cognito. For more information on AWS CloudTrail logs and auditing API calls across AWS services, see AWS CloudTrail.

Q: How do I make an AWS Lambda function respond to changes in an Amazon S3 bucket?

From the AWS Lambda console, you can select a function and associate it with notifications from an Amazon S3 bucket. Alternatively, you can use the Amazon S3 console and configure the bucket's notifications to send to your AWS Lambda function. This same functionality is also available through the AWS SDK and CLI.

Q: How do I make an AWS Lambda function respond to updates in an Amazon DynamoDB table?

You can trigger a Lambda function on DynamoDB table updates by subscribing your Lambda function to the DynamoDB Stream associated with the table. You can associate a DynamoDB Stream with a Lambda function using the Amazon DynamoDB console, the AWS Lambda console or Lambda's registerEventSource API.

Q: How do I use an AWS Lambda function to process records in an Amazon Kinesis stream?

From the AWS Lambda console, you can select a Lambda function and associate it with an Amazon Kinesis stream owned by the same account. This same functionality is also available through the AWS SDK and CLI.

Q: How does AWS Lambda process data from Amazon Kinesis streams and Amazon DynamoDB Streams?

The Amazon Kinesis and DynamoDB Streams records sent to your AWS Lambda function are strictly serialized, per shard. This means that if you put two records in the same shard, Lambda guarantees that your Lambda function will be successfully invoked with the first record before it is invoked with the second record. If the invocation for one record times out, is throttled, or encounters any other error, Lambda will retry until it succeeds (or the record reaches its 24-hour expiration) before moving on to the next record. The ordering of records across different shards is not guaranteed, and processing of each shard happens in parallel.

Q: How do I use an AWS Lambda function to respond to notifications sent by Amazon Simple Notification Service (SNS)?

From the AWS Lambda console, you can select a Lambda function and associate it with an Amazon SNS topic. This same functionality is also available through the AWS SDK and CLI.

Q: How do I use an AWS Lambda function to respond to emails sent by Amazon Simple Email Service (SES)?

From the Amazon SES Console, you can set up your receipt rule to have Amazon SES deliver your messages to an AWS Lambda function. The same functionality is available through the AWS SDK and CLI.

Q: How do I use an AWS Lambda function to respond to Amazon CloudWatch alarms?

First, configure the alarm to send Amazon SNS notifications. Then from the AWS Lambda console, select a Lambda function and associate it with that Amazon SNS topic. See the Amazon CloudWatch Developer Guide for more on setting up Amazon CloudWatch alarms.

Q: How do I use an AWS Lambda function to respond to changes in user or device data managed by Amazon Cognito?

From the AWS Lambda console, you can select a function to trigger when any datasets associated with an Amazon Cognito identity pool are synchronized. This same functionality is also available through the AWS SDK and CLI. Visit Amazon Cognito for more information on using Amazon Cognito to share and synchronize data across a user's devices.

Q: How can my application trigger an AWS Lambda function directly?

You can invoke a Lambda function using a custom event through AWS Lambda's invoke API. Only the function's owner or another AWS account that the owner has granted permission can invoke the function. Visit the Lambda Developers Guide to learn more.

Q: What is the latency of invoking an AWS Lambda function in response to an event?

AWS Lambda is designed to process events within milliseconds. Latency will be higher immediately after a Lambda function is created, updated, or if it has not been used recently.

Q: How do I create a mobile back-end using AWS Lambda?

You upload the code you want AWS Lambda to execute and then invoke it from your mobile app using the AWS Lambda SDK included in the AWS Mobile SDK. You can make both direct (synchronous) calls to retrieve or check data in real time as well as asynchronous calls. You can also define a custom API using Amazon API Gateway and invoke your Lambda functions through any REST compatible client. To learn more about the AWS Mobile SDK, visit the AWS Mobile SDK page. To learn more about Amazon API Gateway, visit the Amazon API Gateway page.

Q: How do I invoke an AWS Lambda function over HTTPS?

You can invoke a Lambda function over HTTPS by defining a custom RESTful API using Amazon API Gateway. This gives you an endpoint for your function which can respond to REST calls like GET, PUT and POST. Read more about using AWS Lambda with Amazon API Gateway.

Q: How can my AWS Lambda function customize its behavior to the device and app making the request?

When called through the AWS Mobile SDK, AWS Lambda functions automatically gain insight into the device and application that made the call through the ‘context’ object.

Q: How can my AWS Lambda function personalize their behavior based on the identity of the end user of an application?

When your app uses the Amazon Cognito identity, end users can authenticate themselves using a variety of public login providers such as Amazon, Facebook, Google, and other OpenID Connect-compatible services. User identity is then automatically and secured presented to your Lambda function in the form of an Amazon Cognito id, allowing it to access user data from Amazon Cognito, or as a key to store and retrieve data in Amazon DynamoDB or other web services.

Q: How do I create an Alexa skill using AWS Lambda?

AWS Lambda is integrated with the Alexa Skills Kit, a collection of self-service APIs, tools, documentation and code samples that make it easy for you to create voice-driven capabilities (or "skills") for Alexa. You simply upload the Lambda function code for the new Alexa skill you are creating, and AWS Lambda does the rest, executing the code in response to Alexa voice interactions and automatically managing the compute resources on your behalf. Read the Alexa Skills Kit documentation for more details.

Q: What happens if my function fails while processing an event?

For Amazon S3 bucket notifications and custom events, AWS Lambda will attempt execution of your function three times in the event of an error condition in your code or if you exceed a service or resource limit.

For ordered event sources that AWS Lambda polls on your behalf, such as Amazon DynamoDB Streams and Amazon Kinesis streams, Lambda will continue attempting execution in the event of a developer code error until the data expires. You can monitor progress through the Amazon Kinesis and Amazon DynamoDB consoles and through the Amazon CloudWatch metrics that AWS Lambda generates for your function. You can also set Amazon CloudWatch alarms based on error or execution throttling rates.

Using AWS Lambda to build applications

Q: What is a serverless application?

Lambda-based applications (also referred to as serverless applications) are composed of functions triggered by events. A typical serverless application consists of one or more functions triggered by events such as object uploads to Amazon S3, Amazon SNS notifications, or API actions. These functions can stand alone or leverage other resources such as DynamoDB tables or Amazon S3 buckets. The most basic serverless application is simply a function.

Q: How do I deploy and manage a serverless application?

You can deploy and manage your serverless applications using the AWS Serverless Application Model (AWS SAM). AWS SAM is a specification that prescribes the rules for expressing serverless applications on AWS. This specification aligns with the syntax used by AWS CloudFormation today and is supported natively within AWS CloudFormation as a set of resource types (referred to as "serverless resources"). These resources make it easier for AWS customers to use CloudFormation to configure and deploy serverless applications, using existing CloudFormation APIs.

Q: How can I discover existing serverless applications developed by the AWS community?

You can choose from a collection of serverless applications published by developers, companies, and partners in the AWS community with the AWS Serverless Application Repository. After finding an application, you can configure and deploy it straight from the Lambda console.

Q: How do I automate deployment for a serverless application?

You can automate your serverless application’s release process using AWS CodePipeline and AWS CodeDeploy. CodePipeline is a continuous delivery service that enables you to model, visualize and automate the steps required to release your serverless application. CodeDeploy provides a deployment automation engine for your Lambda-based applications. CodeDeploy lets you orchestrate deployments according to established best-practice methodologies such as canary and linear deployments, and helps you establish the necessary guardrails to verify that newly-deployed code is safe, stable, and ready to be fully released to production.

To learn more about serverless CI/CD, visit our documentation.

Q: How do I get started on building a serverless application?

To get started, visit the AWS Lambda console and download one of our blueprints. The file you download will contain an AWS SAM file (which defines the AWS resources in your application), and a .ZIP file (which includes your function's code). You can then use AWS CloudFormation commands to package and deploy the serverless application that you just downloaded. For more details, visit our documentation.

Q: How do I coordinate calls between multiple AWS Lambda functions?

You can use AWS Step Functions to coordinate a series of AWS Lambda functions in a specific order. You can invoke multiple Lambda functions sequentially, passing the output of one to the other, and/or in parallel, and Step Functions will maintain state during executions for you.

Q: How do I troubleshoot a serverless application?

You can enable your Lambda function for tracing with AWS X-Ray by adding X-Ray permissions to your Lambda function's execution role and changing your function's "tracing mode" to "active. " When X-Ray is enabled for your Lambda function, AWS Lambda will emit tracing information to X-Ray regarding the Lambda service overhead incurred when invoking your function. This will provide you with insights such as Lambda service overhead, function init time, and function execution time. In addition, you can include the X-Ray SDK in your Lambda deployment package to create your own trace segments, annotate your traces, or view trace segments for downstream calls made from your Lambda function. X-Ray SDKs are currently available for Node.js and Java. Visit Troubleshooting Lambda-based applications to learn more. AWS X-Ray rates will apply.

Q. Can I build serverless applications that connect to relational databases?

Yes. You can build highly scalable, secure, Lambda-based serverless applications that connect to relational databases using Amazon RDS Proxy, a highly available database proxy that manages thousands of concurrent connections to relational databases. Currently, RDS Proxy supports MySQL and Aurora databases. You can begin using RDS Proxy through the Amazon RDS console or the AWS Lambda console. Serverless applications that use fully managed connection pools from RDS Proxy will be billed according to RDS Proxy Pricing.

Q: How is AWS SAM licensed?

The specification is open sourced under Apache 2.0, which allows you and others to adopt and incorporate AWS SAM into build, deployment, monitoring and management tools with a commercial-friendly license. You can access the AWS SAM repository on GitHub here.

Provisioned Concurrency

Q: What is AWS Lambda Provisioned Concurrency?

Provisioned Concurrency gives you greater control over the performance of your serverless applications. When enabled, Provisioned Concurrency keeps functions initialized and hyper-ready to respond in double-digit milliseconds.

Q: How do I set up and manage Provisioned Concurrency?

You can configure concurrency on your function through the AWS Management Console, the Lambda API, the AWS CLI, and AWS CloudFormation. The simplest way to benefit from Provisioned Concurrency is by using AWS Auto Scaling. You can use Application Auto Scaling to configure schedules, or have Auto Scaling automatically adjust the level of Provisioned Concurrency in real time as demand changes. To learn more about Provisioned Concurrency, see the documentation.

Q: Do I need to change my code if I want to use Provisioned Concurrency?

You don’t need to make any changes to your code to use Provisioned Concurrency. It works seamlessly with all existing functions and runtimes. There is no change to the invocation and execution model of Lambda when using Provisioned Concurrency.

Q: How will I be charged for Provisioned Concurrency?

Provisioned Concurrency adds a pricing dimension, of ‘Provisioned Concurrency’, for keeping functions initialized. When enabled, you pay for the amount of concurrency that you configure and for the period of time that you configure it. When your function executes while Provisioned Concurrency is configured on it, you also pay for Requests and execution Duration. To learn more about the pricing of Provisioned Concurrency, see AWS Lambda Pricing.

Q: When should I use Provisioned Concurrency?

Provisioned Concurrency is ideal for building latency-sensitive applications, such as web or mobile backends, synchronously invoked APIs, and interactive microservices. You can easily configure the appropriate amount of concurrency based on your application's unique demand. You can increase the amount of concurrency during times of high demand and lower it, or turn it off completely, when demand decreases.

Q: What happens if a functions receives invocations above the configured level of Provisioned Concurrency?

If the concurrency of a function reaches the configured level, subsequent invocations of the function have the latency and scale characteristics of regular Lambda functions. You can restrict your function to only scale up to the configured level. Doing so prevents the function from exceeding the configured level of Provisioned Concurrency. This is a mechanism to prevent undesired variability in your application when demand exceeds the anticipated amount.

Lambda@Edge

Q: What is Lambda@Edge?

Lambda@Edge allows you to run code across AWS locations globally without provisioning or managing servers, responding to end users at the lowest network latency. You just upload your Node.js or Python code to AWS Lambda and configure your function to be triggered in response to Amazon CloudFront requests (i.e., when a viewer request lands, when a request is forwarded to or received back from the origin, and right before responding back to the end user). The code is then ready to execute across AWS locations globally when a request for content is received, and scales with the volume of CloudFront requests globally. Learn more in our documentation.

Q: How do I use Lambda@Edge?

To use Lambda@Edge, you just upload your code to AWS Lambda and associate a function version to be triggered in response to Amazon CloudFront requests. Your code must satisfy the Lambda@Edge service limits. Lambda@Edge supports Node.js and Python for global invocation by CloudFront events at this time. Learn more in our documentation.

Q: When should I use Lambda@Edge?

Lambda@Edge is optimized for latency sensitive use cases where your end viewers are distributed globally. All the information you need to make a decision should be available at the CloudFront edge, within the function and the request. This means that use cases where you are looking to make decisions on how to serve content based on user characteristics (e.g., location, client device, etc) can now be executed and served close to your users without having to be routed back to a centralized server.

Q: Can I deploy my existing Lambda functions for global invocation?

You can associate existing Lambda functions with CloudFront events for global invocation if the function satisfies the Lambda@Edge service requirements and limits. Read more here on how to update your function properties.

Q: What Amazon CloudFront events can be used to trigger my functions?

Your functions will automatically trigger in response to the following Amazon CloudFront events:

- Viewer Request - This event occurs when an end user or a device on the Internet makes an HTTP(S) request to CloudFront, and the request arrives at the edge location closest to that user.

- Viewer Response - This event occurs when the CloudFront server at the edge is ready to respond to the end user or the device that made the request.
- Origin Request - This event occurs when the CloudFront edge server does not already have the requested object in its cache, and the viewer request is ready to be sent to your backend origin webserver (e.g. Amazon EC2, or Application Load Balancer, or Amazon S3).
- Origin Response - This event occurs when the CloudFront server at the edge receives a response from your backend origin webserver.

Q: How is AWS Lambda@Edge different from using AWS Lambda behind Amazon API Gateway?

The difference is that API Gateway and Lambda are regional services. Using Lambda@Edge and Amazon CloudFront allows you to execute logic across multiple AWS locations based on where your end viewers are located.

Scalability and availability

Q: How available are AWS Lambda functions?

AWS Lambda is designed to use replication and redundancy to provide high availability for both the service itself and for the Lambda functions it operates. There are no maintenance windows or scheduled downtimes for either.

Q: Do my AWS Lambda functions remain available when I change my code or its configuration?

Yes. When you update a Lambda function, there will be a brief window of time, typically less than a minute, when requests could be served by either the old or the new version of your function.

Q: Is there a limit to the number of AWS Lambda functions I can execute at once?

No. AWS Lambda is designed to run many instances of your functions in parallel. However, AWS Lambda has a default safety throttle for number of concurrent executions per account per region (visit here for info on default safety throttle limits). You can also control the maximum concurrent executions for individual AWS Lambda functions which you can use to reserve a subset of your account concurrency limit for critical functions, or cap traffic rates to downstream resources.

If you wish to submit a request to increase the throttle limit you can visit our Support Center, click "Open a new case," and file a service limit increase request.

Q: What happens if my account exceeds the default throttle limit on concurrent executions?

On exceeding the throttle limit, AWS Lambda functions being invoked synchronously will return a throttling error (429 error code). Lambda functions being invoked asynchronously can absorb reasonable bursts of traffic for approximately 15-30 minutes, after which incoming events will be rejected as throttled. In case the Lambda function is being invoked in response to Amazon S3 events, events rejected by AWS Lambda may be retained and retried by S3 for 24 hours. Events from Amazon Kinesis streams and Amazon DynamoDB streams are retried until the Lambda function succeeds or the data expires. Amazon Kinesis and Amazon DynamoDB Streams retain data for 24 hours.

Q: Is the default limit applied on a per function level?

No, the default limit only applies at an account level.

Q: What happens if my Lambda function fails during processing an event?

On failure, Lambda functions being invoked synchronously will respond with an exception. Lambda functions being invoked asynchronously are retried at least 3 times. Events from Amazon Kinesis streams and Amazon DynamoDB streams are retried until the Lambda function succeeds or the data expires. Kinesis and DynamoDB Streams retain data for a minimum of 24 hours.

Q: What happens if my Lambda function invocations exhaust the available policy?

On exceeding the retry policy for asynchronous invocations, you can configure a “dead letter queue” (DLQ) into which the event will be placed; in the absence of a configured DLQ the event may be rejected. On exceeding the retry policy for stream based invocations, the data would have already expired and therefore rejected.

Q: What resources can I configure as a dead letter queue for a Lambda function?

You can configure an Amazon SQS queue or an Amazon SNS topic as your dead letter queue.

Security and access control

Q: How do I allow my AWS Lambda function access to other AWS resources?

You grant permissions to your Lambda function to access other resources using an IAM role. AWS Lambda assumes the role while executing your Lambda function, so you always retain full, secure control of exactly which AWS resources it can use. Visit Setting up AWS Lambda to learn more about roles.

Q: How do I control which Amazon S3 buckets can call which AWS Lambda functions?

When you configure an Amazon S3 bucket to send messages to an AWS Lambda function a resource policy rule will a be created that grants access. Visit the Lambda Developer's Guide to learn more about resource policies and access controls for Lambda functions.

Q: How do I control which Amazon DynamoDB table or Amazon Kinesis stream an AWS Lambda function can poll?

Access controls are managed through the Lambda function’s role. The role you assign to your Lambda function also determines which resource(s) AWS Lambda can poll on its behalf. Visit the Lambda Developer's Guide to learn more.

Q: How do I control which Amazon SQS queue an AWS Lambda function can poll?

Access controls can be managed by the Lambda function's role or a resource policy setting on the queue itself. If both policies are present, the more restrictive of the two permissions will be applied.

Q: Can I access resources behind Amazon VPC with my AWS Lambda function?

Yes. You can access resources behind Amazon VPC.

Q: How do I enable and disable the VPC support for my Lambda function?

To enable VPC support, you need to specify one or more subnets in a single VPC and a security group as part of your function configuration. To disable VPC support, you need to update the function configuration and specify an empty list for the subnet and security group. You can change these settings using the AWS APIs, CLI, or AWS Lambda Management Console.

Q: Can a single Lambda function have access to multiple VPCs?

No. Lambda functions provide access only to a single VPC. If multiple subnets are specified, they must all be in the same VPC. You can connect to other VPCs by peering your VPCs.

Q: Can Lambda functions in a VPC also be able to access the internet and AWS Service endpoints?

Lambda functions configured to access resources in a particular VPC will not have access to the internet as a default configuration. If you need access to external endpoints, you will need to create a NAT in your VPC to forward this traffic and configure your security group to allow this outbound traffic.

AWS Lambda functions in Java

Q: How do I compile my AWS Lambda function Java code?

You can use standard tools like Maven or Gradle to compile your Lambda function. Your build process should mimic the same build process you would use to compile any Java code that depends on the AWS SDK. Run your Java compiler tool on your source files and include the AWS SDK 1.9 or later with transitive dependencies on your classpath. For more details, see our documentation.

Q: What is the JVM environment Lambda uses for execution of my function?

Lambda provides the Amazon Linux build of openjdk 1.8.

AWS Lambda functions in Node.js

Q: Can I use packages with AWS Lambda?

Yes. You can use NPM packages as well as custom packages. Learn more here.

Q: Can I execute other programs from within my AWS Lambda function written in Node.js?

Yes. Lambda's built-in sandbox lets you run batch ("shell") scripts, other language runtimes, utility routines, and executables. Learn more here.

Q: Is it possible to use native modules with AWS Lambda functions written in Node.js?

Yes. Any statically linked native module can be included in the ZIP file you upload, as well as dynamically linked modules compiled with an rpath pointing to your Lambda function root directory. Learn more here.

Q: Can I execute binaries with AWS Lambda written in Node.js?

Yes. You can use Node.js' child_process command to execute a binary that you've included in your function or any executable from Amazon Linux that is visible to your function. Alternatively several NPM packages exist that wrap command line binaries such as node-ffmpeg. Learn more here.

Q: How do I deploy AWS Lambda function code written in Node.js?

To deploy a Lambda function written in Node.js, simply package your Javascript code and dependent libraries as a ZIP. You can upload the ZIP from your local environment, or specify an Amazon S3 location where the ZIP file is located. For more details, see our documentation.

AWS Lambda functions in Python

Q: Can I use Python packages with AWS Lambda?

Yes. You can use pip to install any Python packages needed.

AWS Lambda functions in C#

Q: How do I package and deploy an AWS Lambda function in C#?

You can create a C# Lambda function using the Visual Studio IDE by selecting "Publish to AWS Lambda" in the Solution Explorer. Alternatively, you can directly run the "dotnet lambda publish" command from the dotnet CLI which has the [# Lambda CLI tools patch] installed, which creates a ZIP of your C# source code along with all NuGet dependencies as well as your own published DLL assemblies, and automatically uploads it to AWS Lambda using the runtime parameter “dotnetcore1.0”

AWS Lambda functions in PowerShell

Q: How do I deploy AWS Lambda function code written in PowerShell?

A PowerShell Lambda deployment package is a ZIP file that contains your PowerShell script, PowerShell modules that are required for your PowerShell script, and the assemblies needed to host PowerShell Core. You then use the *AWSLambdaPSCore* PowerShell module that you can install from the PowerShell Gallery to create your PowerShell Lambda deployment package.

AWS Lambda functions in Go

Q: How do I package and deploy an AWS Lambda function in Go?

Upload your Go executable artifact as a ZIP file through the AWS CLI or Lambda console and select the go1.x runtime. With Lambda, you can use Go's native tools to build and package your code. For more details, read our documentation.

AWS Lambda functions in Ruby

Q: How do I deploy AWS Lambda function code written in Ruby?

To deploy a Lambda function written in Ruby, package your Ruby code and gems as a ZIP. You can upload the ZIP from your local environment, or specify an Amazon S3 location where the ZIP file is located.

Other topics

Q: Which versions of Amazon Linux, Node.js, Python, JDK, .NET Core, SDKs, and additional libraries does AWS Lambda support?

You can view the list of supported versions here.

Q: Can I change the version of Amazon Linux or any language runtime?

No. AWS Lambda offers a single version of the operating system and managed language runtime to all users of the service. You can bring your own language runtime to use in Lambda.

Q: How can I record and audit calls made to the AWS Lambda API?

AWS Lambda is integrated with AWS CloudTrail. AWS CloudTrail can record and deliver log files to your Amazon S3 bucket describing the API usage of your account.

Q: How do I coordinate calls between multiple Lambda functions?

You can use Amazon Step Functions to coordinate multiple invoking Lambda functions. You can invoke multiple Lambda functions serially, passing the output of one to the other, or in parallel. See our documentation for more details.

AWS Outposts FAQs

General

Q: Why would I use AWS Outposts instead of operating in an AWS Region?

You can use Outposts to support your applications that have low latency or local data processing requirements. These applications may need to generate near real-time responses to end user applications or need to communicate with other on-premises systems or control on-site equipment. These can include workloads running on factory floors for automated operations in manufacturing, real-time patient diagnosis or medical imaging, and content and media streaming. You can use Outposts to securely store and process customer data that needs to remain on premises or in countries where there is no AWS region. You can run data intensive workloads on Outposts and process data locally when transmitting data to the cloud is expensive and wasteful and for better control on data analysis, back-up and restore.

Q: In which regions is Outposts available?

Outposts is supported in the following regions and customers can connect their Outposts to the following regions:

US East (Ohio)	us-east-2
US East (N. Virginia)	us-east-1
US West (N. California)	us-west-1
US West (Oregon)	us-west-2
Canada (Central)	ca-central-1
EU (Frankfurt)	eu-central-1
EU (Ireland)	eu-west-1

EU (Stockholm)	eu-north-1
EU (Paris)	eu-west-3
EU (London)	eu-west-2
Middle East (Bahrain)	me-south-1
Asia Pacific (Sydney)	ap-southeast-2
Asia Pacific (Tokyo)	ap-northeast-1
Asia Pacific (Seoul)	ap-northeast-2
Asia Pacific (Singapore)	ap-southest-1
Asia Pacific (Hong Kong)	ap-east-1

Q: In which countries will Outposts be available?

A: Outposts can be shipped to and installed in the following countries

- NA - US, Canada
- EMEA - All EU countries, Switzerland, Norway, and Bahrain
- APAC - Australia, Japan, South Korea, Singapore, and Hong Kong Special Administrative Region

Support for more countries coming soon.

Q: Can I order an Outpost to a country where Outposts has not launched and link it back to a supported Region?

A: No, we can deliver and install Outposts only in countries where Outposts can be delivered and supported.

Q: Can I use Outposts when it is not connected to the AWS Region or in a disconnected environment?

An Outpost relies on connectivity to the parent AWS Region. Outposts are not designed for disconnected operations or environments with limited to no connectivity. We recommend that customers have highly available networking connections back to their AWS Region. If interested in leveraging AWS services in disconnected environments such as cruise ships or remote mining locations, learn more about AWS services such as Snowball Edge which is optimized to operate in environments with limited to no connectivity.

Q: Can Outposts be used to meet data sovereignty requirements?

Outposts may address data sovereignty based on customers' requirements. You can ensure that sensitive customer data stays on premises. However control plane data (e.g. instance IDs, monitoring metrics, metering records, tags, etc.) will flow back to the AWS Region. This addresses many of the data residency requirements we hear from our customers. We recommend you confirm with your compliance teams to ensure this addresses your residency and sovereignty requirements.

Q: Can S3 for Outposts support data residency or sovereignty requirements?

S3 for Outposts can be configured to never replicate data to the region. However, it is up to the customer to determine if it meets their business needs based on architecture and durability. For instance, certain control plane data such as bucket names and metrics may be stored in the AWS Region for reporting and management. More detail will be available when S3 for Outposts is generally available.

Q: Can I reuse my existing servers in an Outpost?

No, AWS Outposts leverages AWS designed infrastructure, and is only supported on AWS-designed hardware that is optimized for secure, high-performance, and reliable operations.

Q: Is there a software-only version of AWS Outposts?

No, AWS Outposts is a fully managed service that provides you with native access to AWS services.

Q: Can I order my own hardware that can be installed as part of my Outpost rack?

No, AWS Outposts provides fully integrated AWS designed configurations with built in top-of-rack switches and redundant power supply to ensure an ideal AWS experience. You can order as much compute and storage infrastructure as you need by selecting from the range of available Outpost options, or work with us to create a custom combination with your desired EC2 and EBS capacity. These are pre-validated and tested to ensure that you can get started quickly with no additional effort or configuration required on-site.

AWS services

Q: Can I create EC2 instances using an EBS backed AMI on my Outposts?

A: Yes, you can launch EC2 instances using the AMIs backed with EBS gp2 volume types..

Q: Where are EBS snapshots stored?

A: Any EBS snapshots will be stored using Amazon S3 in the Region associated with your Outpost.

Getting started with ordering & installation

Q: Are there any prerequisites for deploying an Outpost at my location?

A: Your site must support the basic power, networking and space requirements to host an Outpost. Outposts need 5-15 kVA, can support 1/10/40/100 Gbps uplinks, and space for a 42U rack (80” X 24” X 48” dimensions). As Outposts require reliable network connectivity to the AWS Region, you should plan for a public internet connection. Customers must have Enterprise Support, which provides 24x7 remote support within 15 mins.

Security & compliance

Q: Do the same compliance certifications for AWS Services today apply for services on Outposts?

A: No, the existing certifications for AWS Services are applicable to services running entirely in an AWS Region. AWS Outposts and services on Outposts (i.e., launched locally on an Outpost) will go through a separate evaluation for certifications. Compared to certification for other AWS services, with AWS Outposts the customer owns the responsibility for physical security and access controls around the Outpost for compliance certification.

Q: Who is responsible for the physical security of the Outposts at my datacenter?

A: Customers are responsible for attesting to physical security and access controls around the Outpost as part of a shared responsibility model.

Support & maintenance

Q: How does AWS maintain AWS Outposts infrastructure?

A: When your Outpost is installed and is visible in the AWS Management Console, AWS will monitor it as part of the public Region and will automatically execute software upgrades and patches.

If there is a need to perform physical maintenance, AWS will reach out to schedule a time to visit your site. AWS may replace a given module as appropriate but will not perform any host or network switch servicing on customer premises.

Q: What happens when my facility's network connection goes down?

A: EC2 instances and EBS volumes on the Outpost will continue to operate normally and can be accessed locally via the local gateway. Similarly, AWS service resources such as ECS worker nodes continue to run locally. However, API availability will be degraded, for instance run/start/stop/terminate APIs may not work. Instance metrics and logs will continue to be cached locally for a few hours, and will be pushed to the AWS Region when connectivity returns.

Disconnection beyond a few hours however may result in loss of metrics and logs. As Route53 DNS will not resolve when disconnected, an on-premise DNS resolver should be used if network disconnections are expected. If you expect to lose network connectivity, we strongly recommend regularly testing your workload to ensure it behaves properly in this state when an Outpost is disconnected.

Q: What type of control plane information flows back to the parent Region?

A: As an example, information about instance health, instance activity (launched, stopped), and the underlying hypervisor system may be sent back to the parent AWS Region. This information enables AWS to provide alerting on instance health and capacity, and apply patches and updates to the Outpost. Your team does not need to implement your own tooling to manage these elements, or to actively push security updates and patches for your Outpost. When disconnected, this information cannot be sent back to the parent Region.

FAQs and Terms

GENERAL

Q: What are serverless applications?

Q: What is the AWS Serverless Application Repository?

Q: In which regions is the AWS Serverless Application Repository available?

Q: What kinds of applications are available in the AWS Serverless Application Repository?

Q: Does it cost anything to use the AWS Serverless Application Repository?

Q: How are applications in the AWS Serverless Application Repository licensed?

Q: Are applications in the AWS Serverless Application Repository verified by AWS?

Q: Can I use the AWS Serverless Application Repository in combination with GitHub?

DEPLOYING APPLICATIONS

Q: How do I manage serverless applications deployed to my account?

PUBLISHING APPLICATIONS

Q: How do I publish a serverless application to the AWS Serverless Application Repository?

Q: Who can deploy the applications I publish to the AWS Serverless Application Repository?

Q: Can I charge a fee for serverless applications I publish?

USING NESTED APPLICATIONS

Q: What is a nested application?

Q: How are nested applications deployed?

Q: How do I include a nested application in my SAM template?

Q: How do I handle resource name conflicts when nesting applications?

Q: How do I package a nested application?

Q: Can I package a hierarchy of nested applications using SAM CLI?

Consuming Nested Applications

Q: How do I manage serverless applications deployed to my account?

Q: How do I nest applications shared with me via the Serverless Application Repository?

Q: How do I package a nested app that I used from the Serverless Application Repository?

Q: What happens if an application that I nested is no longer available?

Q: How can I tell if an application contains other nested applications?

Publishing Nested Applications

Q: Can I publish nested applications to the Serverless Application Repository?

Sharing Nested Applications

Q: How do I share an application that contains nested applications?

Q: How does sharing work when there is a hierarchy of nested applications?

USER TERMS

1. Publishers, who are AWS customers, may submit their AWS serverless applications and components (“AWS Serverless Applications”) to be made available through the AWS Serverless Application repository (“Repository”) either privately, across specified AWS accounts or to all AWS customers using the Repository pursuant to the Repository console publication process. AWS Serverless Applications to be made available either privately or across specified AWS accounts may be in binary or source code form; AWS Serverless Applications made available to all AWS customers may be in binary or source code form, and must include sufficient details to enable the user access to the source code.

2. Publishers must have all licenses and necessary permissions or rights to submit their AWS Serverless Applications to the Repository. Publisher must submit to the Repository, along with its AWS Serverless Application, the terms of the AWS Serverless Application's license(s), including any open source license attribution requirements. Publisher is responsible for reviewing, evaluating and testing any AWS Serverless Application before submitting it to the Repository.

3. Publisher hereby grants AWS and its affiliates the rights to reproduce, distribute, display publicly or within specified AWS accounts (as applicable), perform, transmit, use and otherwise digitally make available (via all means of online and electronic distribution) its AWS Serverless Applications in the Repository.

4. Publisher represents and warrants it has all rights to submit its AWS Serverless Application to the Repository, has all rights to allow downloading of its AWS Serverless Application from the Repository and has provided all required attributions. Publisher will not submit AWS Serverless Applications with malware, malicious or other harmful content with the intent or purpose to harm others. AWS may remove and take down any AWS Serverless Application in the sole discretion of AWS for this or other reasons.

5. AWS customers will comply with the license(s) (including any attribution or other requirements) for any AWS Serverless Application they download.

6. Any AWS customer who creates a derivative work of any AWS Serverless Application is responsible for determining whether it has the appropriate rights under the AWS Serverless Application's license(s) to do so and must comply with any attribution or other requirements.

7. Any Publisher's AWS Serverless Application license or other agreement is solely between the Publisher and AWS customers. Neither AWS nor any of its affiliates are a party to that license or other agreement and none of them will have any liability or obligations under that license or other agreement. AWS is not responsible and has no liability for ensuring that Publishers or AWS customers comply with licensing (including attribution) or other requirements.

8. AWS Serverless Applications and any other third-party materials available in the Repository are "Repository Content." THE Repository Content IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL PUBLISHERS, COPYRIGHT HOLDERS, AWS OR ITS AFFILIATES BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN

CONNECTION WITH THE Repository Content OR THE USE OR OTHER DEALINGS IN THE Repository Content.

9. Publishers and AWS customers acknowledge they will comply with all of these terms in their use of the Repository and these terms may be updated by AWS from time to time.

AWS Wavelength FAQs

Q: What is Wavelength?

Wavelength combines the high bandwidth and single-digit millisecond latency of 5G networks with AWS compute and storage services to enable developers to innovate and build a whole new class of applications. Wavelength will initially be available in partnership with Verizon starting in 2020. AWS is also working with other carriers like Vodafone, SK Telecom, and KDDI to expand Wavelength Zones to more locations by the end of 2020.

Q: What is a Wavelength Zone?

Wavelength Zones are AWS infrastructure deployments that embed AWS compute and storage services within telecommunications providers' datacenters at the edge of the 5G network, so application traffic can reach application servers running in Wavelength Zones without leaving the mobile providers' network. This prevents the latency that would result from multiple hops to the Internet and enables customers to take full advantage of 5G networks. Wavelength Zones extend AWS to the 5G edge, delivering a consistent developer experience across multiple 5G networks around the world and allowing developers to build the next generation of ultra-low latency applications using the same familiar AWS services, APIs, tools, and functionality they already use today.

Q: Who should use Wavelength?

You should use Wavelength when you need to deploy high performance applications that can be accessed by mobile end-users and devices that require single digit millisecond latency. AWS customers who want to build public applications like game streaming and AR/VR services can use Wavelength to reach end-users with millisecond-level connections, optimizing the user experience and performance of their applications. AWS enterprise customers that build applications to serve their own use-cases such as IoT, live media production, and industrial automation can use Wavelength to deliver low-latency solutions. Customers with edge data processing needs such as image and video recognition, inference, data aggregation, and responsive analytics can use Wavelength to perform low-latency operations and processing right where their data is generated, reducing the need to move large amounts of data to be processed in centralized locations.

Q: Why should I use Wavelength?

Wavelength lets you go from the device on the 5G network to your application’s resources on the AWS cloud with the fewest network hops because the compute and storage are hosted directly within the telco providers’ 5G networks. This reduces latency caused by network congestion or longer routes that would be required to access application resources outside of the 5G network, making possible new classes of applications that are both compute-intensive and very sensitive to latency (e.g. a fleet of autonomous cars interacting with road sensors to prevent crashes, or smart industrial robots assessing and reacting to plant conditions in a dangerous manufacturing environment, or retailers serving personalized promotions to shoppers’ mobile phones in real-time as they pass product displays). Wavelength brings the power of the AWS cloud to the network edge to enable latency sensitive use cases that require near real-time responses, and where processing at the network edge can be used to avoid transmitting large volumes of data over the network provider’s infrastructure, and to offload processing from the hardware on mobile devices.

Q: How should I think about when to use AWS Wavelength, AWS Local Zones, or AWS Outposts for applications that require low latency or local data processing?

AWS is helping customers by delivering a consistent experience to support applications with low latency or local data processing requirements wherever they need to be deployed.

AWS Wavelength is designed to deliver ultra-low latency applications to 5G devices by extending AWS infrastructure, services, APIs, and tools to 5G networks. Wavelength embeds storage and compute inside telco providers' 5G networks to help developers build new applications for 5G end users that require single-digit millisecond latency, like IoT devices, game streaming, autonomous vehicles, and live media production.

AWS Local Zones are a new type of AWS infrastructure designed to run workloads that require single-digit millisecond latency in more locations , like video rendering and graphics intensive, virtual desktop applications. Not every customer wants to operate their own on-premises data center, while others may be interested in getting rid of their local data center entirely. Local Zones allow customers to gain all the benefits of having compute and storage resources closer to end-users, without the need to own and operate their own data center infrastructure.

AWS Outposts is designed for workloads that need to remain on-premises due to latency requirements, where customers want that workload to run seamlessly with the rest of their other workloads in AWS. AWS Outposts are fully managed and configurable compute and storage racks built with AWS-designed hardware that allow customers to run compute and storage on-premises, while seamlessly connecting to AWS’s broad array of services in the cloud.

Check out the product features

Learn more »

Sign up to learn more

Sign up »

VMware Cloud on AWS | FAQs

General

What is VMware Cloud on AWS?

VMware Cloud on AWS brings VMware’s enterprise-class SDDC software to the AWS Cloud with optimized access to native AWS services. Powered by VMware Cloud Foundation, VMware Cloud on AWS integrates VMware's compute, storage and network virtualization products (VMware vSphere, VMware vSAN and VMware NSX) along with VMware vCenter Server management, optimized to run on dedicated, elastic, bare-metal AWS infrastructure.

Why should I use VMware Cloud on AWS?

VMware Cloud on AWS provides you consistent and interoperable infrastructure and services between VMware-based datacenters and the AWS cloud, which minimizes the complexity and associated risks of managing diverse environments. VMware Cloud on AWS offers native access to AWS services and innovation that extends the value of enterprise applications over their lifecycle.

Where is VMware Cloud on AWS available today?

The service is newly available in AWS South America (Sao Paulo) and AWS Asia Pacific (Seoul). In addition, the service is also available in AWS US East (N. Virginia), AWS US East (Ohio), AWS US West (N. California), AWS US West (Oregon), AWS Canada (Central), AWS Europe (Frankfurt), AWS Europe (Ireland), AWS Europe (London), AWS Europe (Paris), AWS Asia Pacific (Singapore), AWS Asia Pacific (Sydney), AWS Asia Pacific (Tokyo), AWS Asia Pacific (Mumbai) Region and AWS GovCloud (US West) regions.

Can workloads running in a VMware Cloud on AWS instance integrate with AWS services?

Yes. VMware Cloud on AWS SDDC is directly connected to customer’s VPC using Elastic Network Interface(ENI) and therefore has access to AWS services. Virtual machine workloads can access public API endpoints for AWS services such as AWS Lambda, Amazon Simple Queue Service (SQS), Amazon S3 and Elastic Load Balancing, as well as private resources in the customer's Amazon VPC such as Amazon EC2, and data and analytics services such as Amazon RDS, Amazon DynamoDB, Amazon Kinesis and Amazon Redshift. Customers can also now enjoy Amazon Elastic File System (EFS) for fully managed file service to scale the file-based storage automatically to petabyte scale with high availability and durability across multiple Availability Zones (AZs) and the newest generation of VPC Endpoints designed to access AWS services while keeping all the traffic within the AWS network.

How do I get started with VMware Cloud on AWS?

With a new purchase agreement in place, customers can now buy VMware Cloud on AWS directly through AWS and AWS Partner Network (APN) Partners in the AWS Solution Provider Program. This allows customers the flexibility to purchase VMware Cloud on AWS either through AWS or VMware, or the AWS Solution Provider or VMware VPN Solution Provider of their choice. Through our partnership, customers can use additional services like Amazon RDS for VMware, AWS Outposts for VMware (available in the second half of 2019), and new Amazon EC2 R5 Instance types.

Can I use my existing VMware licenses when using VMware Cloud on AWS?

You can leverage your existing VMware software investments to secure additional discounts for your VMware Cloud on AWS hybrid environment as part of VMware’s Hybrid Loyalty Program.

Can I use my existing Windows Server licenses in VMware Cloud on AWS?

Yes. Please consult your Microsoft Product Terms for more details and any restrictions.

What is single host SDDC starter configuration?

Single host SDDC starter configuration is a time-bound offering for customers to kickstart their VMware Cloud on AWS on-demand hybrid experience at a low, predictable price. Service life for the single host SDDC is limited to 30-day intervals only. This new consumption option is designed for customers who want to prove the value of VMware Cloud on AWS in their environment before scaling to 3+ host configurations for production environments.

What compliance certifications has VMware Cloud on AWS achieved?

VMware Cloud on AWS has been independently verified to comply with ISO 27001, ISO 27017, ISO 27018, SOC 1, SOC 2, SOC 3, and HIPAA. VMware Cloud on AWS also complies with the General Data Protection Regulation (GDPR). For more information about VMware Cloud on AWS and GDPR compliance visit here.

How is VMware Cloud on AWS deployed?

VMware Cloud on AWS infrastructure runs on dedicated, single tenant hosts provided by AWS in a single account. Each host is equivalent to an Amazon EC2 I3.metal instance (2 sockets with 18 cores per socket, 512 GiB RAM, and 15.2 TB Raw SSD storage). Each host is capable of running many VMware Virtual Machines (tens to hundreds depending on their compute, memory and storage requirements). Clusters can range from a minimum 3 hosts up to a maximum of 16 hosts per cluster. A single VMware vCenter server is deployed per SDDC environment.

What version of VMware vSphere do I need in my on-premises environment?

With vSphere 6.0 or later running in your on-premises environment, you can move workloads to and from VMware Cloud on AWS by doing cold migration of VMs. No conversion or modification is necessary. In order to take advantage of “Hybrid Linked Mode” for single pane of glass management between your on-premises environment and VMware Cloud on AWS, you must have VMware vSphere 6.5 or later.

How do I manage resources on VMware Cloud on AWS?

You can use the same management tools you use today. A vCenter Server instance is deployed as part of every VMware Cloud on AWS SDDC. You may connect to this vCenter Server instance to manage their VMware Cloud on AWS clusters. A VMware Cloud Web Console is provided which allows for common tasks such as the add/remove hosts, configure firewalls and other basic networking settings. It is important to note that tools that require plug-ins or extensive vSphere permissions may not function properly in VMware Cloud on AWS. VMware Cloud on AWS uses a least privilege security model in which you (and therefore their tools) do not have full administrative access.

Can I manage both my existing data center VMware vSphere VMs and my VMware Cloud on AWS instances in a single view?

You will need vSphere version 6.5 and vCenter Server 6.5 or later running in your data center to use vCenter Hybrid Linked Mode for single pane of glass management of resources on-premises and in the cloud. If you do not have VMware vSphere 6.5 or later running in your on-premises environment, you will need to run multiple vCenter instances to manage your environment: one vCenter instance on-premises and one vCenter instance in VMware Cloud on AWS.

Can I migrate existing vSphere VMs to my VMware Cloud on AWS deployment?

Yes. There are multiple ways to migrate existing vSphere VMs to VMware Cloud on AWS. You can perform of a live migration of vSphere VMs via a vMotion or by leveraging VMware Hybrid Cloud Extension (HCX).

Compute

What are the hardware specifications for VMware Cloud on AWS hosts?

The VMware Cloud on AWS minimum standard cluster configuration contains 3 hosts. Each host is an Amazon EC2 I3.metal instance. These hosts have dual 2.3 GHz CPUs (custom-built Intel Xeon Processor E5-2686 v4 CPU package) with 18 cores per socket (36 cores total), 512 GiB RAM, and 15.2 TB Raw NVMe storage.

Will VMware Cloud on AWS be running on a “nested” ESXi architecture?

No. ESXi runs directly on ‘bare metal’ without the use of nested virtualization, while still participating in Amazon VPC networking.

Is the underlying EC2 infrastructure hosting ESXi dedicated to each customer or is it a shared, multi-tenant infrastructure?

VMware Cloud on AWS infrastructure runs on dedicated, single-tenant bare metal infrastructure for each customer.

Can I scale the hosts in my VMware Cloud on AWS cluster?

Yes, additional hosts can be added to a VMware Cloud on AWS cluster using the VMware Cloud Portal user interface or programmatically via the VMware Cloud API.

Can I increase or decrease the size of my cluster after I provision an SDDC on VMware Cloud on AWS?

Yes, you can add and/or remove hosts on-demand as long as the minimum cluster size is 3 hosts.

What is the maximum supported clusters size in VMware Cloud on AWS?

The maximum cluster size is 16 ESXi hosts.

Can I use the vCenter Server in my SDDC to manage my on-premises ESXi hosts?

Yes, with Hybrid Linked Mode, you can connect your vCenter Server running in VMware Cloud on AWS to your on-premises vCenter server to get a single inventory view of both your cloud and on-premises resources.

Storage

What storage options are available for VMware Cloud on AWS?

VMware Cloud on AWS includes VMware's vSAN storage technology that provides a single name space shared datastore (vSAN datastore) for VM storage. Each SDDC cluster will utilize an "all flash" vSAN storage solution built on NVMe backed instance storage that offers high performance, and low latency.

Recently announced in preview, VMware vSAN can also utilize Amazon Elastic Block Store (Amazon EBS) with VMware Cloud on AWS running on new Amazon EC2 R5.metal instances to augment existing SDDC for storage-dense environments. Storage per host ranges from 15 to 35 TB in increments of 5 TB. User chooses amount of storage desired and used on all hosts within the cluster. R5.metal clusters can be added to an existing SDDC with at least one existing provisioned cluster.

Can I use any hybrid storage (Flash + Spinning Disk)?

We currently do not offer a hybrid storage solution. All hosts are equipped with NVME Storage.

Can I expand my storage without adding additional hosts?

You will have to add additional hosts to increase your storage capacity.

Can I use Amazon Elastic Block Store (EBS) volumes as vSphere datastores?

Recently announced in preview, VMware vSAN can also utilize Amazon Elastic Block Store (Amazon EBS) with VMware Cloud on AWS running on new Amazon EC2 R5.metal instances to augment existing SDDC for storage-dense environments. Storage per host ranges from 15 to 35 TB in increments of 5 TB. User chooses amount of storage desired and used on all hosts within the cluster. R5.metal clusters can be added to an existing SDDC with at least one existing provisioned cluster.

Can I use AWS Elastic File System (EFS) volumes as vSphere datastores?

Customers cannot mount EFS volumes as ESXi datastores, but can mount EFS volumes to Linux VMs.

Will I be able to use VMware's Storage Policy Based Management (SPBM) to provision and manage storage for virtual machine applications?

Yes. You have the flexibility to create specific policies catering to your application needs, including RAID levels, checksum, object space reservation, and IOPS limit. You can apply these policies at the individual vdisk level, or you can choose the default vSAN Datastore policy for simplicity.

What Data Protection/Backup solutions are available for VMware Cloud on AWS?

Data protection solutions based on VMware's vStorage API for Data Protection (VADP) framework are being certified by partners now.

Is data encrypted on vSAN storage?

Yes, data is encrypted at rest for each NVMe flash device backing the vSAN datastore in each host.

How does data at rest encryption work in VMware Cloud on AWS?

Starting now, customer data at rest will be natively encrypted by vSAN. vSAN will use AWS Key Management Service (KMS) to generate the Customer Master Key (CMK). While CMK is acquired from AWS, two additional keys are generated by vSAN. Those keys are an intermediate key, referred as Key Encryption Key (KEK) and Disk Encryption Key (DEK).

The CMK wraps the KEK and the KEK in turn wraps the DEK. The CMK never leaves AWS control, encryption and decryption of the KEK is offered via an standard AWS API call.

One CMK and one KEK is required per cluster and one DEK for every disk in the cluster.

What provisions are available to rotate the keys?

Customers have the option to change the KEK (Key Encryption Key) either through vSAN API or through the vSphere UI. This process is called rekey. Note, shallow rekey doesn't change the Disk Encryption Key (DEK) or the CMK (Customer Master Key).

Changing the DEK and CMK is not supported. In rare situations, if there is a need to change the DEK or CMK, users have the option to set up a new cluster with new CMK and can Storage vMotion the data from existing cluster.

Can I turn on or turn off vSAN Encryption selectively?

Similar to D&C (Deduplication & Compression), vSAN encryption at rest cannot be turned on or off for individual clusters; it is a cluster-wide setting that is always on by default when cluster is provisioned in the SDDC.

What is expected of the clusters set up in prior releases with encryption turned off?

All existing clusters in M4 will be migrated to M5. As part of migration, encryption shall be turned on for all existing clusters. All new clusters will be provisioned with encryption turned on by default.

Are there other options for customers to bring their own keys?

The Customer Master Key is sourced from AWS Key Management Service and this is the only option available.

Network

How are VMware Cloud on AWS SDDCs connected to my on-premises environment?

When you deploy an SDDC using VMware Cloud on AWS, it is configured with two networks: a management network and a compute network. The management network handles network traffic for the SDDC hosts, vCenter Server, NSX Manager, and other management functions. The compute network handles network traffic for your workload VMs. Two VMware NSX edge devices serve as gateways for the VMware virtualized networking environment. The management gateway (MGW) connects the SDDC management infrastructure to your on-premises environment. The compute gateway (CGW) provides connectivity for all workload virtual machines. Traffic can be directed to your on-premises environment using a L3 VPN connection or to your AWS VPC via an Elastic Network Interface (ENI).

Will I be able to move an Elastic IP (EIP) from a VM in VMware Cloud on AWS to a standard Amazon EC2 instance and back again?

No. EIPs are reserved and associated from the VMware Cloud on AWS account and routed to the NSX Edge Gateway.

What network bandwidth will be available to the ESXi hosts?

ESXi hosts are connected to an AWS VPC via AWS Elastic Networking Adapter (ENA) that support throughput up to 25 Gbps.

Am I required to run NSX in on-premises installations when interacting with VMware Cloud on AWS?

No. Customers are not required to run NSX on-premises in order to interoperate with VMware Cloud on AWS. VMware Virtual Machines can be cold migrated to VMware Cloud on AWS without any modifications.

How do I connect to the vCenter Server in my SDDC on VMware Cloud on AWS?

By default, there is no external access to the vCenter Server system in your SDDC on VMware Cloud on AWS. Open access to your vCenter Server system by:

- Configuring a firewall rule to allow access to the vCenter Server system.
- Configuring an IPsec VPN between your on-premises data center and your SDDC

Does HCX support NSX-T SDDCs?
HCX supports all capabilities in both NSX-v and NSX-T SDDCs.
NSX-T SDDCs also support the ability to leverage the DX private VIF option for the HCX interconnects. If you are leveraging the internet and would like to shift your HCX interconnects to the private VIF option, please reach out to VMware via support to assist in switching the interconnect configuration.

Stretched Clusters

What is a VMware Cloud on AWS stretched cluster?

A stretched cluster is a deployment model in which two or more VMware Cloud on AWS clusters are part of the same logical cluster but are located in separate Availability Zones.

Can I update a single AZ SDDC to a "Stretched" SDDC?

No. Stretched cluster is a deployment time decision. You cannot upgrade a non-stretched cluster to a stretched cluster.

Can I downgrade a stretched cluster SDDC to a single AZ SDDC?

No. Stretched cluster is a deployment time decision. You cannot downgrade a stretched cluster to a non-stretched cluster.

What is the maximum size I can make a stretched cluster?

The maximum supported size for a stretched cluster is 16 nodes.

Can an SDDC contain both single AZ clusters and stretched clusters?

No. An SDDC can have either single AZ clusters or stretched clusters.

Can I migrate workloads from a single AZ cluster to a stretched cluster?

Yes. You can use all the normal vSphere and Hybrid Cloud Extension (HCX) tools to move workloads between SDDCs.

Can I choose the AZ in which my VMs run?

Yes. When deploying a VM you can choose an ESXi host in the desired AZ. In case of failure, the VM will stay in its original AZ if possible.

Can a stretched cluster span across AWS regions?

No. A stretched cluster spans across 2 AZ's within the same region. If you wish to protect against a regional failure, please use a DR tool such as VMware Site Recovery service.

Is there a performance impact when running VMs in a stretched cluster?

Yes. Because we are performing synchronous writes across two AZ's there is additional overhead in write transactions. This is the case in any stretched cluster implementation.

How many failures can be tolerated in an AZ?

This depends on your Storage Policy Based Management (SPBM) settings. By default, VM's are configured to survive the failure of ALL the hosts in a single AZ without data loss.

What happens when an AZ fails and when it comes back after a failure?

We will automatically re-synch the vSAN datastore. This resync time will depend on how much data you have stored and how long the systems have been segmented.

What is Multi-Cluster support?

Multi-Cluster support is the ability for SDDC administrators to add additional clusters to an existing SDDC. You are able to create multiple clusters in your SDDC and these will share a common set of management VMs and network.

How do I enable multi-cluster support?

Simply select "Add Cluster" from the VMware Cloud on AWS console to add a second cluster to your existing SDDC.

What is the maximum number of clusters supported?

VMware Cloud on AWS supports a maximum of 10 clusters per SDDC. Your organization may have lower "soft" limits set. If you wish to have your limits raised, please contact your customer success team.

Single Host SDDC Starter Configuration

What is the Single Host SDDC offering?

The VMware Cloud on AWS service minimum cluster size is 3 hosts. With the new time-bound Single Host SDDC starter configuration, you can now purchase a single host VMware Cloud on AWS environment with the ability to seamlessly scale the number of hosts up within that time period, while retaining your data. The service life of the Single Host SDDC starter configuration is limited to 30-day intervals. This single host offering applies to customers who want a lower-cost entry point for proving the value of VMware Cloud on AWS in their environments

Where is the Single Host SDDC available today?

The Single host SDDC is available across all the supported regions where VMware Cloud on AWS is available today.

What are the features included in the Single Host SDDC offering?

Features that do not require more than 1 host are included in the Single Host SDDC offering including hybrid operations between on-premises and VMware Cloud on AWS. However, any operations or capabilities that require more than 1 host would not work. For example, High Availability (HA) and Stretched Clusters across two AWS AZ. Due to the nature of single host, the FTT=0, meaning that if your host fails, your data would be lost. VMware does not currently offer patching or upgrades to a Single Host SDDC.

How many Single Host SDDCs can I provision?

You may provision no more than one Single Host SDDC at a time.

Can I run a Single Host SDDC indefinitely?

A Single Host SDDC will be deleted after 30 days. All data on the SDDC will be lost. You may scale-up a Single Host SDDC into a 3+ host SDDC and retain all your data. A 3+ host SDDC is not time bound.

How can I scale up to the standard 3-host service?

You can simply click on the "Scale Up" button to scale up to the 3-host SDDC service.

Can I convert my standard 3-host SDDC into a Single Host SDDC?

No, a Single Host SDDC must be created as a Single Host. You cannot scale down from 3-host SDDC to Single Host SDDC.

What support would I get for this offering?

Single Host SDDC receives the same unlimited 24/7 VMware Global Support Services as well as 24x5 live chat support via the VMware Cloud on AWS console and via vSphere Client.

How is single host SDDC priced?

Single host is available on-demand for $7/node/hour. Please visit the VMware pricing page for the latest information on pricing.

What is the single host SDDC offer?

First time VMware Cloud on AWS customers are eligible for a 20% discount for the first three months (90days) on the $7/node/hour price. The discount offer is valid on single host purchases on or before June 5th, 2019.

How do I take advantage of the single host offer?

You will receive $1,022 worth of VMware Cloud on AWS credits every month for first three months. These credits can be applied towards your bill. These credits expire if not used towards the service within two months of activation.

Add-on Capabilities

VMware Site Recovery

What is VMware Site Recovery?

VMware site recovery is an add-on service that is optimized for VMware Cloud on AWS to deliver simple, verifiable protection of critical applications between private data centers and AWS. Please visit VMware Site Recovery website to learn more

Can I use VMware Site Recovery on the Single Host SDDC offering?

Yes, the full set of capabilities of VMware Site Recovery is available for use as an add-on purchase to the Single Host SDDC starter configuration that serves as a low-cost option for you to jump-start your hybrid cloud disaster recovery solution. However, due to the time-bound nature and lower data durability of the Single Host SDDC offering, we recommend that you do not switch to using this as the primary disaster recovery solution for all of your on-premises workloads until you have successfully scaled up the environment to a 3-host SDDC.

Horizon 7 on VMware Cloud on AWS

What is Horizon 7 on VMware Cloud on AWS?

VMware Horizon 7 on VMware Cloud on AWS delivers a seamlessly integrated hybrid cloud for virtual desktops and applications. It combines the enterprise capabilities of VMware's Software-Defined Data Center, delivered as a service on AWS, with the market leading capabilities of VMware Horizon - for a simple, secure, and scalable solution.

Where can I find more information on Horizon 7 on VMware Cloud on AWS?

You can find overview information on the blog and VMware's website.

Is Horizon 7 part of VMware Cloud on AWS?

No. Horizon 7 is software that can be deployed by customer on the VMware Cloud on AWS. Ultimately the customer will be responsible for their Horizon 7 infrastructure, even though their SDDC infrastructure will be managed by VMware.

In what regions is Horizon 7 on VMware Cloud on AWS available?

Horizon 7 on VMware Cloud on AWS is available in all the same regions that VMware Cloud on AWS is available.

Support, Accounts, and Billing

Who sells and delivers VMware Cloud on AWS?

AWS, APN Partners or VMware can sell and deliver the service. Billing for the service will be between you and VMware. You will only receive a bill from AWS for services used in a separate AWS account.

Who provides support for VMware Cloud on AWS and do I also need to purchase AWS support?

VMware Cloud on AWS is supported by VMware. However, you can choose to purchase AWS support for the additional services you use.

Will I need an AWS account?

Yes, you will need an active AWS customer account that will be linked to the VMware Cloud on AWS service. If you don't have an existing AWS customer account, you will be asked to create one as part of the onboarding process. One of the key benefits of this offering is seamless integration with other AWS services such as Amazon S3, Redshift and other Amazon EC2 instances. VMware will bill you for what you use in the VMware Cloud on AWS and separately, AWS will bill you for any other AWS services you use.

What accounts are created during the process?

You will have a minimum of two accounts: a VMware account and an AWS account.

Can I have SDDCs from different regions in a single org?

Yes, an org can contain SDDCs from different regions.

Do I need to access region specific endpoints to access my SDDCs?

No, you use the same endpoints to access the VMware Cloud on AWS APIs and VMware Cloud on AWS Console regardless of the region your SDDCs are in.

Where can I find pricing for a specific region?

You can find out pricing at: https://cloud.vmware.com/vmc-aws/pricing

What is the VMware Cloud on AWS sizing and assessment tool?

You can use the sizing and assessment tool to size your workloads for VMware Cloud on AWS. The tool enables you to size for factors including Storage, Compute, Memory and IOPS in the logic to provide you with the most optimized Server and SDDC recommendation for VMware Cloud on AWS. Once you have completed sizing your workloads, you can calculate your total cost of ownership (TCO) for these workloads and compare it with an on-premises virtual environment. The tool will calculate the number of nodes and clusters required to support your workload to run on a VMware Cloud on AWS SDDC. Try the tool here.

How do I access the VMware Cloud on AWS sizing and assessment tool?

Customers can access the tool without any credentials. However, to complete the TCO, you must register with an email address and use those credentials to log into the tool.

Where can I find pricing for VMware Hybrid Cloud Extension for VMware Cloud on AWS?

VMware HCX is included with all VMware Cloud on AWS SDDC targets.

Amazon Connect FAQs

General

Q: Can I really set up Amazon Connect in minutes?

Yes. We encourage you to go to the Amazon Connect console and set up an Amazon Connect contact center now.

Q: Do you have examples of how customers are already using Amazon Connect?

Yes. Please see the Amazon Connect customers web page and the AWS contact center blog channel.

Getting started

Q: How can I try Amazon Connect?

Amazon Connect is self-service so you can try before you buy, without talking to sales, completing RFPs, or hiring for consultants or professional services. Simply log in to the Amazon Connect Console using your AWS account and set up an Amazon Connect instance. To learn how to set up Amazon Connect, see the Getting Started documentation, and visit the Amazon Connect Integration Quickstart Portal. To learn how to configure an Amazon Connect instance for your business, see the Administrator Guide.

Q: Is there a free trial for Amazon Connect?

Yes. As part of the AWS Free Tier, the following usage is included with Amazon Connect for free each month for the first 12 months that you use the service:

- 90 minutes per month of Amazon Connect service usage, which includes all minutes that any user is logged in to and using Amazon Connect, or a contact is active in your contact center, such as a customer interacting with an Amazon Lex bot.
- A local direct inward dial (DID) number for the AWS region
- 30 minutes per month of local (to the country in which the AWS region is located) inbound DID calls
- 30 minutes per month of local (to the country in which the AWS region is located) outbound calls
- For instances created in US regions, a US toll-free number for use per month and 30 minutes per month of US inbound toll-free calls
- 500 messages per month of Amazon Connect chat

For any additional usage, you will be charged at the published Amazon Connect pricing. To start your free trial, create an Amazon Connect instance in your AWS account and start using the service.

Communications and telephony

Q: How do end-customers interact with Amazon Connect?

Customers can call into your Amazon Connect contact center using any phone and speak to an agent. You can define other interactions with your customers in contact flows. For example, you can use voice or DTMF input, and text-to speech output using Amazon Polly, which can optionally be combined with Amazon Lex for natural language interactions.

Q: How do agents interact with Amazon Connect?

Agents use the Contact Control Panel (CCP) to interact with customers, such as answering calls, placing calls, or setting their status. For agent voice communications, Amazon Connect includes a web-based softphone for incoming and outgoing telephony, or agents can use a traditional telephone service using the PSTN.

Q: Do I need to bring my own telephony provider to use Amazon Connect?

No. Telephony services are included with Amazon Connect, which is designed to scale to meet your telephony needs.

Q: Does Amazon Connect support high quality audio?

Yes. With Amazon Connect, calls are made over the Internet from a computing device like a PC, using the Amazon Connect softphone. The Amazon Connect softphone delivers high-quality 16kHz audio and, is resistant to packet loss to ensure a high quality call experience.

Q: Does Amazon Connect support call recordings?

Yes. Amazon Connect enables you to store call recordings of customer interactions in Amazon S3. Interactions are not recorded unless an agent is connected. If multiple agents are connected, each will have an associated call recording.

Web and mobile chat

Q: How do my agents use Amazon Connect chat?

Agents chat with your customers using the Amazon Connect Contact Control Panel (CCP) which is the same web-based application agents use for voice engagements. Since it is web based, agents can work from virtually anywhere. The CCP SDK makes it easy to integrate with other apps like agent desktops or CRMs with just a few lines of code.

Q: How do I add Amazon Connect chat to my website or mobile app to enable my customers to contact me?

By leveraging our open source reference implementation and a few lines of code, you can easily add Amazon Connect chat into your existing website or mobile app. Customizing and branding the chat experience is easy using Amazon Connect's APIs and client SDKs. Learn more »

Q: Can I use Amazon Connect for just chat?

Yes. You can use Amazon Connect for just chat, just voice, or both. Existing instances of Amazon Connect are enabled for chat by default but you only pay for what you use.

ML-powered contact center analytics

Q: Does Amazon Connect have integrated machine learning speech-to-text or sentiment analysis?

Yes. Contact Lens for Amazon Connect is a set of machine learning (ML) capabilities integrated into Amazon Connect. With Contact Lens for Amazon Connect, contact center supervisors can better understand the sentiment, trends, and compliance risks of customer conversations to effectively train agents, replicate successful interactions, and identify crucial company and product feedback.

Using AWS machine learning natural language processing (NLP) and speech-to-text, Contact Lens for Amazon Connect transcribes contact center calls to create a fully searchable archive and surface valuable customer insights.

Contact flows

Q: What is a Contact flow?

Contact flows define the experience your customers have when they interact with your contact center. These are similar in concept to Interactive Voice Response (IVR). Contact flows are comprised of blocks, with each block defining a step or interaction in your contact center. For example, there are blocks to play a prompt, get input from a customer, branch based on customer input, or invoke a AWS Lambda function or and Amazon Lex bot.

Q: How do I edit Amazon Connect Contact flows?

Amazon Connect provides a Contact flow editor that lets you drag-and-drop blocks onto the designer canvas, and then use connectors to connect the blocks in the flow. You can configure the settings for each block after adding it to the designer.

Q: Does Amazon Connect offer text-to-speech in Contact flows?

Yes. Amazon Connect has built in text-to-speech leveraging Amazon Polly. You can access all the languages and voices offered by Amazon Polly.

Q: How do I use Amazon Lex chatbots with Amazon Connect?

You can use an Amazon Lex chatbot in your Amazon Connect contact flow to interact with callers. Callers can provide input to the chatbot by either speaking an utterance or pressing digits on their keypads. Amazon Lex interprets the spoken utterance or dual-tone multi-frequency signaling (DTMF) digits entered on a keypad, and uses them to understand the caller's intent, or for collecting slot information like a zip code or account number.

Q: What are contact attributes in Amazon Connect?

Contact attributes are pieces of data about a contact that you can use to personalize the customer experience, make routing decisions about contacts as they progress through your contact center orretrieve real-time metrics about the queues and agents in your contact center to dynamically route contacts based on queue and agent availability.

For more information, see Contact Attributes in the Amazon Connect documentation.

Skills-based routing

Q: Does Amazon Connect support skills-based routing?

Yes. Contacts can be routed based on availability, agent skill set, customer sentiment, or past history.

Q: Do you have to recreate routing rules and profiles for each channel (e.g. voice and chat)?

No. You can build once and use across channels. Amazon Connect has a single user interface, configuration, workflow, and routing engine for calls and chat.

Q: Can I adjust the priority of contacts in queue?

Yes. You can adjust priority of contacts in the queue using the set routing priority block in an Amazon Connect Contact flow.

Metrics and reporting

Q: What type of metrics reporting does Amazon Connect support?
Amazon Connect offers three metrics experiences:

- Historical metrics reports—Generate reports to analyze how your contact center has performed over a specified period of time. You can generate granular or aggregated reports pivoted on queues, individual agents, and phone numbers.
- Real-time metrics reports—Gain insight into how your contact center is performing in real time. You can see reports pivoted on queues, agents, and routing profiles
- Contact search—View detailed individual contact reports with the option to find and play back call recordings.

Q: Can I create a dashboard to view the metrics reports I've defined?

Yes. You can use the comprehensive dashboard to define and monitor the service levels and agent occupancy performance indicators that are most important to you. You can configure the dashboard so that the metrics you care about are always visible. You can configure your dashboard from the home page of your Amazon Connect contact center instance.

Ecosystem and integrations

Q: Does Amazon Connect work with other AWS services?

Yes. Amazon Connect integrates with several AWS services to provide a richer depth of capabilities and customization, including:

- Amazon Connect can leverage AWS Directory Services for identity and access management.
- Amazon Connect stores any call recordings and scheduled metrics reports in Amazon S3 buckets in your account, letting you control the lifecycle management and retention of your data.
- Amazon Connect can invoke AWS Lambda functions to perform data dips, send encrypted customer input, and other external integration in contact flows.
- Amazon Connect can stream metrics and agent event data to Amazon Kinesis Data Stream or Amazon Kinesis Data Firehose. Amazon Elastic Search Service can consume the data from Amazon Kinesis to enable advanced monitoring.
- Amazon Connect can encrypt data from your contact center, such as call recordings and reports, with encryption keys stored with Amazon Key Management Service.
- Amazon Connect leverages Amazon Lex for Natural Language Understanding and automated customer interactions.
- Amazon Connect uses Amazon CloudWatch for operational metrics and alarms.
- Amazon Connect uses Amazon Polly to provide the voice for text-to-speech messages.

Q: Does Amazon Connect integrate with my existing or other third party systems?

Yes. Amazon Connect is an open platform so it is easy to integrate with existing or other third party systems. Amazon Connect provides out-of-the-box integrations for leading customer relationship management (CRM) offerings, such as Salesforce and Zendesk, Workforce Management (WFM), and Analytics tools.

You can also use Amazon Connect with other AWS services like Amazon S3 and AWS Lambda for storing recorded calls or streaming detailed contact records in real-time to a data warehouse to merge with business intelligence systems for further analysis. Amazon Connect provides an API so you can customize the solution to your needs.

Q: Is there a network of consulting partners should I need their help?

Yes. A set of APN Consulting Partners with the requisite AWS and Amazon Connect knowledge have been approved by the Amazon Connect team to help you successfully configure and implement Amazon Connect. Please refer to the Amazon Connect Partner website for a list of these partners, and check back for updates as additional partners are approved.

Billing

Q: How much does it cost to use Amazon Connect?

With Amazon Connect, you only pay for what you use. There is a per minute charge for the Amazon Connect service. In addition to the charge for the service, there are also associated telecom charges for public switched telephone network (PSTN) usage. These charges are referred to as “Contact Center Telecommunications” charges on your AWS invoice.

Contact Center Telecommunications usage is charged by AMCS LLC and includes the following:

- Telephone numbers that are billed by day, per country, including:
  - Direct Inward Dial (DID), also known as local or toll numbers
  - Toll free numbers
- Telephony usage that is billed per minute, rates vary per country:
  - Inbound to DID number

    - Inbound to toll free
    - Outbound calling (to either a customer or agent)
- Chat is billed by messages sent.

Please consult the Amazon Connect pricing page for the latest prices.

Q: How are usage charges calculated?

The Amazon Connect application charge is calculated based on customer contact duration, with a 10-second minimum and per second granularity.

For example, if there is an inbound call to your Amazon Connect phone number and the caller is on the line between 18:00:03 and 18:01:09 whether or not connected to your contact center agent, the application charge will be 1.1 minutes multiplied by the published per minute rate.

The Contact Center Telecommunications charge is calculated based on the aggregate telecom minutes, rounded up to the nearest minute. For example, if a customer calls your Amazon Connect phone number and is on the line for 10 seconds before hanging up, you will be charged for 1 minute of Contact Center Telecommunications. If your agent is configured to receive their calls through PSTN, and your Amazon Connect instance receives an inbound call from a customer with a duration of 2 minutes and 50 seconds, of which 1 minute and 5 seconds were spent with the agent connected, you will be billed for 3 minutes of inbound usage and 2 minutes of outbound telecom usage, as the Amazon Connect instance placed an outbound PSTN call to your agent (rates vary based on the origination/destination of the calls). You are also charged for Contact Center Telecommunications when your agents place outbound calls to customers.

For Amazon Connect chat, you are billed $0.004 for the initiation of a chat, which includes an optional message, and additional messages at $0.004 per message. Messages sent in a chat by the end-customer, agent, or by contact flow are billed. System generated events such a participant joined, participant left, chat ended, and participant typing events are not billed. Note: We may limit contact flow usage for chats based on our acceptable use policy, service terms and documentation.

Support

Q: How do I get support for Amazon Connect?

The answers to most questions about Amazon Connect can be found in the Administrator Guide.

For additional support options, see the AWS Support Center.

Amazon Pinpoint FAQs

General

Q: What is Amazon Pinpoint?

A: Amazon Pinpoint is AWS's Digital User Engagement Service that enables AWS customers to effectively communicate with their end users and measure user engagement across multiple channels including email, Text Messaging (SMS) and Mobile Push Notifications.

Amazon Pinpoint also provides tools that enables audience management and segmentation, campaign management, scheduling, template management, A/B testing, analytics and data integration. It captures data to track deliverability as well as usage and messaging analytics covering a range of dimensions including user, channels and custom attributes.

Amazon Pinpoint is built on a service-based architecture. Developers can extend their applications and backend services in various ways, including: sending messages directly from their applications via the Amazon Pinpoint channels (Email, SMS and Mobile Push), accessing segmentation data to alter their application behavior for segments of users, create and run campaigns directly from their applications, and access deliverability and analytics data to improve the user engagement of their applications. The system empowers customers to send the right message, to the right audience, at the right time and on the most effective channel.

Q: How will developers benefit from Amazon Pinpoint?

A: Amazon Pinpoint offers developers a single API layer, CLI support, and client-side SDK support to be able to extend the communication channels through which their applications engage users. These channels include: email, SMS text messaging, and push notifications, voice messages, and custom channels. Amazon Pinpoint also provides developers with an analytics system that tracks app user behavior and user engagement. With this service, developers can learn how each user prefers to engage and can personalize their end-user's experience to increase the value of the developer's applications.

Amazon Pinpoint also helps developers address multiple messaging use-cases such direct or transactional messaging, targeted or campaign messaging and event-based messaging.

Integrating and enabling all their end-user engagement channels via Amazon Pinpoint, developers can create a 360-degree view of user engagement across all customer touch points.

Q: How will marketers benefit from Amazon Pinpoint?

A: Amazon Pinpoint allows Marketers to create and execute a unified messaging strategy across all engagement channels relevant to their end-users. Pinpoint includes tools and services to let marketers analyze and engage users directly. The console provides marketers with campaign management tools to create, run and manage multi-channel campaigns across their applications, user-base and devices. Campaigns can be scheduled or triggered on user changes and actions. Users and devices can also be grouped through flexibly defined segments which can be used to determine campaign audiences. Marketers can also leverage the multi-channel templating support to personalize end-user messaging and campaign optimization features such as A/B testing, holdout testing and message caps. Marketers can also measure messaging effectiveness using Pinpoint analytics to understand the impact on user behavior.

Q: How will enterprises benefit from Amazon Pinpoint?

A: Enterprises can use Amazon Pinpoint as their Digital User Engagement Service. They can free developers from having to individually integrate different communication channels into their applications and instead focus on leveraging Pinpoint to learn how their end-users and customers are engaging with their applications. It enables them to measure and improve their technology investments by measuring how engaged their digital customers are across all functions of their enterprise.

Q: Why should I use Amazon Pinpoint to run and manage my campaigns?

A: Amazon Pinpoint makes it easy to run targeted campaigns and drive user engagement of applications using different channels: email, SMS and mobile push notifications. Amazon Pinpoint helps you understand user behavior, define which users to target, determine which messages to send, schedule the best time to deliver the messages, and then track the results of your campaign.

Amazon Pinpoint is built to scale, enabling you to collect and process billions of events per day, and send billions of targeted messages to your users.

Marketers can send targeted messages and calls to action when changes occur in their organizations or in a user’s circumstances, like a new product launch to a change in a user’s locale.

Q: If I use another campaign management service how does Amazon Pinpoint help me?

A: Amazon Pinpoint's architecture is services based. Companies can choose which services to use and integrate with their existing systems and processes. Amazon Pinpoint's core services include: engagement analytics, communication channels, deliverability metrics, audience management and segmentation, template management, and campaign management.

The platform also supports data integration services to extend Amazon Pinpoint analytics and segmentation data from external data sources such as S3, as well as data exports to feed external marketing systems via Kinesis Event Streams.

Q: How much does Amazon Pinpoint cost?

A: Amazon Pinpoint has no upfront costs, no minimum charges, and no subscription fees. You pay only for what you use. Visit the Amazon Pinpoint pricing page for more details.

Q: Who owns the data in Amazon Pinpoint?

A: Customers own their data in Amazon Pinpoint. Amazon Web Services does not own or monetize the data customers collect, and does not share it with third parties. We may use the data to improve the service, monitor the health of the service, and provide technical support to you. As with any other AWS service, customers are responsible for how they use the tools we provide; this responsibility includes providing any necessary notice or opt-outs to end users and complying with applicable law.

Product Details

Q: What services and tools does Amazon Pinpoint provide?

A: Amazon Pinpoint includes a console designed for marketers and developers to use. The console provides capabilities to configure communication channels, manage audiences and segmentation, manage and run campaigns, create and manage message templates, create and manage engagement schedules and analyze user engagement. Standard analytics includes: active users, user activities, sessions, user retention, campaign efficacy and user channel engagement metrics. You can create custom analytics to integrate custom attributes and drive analytics for sales conversion, funnel reporting, product adoption by segment and any other metric required to support the business.

Q: I already use Amazon SNS or Amazon SES. What do I gain by switching to Amazon Pinpoint?

A: In typical Amazon SNS and Amazon SES use cases, you have to set up your application to manage each message's audience, content, and delivery schedule. These same features are built in to Amazon Pinpoint. With Amazon Pinpoint, you can create message templates, delivery schedules, highly-targeted segments, and full campaigns.

Q: How does Amazon Pinpoint Voice differ from Amazon Connect?

A: With Amazon Pinpoint Voice, you can engage with your customers by delivering voice messages over the phone. Pinpoint Voice gives customers a great way to deliver transactional messages—such as one-time passwords, appointment reminders, order confirmations, and more. With Pinpoint Voice you can convert a text script to lifelike speech and then deliver the personalized voice message to your customer. Call metrics—such as number of calls completed and number of calls failed—help you to optimize future voice engagements. With both Poinpoint Voice and SMS channels at your disposal, you can send SMS messages to customers who prefer text and deliver voice messages to those who are either unable to receive SMS messages or who would rather interact via a phone call. With the addition of the voice channel, you can now use Pinpoint to seamlessly engage your customers with timely, relevant content through mobile push, email, SMS, and voice calls. To learn more, please see Amazon Pinpoint Voice.

Amazon Connect is a self-service, cloud-based contact center. With only a few clicks in the AWS Management Console agents can begin making or taking calls within minutes. The service makes it possible to design contact flows, similar in concept to Interactive Voice Response (IVR), that adapt the caller experience, changing based on information retrieved by Amazon Connect from AWS services, like Amazon Redshift, or third-party systems, like CRM or analytics solutions. Contact flow integrations with AWS AI services like Amazon Lex enable intelligent conversational bots to turn automated interactions into natural conversations. The self-service graphical interface in Amazon Connect makes it easy for non-technical users to design contact flows, manage agents, and track performance metrics – no specialized skills required. To learn more, please see Amazon Connect.

Q: What data does Amazon Pinpoint store without using SDKs and instrumentation?

A: Amazon Pinpoint can store four different types of data:

Configuration Data from which Amazon Pinpoint services are provided their rules of engagement. This includes:

1. **Communication.** Communication channels can be configured with restrictions per channel or across channels e.g. maximum number of messages a user can receive per day, maximum messages a user can receive for a campaign and quiet times. There are also channel specific configurations e.g. Mobile Push Notification Abbreviated Numbers,

SMS Short Codes, Email Dedicated IP Addresses, permitted message types and lengths configured per channel.

2. **Campaign.** There are different types of campaigns. Scheduled campaigns include schedules, frequency, segments, holdouts, message templates and A/B tests. Event-based campaigns also include trigger or event rules to replace schedules.
3. **Segmentation.** That can be defined through a set of filters driven off user and user engagement data, or they can be defined through data imports and ingested as lists from files extracted from external systems.
4. **Scheduling Configuration.** Scheduling is primarily assigned to campaign-based engagement and defines the frequency and precise time of sending messages.
5. Custom Attributes Configuration. Defines custom attributes and events that applications can capture and return to Amazon Pinpoint Engagement Data.

   User Data which provides Amazon Pinpoint with endpoint information for sending messages across any channel, device or application. User data is extensible, but includes the following per channel:

   - Mobile Push Notifications. This includes userID, appID, appVersion, DeviceID, DeviceModel, DeviceModelVersion, Device OS, OS version, lastTimezone, lastCity, lastCountry, lastLatitude, lastLongitude, lastPostalCode, lastRegion.

   - SMS. Number.

   - Email. PrimaryWorkEmailAddress, SecondaryWorkEmailAddress, PrimaryPersonalEmailAddress, SecondaryPersonalEmailAddress.

   User data can also include:

   - External User Identifiers. Which map users to the IDs in external systems.

   - Custom Attributes. Users can add custom attributes to associate various data that is customized to their utilization of Amazon Pinpoint.

   User Engagement Data which includes default data per channel as well as custom data attributes when configured. Data per channel includes:

   External Data can include any user, segmentation and analytics data.

Q: What are the options for capturing custom application events?

A: You can either use the Mobile SDK within your mobile application to send custom events and attributes for Mobile Push Notifications, or use the Amazon Pinpoint REST API to send events programmatically from any application.

Q: Does Amazon Pinpoint support cross-device/application identity management?

A: Yes. This is captured under User ID.

Q: What OS versions does Amazon Pinpoint support for Mobile Push Notifications?

A: The iOS SDK supports apps running on iOS 7.0 and higher. The Android SDK supports apps running on Android 2.3.3 and higher.

Q: For mobile push notifications is data cached when a user's device is offline?

A: Yes, when using the AWS Mobile SDK, data is cached on the user's device and is uploaded when a network connection is next established.

Q: Are network channels optimized when sending events via the SDK?

A: Yes, the events are batched, and sent once per minute. You can also specify the transport to send the events: cellular and Wi-Fi, or Wi-Fi only.

Analytics

Q: What types of analytics does Amazon Pinpoint provide?

A: Amazon Pinpoint offers several types of standard analytics that provide insight into how customers use your mobile and web applications, how your engagement efforts are performing, and the impact your engagement efforts have on your business outcomes. Standard analytics include metrics for active users, user activities and demographics, sessions, user retention, campaign efficacy, and transactional messages. Using these metrics in combination with the analytics tools on the console, you can perform in-depth analysis by filtering on certain segments, custom attributes, and more. You can also create funnel reports.

Q: Where can I access analytics data?

A: You can view analytics data on the Amazon Pinpoint console. For each of your projects, the console provides detailed charts and metrics that provide insight into areas such as customer demographics, application usage, purchase activity, and delivery and engagement rates for campaigns. You can also access a subset of these metrics programmatically by using the Amazon Pinpoint API.

Q: How long does Amazon Pinpoint store analytics data?

A: Amazon Pinpoint automatically stores your analytics data for 90 days. You can see your data on the console or you can query a subset of data programmatically using the Amazon Pinpoint API. To keep the data for a longer period of time, you can export data from the console to comma-separated values (.csv) files or configure Amazon Pinpoint to stream event data to Amazon Kinesis. Kinesis is an AWS service that can collect, process, and analyze data from other AWS services in real-time. Amazon Pinpoint can send event data to Kinesis Data Firehose, which streams data to AWS data stores such as Amazon S3 or Amazon Redshift. Amazon Pinpoint can also stream data to Kinesis Data Streams, which ingests and stores multiple data streams for processing in analytics applications.

Q: Can Amazon Pinpoint tell if a single user uses the same app on more than one device (for example, on their phone and on a tablet device)?

A: Amazon Pinpoint distinguishes between endpoints and users. An endpoint is a destination that you can send messages to—such as a user's mobile device, email address, or phone number. A user is an individual who has a unique user ID. This ID can be associated with one or more endpoints.

Some of the Amazon Pinpoint analytics charts report on endpoints, and some report on users. To learn more about the individual charts, see Chart Reference for Amazon Pinpoint Analytics in the Amazon Pinpoint User Guide.

Q: How is a "session" defined?

A: A session is one use of an app by the user. A session begins when an app is launched (or brought to the foreground), and ends when the app is terminated (or goes to the background). To accommodate for brief interruptions, like a text message, an inactivity period of up to 5 seconds is not counted as a new session. Total daily sessions shows the number of sessions your app has each day. Average sessions per daily active user shows the mean number of sessions per user per day.

Q: When an app goes to the background does its session end?

A: Yes, the session ends. When the app comes to the foreground, a new session begins.

Q: How are daily and weekly retention defined?

A: Daily retention is measured by determining the number of users that first used your app on a specific day, came back and used your app in the next 7 days (7-day retention), fourteen days (14-day retention), and thirty days (30-day retention).

Q: What is "sticky factor," and how is it calculated?

A: The sticky factor represents the number of monthly users who used the app on a particular day.

Sticky factor is calculated by dividing daily active users (DAU) by monthly active users (MAU). For example, if an app has 100,000 DAU and 300,000 MAU, its sticky factor is .33. A high sticky factor can indicate strong engagement, appeal, and opportunities for monetization.

Q: What are demographics in Amazon Pinpoint?

A: The Demographics charts provide information about the device attributes for your app users. You can also see custom attributes that you define.

Events

Q: What are custom events?

A: Custom events metrics that you define. They help track user actions specific to your app or game. The Events charts provide a view of how often custom events occur. Custom events can be filtered based on attributes and their associated values.

You create custom events by naming them, such as "Item Bought" or "Button Pressed," and then adding context by specifying attributes (for qualitative measures) and metrics (for quantitative measures). For example, if your business goal is to track purchases of items from within the app, you can use "Item Bought" as a custom event, "Item XYZ" as an attribute, and "Item Price" as the metric. The custom events report enables you to search and filter for each attribute or metric. For example, you can find how often "Item XYZ" was purchased or how often “Item Price” was $1.99. You can also review the weighted average of metric values (per session) and track minimum, maximum, or average metric values. As a best practice, we recommend that custom event names be broad and attribute names be specific.

Q: What are the benefits of using custom events?

A: Custom events help you understand the actions that users take when using your app. For example, a game developer might want to understand both how often a level is completed and how much health each player has left at the end of a level. With custom events, you can create an event called "level_complete", with "add_level" as an attribute, and "health" as an attribute value. Each time a level is completed, you can record a "level_complete" event with the name of the level and the player's health. By reviewing the Events charts, you might discover that a level is too easy because players always finish with maximum health. Using this data, you can adjust the level's difficulty to better challenge and engage players, which might improve retention.

You can also use custom events to create event-based campaigns that are sent when your customers take specific actions within your applications. For example, you can set up a campaign to send a message when a customer creates a new account, when they spend a certain dollar amount, or when they add an item to their cart but don’t purchase it.

Event-based campaigns help you send messages that are timely, personalized, and relevant to your customers, which ultimately increases their trust in your brand and gives them a reason to return. You can create event-based campaigns by using the Amazon Pinpoint console, or by using the Amazon Pinpoint API.

Q: Are there limits for using custom events in my app?

A: You can have up to 1,500 unique custom event types per app and up to 40 attributes and metrics per custom event. For more information, see Reporting Events in Your Application in the Amazon Pinpoint Developer Guide.

Campaigns

Q: What are campaigns?

A: Campaigns are messages that you send to a targeted segment of users on a predefined schedule. You can use targeted campaigns to increase customer engagement and retention. You can create campaigns for use cases such as welcoming new customers, informing customers of new features in your apps, and promoting special offers and deals.

Q: What's a standard campaign?

A: Standard campaigns include a target segment, a message, and a schedule for sending the message. You can also reuse previously defined segments or define a new segment when you create a campaign. For every scheduled campaign, Amazon Pinpoint recalculates the current audience size based on the criteria associated with the segment.

Q: What's an A/B test campaign?

A: A/B campaigns are campaigns with more than one treatment. Each treatment differs from the other based on the message or the sending schedule. You can compare the response rates for each treatment to determine which one had a bigger impact on your customers.

Q: What are silent push notifications and how do I use them?

A: Silent and in-app notifications are messages that are delivered to customers' devices, but aren't displayed on the devices. You can use these messages to manage the configuration of your app, or to deliver messages to the notification center within your app.

Q: What metrics does Amazon Pinpoint track for standard campaigns?

A: For standard campaigns, you can track messages sent, messages delivered, direct app opens, sessions per user, purchases per user, delivery rate, open rate, user devices messaged, and campaign sessions by time of day.

Q: What are my scheduling options for campaigns?

A: During campaign set up in Amazon Pinpoint, you can choose when the campaign should be sent. You have two options, you can send the campaign at a specific time, or you can send it when an event occurs. Time-based campaigns can be scheduled to run one time immediately or at a time you designate in the future. They can also be scheduled with multiple runs—hourly, daily, weekly, or monthly. To define your recurring campaigns, choose a start date and an end date, and specify whether or not deliver messages based on each recipient's local time zone.

You can also use Amazon Pinpoint to create campaigns that send messages, such as text messages, push notifications, and emails, to your customers when they take specific actions within your apps. You can create event-based campaigns by using the Amazon Pinpoint console, or by using the Amazon Pinpoint API. Event-based campaigns are an effective way to implement both transactional and targeted campaign use cases. Rather than define a time to send your message to customers, you select specific events, attributes, and metric values that you want to use to trigger your campaigns. For more information about event-based campaigns, see our blog post.

Q: What are event-based campaigns?

A: Event-based campaigns send messages, such as text messages, push notifications, and emails, to your customers when they take specific actions within your applications, such as making purchases or watching a video. For example, you can set up a campaign to send a message when a customer creates a new account or when they add an item to their cart but don’t purchase it. You can create event-based campaigns by using the Amazon Pinpoint console, or by using the Amazon Pinpoint API. Event-based campaigns are an effective way to implement both transactional use cases, such as one-time-password and order confirmation messages, and targeted uses cases, such as marketing promotions. Rather than define a time to send your message to customers, you select specific events, attributes, and metric values that you want to use to trigger your campaigns. For more information about event-based campaigns, please view this blog post.

Q: How do I get started with event-based campaigns?

A: The first step in setting up an event-based campaign is to create a new campaign. On step 4 of the campaign creation process, you choose when the campaign should be sent. You can choose to send the campaign at a specific time, or you can send it when an event occurs. Choose **"When an event occurs"**, and then choose the events, attributes, and metrics that trigger your campaign.

Q: What is the cost of event-based campaigns?

A: There are no additional charges associated with creating event-based campaigns. You pay only for the number of endpoints that you target, the number of messages that you send, and the number of analytics events that you send to Amazon Pinpoint. To learn more, see our Pricing page.

Q: How can I learn more about event-based campaigns and best practices?

A: For more information and best practices, see our blog post and the Amazon Pinpoint User Guide.

Q: What are the limitations of event-based campaigns?

A: There are a few limitations to be aware of when creating an event-based campaign. For more information, see the Amazon Pinpoint User Guide.

Q: Can I create and schedule event-based campaigns on the voice channel?

A: No. Currently, you can only send event-based campaigns in the SMS, push notification, and email channels.

Q: Can I use server-side events to trigger my campaigns?

A: Not yet. Amazon Pinpoint only lets you execute campaigns based on events that are associated with individual customers. Specifically, you can only trigger campaigns based on actions that users take in the applications, and that are reported by the AWS Mobile SDK.

Q: Can I create and schedule a voice message campaign?

A: No, the voice channel is only available for transactional messages. You can use the Amazon Pinpoint API to deliver transactional voice messages—such as new account creation notifications or order confirmations—directly to specific recipients.

Q: How do campaign limits work?

A: On the **General Settings** page of the Amazon Pinpoint console, you can configure the maximum number of messages an endpoint can receive for a campaign. This feature is useful when you want to place strict limits on the number of messages that an endpoint can receive from a campaign. For example, if you create a campaign that's automatically sent to all new customers, you can set the limit to 1. This setting ensures that new customers only receive the message once.

It's important to note that this feature is based on the number of messages that *target* an endpoint, as opposed to the number of messages that are actually *delivered* to an endpoint. If a campaign is configured to automatically send a message when a customer creates a new account, but the endpoint isn't able to receive the message for some reason (for example, if the quiet time setting for your campaign applies to the endpoint), then the endpoint is still counted as having been targeted. In this situation, the endpoint is removed from subsequent runs of the campaign.

Journeys

Q: What are journeys?

A: In Amazon Pinpoint, journeys are fully automated, end-to-end messaging solutions for engaging with your customers. Picture a flowchart: actions lead to other actions, sometimes branching into several paths, each with their own unique sets of activities.

You can use journeys to implement a variety of use cases, including customer onboarding scenarios and churn prevention programs. Journeys are flexibile and extensible, giving your teams the control they need to build powerful customer experiences without writing code using an easy-to-use graphical interface.

Q: Are there any prerequisites for creating a journey?

A: You have to have an active AWS account. You also have to set up the email channel and create a project in Amazon Pinpoint.

Q: What is an activity in a journey?

A: Journey activities are the configurable components that make up a journey. These components have different functionality, and you configure them to create the experience you want to build.

For example, an **Email** activity sends an email to every journey participant who arrives on it. A **Wait** activity prevents journey participants from proceeding to the next activity in the journey until a specific date and time, or until a certain amount of time elapses. A **Multivariate split** activity sends journey participants down one of up to five unique paths based on their segment membership, or based on their interactions with messages that you sent earlier in the journey.

To learn more about these activities, see Journeys in the *Amazon Pinpoint User Guide*.

Q: What metrics does Amazon Pinpoint provide for my journeys?

A: You can view metrics for a journey using the same web-based management console that you use to create a journey. From the console, you can quickly determine how many participants entered the journey, as well as the number that arrived on each individual step.

You can also export all of the metrics for your journeys to your preferred destination by using an Amazon Kinesis Data Stream or an Amazon Kinesis Data Firehose stream. This capability makes it possible to perform in-depth, post-journey analyses, or to store your data for an extended period time.

Additional charges apply to exporting data using Amazon Kinesis. For more information, see Amazon Kinesis Data Stream Pricing and Amazon Kinesis Data Firehose Pricing.

Q: Can I schedule my journeys?

A: You can configure each journey to start and end at a specific time. Each journey can run continuously for up to 18 months.

You can also schedule how often new participants enter the journey. When you create a journey, you specify a segment of customers that participate in it. You can set up your journey so that this segment is updated on a regular basis—hourly, daily, monthly, quarterly, annually, or not at all.

Q: What can I do if I make a mistake in my journey?

A: Journeys includes a built-in review process that checks for show-stopping errors, while also providing recommendations and best practices. You have to complete this review process before you launch each journey.

Journeys also includes a test feature, which makes it easy to send a group of test participants through your journey. By testing your journey, you can ensure that it behaves the way that you expect it to behave.

If you encounter issues with your journey while it's running, you can stop it at any time. When you stop a journey, participants halt on the activity they're currently on, and never proceed to the next activity.

Q: How much does it cost to use journeys?

A: There's no additional cost for using journeys. You pay for the customers that you target, and for the messages that you send. To learn more, see Amazon Pinpoint Pricing.

Q: Can I use server-side events to trigger journeys?

A: Currently, you can only insert customers into a journey based on their membership in a segment.

However, you can create dynamic segments that are updated using Lambda functions. When a server-side event occurs, you can use Lambda to update your endpoints, adding them to the target segment for your journey. For more information, see Customizing Segments with AWS Lambda in the *Amazon Pinpoint Developer Guide*.

Q: I'd like to see a feature or capability added to journeys. How can I provide feedback?

A: Our product development roadmap is driven by customer feedback, so we always love to hear your feedback. You can let us know what you think of journeys by clicking the "How do you like journeys?" button in the bottom right corner of the journeys workspace.

We read all of the messages that we receive. If we have any questions about your feedback, we'll contact you directly for more information. However, note that we might not be able to respond to every message that we receive.

Text Messaging

Q: Can I use dedicated phone numbers to send SMS messages?

You can use dedicated short codes or dedicated long codes to send SMS messages to recipients in several countries and regions. The following sections discuss the benefits and disadvantages of each of these options.

Short codes

- **Benefit**: Capable of sending hundreds of SMS messages per second (the actual rate varies based on the country where the short code is based).

- **Benefit**: Easy for recipients to identify and remember.
- **Disadvantage**: Usually more expensive than a long code.
- **Disadvantage**: Often takes several weeks to provision.

Long codes

- **Benefit**: Usually less expensive than a dedicated short code.
- **Benefit**: Typically takes days to provision (actual time varies by country).
- **Benefit**: Can be used to send both SMS messages and voice calls in the US and Canada.
- **Disadvantage**: Allows you to send fewer messages per second (usually between 1 and 15, depending on the country where the long code is based).
- **Disadvantage**: Can be subject to daily sending limits (varies based on the country and mobile carrier of the recipient).
- **Disadvantage**: Harder for recipients to remember and identify.

For more information about obtaining dedicated short and long codes, see the Amazon Pinpoint User Guide.

Q: What is two-way text messaging?

A: Two-way text messaging enables customers of Amazon Pinpoint to receive text messages from their users. When a user sends a text message to a customer’s leased number, Amazon Pinpoint passes the text message to the customer, and the customer can use this message to trigger an appropriate response. Depending on the country and the local telecommunication regulation requirements, the customer can use long codes (10-digit phone numbers) and short codes (5 to 6-digit phone numbers).

To receive text messages from their users, the customer enables two-way text messaging in the Amazon Pinpoint console and selects an Amazon SNS topic to receive the text messages. Amazon Pinpoint provides the telephone number of the user and the customer’s message ID if the text message was sent to the customer as a reply.

Q: What are the advantages of two-way text messaging?

A: Two-way text messaging enables several customer engagement use cases. For example, a financial services company could use two-way SMS messaging to send confirmation messages when they detect that a customer's account was used to make an unusual purchase. If the customer responds to the message stating that they initiated the purchase, the institution can authorize it.

Appointment confirmations are another common use case for two-way SMS messaging. For example, medical practices, salons, or restaurants can send messages to confirm a customer's appointment or reservation. The customer can respond, indicating whether or not they can keep their appointment. If the customer can't keep their appointment, you can ask if they want to reschedule, and send them the time and date of the next available appointment.

Q: Why do you require a dedicated number for two-way text messaging?

A: Receiving text messages from the same company through multiple numbers makes it hard for users to associate a number with a single business with which they regularly interact. A dedicated number makes it easier for users to participate in two-way text messaging.

Additionally, Amazon Pinpoint delivers text messages using numbers that are shared by multiple customers. Since these numbers are not exclusive to a sender, it is not possible to accurately route the text message to an appropriate customer when a cellular subscriber texts a number that is owned by Amazon Pinpoint. Because of these reasons, Amazon Pinpoint requires companies to lease a dedicated number for two-way communication.

Q: How can I disable two-way text messaging?

A: You can disable two-way text messaging from the Amazon Pinpoint console. When you disable this feature, you stop receiving incoming text messages from your customers.

Q: Can I create automated responses for specific keywords?

A: Yes. If you have a leased phone number (either a long code or a short code), you can use the Amazon Pinpoint console to create responses to specific keywords. When a customer sends a text message that matches the keyword, Amazon Pinpoint sends the corresponding response to the customer. You can also customize the messages that customers receive when they send HELP or STOP messages to your numbers.

Data Privacy

Q. Does Amazon Pinpoint store my customer data?

A: Yes. Amazon Pinpoint stores user, endpoint, and event data. We have to retain this data so that you can create segments, send messages to recipients, and capture application and campaign engagement data.

Q. Who can access the data stored in Amazon Pinpoint?

A: A very limited number of authorized employees have access to the data stored in your Amazon Pinpoint account.

Maintaining your trust is our highest priority. We use sophisticated physical and technical controls to safeguard your privacy and ensure the security of your data.

Your data is encrypted at rest and during transit. Our processes are designed to prevent unauthorized access to or disclosure of your content.

For more information, see the AWS Data Privacy FAQ.

Q: Do I own my content that is processed and stored by Amazon Pinpoint?

A: You always retain ownership of your content. We only use your content with your consent.

Q: How do I delete the data that Amazon Pinpoint stores?

A: You can selectively delete the data stored in your Amazon Pinpoint account. You can also close your entire AWS account, which deletes all of the data stored in Amazon Pinpoint and all other AWS services in every AWS Region. For more information, see Deleting Data from Amazon Pinpoint in the *Amazon Pinpoint Developer Guide*.

Contacting Us

Q: I received spam or other unsolicited email messages from an Amazon Pinpoint user. How do I report these messages?

A: You can report email abuse by sending an email to email-abuse@amazon.com.

To help us handle the issue as quickly and effectively as possible, please include the full headers of the original email. For procedures for obtaining email headers for several common email clients, see How to Get Email Headers on the MxToolbox.com website.

Q: How can I submit feature requests or send other product feedback about Amazon Pinpoint?

A: Your AWS Account Manager can send your feature requests and feedback directly to the appropriate team. If you don't currently have an AWS Account Manager, you can also provide your feedback on the Amazon Pinpoint forum.

Q: How can I get technical support for Amazon Pinpoint?

A: If you have an AWS Support plan, you can create a new support case directly from the web-based AWS management console. AWS Support plans begin at $29 per month. For more information about AWS Support plans, visit https://aws.amazon.com/premiumsupport/.

To open a new technical support case

1. In the console, on the **Support** menu, choose **Support Center**.
2. Next, choose **Create case**.
3. On the **Create case** page, choose **Technical support**.
4. Provide information about the issue you're experiencing, and then submit the ticket.

If you don't have an AWS Support plan, you can also ask questions and get answers on the Amazon Pinpoint forum.

Amazon SES FAQs

General

Q: What is Amazon Simple Email Service (Amazon SES)?

Amazon Simple Email Service (Amazon SES) is a highly scalable and cost-effective service for sending and receiving email. Amazon SES eliminates the complexity and expense of building an in-house email solution or licensing, installing, and operating a third-party email solution.

Q: Who can use Amazon SES?

Amazon SES is a great solution for anyone who needs a reliable, scalable, and inexpensive way to send and receive email. Our users include a diverse range of organizations, such as online retailers, application developers, and digital marketing organizations.

Q: How can I start sending email using Amazon SES?

For information about how to set up email with Amazon SES, see the Quick Start section of the *Amazon SES Developer Guide*.

Pricing and Billing

Q: How much does it cost to use Amazon SES?

With Amazon SES, you pay only for what you use. There are no minimum fees, and no upfront costs.

The amount you pay depends on the number of messages you send and receive, the volume of data you send. For more information, see the Amazon SES Pricing page.

Q: Is there a free usage tier for Amazon SES?

Yes. If you call Amazon SES from an application hosted in an Amazon EC2 instance, you can send 62,000 email messages each month at no charge. This free tier offer doesn't expire.

Q: How am I billed for my use of Amazon SES?

At the end of each month, we bill you for that month's usage. The billing cycle begins on the first day of each month and ends on the last day of each month.

Q: How can I track my Amazon SES usage?

You can view your charges for the current billing period at any time by visiting the Billing Dashboard in the AWS Management Console.

Q: Am I billed for incoming spam messages?

When you set up Amazon SES to receive email, you create receipt rules. In a receipt rule, you define a recipient (or group of recipients), and specify what Amazon SES does when it receives the email. When an incoming message is addressed to at least one email address that matches an address in an active receipt rule, you pay for that message, even if it is unsolicited (spam) email.

You can configure Amazon SES to block messages from certain senders and IP addresses. If you block incoming messages in this way, you're not billed for the blocked messages.

Limits and Restrictions

Q: What is the Amazon SES sandbox?

The Amazon SES sandbox is an area where new users can test the capabilities of Amazon SES. When your account is in the sandbox, you can only send email to verified identities. A verified identity is an email addresses or domain that you've proven that you own.

Additionally, when your account is in the sandbox, there are limits to the volume of email you can send each day, and to the number of messages you can send each second.

When you're ready to start sending email to non-verified recipients, submit an Amazon SES Sending Limit Increase request through the AWS Support Center. For more information, see Moving Out of the Amazon SES Sandbox in the *Amazon SES Developer Guide*.

Q: Can I send emails from any email address?

No. You can only use Amazon SES to send email from addresses or domains that you own.

To prove that you own an email address or domain, you have to verify it. In each AWS Region, you can verify up to 10,000 email addresses and domains, in any combination. For more information about verifying email addresses and domains, see Verifying Identities in Amazon SES in the *Amazon SES Developer Guide*.

Q: Is there a limit on the size of emails Amazon SES can deliver?

Amazon SES accepts email messages up to 10 MB in size. This includes any images and attachments that are part of the message.

Q: Is there a limit on the number of recipients I can specify in a single email message?

You can specify a maximum of 50 recipients for every message you send using Amazon SES. This limit includes all addresses on the "To:," "CC:," and "BCC:" fields. If you need to send an email message to more than 50 recipients, then you have to split your recipient list into groups of 50 recipients or less, and send separate messages to each group.

Q: Are there any limits on how many emails I can send?

Every Amazon SES account has its own set of sending limits. These limits are:

- **Sending quota**—the maximum number of recipients that you can send email to in a 24-hour period.
- **Maximum send rate**—the maximum number of recipients that you can send email per second.

Sending limits are based on recipients rather than on messages. You can check your sending limits at any time by using the Amazon SES console.

> **Note**: If we determine that the email you send is of poor or questionable quality (for example, if it has high bounce or complaint rates, or if it contains unsolicited or malicious content), we reserve the right to pause your ability to send email.

Security and Encryption

Q: Can Amazon access the emails that I send and receive?

We use in-house anti-spam technologies to filter messages that contain poor-quality content. Additionally, we scan all messages that contain attachments to check for viruses and other malicious content. These processes are completely automated with no human involvement. We only access email content in certain circumstances, such as when we're troubleshooting delivery issues or investigating fraudulent or abusive activity.

Q: Can I encrypt email messages that I receive?

Amazon SES integrates with AWS Key Management Service (KMS), which provides the ability to encrypt the mail that it writes to your Amazon S3 bucket. Amazon SES uses client-side encryption to encrypt your mail before it sends the email to Amazon S3. This means that it is necessary for you to decrypt the content on your side after you retrieve the mail from Amazon S3. The AWS Java SDK and AWS Ruby SDK provide a client that is able to handle the decryption for you.

Authentication, Validation, and Configuration

Q: Do I need to set up reverse DNS records in order to use Amazon SES?

You don't need to set up reverse DNS records to use Amazon SES. Amazon Web Services manages the IP addresses that Amazon SES uses, and provides reverse DNS records for these addresses.

Q: Does Amazon SES support Sender Policy Framework (SPF)?

Yes, Amazon SES supports SPF. You may need to publish an SPF record, depending on how you use Amazon SES to send email.

If you don't need to comply with Domain-based Message Authentication, Reporting and Conformance (DMARC) using SPF, you don't need to publish an SPF record, because by default, Amazon SES sends your emails from a MAIL FROM domain that's owned by Amazon Web Services.

If you want to comply with DMARC using SPF, you have to set up Amazon SES to use your own MAIL FROM domain and publish an SPF record.

Q: Does Amazon SES support Domain Keys Identified Mail (DKIM)?

Yes, Amazon SES supports DKIM. If you have enabled and configured Easy DKIM, Amazon SES signs outgoing messages using DKIM on your behalf. If you prefer, you can also sign your email manually. To ensure maximum deliverability, there are a few DKIM headers that you should not sign. For more information, see Manual DKIM Signing in Amazon SES in the *Amazon SES Developer Guide*.

Q: Can emails from Amazon SES comply with DMARC?

With Amazon SES, your emails can comply with DMARC through SPF, DKIM, or both.

Q: Does Amazon SES send email over an encrypted connection using Transport Layer Security (TLS)?

Amazon SES supports TLS 1.2, TLS 1.1 and TLS 1.0 for TLS connections.

By default, Amazon SES uses *opportunistic TLS*. This means that Amazon SES always attempts to make a secure connection to the receiving mail server. If Amazon SES can't establish a secure connection, it sends the message unencrypted.

You can change this behavior so that Amazon SES only sends the message to the receiving email server if it can establish a secure connection. For more information, see Amazon SES and Security Protocols in the Amazon SES Developer Guide.

Sending Capabilities

Q: Can I use Amazon SES to send email from my existing applications?

Amazon SES allows you to create a private SMTP relay for use with any existing SMTP client software, including software that you develop yourself, or any third-party software that can send email using the SMTP protocol.

For more information, see Using the Amazon SES SMTP Interface to Send Email in the *Amazon SES Developers Guide*.

Q: Can Amazon SES send emails with attachments?

Amazon SES supports many popular content formats, including documents, images, audio, and video.

Note: For your own safety and that of your customers, Amazon SES scans every attachment that you send for viruses and malware.

You can use an email client that supports SMTP to send email with attachments. When you configure a client to send outgoing email through Amazon SES, the client constructs the appropriate MIME parts and email headers before sending the message.

You can also send email with attachments programmatically. To include an attachment in your email, construct a new multipart email message. In the message, include a MIME part that contains an appropriate Content-Type header, along with the MIME-encoded content. Next, use the Content-Disposition header to specify whether the content is to be displayed inline or treated as an attachment.

Once you've composed your message, you can use the SendRawEmail API operation to send it.

Q: Can I send email using a dedicated IP address?

For an additional monthly charge, you can lease IP addresses that are dedicated to your own personal use. For more information about leasing dedicated IP addresses, see Requesting and Relinquishing Dedicated IP Addresses in the *Amazon SES Developer Guide*.

Q: Can I specify a dedicated IP address when I send certain types of email?

If you lease several dedicated IP addresses to use with your Amazon SES account, you can use the dedicated IP pools feature to create groups (pools) of those IP addresses. You can then associate each pool with a configuration set; when you send emails using that configuration set, those emails are only sent from the IP addresses in the associated pool. For more information, see Creating Dedicated IP Pools in the *Amazon SES Developer Guide*.

Q: Can I test Amazon SES responses without sending email to real recipients?

You can use the Amazon SES mailbox simulator to test your sending rate and to test your ability to handle events such as bounces and complaints, without sending email to actual recipients. Messages that you send to the mailbox simulator don't count against your bounce and complaint metrics or your daily sending quota. However, we do charge you for each message you send to the mailbox simulator, just as if they were messages you sent to actual customers.

For more information about the Amazon SES mailbox simulator, see Testing Amazon SES Email Sending in the *Amazon SES Developer Guide*.

Q: Can I use Amazon SES for email-to-text SMS delivery?

Many mobile phone carriers offer an SMTP-to-SMS gateway. Amazon SES users can send text-only emails to the emails addresses associated with these gateways, which are then delivered to the recipients' mobile phones as SMS messages.

However, in order to successfully use an SMS gateway, you must know several pieces of information, including:

- The recipient's mobile phone number
- The recipient's mobile phone carrier
- The domain name of the carrier's SMS gateway (such as sms.carriername.com)

If SMS messages are part of your marketing or communications plan, we recommend using Amazon Pinpoint. When you use Amazon Pinpoint to send SMS messages, you only need to know the recipient's mobile phone number.

Q: How do I control the character encoding of my emails with Amazon SES?

The SMTP protocol requires that all data be sent in 7-bit ASCII format. If you want to use a different character encoding with the Amazon SES SMTP interface, you have to apply the encoding to the subject and body of your message, and then convert them to a valid 7-bit ASCII message before sending it to the SMTP endpoint.

The SendEmail API accepts UTF-8 subject and body inputs, transcodes them into whatever format you specify via an optional encoding parameter, and automatically converts the resulting content into 7-bit ASCII using the appropriate encoded-word syntax and content-transfer-encoding headers before sending the message. The SendRawEmail API requires you to apply your desired encoding to your subject and body and then convert them to a valid 7-bit ASCII message before submitting each request.

Receiving Capabilities

Q: How do I configure Amazon SES to receive email?

The first step in setting up Amazon SES to receive email is to verify your domain. If you've already verified your domain for sending email, you don't need to repeat the process to receive email. For more information about verifying your domain, see Verifying Domains in Amazon SES in the *Amazon SES Developer Guide*.

After you verify your domain, you have to publish a Mail eXchanger (MX) record to the DNS configuration of your domain. This record refers to the regional Amazon SES email receiving endpoint that you use to receive email. For more information about publishing an MX record, see Publishing an MX Record for Amazon SES Email Receiving in the *Amazon SES Developer Guide*.

The final step in setting up email receiving in Amazon SES is to create a receipt rule. Receipt rules tell Amazon SES what it should do with email sent to your domain. For example, you can configure Amazon SES to send incoming email to an Amazon S3 bucket, or to send you a notification by using Amazon SNS. For more information about creating receipt rules, see Creating Receipt Rules in the *Amazon SES Developer Guide*.

Q: What happens when Amazon SES receives my mail?

When Amazon SES receives a message, it references your active receipt rule set to determine if any rules apply to the incoming message's recipients. If there aren't any matches, or if the mail was sent from an IP address on your IP address block list, Amazon rejects the mail in the SMTP conversation. Otherwise, Amazon SES accepts the mail.

After Amazon SES accepts the mail, it evaluates your active receipt rules. Amazon SES then applies these rules in the order that they were defined in.

Q: How do I access my mail in Amazon S3?

When you set up a receipt rule that tells Amazon SES to write your messages to an Amazon S3 bucket, you can also set up Amazon SNS notifications. These notifications contain general information about the message and the action taken on it, including the unique ID of the message. You can use this ID to retrieve the corresponding message from the Amazon S3 bucket where Amazon SES sent the message.

Q: How can I process emails I receive?

There are two ways to process incoming mail. You can write an application that listens for Amazon SNS notifications from Amazon SES, retrieves the mail from Amazon S3, and processes it. Alternatively, you can write a custom AWS Lambda function.

The AWS Lambda event contains all of the metadata about the message that was received, but doesn't include the body of the message. If you need to parse the message body from within Lambda, then you need to first write the message to Amazon S3 using an Amazon S3 action before your AWS Lambda action is evaluated.

You can execute AWS Lambda actions synchronously or asynchronously, depending on whether or not the AWS Lambda function needs to return a result that influences how other actions are executed. We recommend that you use asynchronous execution unless your specific application requires you to use synchronous execution.

Q: Can multiple different AWS accounts receive mail on the same domain?

More than one AWS account can receive mail for the same domain. For each email that arrives on the shared domain, a copy of the message is processed by each account's receipt rule set independently.

Q: Is there any size limit to the messages that I can receive through Amazon SES?

If you store your incoming messages in an Amazon S3 bucket, the maximum message size, including headers and attachments, is 30 megabytes (MB).

If you forward messages using Amazon SNS, the maximum message size is 150 kilobytes (KB).

Q: Is there a limited throughput at which I can receive messages through Amazon SES?

There are no throughput restrictions for incoming email received through Amazon SES.

Q: Can I reject email that was sent over an unencrypted connection?

You can configure your receipt rules in Amazon SES to automatically reject incoming messages that were sent without Transport Layer Security (TLS).

Deliverability

Q: How does Amazon SES help ensure reliable email delivery?

Amazon SES uses content filtering technologies to scan outgoing email messages. These content filters help ensure that the content being sent through Amazon SES meets the standards of ISPs. In order to help you further improve the deliverability of your emails, Amazon SES provides a feedback loop that includes bounce, complaint, and delivery notifications.

Q: Does Amazon SES guarantee receipt of my emails?

Amazon SES closely monitors ISP guidelines to help ensure that legitimate, high-quality email is delivered reliably to recipient inboxes. However, neither Amazon SES nor any other email-sending service can guarantee delivery of every single email. ISPs can drop or lose email messages, recipients can accidentally provide the wrong email address, and if recipients do not wish to receive your email messages, ISPs may choose to reject or silently drop them.

Q: How long does it take for emails sent using Amazon SES to arrive in recipients' inboxes?

Amazon SES attempts to deliver emails to the Internet within a few seconds of each request. However, due to a number of factors and the inherent uncertainties of the Internet, we can't predict with certainty when your email will arrive, nor can we predict the exact route the message will take to get to its destination.

For example, an ISP might be unable to deliver the email to the recipient because of a temporary condition such as "mailbox full." In these cases, Amazon attempts to redeliver the message. If the error is permanent, such as "mailbox does not exist," Amazon SES doesn't try to deliver the message again, and you receive a hard bounce notification. You can set up delivery notifications to alert you when Amazon SES successfully delivers one of your emails to a recipient's mail server.

Bounces and Complaints

Q: How does Amazon SES send bounce, complaint, and delivery notifications to me?

Amazon SES forwards bounce and complaint notifications to you by email or sends them to an Amazon SNS topic, depending on how you set up your Amazon SES account.

Delivery notifications, which are triggered when Amazon SES successfully delivers one of your emails to a recipient’s mail server, are sent to you only through Amazon SNS.

Q. Is there an additional cost to use Amazon SNS to receive bounce, complaint, and delivery notifications?

Yes, there are extra charges associated with using Amazon SNS to receive bounce, complaint, and delivery notifications. For more information about the costs associated with using Amazon SNS, see Amazon SNS Pricing.

Q. When can I expect to be notified of bounces, complaints, and deliveries?

After an ISP sends a bounce or complaint to Amazon SES, we try pass it to you within a few seconds via Amazon SNS or email. However, we may not receive the bounce or complaint notification from the recipient's ISP for a period of time ranging from seconds to weeks or longer, depending on how quickly the ISP notifies us.

Delivery notifications are published as soon as Amazon SES delivers an email to a recipient's mail server. In most cases, email sent through Amazon SES is delivered within seconds, but occasionally it might take longer.

Q: How can I monitor the bounce and complaint rates for the email I send using Amazon SES?

Amazon SES includes a reputation dashboard that you can use to keep track of your bounce and complaint rates, as well as other factors that could impact your ability to send email. You can also use Amazon CloudWatch to create dashboards that track your bounce and complaint rates. For more information about monitoring methods, see Monitoring Your Amazon SES Sending Activity in the *Amazon SES Developer Guide*.

Q: Can my email deliverability affected by bounces or complaints that are caused by other Amazon SES users?

Typically, when other Amazon SES users send messages that result in bounces or complaints, your ability to send email remains unchanged.

An exception to this rule occurs when a recipient's email address generates a hard bounce. When a recipient's email address generates a hard bounce, Amazon SES adds that address to a global suppression list. If you try to send an email to an address that is on the global suppression list, the call to Amazon SES succeeds, but Amazon SES treats the email as a hard bounce instead of attempting to send it.

Emails that you send to addresses on the global suppression list count toward your sending quota and your bounce rate. An email address can remain on the suppression list for up to 14 days.

For more information about the global suppression list, see Amazon SES and Deliverability in the *Amazon SES Developer Guide*.

Q: A recipient address was added to the global suppression list, but I am certain it is a valid address. Can I remove that address from the suppression list?

You can submit a suppression list removal request using the Amazon SES console. For more information, see Removing an Email Address from the Amazon SES Suppression List in the *Amazon SES Developer Guide*.

Q: What happens if I try to send a malformed email message or send an email that is disallowed for any other reason?

If Amazon SES is unable to deliver your message, it returns an error message with information about what caused the delivery to fail. In rare cases, Amazon SES may not detect the problem with your email until after accepting your request. In such cases, your email is returned to you as a bounce with a corresponding error code and reason.

Spam and Viruses

Q: How does Amazon SES ensure that incoming mail is free of spam and viruses?

Amazon SES uses a number of spam and virus protection measures. It uses block lists to prevent mail from known spammers from entering the system in the first place. It also performs virus scans on every incoming email that contains an attachment. Amazon SES makes its spam detection verdicts available to you, enabling you to decide if you trust each message. In addition to the spam and virus verdicts, Amazon SES provides the DKIM and SPF check results.

Q: What prevents Amazon SES users from sending spam?

Amazon SES uses in-house content filtering technologies to scan email content for spam and malware.

If we determine that an account is sending spam or malicious content, we will pause that account's ability to send additional email.

Amazon SES and Other AWS Services

Q: How does Amazon SES integrate with Amazon WorkMail?

Amazon WorkMail uses Amazon SES to send and receive mail. When you set up Amazon WorkMail, Amazon WorkMail creates two items within your Amazon SES configuration settings: a sending authorization policy that allows Amazon WorkMail to send mail through your domain, and a receipt rule with a WorkMail action that delivers your domain's incoming mail to Amazon WorkMail. If you remove either of these items, Amazon WorkMail won't function properly.

Q: How is Amazon SES different from Amazon SNS?

Amazon SES is for applications that need to send communications via email. Amazon SES supports custom email header fields, and many MIME types.

By contrast, Amazon Simple Notification Service (Amazon SNS) is for messaging-oriented applications, with multiple subscribers requesting and receiving "push" notifications of time-critical messages via a choice of transport protocols, including HTTP, Amazon SQS, and email. The body of an Amazon SNS notification is limited to 8192 characters of UTF-8 strings, and isn't intended to support multimedia content.

Q: Do I need to sign up for Amazon EC2 or any other AWS services to use Amazon SES?

Amazon SES users do not need to sign up for any other AWS services. Any application with Internet access can use Amazon SES to deliver email, whether that application runs in your own data center, within Amazon EC2, or as a client software solution.

Q: I'm sending email using my own mail servers hosted on Amazon EC2. Do I have to start using Amazon SES instead?

No, using Amazon SES doesn't affect any Amazon EC2-based solution that you currently use. You can continue to use your existing solution, or use Amazon SES, or use both at the same time.

Q: Does Amazon SES put any restrictions on AWS Lambda functions in addition to the restrictions imposed by AWS Lambda?

There is a 30-second timeout on *RequestResponse* invocations.

SMTP Interface

Q: Does Amazon SES provide an SMTP endpoint?

Amazon SES provides an SMTP interface for seamless integration with applications that can send email via SMTP. You can connect directly to this SMTP interface from your applications, or configure your existing email server to use this interface as an SMTP relay.

In order to connect to the Amazon SES SMTP interface, you have to create SMTP credentials. For more information about creating SMTP credentials, see Obtaining Your Amazon SES SMTP Credentials in the *Amazon SES Developer Guide*.

Q: How can I use the Amazon SES SMTP interface?

To use the Amazon SES SMTP interface, all you need are your SMTP username and password, the SMTP endpoint name, and the port number. Using this information, you can connect to the Amazon SES SMTP interface in the same manner as any other SMTP relay.

For example, you can integrate your existing packaged software so that it sends email through Amazon SES. You can add email sending capability to your applications, using a programming language that supports SMTP. You can integrate Amazon SES sending with popular mail transfer agents (MTAs) such as Sendmail, Postfix, and Exim. You can even connect to the SMTP interface from the command line, and send SMTP commands directly. For more information about the SMTP interface, see Using the Amazon SES SMTP Interface to Send Email in the *Amazon SES Developer Guide*.

APIs and SDKs

Q: How do I make requests to Amazon SES?

Amazon SES accepts Query requests over HTTPS. These requests use verbs such as GET or POST, and a parameter named Action to indicate the action being performed. For security reasons, Amazon SES does not support HTTP requests; you must use HTTPS instead.

Q: What are the available API operations for sending email?

In addition to SMTP sending support, Amazon SES provides the following APIs: SendEmail and SendRawEmail. These two APIs provide different levels of control over the composition of the actual email message. Both APIs provide the same level of email sending reliability and performance:

The SendEmail API requires the user to provide only a source address, destination address, message subject, and message body. Upon calling this API, Amazon SES automatically constructs and sends a properly formatted multi-part MIME email message optimized for display by email client software.

The SendRawEmail API provides the advanced user with flexibility to format and send their own raw email message by specifying headers, MIME parts, and content types.

Q: Do the AWS Software Development Kits contain support for Amazon SES?

Yes, all of the AWS Software Development Kits (SDKs) provide methods for accessing the Amazon SES API. The SDKs take care of low-level functions, such as authentication and request signing, allowing you to send email by making a simple call to the API.

Contacting Us

Q: I received spam or other unsolicited email messages from an Amazon SES user. How do I report these messages?

You can report email abuse by sending an email to email-abuse@amazon.com.

To help us handle the issue as quickly and effectively as possible, please include the full headers of the original email. For procedures for obtaining email headers for several common email clients, see How to Get Email Headers on the MxToolbox.com website.

Q: How can I submit feature requests or send other product feedback about Amazon SES?

Your AWS Account Manager can send your feature requests and feedback directly to the appropriate team. If you don't currently have an AWS Account Manager, you can also provide your feedback on the Amazon SES forum.

Q: How can I get technical support for Amazon SES?

If you have an AWS Support plan, you can create a new support case directly from the web-based AWS management console. AWS Support plans begin at $29 per month. For more information about AWS Support plans, visit https://aws.amazon.com/premiumsupport/.

To open a new technical support case

1. In the console, on the **Support** menu, choose **Support Center**.
2. Next, choose **Create case**.
3. On the **Create case** page, choose **Technical support**.
4. Provide information about the issue you're experiencing, and then submit the ticket.

If you don't have an AWS Support plan, you can also ask questions and get answers on the Amazon SES forum.

Contact Lens for Amazon Connect FAQs

Q: How do I get started?

Please click here to sign up for a preview or reach out to your AWS sales partner. We will enable Contact Lens for Amazon Connect for your account, and you can configure which calls to analyze in the contact flow settings by checking the Contact Lens for Amazon Connect option in the “Set Recording Behavior” block. Once this is completed, Contact Lens for Amazon Connect will start analyzing your specified calls automatically.

Q: How do I access data in Contact Lens for Amazon Connect for use outside of Amazon Connect?

Contact Lens for Amazon Connect-generated metadata (including sentiment analysis, call transcript, non-talk time, etc.) along with the call recordings for each contact will be accessible in the customer’s S3 bucket. This data will be linked to Contact Trace Records (CTR) can be easily used in BI tools like Quicksight and Tableau to create custom visualizations by fusing it with data from other systems (such as CRM).

Q: I am not an Amazon Connect customer yet. Can I use Contact Lens for Amazon Connect with my existing contact center software?

Currently, Contact Lens for Amazon Connect is only offered as part of Amazon Connect.

Q: How does Contact Lens for Amazon Connect relate to Amazon Transcribe and Comprehend?

Contact Lens for Amazon Connect uses Amazon Transcribe to generate call transcripts. Contact Lens for Amazon Connect uses Amazon Comprehend to apply natural language processing on these transcripts. This approach allows organizations to evaluate their customer experience without requiring any technical expertise.

Q: What languages does Contact Lens for Amazon Connect support?

Currently, Contact Lens for Amazon Connect supports US-English. We will be adding Spanish in early 2020 and more languages throughout 2020.

Q: Does Contact Lens for Amazon Connect provide real-time analysis?

Currently, Contact Lens for Amazon Connect provides analysis based on call recordings only. We will be expanding the current feature set to include real-time capabilities.

Q: Does Contact Lens for Amazon Connect support chats?

Currently, Contact Lens for Amazon Connect supports only calls but support for chat is coming early 2020.

Q: Does Contact Lens for Amazon Connect support PII/PCI redaction?

Coming early 2020, you will be able to choose from multiple redaction options to protect sensitive information.

Learn more about Amazon Connect

Learn more about Amazon Connect features and capabilities by visiting the product page.

Learn more »

Sign up for a free account

Instantly get access to the AWS Free Tier.

Sign up »

Sign up for the preview

Get started building with Contact Lens for Amazon Connect in the AWS Management Console.

Sign up »

Amazon Aurora FAQs

General

Q: What is Amazon Aurora?

Amazon Aurora is a relational database engine that combines the speed and reliability of high-end commercial databases with the simplicity and cost-effectiveness of open source databases. Amazon Aurora MySQL delivers up to five times the performance of MySQL without requiring any changes to most MySQL applications; similarly, Amazon Aurora PostgreSQL delivers up to three times the performance of PostgreSQL. Amazon RDS manages your Amazon Aurora databases, handling time-consuming tasks such as provisioning, patching, backup, recovery, failure detection and repair. You pay a simple monthly charge for each Amazon Aurora database instance you use. There are no upfront costs or long-term commitments required.

Q: What does "MySQL compatible" mean?

It means that most of the code, applications, drivers and tools you already use today with your MySQL databases can be used with Aurora with little or no change. The Amazon Aurora database engine is designed to be wire-compatible with MySQL 5.6 and 5.7 using the InnoDB storage engine. Certain MySQL features like the MyISAM storage engine are not available with Amazon Aurora.

Q: What does “PostgreSQL compatible” mean?

It means that most of the code, applications, drivers and tools you already use today with your PostgreSQL databases can be used with Aurora with little or no change. The Amazon Aurora database engine is designed to be wire-compatible with PostgreSQL 9.6 and 10, and supports the same set of PostgreSQL extensions that are supported with RDS for PostgreSQL 9.6 and 10, making it easy to move applications between the two engines.

Q: How do I try Amazon Aurora?

To try Amazon Aurora, sign in to the AWS console, select **RDS** under the **Database** category, and choose Amazon Aurora as your database engine.

Q: How much does Amazon Aurora cost?

Please see our pricing page for current pricing information.

Q: Amazon Aurora replicates each chunk of my database volume six ways across three Availability Zones. Does that mean that my effective storage price will be three or six times what is shown on the pricing page?

No. Amazon Aurora's replication is bundled into the price. You are charged based on the storage your database consumes at the database layer, not the storage consumed in Amazon Aurora's virtualized storage layer.

Q: In which AWS regions is Amazon Aurora available?

Please see our pricing page for current information on regions and prices.

Q: How can I migrate from MySQL to Amazon Aurora and vice versa?

You have several options. You can use the standard mysqldump utility to export data from MySQL and mysqlimport utility to import data to Amazon Aurora, and vice-versa. You can also use Amazon RDS's DB Snapshot migration feature to migrate an RDS MySQL DB Snapshot to Amazon Aurora using the AWS Management Console. Migration completes for most customers in under an hour, though the duration depends on format and data set size. For more information see Best Practices for Migrating MySQL Databases to Amazon Aurora.

Q: How can I migrate from PostgreSQL to Amazon Aurora and vice versa?

You have several options. You can use the standard pg_dump utility to export data from PostgreSQL and pg_restore utility to import data to Amazon Aurora, and vice-versa. You can also use Amazon RDS's DB Snapshot migration feature to migrate an RDS PostgreSQL DB Snapshot to Amazon Aurora using the AWS

Management Console. Migration completes for most customers in under an hour, though the duration depends on format and data set size.

Q: Does Amazon Aurora participate in the AWS Free Tier?

Not at this time. The AWS Free Tier for Amazon RDS offers benefits for Micro DB Instances; Amazon Aurora does not currently offer Micro DB Instance support. Please see our pricing page for current pricing information.

Q: What are IOs in Amazon Aurora and how are they calculated?

IOs are input/output operations performed by the Aurora database engine against its SSD-based virtualized storage layer. Every database page read operation counts as one IO. The Aurora database engine issues reads against the storage layer in order to fetch database pages not present in the buffer cache. Each database page is 16KB in Aurora MySQL and 8KB in Aurora PostgreSQL.

Aurora was designed to eliminate unnecessary IO operations in order to reduce costs and to ensure resources are available for serving read/write traffic. Write IOs are only consumed when pushing transaction log records to the storage layer for the purpose of making writes durable. Write IOs are counted in 4KB units. For example, a transaction log record that is 1024 bytes will count as one IO operation. However, concurrent write operations whose transaction log is less than 4KB can be batched together by the Aurora database engine in order to optimize I/O consumption. Unlike traditional database engines Amazon Aurora never pushes modified database pages to the storage layer, resulting in further IO consumption savings.

You can see how many IOs your Aurora instance is consuming by going to the AWS Console. To find your IO consumption, go to the RDS section of the console, look at your list of instances, select your Aurora instances, then look for the “Billed read operations” and “Billed write operations” metrics in the monitoring section.

Q: Do I need to change client drivers to use Amazon Aurora PostgreSQL?

No, Amazon Aurora will work with standard PostgreSQL database drivers.

Performance

Q: What does "five times the performance of MySQL" mean?

Amazon Aurora delivers significant increases over MySQL performance by tightly integrating the database engine with an SSD-based virtualized storage layer purpose-built for database workloads, reducing writes to the storage system, minimizing lock contention and eliminating delays created by database process threads. Our tests with SysBench on r3.8xlarge instances show that Amazon Aurora delivers over 500,000 SELECTs/sec and 100,000 UPDATEs/sec, five times higher than MySQL running the same benchmark on the same hardware. Detailed instructions on this benchmark and how to replicate it yourself are provided in the Amazon Aurora MySQL Performance Benchmarking Guide.

Q: What does "three times the performance of PostgreSQL" mean?

Amazon Aurora delivers significant increases over PostgreSQL performance by tightly integrating the database engine with an SSD-based virtualized storage layer purpose-built for database workloads, reducing writes to the storage system, minimizing lock contention and eliminating delays created by database process threads. Our tests with SysBench on r4.16xlarge instances show that Amazon Aurora delivers SELECTs/sec and UPDATEs/sec over three times higher than PostgreSQL running the same benchmark on the same hardware. Detailed instructions on this benchmark and how to replicate it yourself are provided in the Amazon Aurora PostgreSQL Performance Benchmarking Guide.

Q: How do I optimize my database workload for Amazon Aurora MySQL?

Amazon Aurora is designed to be compatible with MySQL, so that existing MySQL applications and tools can run without requiring modification. However, one area where Amazon Aurora improves upon MySQL is with highly concurrent workloads. In order to maximize your workload's throughput on Amazon Aurora, we recommend building your applications to drive a large number of concurrent queries and transactions.

Q: How do I optimize my database workload for Amazon Aurora PostgreSQL?

Amazon Aurora is designed to be compatible with PostgreSQL, so that existing PostgreSQL applications and tools can run without requiring modification. However, one area where Amazon Aurora improves upon PostgreSQL is with highly concurrent workloads. In order to maximize your workload’s throughput on Amazon Aurora, we recommend building your applications to drive a large number of concurrent queries and transactions.

Hardware and Scaling

Q: What are the minimum and maximum storage limits of an Amazon Aurora database?

The minimum storage is 10GB. Based on your database usage, your Amazon Aurora storage will automatically grow, up to 64 TB, in 10GB increments with no impact to database performance. There is no need to provision storage in advance.

Q: How do I scale the compute resources associated with my Amazon Aurora DB Instance?

You can scale the compute resources allocated to your DB Instance in the AWS Management Console by selecting the desired DB Instance and clicking the Modify button. Memory and CPU resources are modified by changing your DB Instance class.

When you modify your DB Instance class, your requested changes will be applied during your specified maintenance window. Alternatively, you can use the "Apply Immediately" flag to apply your scaling requests immediately. Both of these options will have an availability impact for a few minutes as the scaling operation is performed. Bear in mind that any other pending system changes will also be applied.

Backup and Restore

Q: How do I enable backups for my DB Instance?

Automated backups are always enabled on Amazon Aurora DB Instances. Backups do not impact database performance.

Q: Can I take DB Snapshots and keep them around as long as I want?

Yes, and there is no performance impact when taking snapshots. Note that restoring data from DB Snapshots requires creating a new DB Instance.

Q: If my database fails, what is my recovery path?

Amazon Aurora automatically maintains 6 copies of your data across 3 Availability Zones and will automatically attempt to recover your database in a healthy AZ with no data loss. In the unlikely event your data is unavailable within Amazon Aurora storage, you can restore from a DB Snapshot or perform a point-in-time restore operation to a new instance. Note that the latest restorable time for a point-in-time restore operation can be up to 5 minutes in the past.

Q: What happens to my automated backups and DB Snapshots if I delete my DB Instance?

You can choose to create a final DB Snapshot when deleting your DB Instance. If you do, you can use this DB Snapshot to restore the deleted DB Instance at a later date. Amazon Aurora retains this final user-created DB Snapshot along with all other manually created DB Snapshots after the DB Instance is deleted. Only DB Snapshots are retained after the DB Instance is deleted (i.e., automated backups created for point-in-time restore are not kept).

Q: Can I share my snapshots with another AWS account?

Yes. Aurora gives you the ability to create snapshots of your databases, which you can use later to restore a database. You can share a snapshot with a different AWS account, and the owner of the recipient account can use your snapshot to restore a DB that contains your data. You can even choose to make your snapshots public – that is, anybody can restore a DB containing your (public) data. You can use this feature to share data between your various environments (production, dev/test, staging, etc.) that have different AWS accounts, as well as keep backups of all your data secure in a separate account in case your main AWS account is ever compromised.

Q: Will I be billed for shared snapshots?

There is no charge for sharing snapshots between accounts. However, you may be charged for the snapshots themselves, as well as any databases you restore from shared snapshots. Learn more about Aurora pricing.

Q: Can I automatically share snapshots?

We do not support sharing automatic DB snapshots. To share an automatic snapshot, you must manually create a copy of the snapshot, and then share the copy.

Q: How many accounts can I share snapshots with?

You may share manual snapshots with up to 20 AWS account IDs. If you want to share the snapshot with more than 20 accounts, you can either share the snapshot as public, or contact support for increasing your quota.

Q: In which regions can I share my Aurora snapshots?

You can share your Aurora snapshots in all AWS regions where Aurora is available.

Q: Can I share my Aurora snapshots across different regions?

No. Your shared Aurora snapshots will only be accessible by accounts in the same region as the account that shares them.

Q: Can I share an encrypted Aurora snapshot?

Yes, you can share encrypted Aurora snapshots.

High Availability and Replication

Q: How does Amazon Aurora improve my database’s fault tolerance to disk failures?

Amazon Aurora automatically divides your database volume into 10GB segments spread across many disks. Each 10GB chunk of your database volume is replicated six ways, across three Availability Zones. Amazon Aurora is designed to transparently handle the loss of up to two copies of data without affecting database write availability and up to three copies without affecting read availability. Amazon Aurora storage is also self-healing. Data blocks and disks are continuously scanned for errors and repaired automatically.

Q: How does Aurora improve recovery time after a database crash?

Unlike other databases, after a database crash Amazon Aurora does not need to replay the redo log from the last database checkpoint (typically 5 minutes) and confirm that all changes have been applied, before making the database available for operations. This reduces database restart times to less than 60 seconds in most cases. Amazon Aurora moves the buffer cache out of the database process and makes it available immediately at restart time. This prevents you from having to throttle access until the cache is repopulated to avoid brownouts.

Q: What kind of replicas does Aurora support?

Amazon Aurora MySQL and Amazon Aurora PostgreSQL support Amazon Aurora Replicas, which share the same underlying volume as the primary instance in the same AWS region. Updates made by the primary are visible to all Amazon Aurora Replicas. With Amazon Aurora MySQL, you can also create cross-region MySQL Read Replicas based on MySQL’s binlog-based replication engine. In MySQL Read Replicas, data from your primary instance is replayed on your replica as transactions. For most use cases, including read scaling and high availability, we recommend using Amazon Aurora Replicas.

You have the flexibility to mix and match these two replica types based on your application needs:

Feature	Amazon Aurora Replicas	MySQL Replicas

Number of replicas	Up to 15	Up to 5
Replication type	Asynchronous (milliseconds)	Asynchronous (seconds)
Performance impact on primary	Low	High
Replica location	In-region	Cross-region
Act as failover target	Yes (no data loss)	Yes (potentially minutes of data loss)
Automated failover	Yes	No
Support for user-defined replication delay	No	Yes
Support for different data or schema vs. primary	No	Yes

You have two additional replication options in addition to the ones listed above. You can use Aurora Global Database for much faster physical replication between Aurora clusters in different regions. And for replication between Aurora and non-Aurora MySQL databases (even outside of AWS), you can set up your own, self-managed binlog replication.

Q: Can I have cross-region replicas with Amazon Aurora?

Yes, with Aurora MySQL you can set up cross-region Aurora Replicas using either logical or physical replication.

Logical replication can replicate to up to five secondary AWS regions, and is based on single threaded MySQL binlog replication, so the replication lag will be influenced by the change/apply rate and delays in network communication between the specific regions selected. Physical replication, called Aurora Global Database, uses dedicated infrastructure that leaves your databases entirely available to serve your application, and can replicate to one secondary region with typical latency of under a second. For low-latency global reads and disaster recovery, we recommend using Global Database.

Aurora PostgreSQL does not currently support cross-region replicas.

Q: Can I create Aurora Replicas on the cross-region replica cluster?

Yes, you can add up to 15 Aurora Replicas on each cross-region cluster, and they will share the same underlying storage as the cross-region replica. A cross-region replica acts as the primary on the cluster and the Aurora Replicas on the cluster will typically lag behind the primary by 10s of milliseconds.

Q: Can I fail over my application from my current primary to the cross-region replica?

Yes, you can promote your cross-region replica to be the new primary from the RDS console. For logical (binlog) replication, the promotion process typically takes a few minutes depending on your workload. The cross-region replication will stop once you initiate the promotion process.

With Aurora Global Database, you can promote a secondary region to take full read/write workloads in under a minute.

Q: Can I prioritize certain replicas as failover targets over others?

Yes. You can assign a promotion priority tier to each instance on your cluster. When the primary instance fails, Amazon RDS will promote the replica with the highest priority to primary. If there is contention between 2 or more replicas in the same priority tier, then Amazon RDS will promote the replica that is the same size as the primary instance. For more information on failover logic, read the Amazon Aurora User Guide.

Q: Can I modify priority tiers for instances after they have been created?

Yes, you can modify the priority tier for an instance at any time. Simply modifying priority tiers will not trigger a failover.

Q: Can I prevent certain replicas from being promoted to the primary instance?

You can assign lower priority tiers to replicas that you don’t want promoted to the primary instance. However, if the higher priority replicas on the cluster are unhealthy or unavailable for some reason, then Amazon RDS will promote the lower priority replica.

Q: How can I improve upon the availability of a single Amazon Aurora database?

You can add Amazon Aurora Replicas. Aurora Replicas in the same AWS Region share the same underlying storage as the primary instance. Any Aurora Replica can be promoted to become primary without any data loss and therefore can be used for enhancing fault tolerance in the event of a primary DB Instance failure. To increase database availability, simply create 1 to 15 replicas, in any of 3 AZs, and Amazon RDS will automatically include them in failover primary selection in the event of a database outage.

You can use Aurora Global Database if you want your database to span multiple AWS Regions. This will replicate your data with no impact on database performance, and provide disaster recovery from region-wide outages.

Q: What happens during failover and how long does it take?

Failover is automatically handled by Amazon Aurora so that your applications can resume database operations as quickly as possible without manual administrative intervention.

- If you have an Amazon Aurora Replica, in the same or a different Availability Zone, when failing over, Aurora flips the canonical name record (CNAME) for your DB Instance to point at the healthy replica, which is in turn promoted to become the new primary. Start-to-finish, failover typically completes within 30 seconds.
- If you are running Aurora Serverless and the DB instance or AZ become unavailable, Aurora will automatically recreate the DB instance in a different AZ.
- If you do not have an Amazon Aurora Replica (i.e. single instance) and are not running Aurora Serverless, Aurora will attempt to create a new DB Instance in the same Availability Zone as the original instance. This replacement of the original instance is done on a best-effort basis and may

not succeed, for example, if there is an issue that is broadly affecting the Availability Zone.

Your application should retry database connections in the event of connection loss.

Disaster recovery across regions is a manual process, where you promote a secondary region to take read/write workloads.

Q: If I have a primary database and an Amazon Aurora Replica actively taking read traffic and a failover occurs, what happens?

Amazon RDS will automatically detect a problem with your primary instance and trigger a failover. If you are using the Cluster Endpoint, your read/write connections will be automatically redirected to an Amazon Aurora Replica that will be promoted to primary.

In addition, the read traffic that your Aurora Replicas were serving will be briefly interrupted. If you are using the Cluster Reader Endpoint to direct your read traffic to the Aurora Replica, the read only connections will be directed to the newly promoted Aurora Replica until the old primary node is recovered as a replica.

Q: How far behind the primary will my replicas be?

Since Amazon Aurora Replicas share the same data volume as the primary instance in the same AWS Region, there is virtually no replication lag. We typically observe lag times in the 10s of milliseconds. For MySQL Read Replicas, the replication lag can grow indefinitely based on change/apply rate as well as delays in network communication. However, under typical conditions, under a minute of replication lag is common.

Cross-region replicas using logical replication will be influenced by the change/apply rate and delays in network communication between the specific regions selected. Cross-region replicas using Aurora Global Database will have a typical lag of under a second.

Q: Can I set up replication between my Aurora MySQL database and an external MySQL database?

Yes, you can set up binlog replication between an Aurora MySQL instance and an external MySQL database. The other database can run on Amazon RDS, or as a self-managed database on AWS, or completely outside of AWS.

If you're running Aurora MySQL 5.7, consider setting up GTID-based binlog replication. This will provide complete consistency so your replication won’t miss transactions or generate conflicts, even after failover or downtime.

Q: What is Amazon Aurora Global Database?

Amazon Aurora Global Database is a feature that allows a single Amazon Aurora database to span multiple AWS regions. It replicates your data with no impact on database performance, enables fast local reads in each region with typical latency of less than a second, and provides disaster recovery from region-wide outages. In the unlikely event of a regional degradation or outage, a secondary region can be promoted to full read/write capabilities in less than 1 minute.

This feature is available for Amazon Aurora MySQL.

Q: How do I create an Aurora Global Database?

You can create an Aurora Global Database with just a few clicks in the Amazon RDS Management Console. Alternatively, you can use the SDK or CLI. You need to provision at least one instance per region in your Aurora Global Database.

Q: If I use Aurora Global Database, can I also use logical replication (binlog) on the primary database?

Yes. If your goal is to analyze database activity, consider using Aurora advanced auditing, general logs, and slow query logs instead, to avoid impacting the performance of your database.

Q: Will Aurora automatically fail over to a secondary region of an Aurora Global Database?

No. If your primary region becomes unavailable, you can manually remove a secondary region from an Aurora Global Database and promote it to take full reads and writes. You will also need to point your application to the newly promoted region.

Q: What is Amazon Aurora Multi-Master?

Amazon Aurora Multi-Master is a new feature of the Aurora MySQL-compatible edition that adds the ability to scale out write performance across multiple Availability Zones, allowing applications to direct read/write workloads to multiple instances in a database cluster and operate with higher availability.

Q: How can I get started with Amazon Aurora Multi-Master?

Amazon Aurora Multi-Master is now generally available. You can read the Amazon Aurora documentation to learn more. You can create an Aurora Multi-Master cluster with just a few clicks in the Amazon RDS Management Console or download the latest AWS SDK or CLI.

Security

Q: Can I use Amazon Aurora in Amazon Virtual Private Cloud (Amazon VPC)?

Yes, all Amazon Aurora DB Instances must be created in a VPC. With Amazon VPC, you can define a virtual network topology that closely resembles a traditional network that you might operate in your own datacenter. This gives you complete control over who can access your Amazon Aurora databases.

Q: Does Amazon Aurora encrypt my data in transit and at rest?

Yes. Amazon Aurora uses SSL (AES-256) to secure the connection between the database instance and the application. Amazon Aurora allows you to encrypt your databases using keys you manage through AWS Key Management Service (KMS). On a database instance running with Amazon Aurora encryption, data stored at rest in the underlying storage is encrypted, as are its automated backups, snapshots, and replicas in the same cluster. Encryption and decryption are handled seamlessly. For more information about the use of KMS with Amazon Aurora, see the Amazon RDS User's Guide.

Q: Can I encrypt an existing unencrypted database?

Currently, encrypting an existing unencrypted Aurora instance is not supported. To use Amazon Aurora encryption for an existing unencrypted database, create a new DB Instance with encryption enabled and migrate your data into it.

Q: How do I access my Amazon Aurora database?

Access to Amazon Aurora databases must be done through the database port entered on database creation. This is done to provide an additional layer of security for your data. Step by step instructions on how to connect to your Amazon Aurora database is provided in the Amazon Aurora Connectivity Guide.

Q: Can I use Amazon Aurora with applications that require HIPAA compliance?

Yes, the MySQL- and PostgreSQL-compatible editions of Aurora are HIPAA-eligible, so you can use them to build HIPAA-compliant applications and store healthcare related information, including protected health information (PHI) under an executed Business Associate Agreement (BAA) with AWS. If you already have an executed BAA, no action is necessary to begin using these services in the account(s) covered by your BAA. For more information about building compliant applications on AWS, see Healthcare Providers & Insurers in the Cloud.

Q: Where can I access a list of Common Vulnerabilities and Exposures (CVE) entries for publicly known cybersecurity vulnerabilities for Amazon Aurora releases?

You can currently find a list of CVEs at Amazon Aurora Security Updates.

Serverless

Q: What is Amazon Aurora Serverless?

Amazon Aurora Serverless is an on-demand, autoscaling configuration for the MySQL-compatible and PostgreSQL-compatible editions of Amazon Aurora. An Aurora Serverless DB cluster automatically starts up, shuts down, and scales capacity up or down based on your application's needs. Aurora Serverless provides a relatively simple, cost-effective option for infrequent, intermittent, or unpredictable workloads. Read more in the Amazon Aurora User Guide.

Q: Which versions of Amazon Aurora are supported for Aurora Serverless?

Aurora Serverless is currently available for Aurora with MySQL 5.6 compatibility and for Aurora with PostgreSQL 10.7+ compatibility.

Q: Can I migrate an existing Aurora DB cluster to Aurora Serverless?

Yes, you can restore a snapshot taken from an existing Aurora provisioned cluster into an Aurora Serverless DB Cluster (and vice versa).

Q: How do I connect to an Aurora Serverless DB cluster?

You access an Aurora Serverless DB cluster from within a client application runing in the same Amazon Virtual Private Cloud (VPC). You can't give an Aurora Serverless DB cluster a public IP address.

Q: Can I explicitly set the capacity of an Aurora Serverless cluster?

While Aurora Serverless automatically scales based on the active database workload, in some cases, capacity might not scale fast enough to meet a sudden workload change, such as a large number of new transactions. In these cases, you can set the capacity explicitly to a specific value with the AWS Management Console, the AWS CLI, or the RDS API.

Q: Why isn't my Aurora Serverless DB Cluster automatically scaling?

Once a scaling operation is initiated, Aurora Serverless attempts to find a scaling point, which is is a point in time at which the database can safely complete scaling. Aurora Serverless might not be able to find a scaling point if you have long-running queries or transactions in progress, or temporary tables or table locks in use.

Q: How am I billed for Aurora Serverless?

In Aurora Serverless, database capacity is measured in Aurora Capacity Units (ACUs). You pay a flat rate per second of ACU usage, with a minimum of 5 minutes of usage each time the database is activated. Storage and I/O prices are the same for provisioned and Serverless configurations. View an Aurora Serverless pricing example.

Parallel Query

Q: What is Amazon Aurora Parallel Query?

Amazon Aurora Parallel Query refers to the ability to push down and distribute the computational load of a single query across thousands of CPUs in Aurora's storage layer. Without Parallel Query, a query issued against an Amazon Aurora database would be executed wholly within one instance of the database cluster; this would be similar to how most databases operate.

Q: What's the target use case?

Parallel Query is a good fit for analytical workloads requiring fresh data and good query performance, even on large tables. Workloads of this type are often operational in nature.

Q: What benefits does Parallel Query provide?

Faster performance: Parallel Query can speed up analytical queries by up to 2 orders of magnitude.

Operational simplicity and data freshness: you can issue a query directly over the current transactional data in your Aurora cluster.

Transactional and analytical workloads on the same database: Parallel Query allows Aurora to maintain high transaction throughput alongside concurrent analytical queries.

Q: What specific queries improve under Parallel Query?

Most queries over large data sets that are not already in the buffer pool can expect to benefit. The initial version of Parallel Query can push down and scale out of the processing of more than 200 SQL functions, equijoins, and projections.

Q: What performance improvement can I expect?

The improvement to a specific query’s performance depends on how much of the query plan can be pushed down to the Aurora storage layer. Customers have reported more than an order of magnitude improvement to query latency.

Q: Is there any chance that performance will be slower?

Yes, but we expect such cases to be rare.

Q: What changes do I need to make to my query to take advantage of Parallel Query?

No changes in query syntax are required. The query optimizer will automatically decide whether to use PQ for your specific query. To check if a query is using PQ, you can view the query execution plan by running the EXPLAIN command. If you wish to bypass the heuristics and force Parallel Query for test purposes, use the aurora_pq_force session variable.

Q: How do I turn the feature on or off?

Parallel Query can be enabled and disabled dynamically at both the global and session level using the aurora_pq parameter.

Q: Are there any additional charges associated with using Parallel Query?

No. You aren’t charged for anything other than what you already pay for instances, IO, and storage.

Q: Since Parallel Query reduces IO, will turning it on reduce my Aurora IO charges?

No, IO costs for your query are metered at the storage layer, and will be the same or larger with Parallel Query turned on. Your benefit is the improvement in query performance.

There are two reasons for potentially higher IO costs with Parallel Query. First, even if some of the data in a table is in the buffer pool, PQ requires all data to be scanned at the storage layer, incurring IO. Second, a side effect of avoiding contention in the buffer pool is that running a PQ query does not warm up the buffer pool. As a result, consecutive runs of the same PQ query will incur the full IO cost.

Q: What versions of Amazon Aurora support Parallel Query?

Parallel Query is available for the MySQL 5.6-compatible version of Amazon Aurora, starting with v1.18.0. We plan to extend Parallel Query to Aurora with MySQL 5.7 compatibility, and to Aurora with PostgreSQL compatibility.

Q: Is Parallel Query available with all instance types?

No. At this time, you can use Parallel Query with instances in the R* instance family.

Q: Is Parallel Query compatible with all other Aurora features?

Not initially. At this time, you can only turn it on for database clusters that aren't running the Serverless or Backtrack features. Further, it doesn’t support functionality specific to Aurora with MySQL 5.7 compatibility.

Q: If Parallel Query speeds up queries with only rare performance losses, should I simply turn it on for all all the time?

No. While we expect Parallel Query to improve query latency in most cases, you may incur higher IO costs. We recommend that you thoroughly test your workload with the feature enabled and disabled; once you're convinced that Parallel Query is the right choice, you can rely on the query optimizer to automatically decide which queries will use PQ. In the rare case when the optimizer doesn’t make the optimal decision, you can override the setting.

Q: Can Aurora Parallel Query replace my data warehouse?

Aurora Parallel Query is not a data warehouse, and doesn’t provide the functionality typically found in such products. It’s designed to speed up query performance on your relational database, and is suitable for use cases such as operational analytics, when you need to perform fast analytical queries on fresh data in your database.

Amazon DocumentDB (with MongoDB compatibility) FAQs

General

Q: What is Amazon DocumentDB (with MongoDB compatibility)?

Amazon DocumentDB (with MongoDB compatibility) is a fast, scalable, highly available, and fully managed document database service that supports MongoDB workloads. Developers can use the same MongoDB application code, drivers, and tools as they do today to run, manage, and scale workloads on Amazon DocumentDB and enjoy improved performance, scalability, and availability without having to worry about managing the underlying infrastructure. Customers can use AWS Database Migration Service (DMS) for free (for six months) to easily migrate their on-premises or Amazon Elastic Compute Cloud (EC2) MongoDB databases to Amazon DocumentDB with virtually no downtime. There are no up-front investments required to use Amazon DocumentDB, and customers only pay for the capacity they use.

Q: What does "MongoDB-compatible" mean?

It means that a vast majority of the applications, drivers, and tools you already use today with your MongoDB database can be used with Amazon DocumentDB with little or no change. Amazon DocumentDB emulates the responses that a client expects from a MongoDB server by implementing the Apache 2.0 open source MongoDB 3.6 API on a purpose-built, distributed, fault-tolerant, self-healing storage system that gives customers the performance, scalability, and availability they need when operating mission-critical MongoDB workloads at scale. Learn more about supported MongoDB APIs.

Q: How does Amazon DocumentDB work?

Amazon DocumentDB emulates the responses that a client expects from a MongoDB server by implementing the Apache 2.0 open source MongoDB 3.6 API on a purpose-built, distributed, fault-tolerant, self-healing storage system that gives customers the performance, scalability, and availability they need when operating mission-critical MongoDB workloads at scale.

Q: How can I migrate data from an existing MongoDB database to Amazon DocumentDB?

Customers can use AWS Database Migration Service (DMS) for free (for six months) to easily migrate their on-premises or Amazon Elastic Compute Cloud (EC2) MongoDB databases to Amazon DocumentDB with virtually no downtime. With DMS, you can migration from a MongoDB replica set or from a sharded cluster to Amazon DocumentDB. Additionally, you can use most existing tools to migrate data from a MongoDB database to Amazon DocumentDB, including mongodump/mongorestore, mongoexport/mongoimport, and third-party tools that support Change Data Capture via the oplog. For more information, see Migrating to Amazon DocumentDB.

Q: What is Free DMS and does it apply to Amazon DocumentDB?

AWS Database Migration Service (DMS) offers free use for 6 months per instance if you're migrating Amazon DocumentDB. AWS Database Migration Service helps you migrate databases to AWS quickly and securely. The source database remains fully operational during the migration, minimizing downtime to applications that rely on the database. The AWS Database Migration Service can migrate your data to Amazon DocumentDB from most widely used commercial and open-source databases. For more information, see DMS Free.

Q: Do I need to change client drivers to use Amazon DocumentDB?

No, Amazon DocumentDB works with a vast majority of MongoDB drivers compatible with MongoDB 3.4+.

Q: How do I access my Amazon DocumentDB cluster?

Amazon DocumentDB clusters are deployed within a customer's Amazon VPC and can be accessed directly by EC2 instances or other AWS services that are deployed in the same VPC. Additionally, Amazon DocumentDB can be accessed by EC2 instances or other AWS services in different VPCs in the same region or other regions via VPC peering. Access to Amazon DocumentDB clusters must be done through the mongo shell or with MongoDB drivers. Amazon DocumentDB requires that you authenticate when connecting to a cluster. For additional options, see Connecting to an Amazon DocumentDB Cluster from Outside an Amazon VPC.

Q: Why are Amazon RDS permissions and resources required to use Amazon DocumentDB?

For certain management features such as instance lifecycle management, encryption-at-rest with Amazon Key Management Service (KMS) keys and security groups management, Amazon DocumentDB leverages operational technology that is shared with Amazon RDS.

Q: What instances types does Amazon DocumentDB offer?

Please see the Amazon DocumentDB pricing page for current information on available instant types per region.

Q: How do I try Amazon DocumentDB?

To try Amazon DocumentDB, please see the Getting Started section.

Q: Does Amazon DocumentDB have an SLA?

Yes. For more information, please see Amazon DocumentDB (with MongoDB compatibility) Service Level Agreement.

Performance

Q: What type of performance can I expect from Amazon DocumentDB?

When writing to storage, Amazon DocumentDB only persists a write-ahead logs, and does not need to write full buffer page syncs. As a result of this optimization, which does not compromise durability, Amazon DocumentDB writes are typically faster than traditional databases. Amazon DocumentDB clusters can scale out to millions of reads per second with up to 15-read replicas.

Pricing

Q: How much does Amazon DocumentDB cost?

Please see the Amazon DocumentDB pricing page for current pricing information.

Q: In which AWS regions is Amazon DocumentDB available?

Please see the Amazon DocumentDB pricing page for current information on regions and prices.

Q: Does Amazon DocumentDB have a free tier?

No, Amazon DocumentDB does not have a free tier.

Q: Amazon DocumentDB replicates each chunk of my storage volume six ways across three Availability Zones. Does that mean that my effective storage price will be three or six times what is shown on the pricing page?

No. Amazon DocumentDB's storage replication feature is included in the price. You are charged based on the size of the data in your cluster and you are not charged separately for the replicated storage.

Q: What are IOs in Amazon DocumentDB and how are they calculated?

IOs are input/output operations performed by Amazon DocumentDB against an SSD-based virtualized storage layer. Every database page read operation counts as one IO. Amazon DocumentDB issues reads against the storage layer in order to fetch pages not present in the buffer cache. Each page is 8KB in Amazon DocumentDB.

Amazon DocumentDB was designed to eliminate unnecessary IO operations in order to reduce costs and to ensure resources are available for serving read/write traffic. Write IOs are only consumed when pushing write-ahead log records to the storage layer for the purpose of making writes durable. Write IOs are counted in 4KB units. For example, a log record that is 1KB will count as one IO operation. However, concurrent write operations whose write-ahead log is less than 4KB can be batched together by the Amazon DocumentDB database engine in order to optimize I/O consumption. Unlike traditional database engines Amazon DocumentDB never pushes modified database pages to the storage layer, resulting in further IO consumption savings.

You can see how many IOs your Amazon DocumentDB cluster is consuming by going to the AWS Console. To find your IO consumption, go to the Amazon DocumentDB section of the console, look at your list of clusters, select your Amazon DocumentDB cluster, then look for the “VolumeReadIOPs” and “VolumeWriteIOPs” metrics in the monitoring section.

Q: How does per-second billing work?

Instance pricing is calculated from the time the instance is created to the time the instance is deleted. Instances are billed in one second increments, with a 10 minute minimum charge following a billable status change such as creating, modifying, or deleting an instance.

As an example, if you create an r5.large instance at 1:00:00 PM, modify the instance to an r5.xlarge at 1:30:00 PM and delete the instance at 1:50:00 PM, you will be charged for 1,800 seconds (30 minutes) at r5.large pricing and 1,200 seconds (20 minutes) at r5.xlarge prices. If you create an r5.large instance at 3:00:00 PM and delete it at 3:06:00PM, the 10 minute minimum applies, and you will be charged for 600 seconds (10 minutes) at r5.large pricing.

Hardware, Scaling, and Storage

Q: What are the minimum and maximum storage limits of an Amazon DocumentDB cluster?

The minimum storage is 10GB. Based on your cluster usage, your Amazon DocumentDB storage will automatically grow, up to 64 TB in 10GB increments with no impact on performance. There is no need to provision storage in advance.

Q: How does Amazon DocumentDB scale?

Amazon DocumentDB scales in two dimensions: storage and compute. Amazon DocumentDB's storage automatically scales from 10GB to 64 TB in increments of 10GB. Amazon DocumentDB's compute capacity can be scaled up by creating larger instances and horizontally (for greater read throughput) by adding additional replica instances (up to 15) to the cluster.

Q: How do I scale the compute resources associated with my Amazon DocumentDB cluster?

You can scale the compute resources allocated to your instance in the AWS Management Console by selecting the desired instance and clicking the “modify” button. Memory and CPU resources are modified by changing your instance class.

When you modify your instance class, your requested changes will be applied during your specified maintenance window. Alternatively, you can use the "Apply Immediately" flag to apply your scaling requests immediately. Both of these options will have an availability impact for a few minutes as the scaling operation is performed. Bear in mind that any other pending system changes will also be applied.

Q: Does Amazon DocumentDB support MongoDB sharding?

No. Amazon DocumentDB's distributed storage architecture is a different approach to scaling than MongoDB sharding.

Backup and Restore

Q: How do I enable backups for my cluster?

Automated backups are always enabled on Amazon DocumentDB clusters. You can increase your backup window for point-in-time restores up to 35-days. Backups do not impact database performance.

Q: Can I take cluster snapshots and keep them around as long as I want?

Yes. Manual snapshots can be retained beyond the backup window and there is no performance impact when taking snapshots. Note that restoring data from cluster snapshots requires creating a new cluster.

Q: If my instance fails, what is my recovery path?

Amazon DocumentDB automatically maintains six copies of your data across three Availability Zones and will automatically attempt to recover your instance in a healthy AZ with no data loss. In the unlikely event your data is unavailable within Amazon DocumentDB storage, you can restore from a cluster snapshot or perform a point-in-time restore operation to a new cluster. Note that the latest restorable time for a point-in-time restore operation can be up to five minutes in the past.

Q: What happens to my automated backups and cluster snapshots if I delete my cluster?

You can choose to create a final snapshot when deleting your instance. If you do, you can use this snapshot to restore the deleted instance at a later date. Amazon DocumentDB retains this final user-created snapshot along with all other manually created snapshots after the instance is deleted. Only snapshots are retained after the instance is deleted (i.e., automated backups created for point-in-time restore are not kept).

Q: What happens to my automated backups and cluster snapshots if I delete my account?

Deleting your AWS account will delete all automated backups and snapshot backups contained in the account.

Q: Can I share my snapshots with another AWS account?

Yes. Amazon DocumentDB gives you the ability to create snapshots of your cluster, which you can use later to restore a cluster. You can share a snapshot with a different AWS account, and the owner of the recipient account can use your snapshot to restore a cluster that contains your data. You can even choose to make your snapshots public – that is, anybody can restore a cluster containing your (public) data. You can use this feature to share data between your various environments (production, dev/test, staging, etc.) that have different AWS accounts, as well as keep backups of all your data secure in a separate account in case your main AWS account is ever compromised.

Q: Will I be billed for shared snapshots?

There is no charge for sharing snapshots between accounts. However, you may be charged for the snapshots themselves, as well as any clusters that you restore from shared snapshots.

Q: Can I automatically share snapshots?

We do not support sharing automatic cluster snapshots. To share an automatic snapshot, you must manually create a copy of the snapshot, and then share the copy.

Q: Can I share my Amazon DocumentDB snapshots across different regions?

No. Your shared Amazon DocumentDB snapshots will only be accessible by accounts in the same region as the account that shares them.

Q: Can I share an encrypted Amazon DocumentDB snapshot?

Yes. You can share encrypted Amazon DocumentDB snapshots. The recipient of the shared snapshot must have access to the KMS key that was used to encrypt the snapshot.

Q: Can I use Amazon DocumentDB snapshots outside of the service?

No. Amazon DocumentDB snapshots can only be used inside of the service.

Q: What happens to my backups if I delete my cluster?

You can choose to create a final snapshot when deleting your cluster. If you do, you can use this snapshot to restore the deleted cluster at a later date. Amazon DocumentDB retains this final user-created snapshot along with all other manually created snapshots after the cluster is deleted.

High Availability and Replication

Q: How does Amazon DocumentDB improve my cluster's fault tolerance to disk failures?

Amazon DocumentDB automatically divides your storage volume into 10GB segments spread across many disks. Each 10GB chunk of your storage volume is replicated six ways, across three Availability Zones. Amazon DocumentDB is designed to transparently handle the loss of up to two copies of data without affecting write availability and up to three copies without affecting read availability. Amazon DocumentDB's storage volume is also self-healing. Data blocks and disks are continuously scanned for errors and repaired automatically.

Q: How does Amazon DocumentDB improve recovery time after a database crash?

Unlike other databases, after a database crash Amazon DocumentDB does not need to replay the redo log from the last database checkpoint (typically 5 minutes) and confirm that all changes have been applied, before making the database available for operations. This reduces database restart times to less than 60 seconds in most cases. Amazon DocumentDB moves the cache out of the database process and makes it available immediately at restart time. This prevents you from having to throttle access until the cache is repopulated to avoid brownouts.

Q: What kind of replicas does Amazon DocumentDB support?

Amazon DocumentDB supports read replicas, which share the same underlying storage volume as the primary instance. Updates made by the primary instance are visible to all Amazon DocumentDB replicas.

Feature: Amazon DocumentDB read replicas

Number of replicas: Up to 15

Replication Type: Asynchronous (typically milliseconds)

Performance impact on primary: Low

Act as failover target: Yes (no data loss)

Automated failover: Yes

Q: Can I have cross-region replicas with Amazon DocumentDB?

No. Amazon DocumentDB does not support cross-region replicas.

Q: Can I prioritize certain replicas as failover targets over others?

Yes. You can assign a promotion priority tier to each instance on your cluster. If the primary instance fails, Amazon DocumentDB will promote the replica with the highest priority to primary. If there are inconsistencies between two or more replicas in the same priority tier, then Amazon DocumentDB will promote the replica that is the same size as the primary instance.

Q: Can I modify priority tiers for instances after they have been created?

You can modify the priority tier for an instance at any time. Simply modifying priority tiers will not trigger a failover.

Q: Can I prevent certain replicas from being promoted to the primary instance?

You can assign lower priority tiers to replicas that you do not want promoted to the primary instance. However, if the higher priority replicas on the cluster are unhealthy or unavailable for some reason, then Amazon DocumentDB will promote the lower priority replica.

Q: How does Amazon DocumentDB assure high availability of my cluster?

Amazon DocumentDB can be deployed in a high-availability configuration by using replica instances in multiple AWS Availability Zones as failover targets. In the event of a primary instance failure, a replica instance is automatically promoted to be the new primary with minimal service interruption.

Q: How can I improve upon the availability of a single Amazon DocumentDB instance?

You can add additional Amazon DocumentDB replicas. Amazon DocumentDB replicas share the same underlying storage as the primary instance. Any Amazon DocumentDB replica can be promoted to become primary without any data loss and therefore can be used for enhancing fault tolerance in the event of a primary instance failure. To increase cluster availability, simply create 1 to 15 replicas, in multiple AZs, and Amazon DocumentDB will automatically include them in failover primary selection in the event of an instance outage.

Q: What happens during failover and how long does it take?

Failover is automatically handled by Amazon DocumentDB so that your applications can resume database operations as quickly as possible without manual administrative intervention.

- If you have an Amazon DocumentDB replica, in the same or a different Availability Zone, when failing over, Amazon DocumentDB flips the canonical name record (CNAME) for your cluster endpoint to a healthy replica, which is in turn is promoted to become the new primary. Start-to-finish, failover typically completes within 30 seconds. Additionally, the read replicas endpoint doesn't require any CNAME updates during failover.
- If you do not have an Amazon DocumentDB Replica (i.e. single instance), Amazon DocumentDB will first attempt to create a new instance in the same Availability Zone as the original instance. If unable to do so, Amazon DocumentDB will attempt to create a new instance in a different Availability Zone. From start to finish, failover typically completes in under 15 minutes.

Your application should retry requests in the event of connection loss.

Q: If I have a primary instance and an Amazon DocumentDB replica instance actively taking read traffic and a failover occurs, what happens?

Amazon DocumentDB will automatically detect a problem with your primary instance and begin routing your read/write traffic to an Amazon DocumentDB replica instance. On average, this failover will complete within 30 seconds. In addition, the read traffic that your Amazon DocumentDB replicas instances were serving will be briefly interrupted.

Q: How far behind the primary will my replicas be?

Since Amazon DocumentDB replicas share the same data volume as the primary instance, there is virtually no replication lag. We typically observe lag times in the 10s of milliseconds.

Security and Compliance

Q: Can I use Amazon DocumentDB in Amazon Virtual Private Cloud (Amazon VPC)?

Yes. All Amazon DocumentDB clusters must be created in a VPC. With Amazon VPC, you can define a virtual network topology that closely resembles a traditional network that you might operate in your own datacenter. This gives you complete control over who can access your Amazon DocumentDB clusters.

Q: How do the existing MongoDB authentication modes work with Amazon DocumentDB?

Amazon DocumentDB utilizes VPC's strict network and authorization boundary. Authentication and authorization for Amazon DocumentDB management APIs is provided by IAM users, roles, and policies. Authentication to an Amazon DocumentDB database is done via standard MongoDB tools and drivers with Salted Challenge Response Authentication Mechanism (SCRAM), the default authentication mechanism for MongoDB.

Q: Does Amazon DocumentDB support encrypting my data-at-rest?

Yes. Amazon DocumentDB allows you to encrypt your clusters using keys you manage through AWS Key Management Service (KMS). On a cluster running with Amazon DocumentDB encryption, data stored at rest in the underlying storage is encrypted, as are its automated backups, snapshots, and replicas in the same cluster. Encryption and decryption are handled seamlessly. For more information about the use of KMS with Amazon DocumentDB, see the Encrypting Amazon DocumentDB Data at Rest.

Q: Can I encrypt an existing unencrypted cluster?

Currently, encrypting an existing unencrypted Amazon DocumentDB cluster is not supported. To use Amazon DocumentDB encryption for an existing unencrypted cluster, create a new cluster with encryption enabled and migrate your data into it.

Q: What compliance certifications does Amazon DocumentDB meet?

Amazon DocumentDB was designed to meet the highest security standards and to make it easy for you to verify our security and meet your own regulatory and compliance obligations. Amazon DocumentDB has been assessed to comply with PCI DSS, ISO 9001, 27001, 27017, and 27018, SOC 1, 2 and 3, and Health Information Trust Alliance (HITRUST) Common Security Framework (CSF) certification, in addition to being HIPAA eligible. AWS compliance reports are available for download in AWS Artifact.

Amazon DynamoDB FAQs

What is DynamoDB?

Q: What is Amazon DynamoDB? >>

Q: What does DynamoDB manage on my behalf? >>

Q: What is the consistency model of DynamoDB? >>

Getting started

Q: What kind of query functionality does DynamoDB support? >>

Q: How do I update and query data items with DynamoDB? >>

Q: Can DynamoDB be used by applications running on any operating system? >>

Planning

Q: How am I charged for my use of DynamoDB? >>

Q: What is the maximum throughput I can provision for a single DynamoDB table? >>

Q: What is the minimum throughput I can provision for a single DynamoDB table? >>

How it works

Q: Data models and APIs >>

Q: Scalability, availability, and durability >>

Q: Auto scaling >>

Q: Security and control >>

Amazon ElastiCache FAQs

General

Memcached

Redis

General

The Basics

Q: What is Amazon ElastiCache?

Amazon ElastiCache is a web service that makes it easy to deploy and run Memcached or Redis protocol-compliant server nodes in the cloud. Amazon ElastiCache improves the performance of web applications by allowing you to retrieve information from a fast, managed, in-memory system, instead of relying entirely on slower disk-based databases. The service simplifies and offloads the management, monitoring and operation of in-memory environments, enabling your engineering resources to focus on developing applications. Using Amazon ElastiCache, you can not only improve load and response times to user actions and queries, but also reduce the cost associated with scaling web applications.

Amazon ElastiCache automates common administrative tasks required to operate a distributed in-memory key-value environment. Using Amazon ElastiCache, you can add a caching or in-memory layer to your application architecture in a matter of minutes via a few clicks of the AWS Management Console. Once a cluster is provisioned, Amazon ElastiCache automatically detects and replaces failed nodes, providing a resilient system that mitigates the risk of overloaded databases, which slow website and application load times. Through integration with Amazon CloudWatch monitoring, Amazon ElastiCache provides enhanced visibility into key performance metrics associated with your nodes. Amazon ElastiCache is protocol-compliant with Memcached and Redis, so code, applications, and popular tools that you use today with your existing Memcached or Redis environments will work seamlessly with the service. With the support for clustered configuration in Amazon ElastiCache, you get the benefits of fast, scalable and easy to use managed service that can meet the needs of your most demanding applications. As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources you use.

Q: What is in-memory caching and how does it help my applications?

The in-memory caching provided by Amazon ElastiCache can be used to significantly improve latency and throughput for many read-heavy application workloads (such as social networking, gaming, media sharing and Q&A portals) or compute-intensive workloads (such as a recommendation engine). In-memory caching improves application performance by storing critical pieces of data in memory for low-latency access. Cached information may include the results of I/O-intensive database queries or the results of computationally-intensive calculations.

Q: Can I use Amazon ElastiCache for use cases other than caching?

A: Yes. ElastiCache for Redis can be used as a primary in-memory key-value data store, providing fast, sub millisecond data performance, high availability and scalability up to 250 nodes and 250 shards, giving you up to 170.6 TB of in-memory data. See here for other use cases, such as leaderboards, rate limiting, queues, and chat.

Q: Can I use Amazon ElastiCache through AWS CloudFormation?

AWS CloudFormation simplifies provisioning and management by providing AWS CloudFormation templates for quick and reliable provisioning of the services or applications. AWS CloudFormation provides comprehensive support for Amazon ElastiCache by providing templates to create cluster (both MemCached and Redis) and Replication Groups. The templates are up to date with the latest ElastiCache Redis announcement for clustered Redis configuration and provide flexibility and ease of use to Amazon ElastiCache customers.

Q: What does Amazon ElastiCache manage on my behalf?

Amazon ElastiCache manages the work involved in setting up a distributed in-memory environment, from provisioning the server resources you request to installing the software. Once your environment is up and running, the service automates common administrative tasks such as failure detection and recovery, and software patching. Amazon ElastiCache provides detailed monitoring metrics associated with your nodes, enabling you to diagnose and react to issues very quickly. For example, you can set up thresholds and receive alarms if one of your nodes is overloaded with requests.

Q: What are Amazon ElastiCache nodes, shards and clusters?

A node is the smallest building block of an Amazon ElastiCache deployment. It is a fixed-size chunk of secure, network-attached RAM. Each node runs an instance of the Memcached or Redis protocol-compliant service and has its own DNS name and port. Multiple types of nodes are supported, each with varying amount of associated memory. A Redis shard is a subset of the cluster’s keyspace, that can include a primary node and zero or more read-replicas. For more details on Redis deployments see the Redis section below. The shards add up to form a cluster.

Q: Which engines does Amazon ElastiCache support?

Amazon ElastiCache for Memcached currently supports Memcached 1.5.10, 1.4.34, 1.4.33, 1.4.24, 1.4.14, and 1.4.5.

Amazon ElastiCache for Redis currently supports Redis 5.0.6, 5.0.5, 5.0.4, 5.0.3, 5.0.0, 4.0.10, 3.2.10, 3.2.6, 3.2.4, 2.8.24, 2.8.23, 2.8.22, 2.8.21, 2.8.19, 2.8.6, and 2.6.13.

Q: How do I get started with Amazon ElastiCache?

If you are not already signed up for Amazon ElastiCache, you can click the "Get started" button on the Amazon ElastiCache page and complete the sign-up process. You must have an Amazon Web Services account; if you do not already have one, you will be prompted to create one when you begin the Amazon ElastiCache sign-up process. After you are signed up for ElastiCache, please refer to the Amazon ElastiCache documentation, which includes the Getting Started Guide for Amazon ElastiCache for Redis or Amazon ElastiCache for Memcached.

Once you have familiarized yourself with Amazon ElastiCache, you can launch a cluster within minutes by using the AWS Management Console or Amazon ElastiCache APIs.

Q: How do I create a cluster?

Clusters are simple to create, using the AWS Management Console, Amazon ElastiCache APIs, or Command Line Tools. To launch a cluster using the AWS Management Console, click on the "Create" button in either the “Memcached” or “Redis” tab. From there, all you need to specify is your Cluster Identifier, Node Type, and Number of Nodes to create a cluster with the amount of memory you require. Alternatively, you can create your cluster using the CreateCacheCluster API or elasticache -create-cache-cluster command. If you do not specify an Availability Zone when creating a cluster, AWS will place it automatically based upon your memory requirements and available capacity.

Q: What Node Types can I select?

Amazon ElastiCache supports Nodes of the following types:

Current Generation Nodes:

- cache.m4.large: 6.42 GB
- cache.m4.xlarge: 14.28 GB
- cache.m4.2xlarge: 29.7 GB
- cache.m4.4xlarge: 60.78 GB
- cache.m4.10xlarge: 154.64 GB
- cache.m5.large: 6.38 GB

- cache.m5.xlarge: 12.93 GB
- cache.m5.2xlarge: 26.04 GB
- cache.m5.4xlarge: 52.26 GB
- cache.m5.12xlarge: 157.12 GB
- cache.m5.24xlarge: 314.32 GB
- cache.r4.large: 12.3 GB
- cache.r4.xlarge: 25.05 GB
- cache.r4.2xlarge: 50.47 GB
- cache.r4.4xlarge: 101.38 GB
- cache.r4.8xlarge: 203.26 GB
- cache.r4.16xlarge: 407 GB
- cache.r5.large: 13.07 GB
- cache.r5.xlarge: 26.32 GB
- cache.r5.2xlarge: 52.82 GB
- cache.r5.4xlarge: 105.81 GB
- cache.r5.12xlarge: 317.77 GB
- cache.r5.24xlarge: 635.61 GB
- cache.t2.micro: 555 MB
- cache.t2.small: 1.55 GB
- cache.t2.medium: 3.22 GB

Previous Generation Nodes:

- cache.m1.small: 1.3 GB
- cache.m1.medium: 3.35 GB
- cache.m1.large: 7.1 GB

- cache.m1.xlarge: 14.6 GB
- cache.m2.xlarge: 16.7 GB
- cache.m2.2xlarge: 33.8 GB
- cache.m2.4xlarge: 68 GB
- cache.m3.medium: 2.78 GB
- cache.m3.large: 6.05 GB
- cache.m3.xlarge: 13.3 GB
- cache.m3.2xlarge: 27.9 GB
- cache.r3.large: 13.5 GB
- cache.r3.xlarge: 28.4 GB
- cache.r3.2xlarge: 58.2 GB
- cache.r3.4xlarge: 118 GB
- cache.r3.8xlarge: 237 GB
- cache.t1.micro: 213 MB
- cache.c1.xlarge: 6.6 GB

Each Node Type above lists the memory available to Memcached or Redis after taking Amazon ElastiCache System Software overhead into account. The total amount of memory in a cluster is an integer multiple of the memory available in each shard. For example, a cluster consisting of ten shards of 6 GB each will provide 60 GB of total memory.

Q: How do I access my nodes?

Once your cluster is available, you can retrieve your node endpoints using the following steps on the AWS Management Console:

- Navigate to the "Amazon ElastiCache" tab.
- Click on the "(Number of) Nodes" link and navigate to the "Nodes" tab.
- Click on the "Copy Node Endpoint(s)" button.

Alternatively, you can use the DescribeCacheClusters API to retrieve the Endpoint list.

You can then configure your Memcached or Redis client with this endpoint list and use your favorite programming language to add or delete data from your ElastiCache Nodes. In order to allow network requests to your nodes, you will need to authorize access. For a detailed explanation to get started, please refer to our Getting Started Guide for Amazon ElastiCache for Redis or Amazon ElastiCache for Memcached.

Q: What is a maintenance window? Will my nodes be available during software maintenance?

You can think of the Amazon ElastiCache maintenance window as an opportunity to control when software patching occurs, in the event either are requested or required. If a "maintenance" event is scheduled for a given week, it will be initiated and completed at some point during the 60 minute maintenance window you identify.

Your nodes could incur some downtime during your maintenance window if software patching is scheduled. Please refer to Engine Version Management for more details. Patching can be user requested - for example cache software upgrade, or determined as required (if we identify any security vulnerabilities in the system or caching software). Software patching occurs infrequently (typically once every few months) and should seldom require more than a fraction of your maintenance window. If you do not specify a preferred weekly maintenance window when creating your Cluster, a 60 minute default value is assigned. If you wish to modify when maintenance is performed on your behalf, you can do so by modifying your DB Instance in the AWS Management Console or by using the ModifyCacheCluster API. Each of your Clusters can have different preferred maintenance windows, if you so choose.

Billing

Q: How will I be charged and billed for my use of Amazon ElastiCache?

You pay only for what you use and there is no minimum fee. Pricing is per Node-hour consumed for each Node Type. Partial Node-hours consumed are billed as full hours. There is no charge for data transfer between Amazon EC2 and Amazon ElastiCache within the same Availability Zone. While standard Amazon EC2 Regional Data Transfer charges apply when transferring data between an Amazon EC2 instance and an Amazon ElastiCache Node in different Availability Zones of the same Region, you are only charged for the Data Transfer in or out of the Amazon EC2 instance. There is no Amazon ElastiCache Data Transfer charge for traffic in or out of the Amazon ElastiCache Node itself. For more information, please visit the pricing page.

Q: When does billing of my Amazon ElastiCache Nodes begin and end?

Billing commences for a node as soon as the node is available. Billing continues until the node is terminated, which would occur upon deletion.

Q: What defines billable ElastiCache Node hours?

Node hours are billed for any time your nodes are running in an "Available" state. If you no longer wish to be charged for your node, you must terminate it to avoid being billed for additional node hours.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Reserved Nodes

Q: What are Amazon ElastiCache Reserved Nodes?

Reserved Nodes or Reserved Instance (RI) is an offering that provides you with a significant discount over on-demand usage when you commit to a one-year or three-year term. With Reserved Nodes, you can make a one-time, up-front payment to create a one or three year reservation to run your node in a specific Region and receive a significant discount off of the ongoing hourly usage charge. There are three Reserved Node types All Upfront, No Upfront and Partial Upfront that enable you to balance the amount you pay upfront with your effective hourly price.

Q: How are Reserved Nodes different from On-Demand Nodes?

Functionally, Reserved Nodes and On-Demand Nodes are exactly the same. The only difference is how your Node(s) are billed; with Reserved Nodes, you make a one-time up-front payment and receive a lower ongoing hourly usage rate (compared with On-Demand Nodes) for the duration of the term.

Q: How do I purchase and create Reserved Nodes?

You can use the "Purchase Reserved Nodes" option in the AWS Management Console. Alternatively, you can use the API tools to list the reservations available for purchase with the DescribeReservedCacheNodesOfferings API method and then purchase a cache node reservation by calling the PurchaseReservedCacheNodesOffering method.

Creating a Reserved Node is no different than launching an On-Demand Node. You simply specify the node class and Region for which you made the reservation. So long as your reservation purchase was successful, Amazon ElastiCache will apply the reduced hourly rate for which you are eligible to the new node.

Q: Will there always be reservations available for purchase?

Yes. Reserved Nodes are purchased for the Region rather than for the Availability Zone. This means that even if capacity is limited in one Availability Zone, reservations can still be purchased in that Region and used in a different Availability Zone within that Region.

Q: How many Reserved Cache can I purchase?

You can purchase up to 300 Reserved Nodes. If you wish to run more than 300 Nodes please complete the Amazon ElastiCache Node request form.

Q: What if I have an existing node that I'd like to convert to a Reserved Node?

Simply purchase a node reservation with the same node class, within the same region as the node you are currently running and would like to reserve. If the reservation purchase is successful, Amazon ElastiCache will automatically apply your new hourly usage charge to your existing node.

Q: If I sign up for a Reserved Node, when does the term begin? What happens to my node when the term ends?

Pricing changes associated with a Reserved Node are activated once your request is received while the payment authorization is processed. You can follow the status of your reservation on the AWS Account Activity page or by using the DescribeReservedCacheNodes API. If the one-time payment cannot be successfully authorized by the next billing period, the discounted price will not take effect.

When your reservation term expires, your Reserved Node will revert to the appropriate On-Demand hourly usage rate for your node class and region.

Q: How do I control which nodes are billed at the Reserved Node rate?

The Amazon ElastiCache APIs for creating, modifying, and deleting nodes do not distinguish between On-Demand and Reserved Nodes so that you can seamlessly use both. When computing your bill, our system will automatically apply your Reservation(s), such that all eligible nodes are charged at the lower hourly Reserved Cache Node rate.

Q: Can I move a Reserved Node from one Region or Availability Zone to another?

Each Reserved Node is associated with a specific Region, which is fixed for the lifetime of the reservation and cannot be changed. Each reservation can, however, be used in any of the available AZs within the associated Region.

Q: Can I cancel a reservation?

No, you cannot cancel your reserved DB instance and the one-time payment (if applicable) is not refundable. You will continue to pay for every hour during your Reserved DB instance term regardless of your usage.

Q: How do the payment options impact my bill?

When you purchase an RI under the All Upfront payment option, you pay for the entire term of the RI in one upfront payment. You can choose to pay nothing upfront by choosing the No Upfront option. The entire value of the No Upfront RI is spread across every hour in the term and you will be billed for every hour in the term, regardless of usage. The Partial Upfront payment option is a hybrid of the All Upfront and No Upfront options. You make a small upfront payment, and you are billed a low hourly rate for every hour in the term regardless of usage.

Security

Q: How do I control access to Amazon ElastiCache?

When not using VPC, Amazon ElastiCache allows you to control access to your clusters through Cache Security Groups. A Security Group acts like a firewall, controlling network access to your cluster. By default, network access is turned off to your clusters. If you want your applications to access your cluster, you must explicitly enable access from hosts in specific EC2 security groups. This process is called ingress.

To allow network access to your cluster, create a Security Group and link the desired EC2 security groups (which in turn specify the EC2 instances allowed) to it. The Security Group can be associated with your cluster at the time of creation, or using the "Modify" option on the AWS Management Console.

Please note that IP-range based access control is currently not enabled for clusters. All clients to a cluster must be within the EC2 network, and authorized via security groups as described above.

When using VPC, please see here for more information.

Q: Can programs running on servers in my own data center access Amazon ElastiCache?

Yes. You can access an Amazon ElastiCache cluster from an application running in your data center providing there is connectivity between your VPC and the data center either through VPN or Direct Connect. The details are described here.

Q: Can programs running on EC2 instances in a VPC access Amazon ElastiCache?

Yes, EC2 instances in a VPC can access Amazon ElastiCache if the ElastiCache cluster was created within the VPC. Details on how to create an Amazon ElastiCache cluster within a VPC are given here.

Q: What is Amazon Virtual Private Cloud (VPC) and why may I want to use with Amazon ElastiCache?

Amazon VPC lets you create a virtual networking environment in a private, isolated section of the Amazon Web Services (AWS) cloud, where you can exercise complete control over aspects such as private IP address ranges, subnets, routing tables and network gateways. With Amazon VPC, you can define a virtual network topology and customize the network configuration to closely resemble a traditional IP network that you might operate in your own datacenter.

One of the scenarios where you may want to use Amazon ElastiCache in a VPC is if you want to run a public-facing web application, while still maintaining non-publicly accessible backend servers in a private subnet. You can create a public-facing subnet for your webservers that has access to the Internet, and place your backend infrastructure in a private-facing subnet with no Internet access. Your backend infrastructure could include RDS DB Instances and an Amazon ElastiCache Cluster providing the in-memory layer. For more information about Amazon VPC, refer to the Amazon Virtual Private Cloud User Guide.

Q: How do I create an Amazon ElastiCache Cluster in VPC?

For a walk through example of creating an Amazon ElastiCache Cluster in VPC, refer to the Amazon ElastiCache User Guide.

Following are the pre-requisites necessary to create a cluster within a VPC:

- You need to have a VPC set up with at least one subnet. For information on creating Amazon VPC and subnets refer to the Getting Started Guide for Amazon VPC.
- You need to have a Subnet Group (for Redis or Memcached) defined for your VPC.
- You need to have a VPC Security Group defined for your VPC (or you can use the default provided).
- In addition, you should allocate adequately large CIDR blocks to each of your subnets so that there are enough spare IP addresses for Amazon ElastiCache to use during maintenance activities such as cache node replacement.

Q: How do I create an Amazon ElastiCache Cluster in an existing VPC?

Creating an Amazon ElastiCache Cluster in an existing VPC is the same as that for a newly created VPC. Here are more details for Redis or Memcached.

Q: How do I connect to an ElastiCache Node in VPC?

Amazon ElastiCache Nodes, deployed within a VPC, can be accessed by EC2 Instances deployed in the same VPC. If these EC2 Instances are deployed in a public subnet with associated Elastic IPs, you can access the EC2 Instances via the internet.

If you want to access Amazon ElastiCache Nodes, deployed within a VPC, from the Internet or from EC2 Instances outside the VPC, please see guidelines for Redis or Memcached.

ElastiCache ensures that both the DNS name and the IP address of the cache node remain the same when cache nodes are recovered in case of failure.

Q: What is a Subnet Group and why do I need one?

A Subnet Group is a collection of subnets that you must designate for your Amazon ElastiCache Cluster in a VPC. A Subnet Group is created using the Amazon ElastiCache Console. Each Subnet Group should have at least one subnet. Amazon ElastiCache uses the Subnet Group to select a subnet. The IP Addresses from the selected subnet are then associated with the Node Endpoints. Furthermore, Amazon ElastiCache creates and associates Elastic Network Interfaces to nodes with the previously mentioned IP addresses.

Please note that, we strongly recommend you use the DNS Names to connect to your nodes as the underlying IP addresses can change (e.g., after cache node replacement).

Q: Can I change the Subnet Group of my ElastiCache Cluster?

An existing Subnet Group can be updated to add more subnets either for existing Availability Zones or for new Availability Zones added since the creation of the ElastiCache Cluster. However, changing the Subnet Group of a deployed cluster is not currently allowed.

Q: How is using Amazon ElastiCache inside a VPC different from using it outside?

The basic functionality of Amazon ElastiCache remains the same whether VPC is used or not. Amazon ElastiCache manages automatic failure detection, recovery, scaling, auto discovery, and software patching whether your ElastiCache Cluster is inside or outside a VPC.

Within a VPC, nodes of an ElastiCache cluster only have a private IP address (within a subnet that you define). Outside of a VPC, the access to the ElastiCache cluster can be controlled using Security Groups as described here.

Q: Can I move my existing ElastiCache Cluster from outside VPC into my VPC?

No, you cannot move an existing Amazon ElastiCache Cluster from outside VPC into a VPC. You will need to create a new Amazon ElastiCache Cluster inside the VPC.

Q: Can I move my existing ElastiCache Cluster from inside VPC to outside VPC?

Currently, direct migration of ElastiCache Cluster from inside to outside VPC is not supported. You will need to create a new Amazon ElastiCache Cluster outside VPC.

Q: How do I control network access to my cluster?

Amazon ElastiCache allows you to control access to your cluster and therefore the nodes using Security Groups in non-VPC deployments. A Security Group acts like a firewall controlling network access to your node. By default, network access is turned off to your nodes. If you want your applications to access your node, you can set your Security Group to allow access from EC2 Instances with specific EC2 Security Group membership or IP ranges. This process is called ingress. Once ingress is configured for a Security Group, the same rules apply to all nodes associated with that Security Group. Security Groups can be configured with the “Security Groups” section of the Amazon ElastiCache Console or using the Amazon ElastiCache APIs.

In VPC deployments, access to your nodes is controlled using the VPC Security Group and the Subnet Group. The VPC Security Group is the VPC equivalent of the Security Group.

Q: What precautions should I take to ensure that my ElastiCache Nodes in VPC are accessible by my application?

You are responsible for modifying routing tables and networking ACLs in your VPC to ensure that your ElastiCache Nodes are reachable from your client instances in the VPC. To learn more see the Amazon ElastiCache for Redis or Amazon ElastiCache for Memcached Documentation.

Q: Can I use Security Groups to configure the clusters that are part of my VPC?

No, Security Groups are not used when operating in a VPC. Instead they are used in the non VPC settings. When creating a cluster in a VPC you will need to use VPC Security Groups.

Q: Can I associate a regular EC2 security group with a cluster that is launched within a VPC?

No, you can only associate VPC security groups that are part of the same VPC as your cluster.

Q: Can nodes of an ElastiCache cluster span multiple subnets?

Yes, nodes of an Amazon ElastiCache cluster can span multiple subnets as long as the subnets are part of the same Subnet Group that was associated with the ElastiCache Cluster at creation time.

Parameter Groups

Q: What are Parameter Groups? How are they helpful?

A Parameter Group acts as a "container" for engine configuration values that can be applied to one or more clusters. If you create a cluster without specifying a Parameter Group, a default Parameter Group is used. This default group contains engine defaults and Amazon ElastiCache system defaults optimized for the cluster you are running. However, if you want your cluster to run with your custom-specified engine configuration values, you can simply create a new Parameter Group, modify the desired parameters, and modify the cluster to use the new Parameter Group. Once associated, all clusters that use a particular Parameter Group get all the parameter updates to that Parameter Group. For more information on configuring Parameter Groups, please refer to the Amazon ElastiCache for Redis or Amazon ElastiCache for Memcached User Guide.

Q: How do I choose the right configuration parameters for my Cluster(s)?

Amazon ElastiCache by default chooses the optimal configuration parameters for your cluster taking into account the Node Type's memory/compute resource capacity. However, if you want to change them, you can do so using our configuration management APIs. Please note that changing configuration parameters from recommended values can have unintended effects, ranging from degraded performance to system crashes, and should only be attempted by advanced users who wish to assume these risks. For more information on changing parameters, please refer to the Amazon ElastiCache User Guide.

Q: How do I see the current setting for my parameters for a given Parameter Group?

You can use the AWS Management Console, Amazon ElastiCache APIs, or Command Line Tools to see information about your Parameter Groups and their corresponding parameter settings.

Memcached

Features

Q: What can I cache using Amazon ElastiCache for Memcached?

You can cache a variety of objects using the service, from the content in persistent data stores (such as Amazon RDS, DynamoDB, or self-managed databases hosted on EC2) to dynamically generated web pages (with Nginx for example), or transient session data that may not require a persistent backing store. You can also use it to implement high-frequency counters to deploy admission control in high volume web applications.

Q: Can I use Amazon ElastiCache for Memcached with an AWS persistent data store such as Amazon RDS or Amazon DynamoDB?

Yes, Amazon ElastiCache is an ideal front-end for data stores like Amazon RDS or Amazon DynamoDB, providing a high-performance middle tier for applications with extremely high request rates and/or low latency requirements.

Q: I use Memcached today. How do I migrate to Amazon ElastiCache?

Amazon ElastiCache is protocol-compliant with Memcached. Therefore, you can use standard Memcached operations like get, set, incr and decr in exactly the same way as you would in your existing Memcached deployments. Amazon ElastiCache supports both the text and binary protocols. It also supports most of the standard stats results, which can also be viewed as graphs via CloudWatch. As a result, you can switch to using Amazon ElastiCache without recompiling or re-linking your applications - the libraries you use will continue to work. To configure the cache servers your application accesses, all you will need to do is to update your application's Memcached config file to include the endpoints of the servers (nodes) we provision for you. You can simply use the "Copy Node Endpoints" option on the AWS Management Console or the "DescribeCacheClusters" API to get a list of the endpoints. As with any migration process, we recommend thorough testing of your new Amazon ElastiCache deployment before completing the cut over from your current solution.

Please note that Amazon ElastiCache currently allows access only from the Amazon EC2 network, so in order to use the service, you should have your application servers in Amazon EC2.

Amazon ElastiCache uses DNS entries to allow client applications to locate servers (nodes). The DNS name for a node remains constant, but the IP address of a node can change over time, for example, when nodes are auto replaced after a failure on a non-VPC installation. See this FAQ for recommendations to deal with node failures.

Configuration and Scaling

Q: How do I select an appropriate Node Type for my application?

Though there is no precise answer for this question, with Amazon ElastiCache, you don't need to worry about getting the number of nodes exactly right, as you can very easily add or remove nodes later. The following two inter-related aspects could be considered for the choice of your initial configuration:

- The total memory required for your data to achieve your target cache-hit rate, and
- The number of nodes required to maintaining acceptable application performance without overloading the database backend in the event of node failure(s).

The amount of memory required is dependent upon the size of your data set and the access patterns of your application. To improve fault tolerance, once you have a rough idea of the total memory required, divide that memory into enough nodes such that your application can survive the loss of one or two nodes. For example, if your memory requirement is 13GB, you may want to use two cache.m4.large nodes instead of using one cache.m4.xlarge node. It is important that other systems such as databases will not be overloaded if the cache-hit rate is temporarily reduced during failure recovery of one or more of nodes. Please refer to the Amazon ElastiCache User Guide for more details.

Q: Can a cluster span multiple Availability Zones?

Yes. When creating a cluster or adding nodes to an existing cluster, you can chose the availability zones for the new nodes. You can either specify the requested amount of nodes in each availability zones or select “spread nodes across zones”. If the cluster is in VPC, nodes can only be placed in availability zones that are part of the selected cache subnet group. For additional details please see ElastiCache VPC documentation.

Q: How many nodes can I run per region in Amazon ElastiCache Memcached?

You can run a maximum of 100 nodes per region. If you need more nodes, please fill in the ElastiCache Limit Increase Request form.

Q: How does Amazon ElastiCache respond to node failure?

The service will detect the node failure and react with the following automatic steps:

- Amazon ElastiCache will repair the node by acquiring new service resources, and will then redirect the node's existing DNS name to point to the new service resources. For VPC installations, ElastiCache will ensure that both the DNS name and the IP address of the node remain the same when nodes are recovered in case of failure. For non-VPC installations, ElastiCache will ensure that the DNS name of a node is unchanged; however, the underlying IP address of the node can change.

- If you associated an SNS topic with your cluster, when the new node is configured and ready to be used, Amazon ElastiCache will send an SNS notification to let you know that node recovery occurred. This allows you to optionally arrange for your applications to force the Memcached client library to attempt to reconnect to the repaired nodes. This may be important, as some Memcached libraries will stop using a server (node) indefinitely if they encounter communication errors or timeouts with that server.

Q: If I determine that I need more memory to support my application, how do I increase the total memory with Amazon ElastiCache?

You could add more nodes to your existing Memcached Cluster by using the "Add Node" option on "Nodes" tab for your Cache Cluster on the AWS Management Console or calling the ModifyCacheCluster API.

Compatibility

Q: How does Amazon ElastiCache interact with other Amazon Web Services?

Amazon ElastiCache is ideally suited as a front-end for Amazon Web Services like Amazon RDS and Amazon DynamoDB, providing extremely low latency for high performance applications and offloading some of the request volume while these services provide long lasting data durability. The service can also be used to improve application performance in conjunction with Amazon EC2 and EMR.

Q: Is Amazon ElastiCache better suited to any specific programming language?

Memcached client libraries are available for many, if not all of the popular programming languages. For more information on Memcached clients, please see this. If you encounter any issues with specific Memcached clients when using Amazon ElastiCache, please engage us via the Amazon ElastiCache community forum.

Q: What popular Memcached libraries are compatible with Amazon ElastiCache?

Amazon ElastiCache does not require specific client libraries and works with existing Memcached client libraries without recompilation or application re-linking (Memcached 1.4.5 and later); examples include libMemcached (C) and libraries based on it (e.g. PHP, Perl, Python), spyMemcached (Java) and fauna (Ruby).

Auto Discovery

Q: What is Auto Discovery and what can I do with it?

Auto Discovery is a feature that saves developers time and effort, while reducing complexity of their applications. Auto Discovery enables automatic discovery of cache nodes by clients when they are added to or removed from an Amazon ElastiCache cluster. Until now to handle cluster membership changes, developers must update the list of cache node endpoints manually. Depending on how the client application is architected, typically a client initialization, by shutting down the application and restarting it, is needed resulting in downtime. Through Auto Discovery we are eliminating this complexity. With Auto Discovery, in addition to being backwards protocol-compliant with the Memcached protocol, Amazon ElastiCache provides clients with information on cache cluster membership. A client capable of processing the additional information reconfigures itself, without any initialization, to use the most current nodes of an Amazon ElastiCache cluster.

Q: How does Auto Discovery work?

An Amazon ElastiCache cluster can be created with nodes that are addressable via named endpoints. With Auto Discovery the Amazon ElastiCache cluster is also given a unique Configuration Endpoint which is a DNS Record that is valid for the lifetime of the cluster. This DNS Record contains the DNS Names of the nodes that belong to the cluster. Amazon ElastiCache will ensure that the Configuration Endpoint always points to at least one such “target” node. A query to the target node then returns endpoints for all the nodes of the cluster in question. After this, you can connect to the cluster nodes just as before and use the Memcached protocol commands such as get, set, incr and decr. For more details, see here. To use Auto Discovery, you will need an Auto Discovery capable client. Auto Discovery clients for Java and PHP are available for download from the Amazon ElastiCache console. Upon initialization, the client will automatically determine the current members of the Amazon ElastiCache cluster using the Configuration Endpoint. When you make changes to your cache cluster by adding or removing nodes or if a node is replaced upon failure, the Auto Discovery client automatically determines the changes and you do not need to initialize your clients manually.

Q: How can I get started using Auto Discovery?

To get started, download the Amazon ElastiCache Cluster Client by clicking the “Download ElastiCache Cluster Client” link on the Amazon ElastiCache console. Before you can download, you must have an Amazon ElastiCache account; if you do not already have one, you can sign up from the Amazon ElastiCache detail page. After you download the client, you can begin setting up and activating your Amazon ElastiCache cluster by visiting the Amazon ElastiCache console. More details can be found here.

Q: If I continue to use my own Memcached clients with my ElastiCache cluster – will I be able to get this feature?

No, you will not get the Auto Discovery feature with the existing Memcached clients. To use the Auto Discovery feature a client must be able to use a Configuration Endpoint and determine the cluster node endpoints. You may either use the Amazon ElastiCache Cluster Client or extend your existing Memcached client to include the Auto Discovery command set.

Q: What are the minimum hardware / software requirements for Auto Discovery?

To take advantage of Auto Discovery, an Auto Discovery capable client must be used to connect to an Amazon ElastiCache Cluster. Amazon ElastiCache currently supports Auto Discovery capable clients for both Java and PHP. These can be downloaded from the Amazon ElastiCache console. Our customers can create clients for any other language by building upon the popular Memcached clients available.

Q: How do I modify or write my own Memcached client to support auto-discovery?

You can take any Memcached Client Library and add support for Auto Discovery. If you would like to add or modify your own client to enable Auto Discovery, please refer to the Auto Discovery command set documentation.

Q: Can I continue to work with my existing Memcached client if I don’t need Auto-discovery?

Yes, Amazon ElastiCache is still Memcached protocol compliant and does not require you to change your clients. However, for taking advantage of auto-discovery feature, we had to enhance the Memcached client capabilities. If you choose to not use the Amazon ElastiCache Cluster Client, you can continue to use your own clients or modify your own client library to understand the auto-discovery command set.

Q: Can I have heterogeneous clients when using Auto Discovery?

Yes, the same Amazon ElastiCache cluster can be connected through an Auto Discovery capable Client and the traditional Memcached client at the same time. Amazon ElastiCache remains 100% Memcached compliant.

Q: Can I stop using Auto Discovery?

Yes, you can stop using Auto Discovery anytime. You can disable Auto Discovery by specifying the mode of operation during the Amazon ElastiCache Cluster client initialization. Also, since Amazon ElastiCache continues to support Memcached 100% you may use any Memcached protocol-compliant client as before.

Engine Version Management

Q: Can I control if and when the engine version powering Amazon ElastiCache Cluster is upgraded to new supported versions?

Amazon ElastiCache allows you to control if and when the Memcached protocol-compliant software powering your cluster is upgraded to new versions supported by Amazon ElastiCache. This provides you with the flexibility to maintain compatibility with specific Memcached versions, test new versions with your application before deploying in production, and perform version upgrades on your own terms and timelines. Version upgrades involve some compatibility risk, thus they will not occur automatically and must be initiated by you. This approach to software patching puts you in the driver's seat of version upgrades, but still offloads the work of patch application to Amazon ElastiCache. You can learn more about version management by reading the FAQs that follow. Alternatively, you can refer to the Amazon ElastiCache User Guide. While Engine Version Management functionality is intended to give you as much control as possible over how patching occurs, we may patch your cluster on your behalf if we determine there is any security vulnerability in the system or cache software.

Q: How do I specify which supported Memcached Version my Cluster should run?

You can specify any currently supported version (minor and/or major) when creating a new cluster. If you wish to initiate an upgrade to a supported engine version release, you can do so using the "Modify" option for your cluster. Simply specify the version you wish to upgrade to via the "Cache Engine Version" field. The upgrade will then be applied on your behalf either immediately (if the "Applied Immediately" option is checked) or during the next scheduled maintenance window for your cluster.

Q: Can I test my cluster against a new version before upgrading?

Yes. You can do so by creating a new cluster with the new engine version. You can point your development/staging application to this cluster, test it and decide whether or not to upgrade your original cluster.

Q: Does Amazon ElastiCache provide guidelines for supporting new Memcached version releases and/or deprecating versions that are currently supported?

Over time, we plan to support additional Memcached versions for Amazon ElastiCache, both major and minor. The number of new version releases supported in a given year will vary based on the frequency and content of the Memcached version releases and the outcome of a thorough vetting of the release by our engineering team. However, as a general guidance, we aim to support new Memcached versions within 3-5 months of their General Availability release.

Q: Which version of the Memcached wire protocol does Amazon ElastiCache support?

Amazon ElastiCache supports the Memcached text and binary protocol of versions 1.5.10, 1.4.34, 1.4.33, 1.4.24, 1.4.14, and 1.4.5 of Memcached.

Q: What should I do to upgrade to the latest Memcached version?

You can upgrade your existing Memcached cluster by using the Modify process. When upgrading from an older version of Memcached to Memcached version 1.4.33 or newer, please ensure that your existing parameter max_chunk_size values satisfies conditions needed for slab_chunk_max parameter. Please review upgrade prerequisites here.

Redis

Features

Q: What is Amazon ElastiCache for Redis?

Amazon ElastiCache for Redis is a web service that makes it easy to deploy and run Redis protocol-compliant server nodes in the cloud. The service enables the management, monitoring and operation of a Redis node; creation, deletion and modification of the node can be carried out through the ElastiCache console, the command line interface or the web service APIs. Amazon ElastiCache for Redis supports Redis Master / Slave replication.

Q: Is Amazon ElastiCache for Redis protocol-compliant with open source Redis?

Yes, Amazon ElastiCache for Redis is protocol-compliant with open source Redis. Code, applications, drivers and tools a customer uses today with their existing standalone Redis data store will continue to work with ElastiCache for Redis and no code changes will be required for existing Redis deployments migrating to ElastiCache for Redis unless noted.

We currently support Redis 5.0.6, 5.0.5, 5.0.4, 5.0.3, 5.0.0, 4.0.10, 3.2.10, 3.2.6, 3.2.4, 2.8.24, 2.8.23, 2.8.22, 2.8.21, 2.8.19, 2.8.6, and 2.6.13.

Q: What are Amazon ElastiCache for Redis nodes and shards?

An Amazon ElastiCache node is the smallest building block of an ElastiCache for Redis Cluster deployment. Each node supports the Redis protocol with Amazon’s enhancements and has its own endpoint and port. Multiple types of nodes are supported, each with varying amount of CPU capability, and memory capacity.

A shard is a collection of one or more nodes that is responsible for a partition of the logical key space. Within a shard, a node may exist in isolation or in a primary/replica relationship with other nodes. If there are multiple nodes within a shard, one of the nodes will take on the read/write primary role and all other nodes will take on a read-only replica role.

Q: Does Amazon ElastiCache for Redis support Redis persistence?

Yes, you can achieve persistence by snapshotting your Redis data using the Backup and Restore feature. Please see here for details.

Q: How can I migrate from Amazon ElastiCache for Memcached to Amazon ElastiCache for Redis and vice versa?

We currently do not support automatically migrating from Memcached to Redis or vice versa. You may, however, use a Memcached client to read from a Memcached cluster and use a Redis client to write to a Redis cluster. Similarly, you may read from a Redis cluster using a Redis client and use a Memcached client to write to a Memcached cluster. Make sure to consider the differences in data format, and cluster configuration between the two engines.

Q: Does Amazon ElastiCache for Redis support Multi-AZ operation?

Yes, with Amazon ElastiCache for Redis you can create a read replica in another AWS Availability Zone. Upon a failure of the primary node, we will provision a new primary node. In scenarios where the primary node cannot be provisioned, you can decide which read replica to promote to be the new primary. For more details on how to handle node failures see here.

Q: What options does Amazon ElastiCache for Redis provide for node failures?

Amazon ElastiCache for Redis will repair the node by acquiring new service resources, and will then redirect the node's existing DNS name to point to the new service resources. Thus, the DNS name for a Redis node remains constant, but the IP address of a Redis node can change over time. If you have a replication group with one or more read replicas and Multi-AZ is enabled, then in case of primary node failure ElastiCache will automatically detect the failure, select a replica and promote it to become the new primary. It will also propagate the DNS so that you can continue to use the primary endpoint and after the promotion it will point to the newly promoted primary. For more details see the Multi-AZ section of this FAQ. When Redis replication option is selected with Multi-AZ disabled, in case of primary node failure you will be given the option to initiate a failover to a read replica node. The failover target can be in the same zone or another zone. To failback to the original zone, promote the read replica in the original zone to be the primary. You may choose to architect your application to force the Redis client library to reconnect to the repaired Redis server node. This can help as some Redis libraries will stop using a server indefinitely when they encounter communication errors or timeouts.

Q: How does failover work?

For Multi-AZ enabled replication groups, the failover behavior is described at the Multi-AZ section of this FAQ.

If you choose not to enable Multi-AZ, then if Amazon ElastiCache monitors the primary node, and in case the node becomes unavailable or unresponsive, Amazon ElastiCache for Redis will repair the node by acquiring new service resources, and will then redirect the node's existing DNS name to point to the new service resources. Thus, the DNS name for a Redis node remains constant, but the IP address of a Redis node can change over time. However, if the primary node cannot be healed (and your Multi-AZ is disabled) you will have the choice to promote one of the read replicas to be the new primary. See here for how to select a new primary. The DNS record of the primary's endpoint will be updated to point to the promoted read replica node. A read replica node in the original primary's AZ will then be created to be a read replica in the shard and will follow the new primary.

Q: Are my read replicas available during a primary node failure?

Yes, during a primary node failure, the read replicas continue to service requests. After the primary node is restored, either as a healed node or as a promoted read replica, there is a brief period during which the read replicas will not serve any requests as they sync the cache information from the primary.

Q: How do I configure parameters of my Amazon ElastiCache for Redis nodes?

You can configure your Redis installation using a parameter group, which must be specified for a Redis cluster. All read replica clusters use the parameter group of their primary cluster. A Redis parameter group acts as a “container” for Redis configuration values that can be applied to one or more Redis primary clusters. If you create a Redis primary cluster without specifying a parameter group, a default parameter group is used. This default group contains defaults for the node type you plan to run. However, if you want your Redis primary cluster to run with specified configuration values, you can simply create a new cache parameter group, modify the desired parameters, and modify the primary Redis cluster to use the new parameter group.

Q: Can I access Redis through the Amazon ElastiCache console?

Yes, Redis appears as an Engine option in the ElastiCache console. You can create a new Redis cache cluster with the Launch Wizard by choosing the Redis engine. You can also modify or delete an existing Redis cluster using the ElastiCache console.

Q: Can Amazon ElastiCache for Redis clusters be created in an Amazon VPC?

Yes, just as you can create Memcached clusters within a VPC, you can create Redis clusters within a VPC as well. If your account is a VPC by default account, your Redis clusters will be created within the default VPC associated with your account. Using the ElastiCache console, you can specify a different VPC when you create your cluster.

Q. Is Redis AUTH functionality supported in Amazon ElastiCache for Redis?

Yes, Redis AUTH functionality is available on Amazon ElastiCache for Redis. At the time of Redis cluster creation via the console or command line interface, once you enable encryption in-transit, you can use the Redis AUTH command to provide an authentication token for communication with the Redis cluster.

Q. How do I upgrade to a newer engine version?

You can easily upgrade to a newer engine version by using the ModifyCacheCluster or ModifyReplicationGroup APIs and specifying your preferred engine version for the EngineVersion parameter. On the ElastiCache console, you can select a cluster and click “Modify”. In the “Modify” window select your preferred engine version from the available options. If you are upgrading to Redis version 5.0.5 or higher, and using Redis Cluster configurations with Redis Cluster client or using non-Cluster configuration with auto-failover enabled, then the engine version upgrade completes while the cluster stays online and continues serving incoming requests.The engine upgrade process is designed to make a best effort to retain your existing data and requires Redis replication to succeed. For more details on that see here.

Q. Can I downgrade to an earlier engine version?

No. Downgrading to an earlier engine version is not supported.

Q. How do I scale up to a larger node type?

You can easily scale up to a larger node type by using the ModifyCacheCluster or ModifyReplicationGroup APIs and specifying your preferred node type for the CacheNodeType parameter. On the ElastiCache console, you can select a cache cluster or replication group and click “Modify”. In the “Modify” window select your preferred node type from the available options.

Amazon ElastiCache offers online vertical scaling for Redis Cluster mode as well as non-Redis Cluster mode with auto-failover (on Redis 5.0.5 onwards), allowing you to change your node type while the cluster continues to stay online and serve incoming requests. For the non-Redis Cluster mode on versions prior to Redis 5.0.5, you may notice a brief interruption of a few seconds (associated with DNS updates). The scale up process is designed to make a best effort to retain your existing data and requires Redis replication to succeed. For more details on that see here.

Q. Can I scale down to a smaller node type?

Yes, you can modify your cluster to move to a smaller node type and scale down. You can initiate the scale down in the same manner as scale up. When scaling down, please ensure that you are selecting a node that offers sufficient memory for your application needs. For Redis Cluster configurations and for Redis 5.0.5 and above non-Cluster configurations with autofailover enabled, the scaling completes while the cluster continues to stay online and serve incoming requests.

Q. What is the correct metric to use to measure Redis CPU utilization?

Amazon ElastiCache provides two metrics to measure CPU utilization for ElastiCache for Redis workloads – EngineCPUUtilization and CPUUtilization. The CPUUtilization metric measures the CPU utilization for the instance (node), and EngineCPUUtilization metric measures the utilization at the Redis process level. You need the EngineCPUUtilization metric in addition to the CPUUtilization metric as the main Redis process is single threaded and uses just one CPU of the multiple CPU cores available on an instance. Therefore, the CPUUtilization metric does not provide precise visibility into the CPU utilization rates at the Redis process level. We recommend that you use both the CPUUtilization and EngineCPUUtilization metrics together to get a detailed understanding of CPU Utilization for your Redis clusters. Both the metrics are available in all AWS regions, and you can access these metric using CloudWatch or via the AWS Management Console.

Read Replica

Q: What does it mean to run a Redis node as a Read Replica?

Read Replicas serve two purposes in Redis:

- Failure Handing
- Read Scaling

When you run a node with a Read Replica, the “primary” serves both writes and reads. The Read Replica acts as a “standby” which is “promoted” in failover scenarios. After failover, the standby becomes the primary and accepts your cache operations. Read Replicas also make it easy to elastically scale out beyond the capacity constraints of a single node for read-heavy cache workloads.

Q: When would I want to consider using a Redis read replica?

There are a variety of scenarios where deploying one or more read replicas for a given primary node may make sense. Common reasons for deploying a read replica include:

- Scaling beyond the compute or I/O capacity of a single primary node for read-heavy workloads. This excess read traffic can be directed to one or more read replicas.
- Serving read traffic while the primary is unavailable. If your primary node cannot take I/O requests (e.g. due to I/O suspension for backups or scheduled maintenance), you can direct read traffic to your read replicas. For this use case, keep in mind that the data on the read replica may be “stale” since the primary Instance is unavailable. The read replica can also be used to restart a failed primary warmed up.
- Data protection scenarios; in the unlikely event or primary node failure or that the Availability Zone in which your primary node resides becomes unavailable, you can promote a read replica in a different Availability Zone to become the new primary.

Q: How do I deploy a read replica node for a given primary node?

You can create a read replica in minutes using a CreateReplicationGroup API or a few clicks of the Amazon ElastiCache Management Console. When creating a cluster, you specify the MasterCacheClusterIdentifier. The MasterCacheClusterIdentifier is the cache cluster Identifier of the “primary” node from which you wish to replicate. You then create the read replica cluster within the shard by calling the CreateCacheCluster API specifying the ReplicationGroupIdentifier and the CacheClusterIdentifier of the master node. As with a standard cluster, you can also specify the Availability Zone. When you initiate the creation of a read replica, Amazon ElastiCache takes a snapshot of your primary node in a shard and begins replication. As a result, you will experience a brief I/O suspension on your primary node as the snapshot occurs. The I/O suspension typically lasts on the order of one minute.

The read replicas are as easy to delete as they are to create; simply use the Amazon ElastiCache Management Console or call the DeleteCacheCluster API (specifying the CacheClusterIdentifier for the read replica you wish to delete).

Q: Can I create both a primary and read replicas at the same time?

Yes. You can create a new cache cluster along with read replicas in minutes using the CreateReplicationGroup API or using the “Create” wizard at the Amazon ElastiCache Management Console and selecting “Multi-AZ Replication”. When creating the cluster, specify an identifier, the total number of desired shard in a cluster a read replicas per shard, along with cahe creation parameters such as node type, engine version, etc. You can also specify the Availability Zone for each shard in the cluster.

Q: How do I connect to my read replica(s)?

You can connect to a read replica just as you would connect to a primary cache node, using the DescribeCacheClusters API or AWS Management Console to retrieve the endpoint(s) for you read replica(s). If you have multiple read replicas, it is up to your application to determine how read traffic will be distributed amongst them. Here are more details:

- Redis (cluster mode disabled) clusters, either use Reader Endpoint or use the individual Node Endpoints for read operations (In the API/CLI these are referred to as Read Endpoints).
- Redis (cluster mode enabled) clusters, use the cluster's Configuration Endpoint for all operations. You must use a client that supports Redis Cluster (Redis 3.2). You can still read from individual node endpoints (In the API/CLI these are referred to as Read Endpoints).

Q: How many read replicas can I create for a given primary node?

At this time, Amazon ElastiCache allows you to create up to five (5) read replicas for a given primary node.

Q: What happens to read replicas if failover occurs?

In the event of a failover, any associated and available read replicas should automatically resume replication once failover has completed (acquiring updates from the newly promoted read replica).

Q: Can I create a read replica of another read replica?

Creating a read replica of another read replica is not supported.

Q: Can I promote my read replica into a “standalone” primary node?

No, this is not supported. Instead, you may snapshot your ElastiCache for Redis node (you may select the primary or any of the read-replicas). You can then use the snapshot to seed a new ElastiCache for Redis primary.

Q: Will my read replica be kept up-to-date with its primary node?

Updates to a primary node will automatically be replicated to any associated read replicas. However, with Redis’s asynchronous replication technology, a read replica can fall behind its primary cache node for a variety of reasons. Typical reasons include:

- Write I/O volume to the primary cache node exceeds the rate at which changes can be applied to the read replica
- Network partitions or latency between the primary cache node and a read replica

Read replicas are subject to the strengths and weaknesses of Redis replication. If you are using read replicas, you should be aware of the potential for lag between a read replica and its primary cache node, or “inconsistency”. You can monitor such lag potentially occuring via the "Replication Lag" CloudWatch metric, accessible through both the ElastiCache console and API, as well as those of the CloudWatch service.

Q: How do I gain visibility into active read replica(s)?

You can use the standard DescribeCacheClusters API to return a list of all the cache clusters you have deployed (including read replicas), or simply click on the "Redis" tab of the Amazon ElastiCache Management Console.

Amazon ElastiCache monitors the replication status of your read replicas and updates the Replication State field to Error if replication stops for any reason. You can review the details of the associated error thrown by the Redis engine by viewing the Replication Error field and take an appropriate action to recover from it. You can learn more about troubleshooting replication issues in the Troubleshooting a Read Replica problem section of the Amazon ElastiCache User Guide. If a replication error is fixed, the Replication State changes to Replicating.

Amazon ElastiCache allows you to gain visibility into how far a read replica has fallen behind its primary through the Amazon CloudWatch metric ("Replica Lag") available via the AWS Management Console or Amazon CloudWatch APIs.

Q: My read replica has fallen significantly behind its primary node. What should I do?

As discussed in the previous questions, “inconsistency” or lag between a read replica and its primary node is common with Redis asynchronous replication. If an existing read replica has fallen too far behind to meet your requirements, you can reboot it. Keep in mind that replica lag may naturally grow and shrink over time, depending on your primary node’s steady-state usage pattern.

Q: How do I delete a read replica? Will it be deleted automatically if its primary node is deleted?

You can easily delete a read replica with a few clicks of the AWS Management Console or by using DeleteCacheCluster, or DecreaseReplicaCount APIs. If you want to delete the read replica in addition to the primary cache node, you must use the DeleteReplicationGroup API or AWS Management Console.

Q: How much do read replicas cost? When does billing begin and end?

A read replica is billed as a standard node and at the same rates. Just like a standard node, the rate per “Node hour” for a read replica is determined by the node class of the read replica – please see Amazon ElastiCache detail page for up-to-date pricing. You are not charged for the data transfer incurred in replicating data between your primary cache node and read replica. Billing for a read replica begins as soon as the read replica has been successfully created (i.e. when status is listed as “active”). The read replica will continue being billed at standard Amazon ElastiCache cache node hour rates until you issue a command to delete it.

Q: What happens during failover and how long does it take?

Initiated failover is supported by Amazon ElastiCache so that you can resume operations as quickly as possible. When failing over, Amazon ElastiCache simply flips the DNS record for your node to point at the read replica, which is in turn promoted to become the new primary. We encourage you to follow best practices and implement cache node connection retry at the application layer. Start-to-finish, failover typically completes within three to six minutes.

Q: Can I create a read replica in another region as my primary?

No. Your read replica may only be provisioned in the same or different Availability Zone of the same Region as your cache node primary.

Q: Can I see which Availability Zone my primary is currently located in?

Yes, you can gain visibility into the location of the current primary by using the AWS Management Console or DescribeCacheClusters API.

After failover, my primary is now located in a different Availability Zone than my other AWS resources (e.g. EC2 instances).

Q: Should I be concerned about latency?

Availability Zones are engineered to provide low latency network connectivity to other Availability Zones in the same Region. In addition, you may want to consider architecting your application and other AWS resources with redundancy across multiple Availability Zones so your application will be resilient in the event of an Availability Zone failure.

Q: Can I add and remove read replica nodes for my Redis Cluster environment?

Yes. You can add a remove replica across one or more shards in a Redis Cluster environment. The cluster continues to stay online and serve incoming I/O during this operation.

Q: Is reader endpoint available for ElastiCache for Redis (cluster mode enabled) clusters?

No. Reader endpoint is only available for ElastiCache for Redis (cluster mode disabled) clusters.

Q: If I have a single node in my Redis (cluster mode disabled) cluster, will my reader endpoint point to the master node?

Yes, your cluster's reader endpoint will point to the master node. When you add a new read replica to your cluster, the reader endpoint will be updated to point to the read replica instead.

Q: Can I perform writes using my reader endpoint?

No, you cannot perform writes using a reader endpoint.

Q: Will my existing connections to a read replica that is promoted to a master node after a failover get dropped if I'm connected to a reader endpoint?

No. During new master node election at failovers, existing connections to a newly promoted master node are not automatically dropped. New connections to the new master node will not be established when you are using a reader endpoint.

Multi-AZ

Q: What is Multi-AZ for ElastiCache for Redis?

An ElastiCache for Redis shard consists of a primary and up to five read replicas. Redis asynchronously replicates the data from the primary to the read replicas. During certain types of planned maintenance, or in the unlikely event of ElastiCache node failure or Availability Zone failure, Amazon ElastiCache will automatically detect the failure of a primary, select a read replica, and promote it to become the new primary. ElastiCache also propagates the DNS changes of the promoted read replica, so if your application is writing to the primary node endpoint, no endpoint change will be needed.

Q: What are the benefits of using Multi-AZ?

The main benefits of running your ElastiCache for Redis in Multi-AZ mode are enhanced availability and smaller need for administration. If an ElastiCache for Redis primary node failure occurs, the impact on your ability to read/write to the primary is limited to the time it takes for automatic failover to complete. When Multi-AZ is enabled, ElastiCache node failover is automatic and requires no administration. You no longer need to monitor your Redis nodes and manually initiate a recovery in the event of a primary node disruption.

Q: How does Multi-AZ work?

You can use Multi-AZ if you are using ElastiCache for Redis and have a shard consisting of a primary node and one or more read replicas. If the primary node fails, ElastiCache will automatically detect the failure, select one from the available read replicas, and promote it to become the new primary. When cluster_mode parameter is disabled, ElastiCache will propagate the DNS changes of the promoted replica so that your application can keep writing to the primary endpoint. For cluster_mode enabled, ElastiCache will update the node map of the cluster. ElastiCache will also spin up a new node to replace the promoted read replica in the same Availability Zone of the failed primary. In case the primary failed due to temporary Availability Zone disruption, the new replica will be launched once that Availability Zone has recovered.

Q: Can I have replicas in the same Availability Zone as the primary?

Yes. Note that placing both the primary and the replica(s) in the same Availability Zone will not make your ElastiCache for Redis replication group resilient to an Availability Zone disruption.

Q: What events would cause Amazon ElastiCache to fail over to a read replica?

Amazon ElastiCache will failover to a read replica in the event of any of the following:

- Loss of availability in primary's Availability Zone
- Loss of network connectivity to primary
- Compute unit failure on primary

Q: When should I use Multi-AZ?

Using Redis replication in conjunction with Multi-AZ provides increased availability and fault tolerance. Such deployments are a natural fit for use in production environments. When running ElastiCache for Redis Cluster with cluster mode enabled, if your shards have one or more read replicas, Multi-AZ will automatically be enabled.

Q: How do I create an ElastiCache for Redis replication group with Multi-AZ enabled?

You can create an ElastiCache for Redis primary and read replicas by clicking "Create" on the ElastiCache Management Console. You can also do so by calling the CreateReplicationGroup API. For existing clusters (Redis 5.0.6, 5.0.5, 5.0.4, 5.0.3, 5.0.0, 4.0.10, 3.2.10, 3.2.6, 3.2.4, 2.8.24, 2.8.23, 2.8.22, 2.8.21, 2.8.19, and 2.8.6 with cluster_mode=disabled), you can enable Multi-AZ by choosing a cluster and clicking Modify on the ElastiCache Management Console or by using the ModifyReplicationGroup API. Switching a replication group to Multi-AZ is not disruptive to your Redis data and does not interfere your nodes' ability to serve requests.

Q: Which read replica will be promoted in case of primary node failure?

If there are more than one read replicas, the read replica with the smallest asynchronous replication lag to the primary will be promoted.

Q: How much does it cost to use Multi-AZ?

Multi-AZ is free of charge. You only pay for the ElastiCache nodes that you use.

Q: What are the performance implications of Multi-AZ?

ElastiCache currently uses the Redis engine's native, asynchronous replication and is subject to its strengths and limitations. In particular, when a read replica connects to a primary for the first time, or if the primary changes, the read replica does a full synchronization of the data from the primary, imposing load on itself and the primary. For additional details regarding Redis replication please see here.

Q: What node types support Multi-AZ?

All available node types in ElastiCache support Multi-AZ. Only t1.micro does not support this feature at this time.

Q: Will I be alerted when automatic failover occurs?

Yes, Amazon ElastiCache will create an event to inform you that automatic failover occurred. You can use the DescribeEvents API to return information about events related to your ElastiCache node, or click the Events section of the ElastiCache Management Console.

Q: After failover, my primary is now located in a different Availability Zone than my other AWS resources (for example, EC2 instances). Should I be concerned about latency?

Availability Zones are engineered to provide low latency network connectivity to other Availability Zones in the same region. You may consider architecting your application and other AWS resources with redundancy across multiple Availability Zones so your application will be resilient in the event of an Availability Zone disruption.

Q: Where can I get more information about Multi-AZ?

For more information about Multi-AZ, see ElastiCache documentation.

Q: Can I test the Multi-AZ functionality?

Yes. If you have the “Multi-AZ” feature enabled on a cluster or replication group with one or more read replicas, you can trigger a failover. ElastiCache will respond in the same way as a real failure scenario – by detecting the failure, promoting the most up-to-date read replica to become the new primary, and then replacing the failed primary, attaching it as a new read replica in place of the one promoted. For more details on testing failover, please see documentation.

Backup and Restore

Q: What is Backup and Restore?

Backup and Restore is a feature that allows customers to create snapshots of their ElastiCache for Redis clusters. ElastiCache stores the snapshots, allowing users to subsequently use them to restore Redis clusters.

Q: What is a snapshot?

A snapshot is a copy of your entire Redis cluster at a specific moment.

Q: Why would I need snapshots?

Creating snapshots can be useful in case of data loss caused by node failure, as well as the unlikely event of a hardware failure. Another common reason to use backups is for archiving purposes. Snapshots are stored in Amazon S3, which is a durable storage, meaning that even a power failure won't erase your data.

Q: What can I do with a snapshot?

You can use snapshots to warm start an ElastiCache for Redis cluster with preloaded data.

Q: How does Backup and Restore work?

When a backup is initiated, ElastiCache will take a snapshot of a specified Redis cluster that can later be used for recovery or archiving. You can initiate a backup anytime you choose or set a recurring daily backup with retention period of up to 35 days.

When you choose a snapshot to restore, a new ElastiCache for Redis cluster will be created and populated with the snapshot's data. This way you can create multiple ElastiCache for Redis clusters from a specified snapshot.

Currently, ElastiCache uses Redis' native mechanism to create and store an RDB file as the snapshot.

Q: Where are my snapshots stored?

The snapshots are stored in S3.

Q: How can I get started using Backup and Restore?

You can select to use the Backup and Restore feature through the AWS Management Console, through the ElastiCache APIs (CreateCacheCluster, ModifyCacheCluster and ModifyReplicationGroup API's) and CLI. You can deactivate and reactivate the feature anytime you choose.

Q: How do I specify which Redis cluster and node to backup?

Backup and Restore creates snapshots on a cluster basis. Users can specify which ElastiCache for Redis cluster to backup through the AWS Management Console, CLI or through the CreateSnapshot API. In a Replication Group, you can choose to backup the primary or any of the read-replica clusters. We recommend users enable backup on one of the read-replicas, mitigating any latency effect on the Redis primary.

Q: Does ElastiCache for Memcached support Backup and Restore?

No, snapshots are available only for ElastiCache for Redis.

Q: How can I specify when a backup will take place?

Through the AWS Management Console, CLI or APIs you can specify when to start a single backup or a recurring backup. Users are able to:

- Take a snapshot right now (through “Backup” console button in the "Redis" tab, or CreateSnapshot API)
- Set up an automatic daily backup. The backup will take place during your preferred backup window. You can set that up through Creating/Modifying cluster via console or the CreateCacheCluster, ModifyCacheCluster or ModifyReplicationGroup API’s.

Q: What is a backup window and why do I need it?

The preferred backup window is the user-defined period of time during which your ElastiCache for Redis cluster backup will start. This is helpful if you want to backup at a certain time of day or to refrain from backups during a particularly high-utilization period.

Q: What is the performance impact of taking a snapshot?

While taking a snapshot, you may encounter increased latencies for a brief period at the node. Snapshots use Redis’s built-in BGSAVE and are subject to its strengths and limitations. In particular, the Redis process forks and the parent continues to serve requests while the child saves the data on disk and then exits. The forking increases the memory usage for the duration of the snapshot generation. When this memory usage exceeds that of the available memory of the node, swapping can get triggered, further slowing down the node. For this reason, we recommend generating snapshots on one of the read replicas (instead of the primary). Also, we suggest setting the reserved-memory parameter to minimize swap usage. See here for more details.

Q: Can I create a snapshot from an ElastiCache for Redis read replica?

Yes. Creating a snapshot from a read replica is the best way to backup your data while minimizing performance impact.

Q: In what regions is the Backup and Restore feature available?

Backup and Restore feature is available in all regions where ElastiCache service is available.

Q: Can I export ElastiCache for Redis snapshots to an S3 bucket owned by me?

Yes. You can export your ElastiCache for Redis snapshots to an authorized S3 bucket in the same region as your cluster. For more details on exporting snapshots and setting the required permissions, please refer to this.

Q: Can I copy snapshots from one region to another?

Yes. You must first copy your snapshot into an authorized S3 bucket of your choice in the same region and then use the S3 PUT object- Copy API to copy it to a bucket in another region. For more details on copying S3 objects, please see this.

Q: I have multiple AWS accounts using ElastiCache for Redis. Can I use ElastiCache snapshots from one account to warm start an ElastiCache for Redis cluster in a different

one?

Yes. You must first copy your snapshot into an authorized S3 bucket of your choice in the same region and then grant cross-account bucket permissions to the other account. For more details on S3 cross-account permissions, please see this. Finally, specify the S3 location of your RDB file during cluster creation through the Launch Cache Cluster Wizard in the console or through the CreateCacheCluster API.

Q: How much does it cost to use Backup and Restore?

Amazon ElastiCache provides storage space for one snapshot free of charge for each active ElastiCache for Redis cluster. Additional storage will be charged based on the space used by the snapshots with $0.085/GB every month (same price in all regions). Data transfer for using the snapshots is free of charge.

Q: What is the retention period?

Retention period is the time span during which the automatic snapshots are retained. For example, if a retention period is set for 5, a snapshot that was taken today will be retained for 5 days before being deleted. You can choose to copy one or more automatic snapshots to store them as manual so that they won’t be deleted after the retention period is over.

Q: How do I manage the retention of my automated snapshots?

You can use the AWS Management Console or ModifyCluster API to manage the period of time your automated backups are retained by modifying the RetentionPeriod parameter. If you desire to turn off automated backups altogether, you can do so by setting the retention period to 0 (not recommended).

Q: What happens to my snapshots if I delete my ElastiCache for Redis cluster?

When you delete an ElastiCache for Redis cluster, your manual snapshots are retained. You will also have an option to create a final snapshot before the cluster is deleted. Automatic snapshots are not retained.

Q: What nodes types support backup and restore capability?

All ElastiCache for Redis instance node types besides t1.micro family support backup and restore:

Current Generation Nodes:

- cache.m4.large
- cache.m4.xlarge
- cache.m4.2xlarge
- cache.m4.4xlarge

- cache.m4.10xlarge
- cache.m5.large
- cache.m5.xlarge
- cache.m5.2xlarge
- cache.m5.4xlarge
- cache.m5.12xlarge
- cache.m5.24xlarge
- cache.r4.large
- cache.r4.xlarge
- cache.r4.2xlarge
- cache.r4.4xlarge
- cache.r4.8xlarge
- cache.r4.16xlarge
- cache.r5.large
- cache.r5.xlarge
- cache.r5.2xlarge
- cache.r5.4xlarge
- cache.r5.12xlarge
- cache.r5.24xlarge
- cache.t2.medium
- cache.t2.small
- cache.t2.micro

Previous Generation Nodes:

- cache.m1.small
- cache.m1.medium

- cache.m1.large
- cache.m1.xlarge
- cache.m2.xlarge
- cache.m2.2xlarge
- cache.m2.4xlarge
- cache.m3.medium
- cache.m3.large
- cache.m3.xlarge
- cache.m3.2xlarge
- cache.r3.large
- cache.r3.xlarge
- cache.r3.2xlarge
- cache.r3.4xlarge
- cache.r3.8xlarge
- cache.c1.xlarge

Q: Can I use my own RDB snapshots stored in S3 to warm start an ElastiCache for Redis cluster?

Yes. You can specify the S3 location of your RDB file during cluster creation through the "Create Cluster" Wizard in the console or through the CreateCacheCluster API.

Q: Can I use the Backup and Restore feature if I am running ElastiCache in a VPC?

Yes.

Redis Cluster

Q: What is ElastiCache for Redis Cluster?

ElastiCache for Redis Cluster allows customers to create and run managed Redis Clusters with multiple shards. It is compatible with open source Redis 3.2.4 onwards and comes with a number of enhancements for a more stable and robust experience (see the “enhanced engine” section below for additional details on these enhancements).

Q: Why would I need a scale out Redis environment?

There are three main scenarios for running a scale out Redis environment. First, if the total memory size of your Redis data exceeds or is projected to exceed the memory capacity of a single VM. Second, if the write throughput of your application to Redis exceeds the capacity of a single VM. Third, if you would like to spread the data across multiple shards so that any potential issue that comes up with a single node will have a smaller impact on the overall Redis environment.

Q: Why would I run my Redis Cluster workload on Amazon ElastiCache?

Amazon ElastiCache provides a fully managed distributed in-memory Redis environment, from provisioning server resources to installing the engine software and applying any configuration parameters you choose. It uses enhancements to the Redis engine developed by Amazon, which results in a more robust and stable experience (see “enhanced engine” section for more details). Once your Redis environment is up and running, the service automates common administrative tasks such as failure detection and recovery, backups and software patching. It also provides a robust Multi-AZ solution with automatic failover. In case of a failure of one or more primary nodes in your cluster, Amazon ElastiCache will automatically detect the failure and respond by promoting the most up to date replica to primary. This process is automated and does not mandate any manual work on your behalf. Amazon ElastiCache also provides detailed monitoring metrics associated with your ElastiCache nodes, enabling you to diagnose and respond to issues very quickly.

Q: Is ElastiCache for Redis Cluster compatible with open source Redis?

Yes, Amazon ElastiCache for Redis Cluster is compatible with open source Redis 3.2.4 onwards. You can use the open source Redis Cluster clients to access scale-out clusters on ElastiCache for Redis.

Q: What is the upgrade path from current ElastiCache for Redis 2.8.x to ElastiCache for Redis Cluster (version 3.2.4)?

If you are using Redis 3.2 with cluster_mode parameter disabled, you can simply choose the node or cluster you wish to upgrade and modify the engine version. ElastiCache will provision a Redis 3.2.4 cluster and migrate your data to it, while maintaining the endpoint.

If you are using Redis 3.2 with cluster_mode enabled, you can migrate to Redis Cluster by first creating a snapshot of your data using the backup and restore feature. Then, select the created snapshot and click on “Restore Snapshot” to create a Redis 3.2 cluster using the snapshotted data. Finally, update the new endpoint in your client. Note that to use Redis 3.2 in cluster mode you would need to switch to a Redis Cluster client.

Q: Is the pricing for clustered configuration different from non-clustered configuration?

No. Amazon ElastiCache for Redis provides the flexibility of clustered and non-clustered configuration at the same price. Customers can now enjoy enhanced engine functionality within Amazon ElastiCache for Redis and use full feature support for clustered configuration and scalability at the same price.

Q: What is Multi-AZ for ElastiCache for Redis Cluster?

Each shard of an ElastiCache for Redis cluster consists of a primary and up to five read replicas. Redis asynchronously replicates the data from the primary to the read replicas. During certain types of planned maintenance, or in the unlikely event of ElastiCache node failure or Availability Zone failure, Amazon ElastiCache will automatically detect the failure of a primary, select a read-replica, and promote it to become the new primary.

ElastiCache for Redis Cluster provides enhancements and management for Redis 3.x environments. When running an unmanaged Redis environment, in a case of primary node failure, the cluster relies on a majority of masters to determine and execute a failover. If such majority doesn’t exist, the cluster will go into failed state, rejecting any further reads and writes. This could lead to major availability impact on the application, as well as requiring human intervention to manually salvage the cluster. ElastiCache for Redis Multi-AZ capability is built to handle any failover case for Redis Cluster with robustness and efficiency.

Q: How is Multi-AZ in ElastiCache for Redis Cluster different than in ElastiCache for Redis versions 2.8.x?

Redis 3.x works with intelligent clients that store a node map with all the cluster nodes’ endpoints. During a failover, the client updates the node map with the IP endpoint for the new primary. This provides up to 4x faster failover time than with ElastiCache for Redis 2.8.x.

Q: How does Multi-AZ work for Redis Cluster?

You can use Multi-AZ if you are using an ElastiCache for Redis Cluster with each shard having 1 or more read-replicas. If a primary node of a shard fails, ElastiCache will automatically detect the failure, select one of the available read-replicas, and promote it to become the new primary. The Redis 3.x client will update the promoted replica as primary, no application change is required. ElastiCache will also spin up a new node to replace the promoted read-replica in the same Availability Zone of the failed primary. In case the primary failed due to a temporary Availability Zone failure, the new replica will be launched once that Availability Zone has recovered.

Q: What is a backup in ElastiCache for Redis Cluster?

An ElastiCache for Redis Cluster backup is a series of snapshots of the cluster’s shards, stored together to keep a copy of your entire Redis data around a certain time frame.

Q: How is a backup in ElastiCache for Redis Cluster different from a snapshot in ElastiCache for Redis?

Since a non-clustered ElastiCache for Redis environment has a single primary node, a backup is a single file which contains a copy of the Redis data. ElastiCache for Redis Cluster can have one or more shards, thus a backup might contain multiple files.

Q: How do I specify which ElastiCache for Redis nodes to backup in each shard?

You cannot manually specify a node to backup within each shard. When initiating a backup, ElastiCache will automatically select the most up-to-date read replica in each shard and take a snapshot of its data.

Q: How does ElastiCache for Redis Cluster Backup and Restore work?

When a backup is initiated, ElastiCache will take a backup of a specified cluster; that backup can later be used for recovery or archiving. The backup will include a copy of each of the cluster's shards, thus a full backup contains a series of files. You can initiate a backup anytime you choose or set a recurring daily backup with retention period of up to 35 days.

When you choose a backup to restore, a new ElastiCache for Redis cluster will be created and populated with the backup's data. Also, you can use this feature for an easy migration path to a managed Redis Cluster experience on ElastiCache. If you are running self-managed Redis on EC2, you can take RDB snapshots or your existing workloads (both Redis Cluster and single-shard Redis) and store them in S3. Then simply provide them as input for creating a sharded Redis Cluster on ElastiCache, and the desired number of shards. ElastiCache will do the rest.

Currently, ElastiCache uses Redis' native mechanism to create and store an RDB file for each shard as the backup.

Q: Is the backup in ElastiCache for Redis Cluster a point-in-time snapshot?

When you initiate a backup, ElastiCache will trigger backups of all of the shards of your cluster at the same time. In rare cases there might be a need to retake a snapshot of one or more nodes that did not complete successfully the first time. ElastiCache does that automatically and no user intervention is required. But in such a case, while each individual snapshot is a point-in-time representation of the node it was taken from, not all the cluster's snapshots would be taken at the same time.

Q: How can I specify when a backup will take place?

Through the AWS Management Console, CLI or APIs you can specify when to start a single backup or a recurring backup. Users are able to:

- Take a backup right now (through "Create Snapshot" console button or CreateSnapshot API)

- Set up an automatic daily backup. The backup will take place during your preferred backup window. You can set that up through Creating/Modifying cluster via console or the CreateReplicationGroup and ModifyReplicationGroup API's.

Q: Can I use my own RDB snapshots stored in S3 to pre-seed a scale out ElastiCache for Redis Cluster environment?

Yes. You can specify the S3 location of your RDB files during cluster creation through the Create Cluster Wizard in the console or through the CreateReplicationGroup API. ElastiCache will automatically parse the Redis key-space of the RDB snapshot and redistribute it among the shards of the new cluster.

Enhanced Engine

Q: How is the engine within ElastiCache for Redis different from open-source Redis?

The engine within ElastiCache for Redis is fully compatible with open source Redis but also comes with enhancements that improve robustness and stability. Some of the enhancements are:

- More usable memory: You can now safely allocate more memory for your application without risking increased swap usage during syncs and snapshots.
- Improved synchronization: More robust synchronization under heavy load and when recovering from network disconnections. Additionally, syncs are faster as both the primary and replicas no longer use the disk for this operation.
- Smoother failovers: In the event of a failover, your shard now recovers faster as replicas no longer flush their data to do a full re-sync with the primary.

Q: How do I use the enhanced engine?

To use the enhanced engine from the Amazon ElastiCache management console, just select an engine compatible with Redis engine version 2.8.22 or higher when creating a cluster. From that point on you will be using the enhanced engine. You can also use the enhanced engine through the ElastiCache API or AWS CLI by specifying the engine version when running the CreateCacheCluster API.

Q: Do I need to change my application code to use the enhanced engine on ElastiCache?

No. The enhanced engine is fully compatible with open-source Redis, thus you can enjoy its improved robustness and stability without the need to make any changes to your application code.

Q: How much does it cost to use the enhanced engine?

There is no additional charge for using the enhanced engine. As always, you will only be charged for the nodes you use.

Online Cluster Resizing

Q: What is Online Cluster Resizing?

Amazon ElastiCache for Redis provides the ability to add and remove shards from a running cluster. You can dynamically scale-out or scale-in your Redis cluster workloads to adapt to changes in demand. ElastiCache will resize the cluster by adding or removing shards and redistributing hash slots uniformly across the new shard configuration, all while the cluster continues to stay online and serve requests.

Amazon ElastiCache also supports online vertical scaling for sharded Redis Clusters. Please see modifying node type for more details.

Q: What are the benefits of using Online Cluster Resizing?

The ability to dynamically scale-out and scale-in a cluster can help you manage application variability and meet oscillating demands. You can right-size your clusters by adding or removing shards to scale performance and in-memory capacity. The feature eliminates the need to overprovision clusters based on peak demand, helps improve efficiency, and reduces cost.

Q: How can I use Online Cluster Resizing?

Online Cluster Resizing is available with Redis engine version 3.2.10 and up. To reshard your cluster, select the cluster and specify whether you want to add or remove shards. When you resize the cluster to scale-out, ElastiCache adds shards and migrates slots from existing shards to new shards, in a way such that the slots are uniformly distributed (by count) across shards. Similarly, when resizing the cluster to scale-in, ElastiCache migrates slots to the remaining shards to uniformly distribute the slots and deletes specified shards.

Q: How long does the Online Cluster Resizing take?

The time taken to resize a cluster depends on multiple factors, such as number of slots that need to be migrated across shards, size of data and incoming request rate on the cluster. However, the workflow is optimized to parallelize slot migration, which improves the time taken as you add more shards to scale out the cluster.

Q: Can the cluster be used while cluster resizing is in progress?

Yes, the cluster continues to stay online and serve incoming requests, while resharding is in progress. However, snapshotting a cluster while resharding is not supported to prevent increased load on the cluster.

Q: Is there any performance impact of this operation on the cluster?

While Online Cluster Resizing provides the benefits to scale out/in with zero downtime, it is a compute-intensive operation and can increase the latency of your client connection. To reduce the load on the cluster during the operation, we recommend that you follow the best practices (described in the documentation).

Q: How can I track the progress of an online resharding operation?

You can track the progress of the operation by watching the status of the cluster, shards and nodes. During the operation, the cluster, shards and nodes will stay in "modifying" status. Similarly, when shards are being created, deleted or participating in slot migration, the individual shard status will reflect these statuses to show progress. Additionally, the status of end-to-end operation can also be tracked using the progress indicator for the resharding operation, which indicates percentage completed and provides insight into the remaining time for the operation. Lastly, event messages indicate the progress by describing actions being taken (shard creation, slot migration, etc.) during this operation.

Q: What is the rebalance operation for ElastiCache for Redis cluster?

The rebalance operation can be used to redistribute slots amongst existing shards to achieve a uniform distribution. This is useful when a cluster is created with manually specified uneven slot distribution or a scale-out/in operation leaves the cluster with uneven distribution. Assuming the slots are identical in their memory and I/O requirements, uniform slot distribution by count is an easy way to load balance across shards.

Q: How does tagging work when a cluster scales-out?

When new nodes are added to scale-out a cluster, the nodes carry the same set of tags that are common across all existing nodes. Additionally, users can modify tags on all nodes and continue to use tagging as before.

Q: Are there any client or application side changes needed to use online cluster resizing?

No. The enhanced slot distribution used in cluster resizing workflow is compliant with Redis cluster client behavior and does not require any application changes. ElastiCache retains cluster endpoints, enabling you to continue using existing clients without any changes.

Q: How much does it cost to use the enhanced Redis engine?

There is no additional charge for using the enhanced Redis engine. As always, you will only be charged for the nodes you use.

Encryption

Q: What does encryption in-transit for ElastiCache for Redis provide?

The encryption in-transit feature enables you to encrypt all communications between clients and Redis server as well as between the Redis servers (primary and read replica nodes).

Q: What does encryption at-rest for ElastiCache for Redis provide?

Encryption at-rest allows for encryption of all data on disk during backups, restore and swap, as well as backups created and stored in Amazon S3.

Q: How can I use encryption in-transit, at-rest, and Redis AUTH?

Encryption in-transit, encryption at-rest, and Redis AUTH are all opt-in features. At the time of Redis cluster creation via the console or command line interface, you can specify if you want to enable encryption and Redis AUTH and can proceed to provide an authentication token for communication with the Redis cluster. Once the cluster is setup with encryption enabled, ElastiCache seamlessly manages certificate expiration and renewal without requiring any additional action from the application. Additionally, the Redis clients need to support TLS to avail of the encrypted in-transit traffic.

Q: Does Amazon ElastiCache for Redis support AWS Key Management Service (KMS)?

Yes, Amazon ElastiCache for Redis supports AWS KMS to provide encryption at rest using customer managed customer master keys (CMKs) in AWS KMS. You can use your own CMKs to encrypt data at rest in ElastiCache for Redis. Customer managed CMKs are CMKs in your AWS account that you create, own and manage. For more information, see Amazon ElastiCache User Guide.

Q: Is there an Amazon ElastiCache for Redis client that I need to use when using encryption in-transit, or at-rest?

No. Encryption in-transit requires clients to support TLS. Most of the popular Redis clients (such as Lettuce, Predis, go-Redis) provide support for TLS with some configuration settings. You have to make sure that your Redis client of choice is configured to support TLS and continue to use ElastiCache for Redis as before.

Q: Can I enable encryption in-transit and encryption at-rest on my existing ElastiCache for Redis clusters?

No. Encryption in-transit and encryption at-rest support is only available for new clusters and is not supported on existing ElastiCache for Redis clusters. ElastiCache for Redis versions 5.0.3, 5.0.0, 4.0.10, and 3.2.6 support these features.

Q: Is there any action needed to renew certificates?

No. ElastiCache manages certification expiration and renewal behind the scene. No user action is necessary for ongoing certificate maintenance.

Q: Can I use my certificates for encryption?

No. Currently, ElastiCache does not provide the ability for you to use your certificates. ElastiCache manages certificates transparently for you.

Q: Which node types are supported for encryption in transit and encryption at rest?

Encryption in transit and encryption at rest are supported on R5, R4, R3, M5, M4, M3 and T2 node families

Q: Are there additional costs for using encryption?

There are no additional costs for using encryption.

Compliance

Q: Which compliance programs does ElastiCache for Redis support?

ElastiCache for Redis supports compliance programs such as SOC 1, SOC 2, SOC 3, ISO, MTCS, C5, PCI, HIPAA, and FedRAMP. See AWS Services in Scope by Compliance Program for current list of supported complaince programs.

Q: Is Amazon ElastiCache for Redis PCI compliant?

Yes, the AWS PCI compliance program includes Amazon ElastiCache for Redis as a PCI compliant Service.

Amazon ElastiCache for Memcached is currently not included in the list of PCI compliant services.

To learn more, see the following resources:

- Amazon ElastiCache for Redis Compliance page
- AWS PCI Compliance page

To see the current list of compliance programs that Amazon ElastiCache for Redis is in scope for, see AWS Services in Scope by Compliance Program.

Q: Is Amazon ElastiCache for Redis HIPAA eligible?

Yes, Amazon ElastiCache for Redis is a HIPAA Eligible Service and has been added to the AWS Business Associate Addendum (BAA). This means you can use ElastiCache for Redis to help you process, maintain, and store protected health information (PHI) and power healthcare applications.

Q: What do I have to do to use HIPAA eligible ElastiCache for Redis?

If you have an executed Business Associate Agreement (BAA) with AWS, you can use ElastiCache for Redis to build HIPAA-compliant applications. If you do not have a BAA or have other questions about using AWS for your HIPAA-compliant applications, contact us for more information. See Architecting for HIPAA Security and Compliance on Amazon Web Services for information about how to configure Amazon HIPAA Eligible Services to store, process, and transmit PHI.

Q: Is Amazon ElastiCache for Redis FedRAMP authorized?

The AWS FedRAMP compliance program includes Amazon ElastiCache for Redis as a FedRAMP authorized service. United States government customers and their partners can now use the latest version of ElastiCache for Redis to process and store their FedRAMP systems, data, and mission-critical, high-impact workloads in the AWS GovCloud (US) Region, and at moderate impact level in AWS US East/West Regions.

Amazon ElastiCache for Memcached is currently not included in the list of FedRAMP-authorized services.

To learn more, see the following resources:

- Amazon ElastiCache for Redis Compliance page
- AWS FedRAMP Compliance page

To see the current list of compliance programs that Amazon ElastiCache for Redis is in scope for, see AWS Services in Scope by Compliance Program.

Q: Does it cost extra to use compliance features?

No, there is no additional cost for using compliance features.

Q: How is Service Updates feature related to Compliance?

ElastiCache for Redis provides this as a self-service feature to apply the latest security updates to your Redis clusters that are in scope of supported compliance programs – HIPAA, PCI and FedRAMP. Using this feature gives you more control to plan ahead on when to apply the service updates and mitigate any unintended impact to your business applications such as data loss, and cluster unavailability at peak times. The feature also updates you on the progress as it applies the updates to your selected Redis clusters in real time, giving you more control of when to stop or continue the updates. For more details on this feature see Self-Service Updates in Amazon ElastiCache.
To see the requirements of each compliance regime on Amazon ElastiCache for Redis, see ElastiCache for Redis Compliance.

Q: Can I use Service Updates feature even if I am not interested in Compliance?

Yes, this feature is available to everyone and benefits all customers who want to keep their Redis replication groups up-to-date with latest security updates at all times.

Q: Does Service Updates feature replace ElastiCache service maintenance activities?

No, this feature enables you to apply specific updates to your Redis fleet in ElastiCache. Amazon ElastiCache service will continue to apply upgrades that strengthen security, reliability, and operational performance as necessary in your maintenance windows.

Amazon Managed Apache Cassandra Service FAQs

General

Q: What is Amazon Managed Apache Cassandra service?

Amazon Managed Apache Cassandra Service is a scalable, highly available, and managed Apache Cassandra–compatible database service. With Amazon Managed Cassandra Service, you can run your Cassandra workloads on AWS using the same Cassandra application code, Apache 2.0–licensed drivers, and tools that you use today.

Q: What does “Cassandra-compatible” mean?

Amazon Managed Cassandra Service is compatible with the Cassandra Query Language (CQL) 3.11 API (backward-compatible with version 2.x). Amazon Managed Cassandra Service enables you to run your Cassandra workloads on AWS using the same Cassandra application code, Apache 2.0–licensed drivers, and tools that you use today. Amazon Managed Cassandra Service supports all of the most commonly used Cassandra data-plane operations, such as creating keyspaces and tables, reading data, and writing data. Amazon Managed Cassandra Service is serverless, so you don’t have to provision, patch, or manage servers, and you don’t have to install, maintain, or operate software. As a result, Cassandra’s control plane APIs are not required to use Amazon Managed Cassandra Service. Settings such as replication factor and consistency level are configured automatically to provide you with high availability, durability, and single-digit-millisecond performance.

Getting started

Q: How do I get started with Amazon Managed Cassandra Service?

With Amazon Managed Cassandra Service, you can run your Cassandra workloads on AWS using the same Cassandra application code, Apache 2.0–licensed drivers, and tools that you use today. To get started with Amazon Managed Cassandra Service, create a keyspace and table by using either the AWS Management Console or an Apache 2.0–licensed Cassandra driver.

Q: Do I need to change client drivers to use Amazon Managed Cassandra Service?

No. Amazon Managed Cassandra Service works with existing Apache 2.0–licensed Cassandra drivers.

Q: How do I access Amazon Managed Cassandra Service?

You can use your existing Apache 2.0–licensed Cassandra drivers and developer tools with Amazon Managed Cassandra Service.

Planning

Q: What is the maximum throughput an Amazon Managed Cassandra Service table can support? What is the maximum size of a table? What is the maximum number of items that can be stored in a table?

Amazon Managed Cassandra Service tables can scale up and down with virtually unlimited throughput and storage. There is no limit on the size of a table or the number of items you can store in a table. By default, you can create up to 256 tables per account, per AWS Region.

Q: What kind of performance can I expect from Amazon Managed Cassandra Service?

Amazon Managed Cassandra Service offers consistent single-digit-millisecond server-side read and write performance, while also providing high availability and data durability.

Q: How does pricing for Amazon Managed Cassandra Service work?

Amazon Managed Cassandra Service offers on-demand capacity mode for throughput, so you pay for only the reads and writes you use. Additionally, Amazon Managed Cassandra Service charges for data storage and standard internet data transfer fees. For more information, see Amazon Managed Cassandra Service pricing.

Amazon Neptune FAQs

General

Q: What is Amazon Neptune? >>

Q: What popular graph query languages does Amazon Neptune support? >>

Q: Can I use Apache TinkerPop Gremlin and RDF/SPARQL on the same Neptune instance? >>

Q: How can I migrate from an existing Apache TinkerPop Gremlin application to Amazon Neptune? >>

Q: Do I need to change client drivers to use Amazon Neptune’s Gremlin Server? >>

Q: How can I migrate from a triple store with a SPARQL endpoint to Amazon Neptune? >>

Q: Do I need to change client drivers to use Amazon Neptune’s SPARQL Endpoint? >>

Q: Is Neptune ACID (Atomicity, Consistency, Isolation, Durability) compliant? >>

Q: Why are Amazon RDS permissions and resources required to use Amazon Neptune? >>

Q: Does Amazon Neptune have a service level agreement (SLA)? >>

Performance

Q: Do I need to create indices on my data with Amazon Neptune? >>

Q: What types of graph query workloads are optimized to work with Amazon Neptune? >>

Q: Does Amazon Neptune perform query optimization? >>

Q: Is Amazon Neptune built on a relational database? >>

Pricing

Q: How much does Amazon Neptune cost? >>

Q: In which AWS regions is Amazon Neptune available? >>

Q. Amazon Neptune replicates each chunk of my database volume six ways across three Availability Zones. Does that mean that my effective storage price will be three or six times what is shown on the pricing page? >>

Q: What are IOs in Amazon Neptune and how are they calculated? >>

Hardware and Scaling

Q: What are the minimum and maximum storage limits of an Amazon Neptune database? >>

Q: How do I scale the compute resources associated with my Amazon Neptune DB Instance? >>

Backup and Restore

Q: How do I enable backups for my DB Instance? >>

Q: Can I take DB Snapshots and keep them around as long as I want? >>

Q: If my database fails, what is my recovery path? >>

Q: What happens to my automated backups and DB Snapshots if I delete my DB Instance? >>

Q: Can I share my snapshots with another AWS account? >>

Q: Will I be billed for shared snapshots? >>

Q: Can I automatically share snapshots? >>

Q: How many accounts can I share snapshots with? >>

Q: In which regions can I share my Amazon Neptune snapshots? >>

Q. Can I share my Amazon Neptune snapshots across different regions? >>

Q: Can I share an encrypted Amazon Neptune snapshot? >>

Q: Can I use Amazon Neptune snapshots outside of the service? >>

High Availability and Replication

Q: How does Amazon Neptune improve my database’s fault tolerance to disk failures? >>

Q: How does Amazon Neptune improve recovery time after a database crash? >>

Q: What kind of replicas does Neptune support? >>

Q: Can I have cross-region replicas with Amazon Neptune? >>

Q: Can I prioritize certain replicas as failover targets over others? >>

Q: Can I modify priority tiers for instances after they have been created? >>

Q: Can I prevent certain replicas from being promoted to the primary instance? >>

Q: How can I improve upon the availability of a single Amazon Neptune database? >>

Q: What happens during failover and how long does it take? >>

Q: If I have a primary database and an Amazon Neptune Replica actively taking read traffic and a failover occurs, what happens? >>

Q: How far behind the primary will my replicas be? >>

Security

Q: Can I use Amazon Neptune in Amazon Virtual Private Cloud (Amazon VPC)? >>

Q: Does Amazon Neptune support encrypting my data in transit and at rest? >>

Q: Can I encrypt an existing unencrypted database? >>

Q: How do I access my Amazon Neptune database? >>

Amazon Quantum Ledger Database (QLDB) FAQs

General

Q: What is Amazon Quantum Ledger Database?

Amazon Quantum Ledger Database (QLDB) is a purpose-built ledger database that provides a complete and cryptographically verifiable history of all changes made to your application data.

Q: How is a ledger database different from other databases?

Traditional databases allow you to overwrite or delete data, so developers use techniques such as audit tables and audit trails to help track data lineage. While these approaches can work, they require custom development, can be difficult to scale, and put the onus on the application developer to ensure all the right data is being recorded. Data in Amazon QLDB is written to an append-only journal, providing the developer with full data lineage. Moreover, data in Amazon QLDB's journal is immutable and verifiable, meaning you can trust the data in your ledger.

Q: What data should I store in a ledger database?

Amazon QLDB's features make it a natural fit for system-of-record applications – those for which data integrity, completeness, and verifiability are critical. For example, in the supply chain and logistics space, an application built on Amazon QLDB would have the entire history of changes, such as movement between carriers and across borders, available for query and analysis. In finance, system-of-record applications track critical data, such as credit and debit transactions. Instead of building complex record keeping functionality within their application, banks can use QLDB to easily store a permanent and complete record of all financial transactions.

Q: Is Amazon Quantum Ledger Database a distributed ledger or blockchain service?

Amazon QLDB is not a blockchain or distributed ledger technology. Blockchain and distributed ledger technologies focus on solving the problem of decentralized applications involving multiple parties where there can be no single entity that owns the application, and the parties do not necessarily trust each other fully. On the other hand, QLDB is a ledger database purpose-built for customers who need to maintain a complete and verifiable history of data changes in an application that they own. Amazon QLDB offers history, immutability and verifiability combined with the familiarity, scalability and ease of use of a fully managed AWS database. If your application requires decentralization and involves multiple, untrusted parties, a blockchain solution may be appropriate. If your application requires a complete and verifiable history of all application data changes, but does not involve multiple, untrusted parties, Amazon QLDB is a great fit. If you have a use case for distributed ledgers or blockchain, please see Amazon Managed Blockchain.

Q: What kind of functionality does Amazon QLDB support?

In addition to providing a complete and verifiable history of application data changes, Amazon QLDB supports transactions with ACID semantics, a flexible document data model, and a familiar SQL-like API. QLDB is also fully managed and automatically scales to meet the needs of your application with no provisioning required.

Q: How do I connect to Amazon QLDB from my application?

In order to connect to Amazon QLDB and transact with the data in the ledger, you need to use the AWS-provided QLDB driver. Follow the steps in this link to download the driver and configure a connection.

Q: How do I try Amazon QLDB?

Getting started with Amazon QLDB is easy as there are no servers to manage or capacity to provision. You can create a new ledger in minutes using the AWS Management Console, AWS Command Line Interface (CLI), an AWS CloudFormation template, or by making calls to the QLDB API.

Performance

Q: What type of performance can I expect from Amazon QLDB?

Amazon QLDB can execute 2 – 3X as many transactions than ledgers in common blockchain frameworks. Blockchain frameworks are decentralized so to execute a transaction, they require a majority of members of the network to reach consensus on the validity of the transaction. On the other hand, QLDB has a centralized design, allowing its transactions to execute without the need for multi-party consensus.

Querying

Q: What is PartiQL? How does Amazon QLDB support it?

Amazon QLDB allows you to access and manipulate your data using PartiQL, which is a new open standard query language that supports SQL-compatible access to QLDB's document-oriented data model that includes semi-structured and nested data while remaining independent of any particular data source. To learn more about PartiQL read here.

Pricing

Q. How much does Amazon QLDB cost?

For Amazon QLDB pricing, please refer to our pricing page.

Q. In which AWS regions is Amazon QLDB available?

Amazon QLDB is available today in US East (Ohio), US East (N. Virginia), US West (Oregon), Asia Pacific (Tokyo), and EU (Ireland), with additional regions coming soon.

Scalability

Q. How does Amazon QLDB scale?

With Amazon QLDB, you don’t have to worry about provisioning capacity or configuring read and write limits. You create a ledger, define your tables, and QLDB automatically scales to support the demands of your application.

Q. What are the limits associated with an Amazon QLDB?

You can see the limits associated with Amazon QLDB on the AWS Service Limits page.

Backup and Restore

Q. Can I take a snapshot or backup of my ledger?

Amazon QLDB does not support a backup and restore feature as of now. At present, an export to S3 functionality is available. Using this functionality you can export the contents of your QLDB journal to S3.

Q. Can I restore my ledger to a particular point in time?

Amazon QLDB does not support a point-in-time restore feature as of now.

Availability, Durability, and Replication

Q. Is Amazon QLDB durable?

Amazon QLDB's ledger is deployed across multiple AZs with multiple copies per AZ. We maintain redundancy within the region and ensure full recovery from availability zone failures. A write is acknowledged only after being written to a durable storage in multiple AZs, and hence, QLDB is strongly durable.

Q. How does high availability work in Amazon QLDB?

Amazon QLDB is a highly available service. By default, multiple copies of your QLDB ledger are replicated across availability zones in a region. So, in the case of a zone failure you can still continue to operate QLDB.

Q. Does Amazon QLDB have cross-region replication?

Amazon QLDB does not support cross-region replication as of now. QLDB's export to S3 feature enables customers to export the contents of the QLDB journal to a S3 bucket. The S3 buckets can be configured for cross-region replication.

Security

Q: Can I use Amazon QLDB in Amazon Virtual Private Cloud (Amazon VPC)?

Amazon QLDB is integrated with AWS Private Link. Customers can create a VPC endpoint, which enables them to privately connect a VPC to supported AWS services powered by PrivateLink without requiring an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection.

Q: How does authentication work with Amazon QLDB?

Amazon QLDB uses the same authentication mechanism as other AWS services. The mechanism requires a request signature to be attached to the HTTP requests (header or query string). The signature is computed using other requests fields and the AWS credentials (Access Key ID and Secret Access Key).

Q. How does encryption work in Amazon QLDB?

By default, all data in transit and at rest is encrypted. Today, Amazon QLDB does not support customer managed CMKs (Customer Master Keys). Amazon QLDB uses AWS-owned keys to encrypt customer data.

Amazon RDS FAQs

General

Q: What is Amazon RDS?

Amazon Relational Database Service (Amazon RDS) is a managed service that makes it easy to set up, operate, and scale a relational database in the cloud. It provides cost-efficient and resizable capacity, while managing time-consuming database administration tasks, freeing you up to focus on your applications and business.

Amazon RDS gives you access to the capabilities of a familiar MySQL, MariaDB, Oracle, SQL Server, or PostgreSQL database. This means that the code, applications, and tools you already use today with your existing databases should work seamlessly with Amazon RDS. Amazon RDS can automatically back up your database and keep your database software up to date with the latest version. You benefit from the flexibility of being able to easily scale the compute resources or storage capacity associated with your relational database instance. In addition, Amazon RDS makes it easy to use replication to enhance database availability, improve data durability, or scale beyond the capacity constraints of a single database instance for read-heavy database workloads. As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources you use.

Q: Which relational database engines does Amazon RDS support?

Amazon RDS supports Amazon Aurora, MySQL, MariaDB, Oracle, SQL Server, and PostgreSQL database engines.

Q: What does Amazon RDS manage on my behalf?

Amazon RDS manages the work involved in setting up a relational database: from provisioning the infrastructure capacity you request to installing the database software. Once your database is up and running, Amazon RDS automates common administrative tasks such as performing backups and patching the software that powers your database. With optional Multi-AZ deployments, Amazon RDS also manages synchronous data replication across Availability Zones with automatic failover.

Since Amazon RDS provides native database access, you interact with the relational database software as you normally would. This means you're still responsible for managing the database settings that are specific to your application. You'll need to build the relational schema that best fits your use case and are responsible for any performance tuning to optimize your database for your application's workflow.

Q: When would I use Amazon RDS vs. Amazon EC2 Relational Database AMIs?

Amazon Web Services provides a number of database alternatives for developers. Amazon RDS enables you to run a fully featured relational database while offloading database administration. Using one of our many relational database AMIs on Amazon EC2 allows you to manage your own relational database in the cloud. There are important differences between these alternatives that may make one more appropriate for your use case. See Cloud Databases with AWS for guidance on which solution is best for you.

Q: How do I get started with Amazon RDS?

To sign up for Amazon RDS, you must have an Amazon Web Services account. Create an account if you do not already have one. After you are signed up, please refer to the Amazon RDS documentation, which includes our Getting Started Guide.

Amazon RDS is part of the AWS Free Tier so that new AWS customers can get started with a managed database service in the cloud for free.

Q: Are there hybrid or on-premises deployment options for Amazon RDS?

Yes, you can run RDS on premises using Amazon RDS on Outposts and Amazon RDS on VMware. Please see the Amazon RDS on Outposts and Amazon RDS on VMware FAQs for additional information.

Database Instances

Q: What is a database instance (DB instance)?

You can think of a DB instance as a database environment in the cloud with the compute and storage resources you specify. You can create and delete DB instances, define/refine infrastructure attributes of your DB instance(s), and control access and security via the AWS Management Console, Amazon RDS APIs, and AWS Command Line Interface. You can run one or more DB instances, and each DB instance can support one or more databases or database schemas, depending on engine type.

Q: How do I create a DB instance?

DB instances are simple to create, using either the AWS Management Console, Amazon RDS APIs, or AWS Command Line Interface. To launch a DB instance using the AWS Management Console, click "RDS," then the **Launch DB Instance** button on the **Instances** tab. From there, you can specify the parameters for your DB instance including DB engine and version, license model, instance type, storage type and amount, and master user credentials.

You also have the ability to change your DB instance's backup retention policy, preferred backup window, and scheduled maintenance window. Alternatively, you can create your DB instance using the CreateDBInstance API or create-db-instance command.

Q: How do I access my running DB instance?

Once your DB instance is available, you can retrieve its endpoint via the DB instance description in the AWS Management Console, DescribeDBInstances API or describe-db-instances command. Using this endpoint you can construct the connection string required to connect directly with your DB instance using your favorite database tool or programming language. In order to allow network requests to your running DB instance, you will need to authorize access. For a detailed explanation of how to construct your connection string and get started, please refer to our Getting Started Guide.

Q: How many DB instances can I run with Amazon RDS?

By default, customers are allowed to have up to a total of 40 Amazon RDS DB instances. Of those 40, up to 10 can be Oracle or SQL Server DB instances under the "License Included" model. All 40 can be used for Amazon Aurora, MySQL, MariaDB, PostgreSQL and Oracle under the "BYOL" model. Note that RDS for SQL Server has a limit of up to 100 databases on a single DB instance to learn more see the Amazon RDS SQL Server User Guide.

If your application requires more DB instances, you can request additional DB instances via this request form.

Q: How many databases or schemas can I run within a DB instance?

- RDS for Amazon Aurora: No limit imposed by software
- RDS for MySQL: No limit imposed by software
- RDS for MariaDB: No limit imposed by software
- RDS for Oracle: 1 database per instance; no limit on number of schemas per database imposed by software
- RDS for SQL Server: Up to 100 databases per instance see here: Amazon RDS SQL Server User Guide
- RDS for PostgreSQL: No limit imposed by software

Q: How do I import data into an Amazon RDS DB instance?

There are a number of simple ways to import data into Amazon RDS, such as with the mysqldump or mysqlimport utilities for MySQL; Data Pump, import/export or SQL Loader for Oracle; Import/Export wizard, full backup files (.bak files) or Bulk Copy Program (BCP) for SQL Server; or pg_dump for PostgreSQL. For more information on data import and export, please refer to the Data Import Guide for MySQL or the Data Import Guide for Oracle or the Data Import Guide for SQL Server or the Data Import Guide for PostgreSQL.

In addition, AWS Database Migration Service can help you migrate databases to AWS easily and securely.

Q: What is a maintenance window? Will my DB instance be available during maintenance events?

The Amazon RDS maintenance window is your opportunity to control when DB instance modifications, database engine version upgrades, and software patching occurs, in the event they are requested or required. If a maintenance event is scheduled for a given week, it will be initiated during the maintenance window you identify.

Maintenance events that require Amazon RDS to take your DB instance offline are scale compute operations (which generally take only a few minutes from start-to-finish), database engine version upgrades, and required software patching. Required software patching is automatically scheduled only for patches that are security and durability related. Such patching occurs infrequently (typically once every few months) and should seldom require more than a fraction of your maintenance window.

If you do not specify a preferred weekly maintenance window when creating your DB instance, a 30 minute default value is assigned. If you wish to modify when maintenance is performed on your behalf, you can do so by modifying your DB instance in the AWS Management Console, the ModifyDBInstance API or the modify-db-instance command. Each of your DB instances can have different preferred maintenance windows, if you so choose.

Running your DB instance as a Multi-AZ deployment can further reduce the impact of a maintenance event. Please refer to the Amazon RDS User Guide for more information on maintenance operations.

Q: What should I do if my queries seem to be running slowly?

For production databases we encourage you to enable Enhanced Monitoring, which provides access to over 50 CPU, memory, file system, and disk I/O metrics. You can enable these features on a per-instance basis and you can choose the granularity (all the way down to 1 second). High levels of CPU utilization can reduce query performance and in this case you may want to consider scaling your DB instance class. For more information on monitoring your DB instance, refer to the Amazon RDS User Guide.

If you are using RDS for MySQL or MariaDB, you can access the slow query logs for your database to determine if there are slow-running SQL queries and, if so, the performance characteristics of each. You could set the "slow_query_log" DB

Parameter and query the mysql.slow_log table to review the slow-running SQL queries. Please refer to the Amazon RDS User Guide to learn more.

If you are using RDS for Oracle, you can use the Oracle trace file data to identify slow queries. For more information on accessing trace file data, please refer to Amazon RDS User Guide.

If you're using RDS for SQL Server, you can use the client side SQL Server traces to identify slow queries. For information on accessing server side trace file data, please refer to Amazon RDS User Guide.

Database Engine Versions

Q: Which relational database engine versions does Amazon RDS support?

For the list of supported database engine versions, please refer to the documentation for each engine:

- Amazon RDS for MySQL
- Amazon RDS for MariaDB
- Amazon RDS for PostgreSQL
- Amazon RDS for Oracle
- Amazon RDS for SQL Server
- Amazon Aurora

Q: How does Amazon RDS distinguish between “major” and “minor” DB engine versions?

Refer to the FAQs page for each Amazon RDS database engine for specifics on version numbering.

Q: Does Amazon RDS provide guidelines for support of new DB engine versions?

Over time, Amazon RDS adds support for new major and minor database engine versions. The number of new versions supported will vary based on the frequency and content of releases and patches from the engine's vendor or development organization, and the outcome of a thorough vetting of these releases and patches by our database engineering team. However, as a general guidance, we aim to support new engine versions within 5 months of their general availability.

Q: How do I specify which supported DB engine version I would like my DB instance to run?

You can specify any currently supported version (major and minor) when creating a new DB instance via the **Launch DB Instance** operation in the AWS Management Console or the CreateDBInstance API. Please note that not every database engine version is available in every AWS region.

Q: How do I control if and when the engine version of my DB instance is upgraded to new supported versions?

Amazon RDS strives to keep your database instance up to date by providing you newer versions of the supported database engines. After a new version of a database engine is released by the vendor or development organization, it is thoroughly tested by our database engineering team before it is made available in Amazon RDS.

We recommend that you keep your database instance upgraded to the most current minor version as it will contain the latest security and functionality fixes. Unlike major version upgrades, minor version upgrades only include database changes that are backward-compatible with previous minor versions (of the same major version) of the database engine.

If a new minor version does not contain fixes that would benefit RDS customers, we may choose not to make it available in RDS. Soon after a new minor version is available in RDS, we will set it to be the preferred minor version for new DB instances.

To manually upgrade a database instance to a supported engine version, use the **Modify DB Instance** command on the AWS Management Console or the

ModifyDBInstance API and set the **DB Engine Version** parameter to the desired version. By default, the upgrade will be applied or during your next maintenance window. You can also choose to upgrade immediately by selecting the **Apply Immediately** option in the console API.

If we determine that a new engine minor version contains significant bug fixes compared to a previously released minor version, we will schedule automatic upgrades for DB instances which have the **Auto Minor Version Upgrade** setting to “Yes”. These upgrades will be scheduled to occur during customer-specified maintenance windows.

We schedule them so you can plan around them, because downtime is required to upgrade a DB engine version, even for Multi-AZ instances. If you wish to turn off automatic minor version upgrades, you can do so by setting the Auto Minor Version Upgrade setting to “No”.

In the case of RDS for Oracle and RDS for SQL Server, if the upgrade to the next minor version requires a change to a different edition, then we may not schedule automatic upgrades even if you have enabled the **Auto Minor Version Upgrade** setting. The determination on whether to schedule automatic upgrades in such situations will be made on a case-by-case basis.

Since major version upgrades involve some compatibility risk, they will not occur automatically and must be initiated by you (except in the case of major version deprecation, see below).

For more information about upgrading a DB instance to a new DB engine version, refer to the Amazon RDS User Guide.

Q: Can I test my DB instance with a new version before upgrading?

Yes. You can do so by creating a DB snapshot of your existing DB instance, restoring from the DB snapshot to create a new DB instance, and then initiating a version upgrade for the new DB instance. You can then experiment safely on the upgraded copy of your DB instance before deciding whether or not to upgrade your original DB instance.

For more information about restoring a DB snapshot, refer to the Amazon RDS User Guide.

Q: Does Amazon RDS provide guidelines for deprecating database engine versions that are currently supported?

- We intend to support major version releases (e.g., MySQL 5.6, PostgreSQL 9.6) for at least 3 years after they are initially supported by Amazon RDS.
- We intend to support minor versions (e.g., MySQL 5.6.37, PostgreSQL 9.6.1) for at least 1 year after they are initially supported by Amazon RDS.

Periodically, we will deprecate major or minor engine versions. For major versions, this is typically when the version has moved to extended support or is no longer receiving software fixes or security updates. For minor versions, this is when a minor version has significant bugs or security issues that have been resolved in a later minor version.

While we strive to meet these guidelines, in some cases we may deprecate specific major or minor versions sooner, such as when there are security issues. In the unlikely event that such cases occur, Amazon RDS will automatically upgrade your database engine to address the issue. Specific circumstances may dictate different timelines depending on the issue being addressed.

Q: What happens when an RDS DB engine version is deprecated?

When a minor version of a database engine is deprecated in Amazon RDS, we will provide a three (3) month period after the announcement before beginning automatic upgrades. At the end of the this period, all instances still running the deprecated minor version will be scheduled for automatic upgrade to the latest supported minor version during their scheduled maintenance windows.

When a major version of database engine is deprecated in Amazon RDS, we will provide a minimum six (6) month period after the announcement of a deprecation for you to initiate an upgrade to a supported major version. At the end of this period, an automatic upgrade to the next major version will be applied to any instances still running the deprecated version during their scheduled maintenance windows.

Once a major or minor database engine version is no longer supported in Amazon RDS, any DB instance restored from a DB snapshot created with the unsupported version will automatically and immediately be upgraded to a currently supported version.

Billing

Q: How will I be charged and billed for my use of Amazon RDS?

You pay only for what you use, and there are no minimum or setup fees. You are billed based on:

- DB instance hours – Based on the class (e.g. db.t2.micro, db.m4.large) of the DB instance consumed. Partial DB instance hours consumed are billed as full hours.
- Storage (per GB per month) – Storage capacity you have provisioned to your DB instance. If you scale your provisioned storage capacity within the month, your bill will be pro-rated.
- I/O requests per month – Total number of storage I/O requests you have *(for Amazon RDS Magnetic Storage and Amazon Aurora only)*
- Provisioned IOPS per month – Provisioned IOPS rate, regardless of IOPS consumed *(for Amazon RDS Provisioned IOPS (SSD) Storage only)*
- Backup Storage – Backup storage is the storage associated with your automated database backups and any customer-initiated database snapshots. Increasing your backup retention period or taking additional database snapshots increases the backup storage consumed by your database.
- Data transfer – Internet data transfer in and out of your DB instance.

For Amazon RDS pricing information, please visit the pricing section on the Amazon RDS product page.

Q: When does billing of my Amazon RDS DB instances begin and end?

Billing commences for a DB instance as soon as the DB instance is available. Billing continues until the DB instance terminates, which would occur upon deletion or in the event of instance failure.

Q: What defines billable Amazon RDS instance hours?

DB instance hours are billed for each hour your DB instance is running in an available state. If you no longer wish to be charged for your DB instance, you must stop or delete it to avoid being billed for additional instance hours. Partial DB instance hours consumed are billed as full hours.

Q: How will I be billed for a stopped DB instance?

While your database instance is stopped, you are charged for provisioned storage (including Provisioned IOPS) and backup storage (including manual snapshots and automated backups within your specified retention window), but not for DB instance hours.

Q: Why does my additional backup storage cost more than allocated DB instance storage?

The storage provisioned to your DB instance for your primary data is located within a single Availability Zone. When your database is backed up, the backup data (including transactions logs) is geo-redundantly replicated across multiple Availability Zones to provide even greater levels of data durability. The price for backup storage beyond your free allocation reflects this extra replication that occurs to maximize the durability of your critical backups.

Q: How will I be billed for Multi-AZ DB instance deployments?

If you specify that your DB instance should be a Multi-AZ deployment, you will be billed according to the Multi-AZ pricing posted on the Amazon RDS pricing page. Multi-AZ billing is based on:

- Multi-AZ DB instance hours – Based on the class (e.g. db.t2.micro, db.m4.large) of the DB instance consumed. As with standard deployments in a single Availability Zone, partial DB instance hours consumed are billed as full hours. If you convert your DB instance deployment between standard and Multi-AZ within a given hour, you will be charged both applicable rates for that hour.

- Provisioned storage (for Multi-AZ DB instance) – If you convert your deployment between standard and Multi-AZ within a given hour, you will be charged the higher of the applicable storage rates for that hour.
- I/O requests per month – Total number of storage I/O requests you have. Multi-AZ deployments consume a larger volume of I/O requests than standard DB instance deployments, depending on your database write/read ratio. Write I/O usage associated with database updates will double as Amazon RDS synchronously replicates your data to the standby DB instance. Read I/O usage will remain the same.
- Backup Storage – Your backup storage usage will not change whether your DB instance is a standard or Multi-AZ deployment. Backups will simply be taken from your standby to avoid I/O suspension on the DB instance primary.
- Data transfer – You are not charged for the data transfer incurred in replicating data between your primary and standby. Internet data transfer in and out of your DB instance is charged the same as with a standard deployment.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Free Tier

Q: What does the AWS Free Tier for Amazon RDS offer?

The AWS Free Tier for Amazon RDS offer provides free use of Single-AZ Micro DB instances running MySQL, MariaDB, PostgreSQL, Oracle ("Bring-Your-Own-License (BYOL)" licensing model) and SQL Server Express Edition. The free usage tier is capped at 750 instance hours per month. Customers also receive 20 GB of

General Purpose (SSD) database storage and 20 GB of backup storage for free per month.

Q: For what time period will the AWS Free Tier for Amazon RDS be available to me?

New AWS accounts receive 12 months of AWS Free Tier access. Please see the AWS Free Tier FAQs for more information.

Q: Can I run more than one DB instance under the AWS Free Usage Tier for Amazon RDS?

Yes. You can run more than one Single-AZ Micro DB instance simultaneously and be eligible for usage counted under the AWS Free Tier for Amazon RDS. However, any use exceeding 750 instance hours, across all Amazon RDS Single-AZ Micro DB instances, across all eligible database engines and regions, will be billed at standard Amazon RDS prices.

For example: if you run two Single-AZ Micro DB instances for 400 hours each in a single month, you will accumulate 800 instance hours of usage, of which 750 hours will be free. You will be billed for the remaining 50 hours at the standard Amazon RDS price.

Q: Do I have access to 750 instance hours each of the MySQL, MariaDB, PostgreSQL, Oracle and SQL Server Micro DB instances under the AWS Free Tier?

No. A customer with access to the AWS Free Tier can use up to 750 instance hours of Micro instances running either MySQL, PostgreSQL, Oracle or SQL Server Express Edition. Any use exceeding 750 instance hours, across all Amazon RDS Single-AZ Micro DB instances, across all eligible database engines and regions, will be billed at standard Amazon RDS prices.

Q: How am I billed when my instance-hour usage exceeds the Free Tier benefit?

You are billed at standard Amazon RDS prices for instance hours beyond what the Free Tier provides. See the Amazon RDS pricing page for details.

Reserved Instances

Q: What is a reserved instance (RI)?

Amazon RDS reserved instances give you the option to reserve a DB instance for a one or three year term and in turn receive a significant discount compared to the on-demand instance pricing for the DB instance. There are three RI payment options -- No Upfront, Partial Upfront, All Upfront -- which enable you to balance the amount you pay upfront with your effective hourly price.

Q: How are reserved instances different from on-demand DB instances?

Functionally, reserved instances and on-demand DB instances are exactly the same. The only difference is how your DB instance(s) are billed: With Reserved Instances, you purchase a one or three year reservation and in return receive a lower effective hourly usage rate (compared with on-demand DB instances) for the duration of the term. Unless you purchase reserved instances in a Region, all DB instances will be billed at on-demand hourly rates.

Q: How do I purchase and create reserved instances?

You can purchase a reserved instance in the "Reserved Instance" section of the AWS Management Console for Amazon RDS. Alternatively, you can use the Amazon RDS API or AWS Command Line Interface to list the reservations available for purchase then purchase a DB instance reservation.

Once you have made a reserved purchase, using a reserved DB instance is no different than an On-Demand DB instance. Launch a DB instance using the same instance class, engine and region for which you made the reservation. As long as your reservation purchase is active, Amazon RDS will apply the reduced hourly rate for which you are eligible to the new DB instance.

Q: Do reserved instances include a capacity reservation?

Amazon RDS reserved instances are purchased for a Region rather than for a specific Availability Zone. As RIs are not specific to an Availability Zone, they are not capacity reservations. This means that even if capacity is limited in one Availability Zone, reservations can still be purchased in the Region and the discount will apply to matching usage in any Availability Zone within that Region.

Q: How many reserved instances can I purchase?

You can purchase up to 40 reserved DB instances. If you wish to run more than 40 DB instances, please complete the Amazon RDS DB Instance request form.

Q: What if I have an existing DB instance that I'd like to cover with a reserved instance?

Simply purchase a DB instance reservation with the same DB instance class, DB engine, Multi-AZ option and License Model within the same Region as the DB instance you are currently running and would like to reserve. If the reservation purchase is successful, Amazon RDS will automatically apply your new hourly usage charge to your existing DB instance.

Q: If I sign up for a reserved instance, when does the term begin? What happens to my DB instance when the term ends?

Pricing changes associated with a reserved instance are activated once your request is received while the payment authorization is processed. You can follow the status of your reservation on the AWS Account Activity page or by using the DescribeReservedDBInstances API or describe-reserved-db-instances command. If the one-time payment cannot be successfully authorized by the next billing period, the discounted price will not take effect.

When your reservation term expires, your reserved instance will revert to the appropriate On-Demand hourly usage rate for your DB instance class and Region.

Q: How do I control which DB instances are billed at the reserved instance rate?

The Amazon RDS operations for creating, modifying, and deleting DB instances do not distinguish between On-Demand and reserved instances. When computing your bill, our system will automatically apply your Reservation(s) such that all eligible DB instances are charged at the lower hourly reserved DB instance rate.

Q: If I scale my DB instance class up or down, what happens to my reservation?

Each reservation is associated with the following set of attributes: DB engine, DB instance class, Multi-AZ deployment option, license model and Region.

A reservation for a DB engine and license model that is eligible for size-flexibility (MySQL, MariaDB, PostgreSQL, Amazon Aurora or Oracle "Bring Your Own License") will automatically apply to a running DB instance of any size within the same instance family (e.g. M4, T2, or R3) for the same database engine and Region. In addition, the reservation will also apply to DB instances running in either Single-AZ or Multi-AZ deployment options.

For example, let’s say you purchased a db.m4.2xlarge MySQL reservation. If you decide to scale up the running DB instance to a db.m4.4xlarge, the discounted rate of this RI will cover 1/2 of the usage of the larger DB instance.

If you are running a DB engine or license model that is not eligible for size-flexibility (Microsoft SQL Server or Oracle "License Included"), each reservation can only be applied to a DB instance with the same attributes for the duration of the term. If you decide to modify any of these attributes of your running DB instance before the end of the reservation term, your hourly usage rates for that DB instance will revert to on demand hourly rates.

For more details on about size flexibility, see the Amazon RDS User Guide.

Q: Can I move a reserved instance from one Region or Availability Zone to another?

Each reserved instance is associated with a specific Region, which is fixed for the lifetime of the reservation and cannot be changed. Each reservation can, however, be used in any of the available AZs within the associated Region.

Q: Are reserved instances available for Multi-AZ deployments?

Yes. When you purchase a reserved instance, you can select the Multi-AZ option in the DB instance configuration available for purchase. In addition, if you are using a DB engine and license model that supports reserved instance size-flexibility, a Multi-AZ reserved instance will cover usage for two Single-AZ DB instances.

Q: Are reserved instances available for read replicas?

A DB instance reservation can be applied to a read replica, provided the DB instance class and Region are the same. When computing your bill, our system will automatically apply your Reservation(s), such that all eligible DB instances are charged at the lower hourly reserved instance rate.

Q: Can I cancel a reservation?

No, you cannot cancel your reserved DB instance and the one-time payment (if applicable) is not refundable. You will continue to pay for every hour during your Reserved DB instance term regardless of your usage.

Q: How do the payment options impact my bill?

When you purchase an RI under the All Upfront payment option, you pay for the entire term of the RI in one upfront payment. You can choose to pay nothing upfront by choosing the No Upfront option. The entire value of the No Upfront RI is spread across every hour in the term and you will be billed for every hour in the term, regardless of usage. The Partial Upfront payment option is a hybrid of the All Upfront and No Upfront options. You make a small upfront payment, and you are billed a low hourly rate for every hour in the term regardless of usage.

Hardware and Scaling

Q: How do I determine which initial DB instance class and storage capacity are appropriate for my needs?

In order to select your initial DB instance class and storage capacity, you will want to assess your application's compute, memory and storage needs. For information the about the DB instance classes available, please refer to the Amazon RDS User Guide.

Q: How do I scale the compute resources and/or storage capacity associated with my Amazon RDS Database Instance?

You can scale the compute resources and storage capacity allocated to your DB instance with the AWS Management Console (selecting the desired DB instance and clicking the **Modify** button), the RDS API, or the AWS Command Line Interface. Memory and CPU resources are modified by changing your DB Instance class, and storage available is changed when you modify your storage allocation. Please note that when you modify your DB Instance class or allocated storage, your requested changes will be applied during your specified maintenance window. Alternately, you can use the "apply-immediately" flag to apply your scaling requests immediately. Bear in mind that any other pending system changes will be applied as well.

Some older RDS for SQL Server instances may not be eligible for scaled storage. See the RDS for SQL Server FAQ for more information.

Q: What is the hardware configuration for Amazon RDS storage?

Amazon RDS uses EBS volumes for database and log storage. Depending on the size of storage requested, Amazon RDS automatically stripes across multiple EBS volumes to enhance IOPS performance. For MySQL and Oracle, for an existing DB instance, you may observe some I/O capacity improvement if you scale up your storage. You can scale the storage capacity allocated to your DB Instance using the AWS Management Console, the ModifyDBInstance API, or the modify-db-instance command.

For more information, see Storage for Amazon RDS.

Q: Will my DB instance remain available during scaling?

The storage capacity allocated to your DB Instance can be increased while maintaining DB Instance availability. However, when you decide to scale the compute resources available to your DB instance up or down, your database will be temporarily unavailable while the DB instance class is modified. This period of unavailability typically lasts only a few minutes, and will occur during the maintenance window for your DB Instance, unless you specify that the modification should be applied immediately.

Q: How can I scale my DB instance beyond the largest DB instance class and maximum storage capacity?

Amazon RDS supports a variety of DB instance classes and storage allocations to meet different application needs. If your application requires more compute resources than the largest DB instance class or more storage than the maximum allocation, you can implement partitioning, thereby spreading your data across multiple DB instances.

Q: What is Amazon RDS General Purpose (SSD) storage?

Amazon RDS General Purpose (SSD) Storage is suitable for a broad range of database workloads that have moderate I/O requirements. With the baseline of 3 IOPS/GB and ability to burst up to 3,000 IOPS, this storage option provides predictable performance to meet the needs of most applications.

Q: What is Amazon RDS Provisioned IOPS (SSD) storage?

Amazon RDS Provisioned IOPS (SSD) Storage is an SSD-backed storage option designed to deliver fast, predictable, and consistent I/O performance. With Amazon RDS Provisioned IOPS (SSD) Storage, you specify an IOPS rate when creating a DB instance, and Amazon RDS provisions that IOPS rate for the lifetime of the DB instance. Amazon RDS Provisioned IOPS (SSD) Storage is optimized for I/O-intensive, transactional (OLTP) database workloads. For more details, please see the Amazon RDS User Guide.

Q: What is Amazon RDS Magnetic storage?

Amazon RDS magnetic storage is useful for small database workloads where data is accessed less frequently. Magnetic storage is not recommended for production database instances.

Q: How do I choose among the Amazon RDS storage types?

Choose the storage type most suited for your workload.

- High-performance OLTP workloads: Amazon RDS Provisioned IOPS (SSD) Storage

- Database workloads with moderate I/O requirements: Amazon RDS General Purpose (SSD) Storage

Q: What are the minimum and maximum IOPS supported by Amazon RDS?

The IOPS supported by Amazon RDS varies by database engine. For more details, please see the Amazon RDS User Guide.

Automatic Backups and Database Snapshots

Q: What is the difference between automated backups and DB Snapshots?

Amazon RDS provides two different methods for backing up and restoring your DB instance(s) automated backups and database snapshots (DB Snapshots).

The automated backup feature of Amazon RDS enables point-in-time recovery of your DB instance. When automated backups are turned on for your DB Instance, Amazon RDS automatically performs a full daily snapshot of your data (during your preferred backup window) and captures transaction logs (as updates to your DB Instance are made). When you initiate a point-in-time recovery, transaction logs are applied to the most appropriate daily backup in order to restore your DB instance to the specific time you requested. Amazon RDS retains backups of a DB Instance for a limited, user-specified period of time called the retention period, which by default is 7 days but can be set to up to 35 days. You can initiate a point-in-time restore and specify any second during your retention period, up to the Latest Restorable Time. You can use the DescribeDBInstances API to return the latest restorable time for you DB instance, which is typically within the last five minutes. Alternatively, you can find the Latest Restorable Time for a DB instance by selecting it in the AWS Management Console and looking in the “Description” tab in the lower panel of the Console.

DB Snapshots are user-initiated and enable you to back up your DB instance in a known state as frequently as you wish, and then restore to that specific state at any time. DB Snapshots can be created with the AWS Management Console, CreateDBSnapshot API, or create-db-snapshot command and are kept until you explicitly delete them.

The snapshots which Amazon RDS performs for enabling automated backups are available to you for copying (using the AWS console or the copy-db-snapshot command) or for the snapshot restore functionality. You can identify them using the "automated" Snapshot Type. In addition, you can identify the time at which the snapshot has been taken by viewing the "Snapshot Created Time" field. Alternatively, the identifier of the "automated" snapshots also contains the time (in UTC) at which the snapshot has been taken.

Please note: When you perform a restore operation to a point in time or from a DB Snapshot, a new DB Instance is created with a new endpoint (the old DB Instance can be deleted if so desired). This is done to enable you to create multiple DB Instances from a specific DB Snapshot or point in time.

Q: Do I need to enable backups for my DB Instance or is it done automatically?

By default, Amazon RDS enables automated backups of your DB Instance with a 7 day retention period. Free backup storage is limited to the size of your provisioned database and only applies to active DB Instances. For example, if you have 100 GB of provisioned database storage over the month, we will provide 100 GB-months of backup storage at no additional charge.

If you would like to modify your backup retention period, you can do so using the console or the CreateDBInstance API (when creating a new DB Instance) or the the ModifyDBInstance API (for existing instances). You can use these APIs to change the RetentionPeriod parameter to any number from 0 (which will disable automated backups) to the desired number of days. The value can't be set to 0 if the DB instance is a source to Read Replicas. For more information on automated backups, please refer to the Amazon RDS User Guide.

Q: What is a backup window and why do I need it? Is my database available during the backup window?

The preferred backup window is the user-defined period of time during which your DB Instance is backed up. Amazon RDS uses these periodic data backups in conjunction with your transaction logs to enable you to restore your DB Instance to any second during your retention period, up to the LatestRestorableTime (typically up to the last few minutes). During the backup window, storage I/O may be briefly suspended while the backup process initializes (typically under a few seconds) and you may experience a brief period of elevated latency. There is no I/O suspension for Multi-AZ DB deployments, since the backup is taken from the standby.

Q: Where are my automated backups and DB snapshots stored and how do I manage their retention?

Amazon RDS DB snapshots and automated backups are stored in S3.

You can use the AWS Management Console, the ModifyDBInstance API, or the modify-db-instance command to manage the period of time your automated backups are retained by modifying the RetentionPeriod parameter. If you desire to turn off automated backups altogether, you can do so by setting the retention period to 0 (not recommended). You can manage your user-created DB Snapshots via the "Snapshots" section of the Amazon RDS Console. Alternatively, you can see a list of the user-created DB Snapshots for a given DB Instance using the DescribeDBSnapshots API or describe-db-snapshots command and delete snapshots with the DeleteDBSnapshot API or delete-db-snapshot command.

Q: Why do I have more automated DB snapshots than the number of days in the retention period for my DB instance?

It is normal to have 1 or 2 more automated DB snapshots than the number of days in your retention period. One extra automated snapshot is retained to ensure the ability to perform a point in time restore to any time during the retention period. For example, if your backup window is set to 1 day, you will require 2 automated snapshots to support restores to any within previous 24 hours. You may also see an additional automated snapshot as a new automated snapshot is always created before the oldest automated snapshot is deleted.

Q: What happens to my backups and DB snapshots if I delete my DB instance?

When you delete a DB instance, you can create a final DB snapshot upon deletion; if you do, you can use this DB snapshot to restore the deleted DB instance at a later date. Amazon RDS retains this final user-created DB snapshot along with all other manually created DB snapshots after the DB instance is deleted. Refer to the pricing page for details of backup storage costs.

Automated backups are deleted when the DB instance is deleted. Only manually created DB Snapshots are retained after the DB Instance is deleted.

Security

Q: What is Amazon Virtual Private Cloud (VPC) and how does it work with Amazon RDS?

Amazon VPC lets you create a virtual networking environment in a private, isolated section of the AWS cloud, where you can exercise complete control over aspects such as private IP address ranges, subnets, routing tables and network gateways. With Amazon VPC, you can define a virtual network topology and customize the network configuration to closely resemble a traditional IP network that you might operate in your own datacenter.

One way that you can take advantage of VPC is when you want to run a public-facing web application while still maintaining non-publicly accessible backend servers in a private subnet. You can create a public-facing subnet for your webservers that has access to the Internet, and place your backend RDS DB Instances in a private-facing subnet with no Internet access. For more information about Amazon VPC, refer to the Amazon Virtual Private Cloud User Guide.

Q: How is using Amazon RDS inside a VPC different from using it on the EC2-Classic platform (non-VPC)?

If your AWS account was created before 2013-12-04, you may be able to run Amazon RDS in an Amazon Elastic Compute Cloud (EC2)-Classic environment. The basic functionality of Amazon RDS is the same regardless of whether EC2-Classic or EC2-VPC is used. Amazon RDS manages backups, software patching, automatic failure detection, read replicas and recovery whether your DB Instances are deployed inside or outside a VPC. For more information about the differences between EC2-Classic and EC2-VPC, see the EC2 documentation.

Q: What is a DB Subnet Group and why do I need one?

A DB Subnet Group is a collection of subnets that you may want to designate for your RDS DB Instances in a VPC. Each DB Subnet Group should have at least one subnet for every Availability Zone in a given Region. When creating a DB Instance in VPC, you will need to select a DB Subnet Group. Amazon RDS then uses that DB Subnet Group and your preferred Availability Zone to select a subnet and an IP address within that subnet. Amazon RDS creates and associates an Elastic Network Interface to your DB Instance with that IP address.

Please note that, we strongly recommend you use the DNS Name to connect to your DB Instance as the underlying IP address can change (e.g., during failover).

For Multi-AZ deployments, defining a subnet for all Availability Zones in a Region will allow Amazon RDS to create a new standby in another Availability Zone should the need arise. You need to do this even for Single-AZ deployments, just in case you want to convert them to Multi-AZ deployments at some point.

Q: How do I create an Amazon RDS DB Instance in VPC?

For a procedure that walks you through this process, refer to Creating a DB Instance in a VPC in the Amazon RDS User Guide.

Q: How do I control network access to my DB Instance(s)?

Visit the Security Groups section of the Amazon RDS User Guide to learn about the different ways to control access to your DB Instances.

Q: How do I connect to an RDS DB Instance in VPC?

DB Instances deployed within a VPC can be accessed by EC2 Instances deployed in the same VPC. If these EC2 Instances are deployed in a public subnet with associated Elastic IPs, you can access the EC2 Instances via the internet.

DB Instances deployed within a VPC can be accessed from the Internet or from EC2 Instances outside the VPC via VPN or bastion hosts that you can launch in your public subnet, or using Amazon RDS's Publicly Accessible option:

- To use a bastion host, you will need to set up a public subnet with an EC2 instance that acts as a SSH Bastion. This public subnet must have an internet gateway and routing rules that allow traffic to be directed via the SSH host, which must then forward requests to the private IP address of your RDS DB instance.
- To use public connectivity, simply create your DB Instances with the Publicly Accessible option set to yes. With Publicly Accessible active, your DB Instances within a VPC will be fully accessible outside your VPC by default. This means you do not need to configure a VPN or bastion host to allow access to your instances.

You can also set up a VPN Gateway that extends your corporate network into your VPC, and allows access to the RDS DB instance in that VPC. Refer to the Amazon VPC User Guide for more details.

We strongly recommend you use the DNS Name to connect to your DB Instance as the underlying IP address can change (e.g., during failover).

Q: Can I move my existing DB instances outside VPC into my VPC?

If your DB instance is not in a VPC, you can use the AWS Management Console to easily move your DB instance into a VPC. See the Amazon RDS User Guide for more details. You can also take a snapshot of your DB Instance outside VPC and restore it to VPC by specifying the DB Subnet Group you want to use. Alternatively, you can perform a “Restore to Point in Time” operation as well.

Q: Can I move my existing DB instances from inside VPC to outside VPC?

Migration of DB Instances from inside to outside VPC is not supported. For security reasons, a DB Snapshot of a DB Instance inside VPC cannot be restored to outside VPC. The same is true with “Restore to Point in Time” functionality.

Q: What precautions should I take to ensure that my DB Instances in VPC are accessible by my application?

You are responsible for modifying routing tables and networking ACLs in your VPC to ensure that your DB instance is reachable from your client instances in the VPC.

For Multi-AZ deployments, after failover, your client EC2 instance and RDS DB Instance may be in different Availability Zones. You should configure your networking ACLs to ensure that cross-AZ communication is possible.

Q: Can I change the DB Subnet Group of my DB Instance?

An existing DB Subnet Group can be updated to add more subnets, either for existing Availability Zones or for new Availability Zones added since the creation of the DB Instance. Removing subnets from an existing DB Subnet Group can cause unavailability for instances if they are running in a particular AZ that gets removed from the subnet group. View the Amazon RDS User Guide for more information.

Q: What is an Amazon RDS master user account and how is it different from an AWS account?

To begin using Amazon RDS you will need an AWS developer account. If you do not have one prior to signing up for Amazon RDS, you will be prompted to create one when you begin the sign-up process. A master user account is different from an AWS developer account and used only within the context of Amazon RDS to control access to your DB Instance(s). The master user account is a native database user account which you can use to connect to your DB Instance. You can specify the master user name and password you want associated with each DB Instance when you create the DB Instance. Once you have created your DB Instance, you can connect to the database using the master user credentials. Subsequently, you may also want to create additional user accounts so that you can restrict who can access your DB Instance.

Q: What privileges are granted to the master user for my DB Instance?

For MySQL, the default privileges for the master user include: create, drop, references, event, alter, delete, index, insert, select, update, create temporary tables, lock tables, trigger, create view, show view, alter routine, create routine, execute, trigger, create user, process, show databases, grant option.

For Oracle, the master user is granted the "dba" role. The master user inherits most of the privileges associated with the role. Please refer to the Amazon RDS User Guide for the list of restricted privileges and the corresponding alternatives to perform administrative tasks that may require these privileges.

For SQL Server, a user that creates a database is granted the "db_owner" role. Please refer to the Amazon RDS User Guide for the list of restricted privileges and the corresponding alternatives to perform administrative tasks that may require these privileges.

Q: Is there anything different about user management with Amazon RDS?

No, everything works the way you are familiar with when using a relational database you manage yourself.

Q: Can programs running on servers in my own data center access Amazon RDS databases?

Yes. You have to intentionally turn on the ability to access your database over the internet by configuring Security Groups. You can authorize access for only the specific IPs, IP ranges, or subnets corresponding to servers in your own data center.

Q: Can I encrypt connections between my application and my DB Instance using SSL/TLS?

Yes, this option is supported for all Amazon RDS engines.

Amazon RDS generates an SSL/TLS certificate for each DB Instance. Once an encrypted connection is established, data transferred between the DB Instance and your application will be encrypted during transfer.

While SSL offers security benefits, be aware that SSL/TLS encryption is a compute-intensive operation and will increase the latency of your database connection. SSL/TLS support within Amazon RDS is for encrypting the connection between your application and your DB Instance; it should not be relied on for authenticating the DB Instance itself.

For details on establishing an encrypted connection with Amazon RDS, please visit Amazon RDS's MySQL User Guide, MariaDB User Guide, SQL Server User Guide, PostgreSQL User Guide or Oracle User Guide. To learn more about how SSL/TLS works with these engines, you can refer directly to the MySQL documentation, the MariaDB documentation, the MSDN SQL Server documentation, the PostgreSQL documentation, or the Oracle Documentation.

Q: Can I encrypt data at rest on my Amazon RDS databases?

Amazon RDS supports encryption at rest for all database engines, using keys you manage using AWS Key Management Service (KMS). On a database instance running with Amazon RDS encryption, data stored at rest in the underlying storage is encrypted, as are its automated backups, read replicas, and snapshots. Encryption and decryption are handled transparently. For more information about the use of KMS with Amazon RDS, see the Amazon RDS User's Guide.

You can also add encryption to a previously unencrypted DB instance or DB cluster by creating a DB snapshot and then creating a copy of that snapshot and specifying a KMS encryption key. You can then restore an encrypted DB instance or DB cluster from the encrypted snapshot.

Amazon RDS for Oracle and SQL Server support those engines' Transparent Data Encryption (TDE) technologies. For more information, see the Amazon RDS User's Guide for Oracle and SQL Server.

Q: How do I control the actions that my systems and users can take on specific RDS resources?

You can control the actions that your AWS IAM users and groups can take on RDS resources. You do this by referencing the RDS resources in the AWS IAM policies that you apply to your users and groups. RDS resources that can be referenced in an AWS IAM policy includes DB instances, DB snapshots, read replicas, DB security groups, DB option groups, DB parameter groups, event subscriptions and DB subnet groups. In addition, you can tag these resources to add additional metadata to your resources. By using tagging, you can categorize your resources (e.g. "Development" DB instances, "Production" DB instances, and "Test" DB instances), and write AWS IAM policies that list the permissions (i.e.

actions) that can be taken on resources with the same tags. For more information, refer to Managing Access to Your Amazon RDS Resources and Databases and Tagging Amazon RDS Resources

Q: I wish to perform security analysis or operational troubleshooting on my RDS deployment. Can I get a history of all RDS API calls made on my account?

Yes. AWS CloudTrail is a web service that records AWS API calls for your account and delivers log files to you. The AWS API call history produced by CloudTrail enables security analysis, resource change tracking, and compliance auditing. Learn more about CloudTrail at the AWS CloudTrail detail page, and turn it on via CloudTrail's AWS Management Console home page.

Q: Can I use Amazon RDS with applications that require HIPAA compliance?

A: Yes, all RDS database engines are HIPAA-eligible, so you can use them to build HIPAA-compliant applications and store healthcare related information, including protected health information (PHI) under an executed Business Associate Agreement (BAA) with AWS. If you already have an executed BAA, no action is necessary to begin using these services in the account(s) covered by your BAA. If you do not have an executed BAA with AWS, or have any other questions about HIPAA-compliant applications on AWS, please contact your account manager.

Database Configuration

Q: How do I choose the right configuration parameters for my DB Instance(s)?

By default, Amazon RDS chooses the optimal configuration parameters for your DB Instance taking into account the instance class and storage capacity. However, if you want to change them, you can do so using the AWS Management Console, the Amazon RDS APIs, or the AWS Command Line Interface. Please note that changing configuration parameters from recommended values can have unintended effects, ranging from degraded performance to system crashes, and should only be attempted by advanced users who wish to assume these risks.

Q: What are DB Parameter groups? How are they helpful?

A database parameter group (DB Parameter Group) acts as a “container” for engine configuration values that can be applied to one or more DB Instances. If you create a DB Instance without specifying a DB Parameter Group, a default DB Parameter Group is used. This default group contains engine defaults and Amazon RDS system defaults optimized for the DB Instance you are running. However, if you want your DB Instance to run with your custom-specified engine configuration values, you can simply create a new DB Parameter Group, modify the desired parameters, and modify the DB Instance to use the new DB Parameter Group. Once associated, all DB Instances that use a particular DB Parameter Group get all the parameter updates to that DB Parameter Group.

For more information on configuring DB Parameter Groups, please read the Amazon RDS User Guide.

Q: How can I monitor the configuration of my Amazon RDS resources?

You can use AWS Config to continuously record configurations changes to Amazon RDS DB Instances, DB Subnet Groups, DB Snapshots, DB Security Groups, and Event Subscriptions and receive notification of changes through Amazon Simple Notification Service (SNS). You can also create AWS Config Rules to evaluate whether these RDS resources have the desired configurations.

Multi-AZ Deployments

Q: What does it mean to run a DB instance as a Multi-AZ deployment?

When you create or modify your DB instance to run as a Multi-AZ deployment, Amazon RDS automatically provisions and maintains a synchronous “standby” replica in a different Availability Zone. Updates to your DB Instance are synchronously replicated across Availability Zones to the standby in order to keep both in sync and protect your latest database updates against DB instance failure. During certain types of planned maintenance, or in the unlikely event of

DB instance failure or Availability Zone failure, Amazon RDS will automatically fail over to the standby so that you can resume database writes and reads as soon as the standby is promoted. Since the name record for your DB instance remains the same, your application can resume database operation without the need for manual administrative intervention. With Multi-AZ deployments, replication is transparent: you do not interact directly with the standby, and it cannot be used to serve read traffic. More information about Multi-AZ deployments is in the Amazon RDS User Guide.

Q: What is an Availability Zone?

Availability Zones are distinct locations within a Region that are engineered to be isolated from failures in other Availability Zones. Each Availability Zone runs on its own physically distinct, independent infrastructure, and is engineered to be highly reliable. Common points of failures like generators and cooling equipment are not shared across Availability Zones. Additionally, they are physically separate, such that even extremely uncommon disasters such as fires, tornados or flooding would only affect a single Availability Zone. Availability Zones within the same Region benefit from low-latency network connectivity.

Q: What do “primary” and “standby” mean in the context of a Multi-AZ deployment?

When you run a DB instance as a Multi-AZ deployment, the “primary” serves database writes and reads. In addition, Amazon RDS provisions and maintains a “standby” behind the scenes, which is an up-to-date replica of the primary. The standby is “promoted” in failover scenarios. After failover, the standby becomes the primary and accepts your database operations. You do not interact directly with the standby (e.g. for read operations) at any point prior to promotion. If you are interested in scaling read traffic beyond the capacity constraints of a single DB instance, please see the FAQs on Read Replicas.

Q: What are the benefits of a Multi-AZ deployment?

The chief benefits of running your DB instance as a Multi-AZ deployment are enhanced database durability and availability. The increased availability and fault tolerance offered by Multi-AZ deployments make them a natural fit for production environments.

Running your DB instance as a Multi-AZ deployment safeguards your data in the unlikely event of a DB instance component failure or loss of availability in one Availability Zone. For example, if a storage volume on your primary fails, Amazon RDS automatically initiates a failover to the standby, where all of your database updates are intact. This provides additional data durability relative to standard deployments in a single AZ, where a user-initiated restore operation would be required and updates that occurred after the latest restorable time (typically within the last five minutes) would not be available.

You also benefit from enhanced database availability when running your DB instance as a Multi-AZ deployment. If an Availability Zone failure or DB instance failure occurs, your availability impact is limited to the time automatic failover takes to complete. The availability benefits of Multi-AZ also extend to planned maintenance. For example, with automated backups, I/O activity is no longer suspended on your primary during your preferred backup window, since backups are taken from the standby. In the case of patching or DB instance class scaling, these operations occur first on the standby, prior to automatic fail over. As a result, your availability impact is limited to the time required for automatic failover to complete.

Another implied benefit of running your DB instance as a Multi-AZ deployment is that DB instance failover is automatic and requires no administration. In an Amazon RDS context, this means you are not required to monitor DB instance events and initiate manual DB instance recovery (via the RestoreDBInstanceToPointInTime or RestoreDBInstanceFromSnapshot APIs) in the event of an Availability Zone failure or DB instance failure.

Q: Are there any performance implications of running my DB instance as a Multi-AZ deployment?

You may observe elevated latencies relative to a standard DB instance deployment in a single Availability Zone as a result of the synchronous data replication performed on your behalf.

Q: When running my DB instance as a Multi-AZ deployment, can I use the standby for read or write operations?

No, a Multi-AZ standby cannot serve read requests. Multi-AZ deployments are designed to provide enhanced database availability and durability, rather than read scaling benefits. As such, the feature uses synchronous replication between primary and standby. Our implementation makes sure the primary and the standby are constantly in sync, but precludes using the standby for read or write operations. If you are interested in a read scaling solution, please see the FAQs on Read Replicas.

Q: How do I set up a Multi-AZ DB instance deployment?

In order to create a Multi-AZ DB instance deployment, simply click the "Yes" option for "Multi-AZ Deployment" when launching a DB Instance with the AWS Management Console. Alternatively, if you are using the Amazon RDS APIs, you would call the CreateDBInstance API and set the "Multi-AZ" parameter to the value "true." To convert an existing standard (single AZ) DB instance to Multi-AZ, modify the DB instance in the AWS Management Console or use the ModifyDBInstance API and set the Multi-AZ parameter to true.

Q: What happens when I convert my RDS instance from Single-AZ to Multi-AZ?

For the RDS for MySQL, MariaDB, PostgreSQL and Oracle database engines, when you elect to convert your RDS instance from Single-AZ to Multi-AZ, the following happens:

- A snapshot of your primary instance is taken
- A new standby instance is created in a different Availability Zone, from the snapshot
- Synchronous replication is configured between primary and standby instances

As such, there should be no downtime incurred when an instance is converted from Single-AZ to Multi-AZ. However, you may see increased latency while the data on the standby is caught up to match to the primary.

Q: What events would cause Amazon RDS to initiate a failover to the standby replica?

Amazon RDS detects and automatically recovers from the most common failure scenarios for Multi-AZ deployments so that you can resume database operations as quickly as possible without administrative intervention. Amazon RDS automatically performs a failover in the event of any of the following:

- Loss of availability in primary Availability Zone
- Loss of network connectivity to primary
- Compute unit failure on primary
- Storage failure on primary

Note: When operations such as DB instance scaling or system upgrades like OS patching are initiated for Multi-AZ deployments, for enhanced availability, they are applied first on the standby prior to automatic failover. As a result, your availability impact is limited only to the time required for automatic failover to complete. Note that Amazon RDS Multi-AZ deployments do not fail over automatically in response to database operations such as long running queries, deadlocks or database corruption errors.

Q: Will I be alerted when automatic failover occurs?

Yes, Amazon RDS will emit a DB instance event to inform you that automatic failover occurred. You can click the “Events” section of the Amazon RDS Console or use the DescribeEvents API to return information about events related to your DB instance. You can also use Amazon RDS Event Notifications to be notified when specific DB events occur.

Q: What happens during Multi-AZ failover and how long does it take?

Failover is automatically handled by Amazon RDS so that you can resume database operations as quickly as possible without administrative intervention. When failing over, Amazon RDS simply flips the canonical name record (CNAME) for your DB instance to point at the standby, which is in turn promoted to become the new primary. We encourage you to follow best practices and implement database connection retry at the application layer.

Failovers, as defined by the interval between the detection of the failure on the primary and the resumption of transactions on the standby, typically complete within one to two minutes. Failover time can also be affected by whether large uncommitted transactions must be recovered; the use of adequately large instance types is recommended with Multi-AZ for best results. AWS also recommends the use of Provisioned IOPS with Multi-AZ instances, for fast, predictable, and consistent throughput performance.

Q: Can I initiate a “forced failover” for my Multi-AZ DB instance deployment?

Amazon RDS will automatically fail over without user intervention under a variety of failure conditions. In addition, Amazon RDS provides an option to initiate a failover when rebooting your instance. You can access this feature via the AWS Management Console or when using the RebootDBInstance API call.

Q: How do I control/configure Multi-AZ synchronous replication?

With Multi-AZ deployments, you simply set the “Multi-AZ” parameter to true. The creation of the standby, synchronous replication, and failover are all handled automatically. This means you cannot select the Availability Zone your standby is deployed in or alter the number of standbys available (Amazon RDS provisions one dedicated standby per DB instance primary). The standby also cannot be configured to accept database read activity. Learn more about Multi-AZ configurations.

Q: Will my standby be in the same Region as my primary?

Yes. Your standby is automatically provisioned in a different Availability Zone of the *same Region* as your DB instance primary.

Q: Can I see which Availability Zone my primary is currently located in?

Yes, you can gain visibility into the location of the current primary by using the AWS Management Console or DescribeDBInstances API.

Q: After failover, my primary is now located in a different Availability Zone than my other AWS resources (e.g. EC2 instances). Should I be concerned about latency?

Availability Zones are engineered to provide low latency network connectivity to other Availability Zones in the same Region. In addition, you may want to consider architecting your application and other AWS resources with redundancy across multiple Availability Zones so your application will be resilient in the event of an Availability Zone failure. Multi-AZ deployments address this need for the database tier without administration on your part.

Q: How do DB Snapshots and automated backups work with my Multi-AZ deployment?

You interact with automated backup and DB Snapshot functionality in the same way whether you are running a standard deployment in a Single-AZ or Multi-AZ deployment. If you are running a Multi-AZ deployment, automated backups and DB Snapshots are simply taken from the standby to avoid I/O suspension on the primary. Please note that you may experience increased I/O latency (typically lasting a few minutes) during backups for both Single-AZ and Multi-AZ deployments.

Initiating a restore operation (point-in-time restore or restore from DB Snapshot) also works the same with Multi-AZ deployments as standard, Single-AZ deployments. New DB instance deployments can be created with either the RestoreDBInstanceFromSnapshot or RestoreDBInstanceToPointInTime APIs. These new DB instance deployments can be either standard or Multi-AZ, regardless of whether the source backup was initiated on a standard or Multi-AZ deployment.

Read Replicas

Q: What does it mean to run a DB Instance as a read replica?

Read replicas make it easy to take advantage of supported engines' built-in replication functionality to elastically scale out beyond the capacity constraints of a single DB instance for read-heavy database workloads. You can create a read replica with a few clicks in the AWS Management Console or using the CreateDBInstanceReadReplica API. Once the read replica is created, database updates on the source DB instance will be replicated using a supported engine's native, asynchronous replication. You can create multiple read replicas for a given source DB Instance and distribute your application’s read traffic amongst them.

Since read replicas use supported engines' built-in replication, they are subject to its strengths and limitations. In particular, updates are applied to your read replica(s) after they occur on the source DB instance, and replication lag can vary significantly. Read replicas can be associated with Multi-AZ deployments to gain read scaling benefits in addition to the enhanced database write availability and data durability provided by Multi-AZ deployments.

Q: When would I want to consider using an Amazon RDS read replica?

There are a variety of scenarios where deploying one or more read replicas for a given source DB instance may make sense. Common reasons for deploying a read replica include:

- Scaling beyond the compute or I/O capacity of a single DB instance for read-heavy database workloads. This excess read traffic can be directed to one or more read replicas.
- Serving read traffic while the source DB instance is unavailable. If your source DB Instance cannot take I/O requests (e.g. due to I/O suspension for backups or scheduled maintenance), you can direct read traffic to your read replica(s). For this use case, keep in mind that the data on the read replica may be “stale” since the source DB Instance is unavailable.
- Business reporting or data warehousing scenarios; you may want business reporting queries to run against a read replica, rather than your primary, production DB Instance.
- You may use a read replica for disaster recovery of the source DB instance, either in the same AWS Region or in another Region.

Q: Do I need to enable automatic backups on my DB instance before I can create read replicas?

Yes. Enable automatic backups on your source DB Instance before adding read replicas, by setting the backup retention period to a value other than 0. Backups must remain enabled for read replicas to work.

Q: Which versions of database engines support Amazon RDS read replicas?

Amazon Aurora: All DB clusters.

Amazon RDS for MySQL: All DB instances support creation of read replicas. Automatic backups must be and remain enabled on the source DB instance for read replica operations. Automatic backups on the replica are supported only for Amazon RDS read replicas running MySQL 5.6 and later, not 5.5.

Amazon RDS for PostgreSQL: DB instances with PostgreSQL version 9.3.5 or newer support creation of read replicas. Existing PostgreSQL instances prior to version 9.3.5 need to be upgraded to PostgreSQL version 9.3.5 to take advantage of Amazon RDS read replicas.

Amazon RDS for MariaDB: All DB instances support creation of read replicas. Automatic backups must be and remain enabled on the source DB Instance for read replica operations.

Amazon RDS for Oracle: Supported for Oracle version 12.1.0.2.v12 and higher and for all 12.2 versions using the Bring Your Own License model with Oracle Database Enterprise Edition and licensed for the Active Data Guard Option.

Q: How do I deploy a read replica for a given DB instance?

You can create a read replica in minutes using the standard CreateDBInstanceReadReplica API or a few clicks on the AWS Management Console. When creating a read replica, you can identify it as a read replica by specifying a SourceDBInstanceIdentifier. The SourceDBInstanceIdentifier is the DB Instance Identifier of the “source” DB Instance from which you wish to replicate. As with a standard DB Instance, you can also specify the Availability Zone, DB instance class, and preferred maintenance window. The engine version (e.g., PostgreSQL 9.3.5) and storage allocation of a read replica is inherited from the source DB instance. When you initiate the creation of a read replica, Amazon RDS takes a snapshot of your source DB instance and begins replication. As a result, you will experience a brief I/O suspension on your source DB instance as the snapshot occurs. The I/O suspension typically lasts on the order of one minute, and is avoided if the source DB instance is a Multi-AZ deployment (in the case of Multi-AZ deployments, snapshots are taken from the standby).

Amazon RDS is also currently working on an optimization (to be released shortly) such that if you create multiple Read Replicas within a 30 minute window, all of them will use the same source snapshot to minimize I/O impact (“catch-up” replication for each Read Replica will begin after creation).

Q: How do I connect to my read replica(s)?

You can connect to a read replica just as you would connect to a standard DB instance, using the DescribeDBInstance API or AWS Management Console to retrieve the endpoint(s) for you read replica(s). If you have multiple read replicas, it is up to your application to determine how read traffic will be distributed amongst them.

Q: How many read replicas can I create for a given source DB instance?

Amazon RDS for MySQL, MariaDB, PostgreSQL, and Oracle allow you to create up to 5 read replicas for a given source DB instance.

Q: Can I create a read replica in an AWS Region different from that of the source DB instance?

Yes, Amazon RDS supports cross-region read replicas. The amount of time between when data is written to the source DB instance and when it is available in the read replica will depend on the network latency between the two regions.

Q: Do Amazon RDS read replicas support synchronous replication?

No. Read replicas in Amazon RDS for MySQL, MariaDB, PostgreSQL, and Oracle are implemented using those engines' native asynchronous replication. Amazon Aurora uses a different, but still asynchronous, replication mechanism.

Q: Can I use a read replica to enhance database write availability or protect the data on my source DB instance against failure scenarios?

If you are looking to use replication to increase database write availability and protect recent database updates against various failure conditions, we recommend you run your DB instance as a Multi-AZ deployment. With Amazon RDS Read Replicas, which employ supported engines' native, asynchronous replication, database writes occur on a read replica after they have already occurred on the source DB instance, and this replication “lag” can vary significantly. In contrast, the replication used by Multi-AZ deployments is synchronous, meaning that all database writes are concurrent on the primary and standby. This protects your latest database updates, since they should be available on the standby in the event failover is required. In addition, with Multi-AZ deployments replication is fully managed. Amazon RDS automatically monitors for DB instance failure conditions or Availability Zone failure and initiates automatic failover to the standby (or to a read replica, in the case of Amazon Aurora) if an outage occurs.

Q: Can I create a read replica with a Multi-AZ DB instance deployment as its source?

Yes. Since Multi-AZ DB instances address a different need than read replicas, it makes sense to use the two in conjunction for production deployments and to associate a read replica with a Multi-AZ DB Instance deployment. The “source” Multi AZ-DB instance provides you with enhanced write availability and data durability, and the associated read replica would improve read traffic scalability.

Q: Can I configure my Amazon RDS read replicas themselves Multi-AZ?

Yes. Amazon RDS for MySQL, MariaDB, PostgreSQL, and Oracle allow you to enable Multi-AZ configuration on read replicas to support disaster recovery and minimize downtime from engine upgrades.

Q: If my read replica(s) use a Multi-AZ DB instance deployment as a source, what happens if Multi-AZ failover occurs?

In the event of Multi-AZ failover, any associated and available read replicas will automatically resume replication once failover has completed (acquiring updates from the newly promoted primary).

Q: Can I create a read replica of another read replica?

Amazon Aurora, Amazon RDS for MySQL and MariaDB: You can create a second-tier read replica from an existing first-tier read replica. By creating a second-tier read replica, you may be able to move some of the replication load from the master database instance to a first-tier Read Replica. Please note that a second-tier Read Replica may lag further behind the master because of additional replication latency introduced as transactions are replicated from the master to the first tier replica and then to the second-tier replica.

Amazon RDS for PostgreSQL, Amazon RDS for Oracle: Read Replicas of Read Replicas are not currently supported.

Q: Can my read replicas only accept database read operations?

Read replicas are designed to serve read traffic. However, there may be use cases where advanced users wish to complete Data Definition Language (DDL) SQL statements against a read replica. Examples might include adding a database index to a read replica that is used for business reporting, without adding the same index to the corresponding source DB instance.

Amazon RDS for MySQL can be configured to permit DDL SQL statements against a read replica. If you wish to enable operations other than reads for a given read replica, modify the active DB parameter group for the read replica, setting the “read_only” parameter to “0.”

Amazon RDS for PostgreSQL does not currently support the execution of DDL SQL statements against a read replica.

Q: Can I promote my read replica into a “standalone” DB Instance?

Yes. Refer to the Amazon RDS User Guide for more details.

Q: Will my read replica be kept up-to-date with its source DB instance?

Updates to a source DB instance will automatically be replicated to any associated read replicas. However, with supported engines' asynchronous replication technology, a read replica can fall behind its source DB instance for a variety of reasons. Typical reasons include:

- Write I/O volume to the source DB instance exceeds the rate at which changes can be applied to the read replica (this problem is particularly likely to arise if the compute capacity of a read replica is less than the source DB Instance)

- Complex or long-running transactions to the source DB Instance hold up replication to the read replica
- Network partitions or latency between the source DB instance and a read replica

Read Replicas are subject to the strengths and weaknesses of supported engines' native replication. If you are using Read Replicas, you should be aware of the potential for lag between a Read Replica and its source DB Instance, or “inconsistency”.

Q: How do I see the status of my active read replica(s)?

You can use the standard DescribeDBInstances API to return a list of all the DB Instances you have deployed (including Read Replicas), or simply click on the "Instances" tab of the Amazon RDS Console.

Amazon RDS allows you to gain visibility into how far a read replica has fallen behind its source DB instance. The number of seconds that the read replica is behind the master is published as an Amazon CloudWatch metric ("Replica Lag") available via the AWS Management Console or Amazon CloudWatch APIs. For Amazon RDS for MySQL, the source of this information is the same as that displayed by issuing a standard "Show Slave Status" MySQL command against the read replica. For Amazon RDS for PostgreSQL, you can use the pg_stat_replication view on the source DB instance to explore replication metrics.

Amazon RDS monitors the replication status of your Read Replicas and updates the Replication State field in the AWS Management console to "Error" if replication stops for any reason (e.g., attempting DML queries on your replica that conflict with the updates made on the master database instance could result in a replication error). You can review the details of the associated error thrown by the MySQL engine by viewing the Replication Error field and take an appropriate action to recover from it. You can learn more about troubleshooting replication issues in the Troubleshooting a Read Replica Problem section of the User Guide for Amazon RDS for MySQL or PostgreSQL.

If a replication error is fixed, the Replication State changes to Replicating.

Q: I scaled the compute and/or storage capacity of my source DB instance. Should I scale the resources for associated read replica(s) as well?

For replication to work effectively, we recommend that read replicas have as much or more compute and storage resources as their respective source DB instances. Otherwise replication lag is likely to increase or your read replica may run out of space to store replicated updates.

Q: How do I delete a read replica? Will it be deleted automatically if its source DB Instance is deleted?

You can easily delete a read replica with a few clicks of the AWS Management Console or by passing its DB Instance identifier to the DeleteDBInstance API.

An Amazon Aurora replica will stay active and continue accepting read traffic even after its corresponding source DB Instance has been deleted. One of the replicas in the cluster will automatically be promoted as the new master, and will start accepting write traffic.

An Amazon RDS for MySQL or MariaDB read replica will stay active and continue accepting read traffic even after its corresponding source DB instance has been deleted. If you desire to delete the Read Replica in addition to the source DB instance, you must explicitly do so using the DeleteDBInstance API or AWS Management Console.

If you delete an Amazon RDS for PostgreSQL DB Instance that has read replicas, all Read Replicas will be promoted to standalone DB Instances and will be able to accept both read and write traffic. The newly promoted DB Instances will operate independently of one another. If you desire to delete these DB Instances in addition to the original source DB Instance, you must explicitly do so using the DeleteDBInstance API or AWS Management Console.

Q: How much do read replicas cost? When does billing begin and end?

A read replica is billed as a standard DB Instance and at the same rates. Just like a standard DB instance, the rate per “DB Instance hour” for a read replica is determined by the DB instance class of the read replica – please see pricing page for up-to-date pricing. You are not charged for the data transfer incurred in replicating data between your source DB instance and read replica within the same AWS Region.

Billing for a read replica begins as soon as the replica has been successfully created (i.e. when status is listed as “active”). The read replica will continue being billed at standard Amazon RDS DB instance hour rates until you issue a command to delete it.

Enhanced Monitoring

Q: What is Enhanced Monitoring for RDS?

Enhanced Monitoring for RDS gives you deeper visibility into the health of your RDS instances. Just turn on the “Enhanced Monitoring” option for your RDS DB Instance and set a granularity and Enhanced Monitoring will collect vital operating system metrics and process information, at the defined granularity.

For an even deeper level of diagnostics and visualization of your database load, and a longer data retention period, you can try Performance Insights.

Q: Which metrics and processes can I monitor in Enhanced Monitoring?

Enhanced Monitoring captures your RDS instance system level metrics such as the CPU, memory, file system and disk I/O among others. The complete list of metrics can be found in the documentation.

Q: Which engines are supported by Enhanced Monitoring?

Enhanced Monitoring supports all RDS database engines.

Q: Which instance types are supported by Enhanced Monitoring?

Enhanced Monitoring supports every instance type except t1.micro and m1.small. The software uses a small amount of CPU, memory and I/O and for general purpose monitoring, we recommend switching on higher granularities for instances that are medium or larger. For non-production DB Instances, the default setting for Enhanced Monitoring is “off”, and you have the choice of leaving it disabled or modifying the granularity when it is on.

Q: What information can I view on the RDS dashboard?

You can view all the system metrics and process information for your RDS DB Instances in a graphical format on the console. You can manage which metrics you want to monitor for each instance and customize the dashboard according to your requirements.

Q: Will all the instances in my RDS account sample metrics at the same granularity?

No. You can set different granularities for each DB Instance in your RDS account. You can also choose the instances on which you want to enable Enhanced Monitoring as well as modify the granularity of any instance whenever you want.

Q: How far back can I see the historical metrics on the RDS console?

You can see the performance values for all the metrics up to 1 hour back, at a granularity of up to 1 second, based on your settings.

Q: How can I visualize the metrics generated by RDS Enhanced Monitoring in CloudWatch?

The metrics from RDS Enhanced Monitoring are delivered into your CloudWatch Logs account. You can create metrics filters in CloudWatch from CloudWatch Logs and display the graphs on the CloudWatch dashboard. For more details, please visit the Amazon CloudWatch page.

Q: When should I use CloudWatch instead of the RDS console dashboard?

You should use CloudWatch if you want to view historical data beyond what is available on the RDS console dashboard. You can monitor your RDS instances in CloudWatch to diagnose the health of your entire AWS stack in a single location. Currently, CloudWatch supports granularities of up to 1 minute and the values will be averaged out for granularities less than that.

Q: Can I set up alarms and notifications based on specific metrics?

Yes. You can create an alarm in CloudWatch that sends a notification when the alarm changes state. The alarm watches a single metric over a time period that you specify, and performs one or more actions based on the value of the metric relative to the specified threshold over a number of time periods. For more details on CloudWatch alarms, please visit the Amazon CloudWatch Developer Guide.

Q: How do I integrate Enhanced Monitoring with my tool that I currently use?

RDS Enhanced Monitoring provides a set of metrics formed as JSON payloads which are delivered into your CloudWatch Logs account. The JSON payloads are delivered at the granularity last configured for the RDS instance.

There are two ways you can consume the metrics via a third-party dashboard or application. Monitoring tools can use CloudWatch Logs Subscriptions to set up a near real time feed for the metrics. Alternatively, you can use filters in CloudWatch Logs to bridge metrics across to CloudWatch to and integrate your application with CloudWatch. Please visit Amazon CloudWatch Documentation for more details.

Q: How can I delete historical data?

Since Enhanced Monitoring delivers JSON payloads into a log in your CloudWatch Logs account, you can control its retention period just like any other CloudWatch Logs stream. The default retention period configured for Enhanced Monitoring in CloudWatch Logs is 30 days. For details on how to change retention settings, please visit Amazon CloudWatch Developer Guide.

Q: What impact does Enhanced Monitoring have on my monthly bills?

Since the metrics are ingested into CloudWatch Logs, your charges will be based on CloudWatch Logs data transfer and storage rates once you exceed CloudWatch Logs free tier. Pricing details can be found here. The amount of information transferred for an RDS instance is directly proportional to the defined granularity for the Enhanced Monitoring feature. Administrators can set different granularities for different instances in their accounts to manage costs.

The approximate volume of data ingested into CloudWatch Logs by Enhanced Monitoring for an instance is as shown below:

Granularity	60 seconds	30 seconds	15 seconds	10 seconds	5 seconds	1 second
Data ingested in CloudWatch Logs* (GB per month)	0.27	0.53	1.07	1.61	3.21	16.07

Amazon RDS Proxy

Q: What is Amazon RDS Proxy?

Amazon RDS Proxy is a fully managed, highly available database proxy feature for Amazon RDS. RDS Proxy makes applications more scalable, more resilient to database failures, and more secure.

Q: Why would I use Amazon RDS Proxy?

Amazon RDS Proxy is a fully managed, highly available, and easy-to-use database proxy feature of Amazon RDS that enables your applications to: 1) improve scalability by pooling and sharing database connections; 2) improve availability by reducing database failover times by up to 66% and preserving application connections during failovers; and 3) improve security by optionally enforcing AWS IAM authentication to databases, and securely storing credentials in AWS Secrets Manager.

Q: What use cases does Amazon RDS Proxy address?

Amazon RDS Proxy addresses a number of use cases related to scalability, availability, and security of your applications, including:

Applications with unpredictable workloads: Applications that support highly variable workloads may attempt to open a burst of new database connections.

Amazon RDS Proxy’s connection governance allows you to gracefully scale applications dealing with unpredictable workloads by efficiently reusing database connections. First, RDS proxy enables multiple application connections to share a database connection for efficient use of database resources. Second, RDS Proxy allows you to maintain predictable database performance by regulating the number of database connections that are opened. Third, RDS Proxy removes requests that cannot be served to preserve overall performance and availability of the application.

Applications that frequently open and close database connections: Applications built on technologies such as Serverless, PHP, or Ruby on Rails may open and close database connections frequently to serve application requests. Amazon RDS Proxy maintains a pool of database connections to avoid unnecessary stress on database compute and memory for establishing new connections.

Applications that keep connections open but idle: Applications in industries such as SaaS or eCommerce may keep database connections idling to minimize the response time when a customer reengages. Instead of overprovisioning databases to support mostly idling connections, you can use Amazon RDS Proxy to hold idling connections while only establishing database connections as required to optimally serve active requests.

Applications requiring availability through transient failures: With Amazon RDS Proxy, you can build applications that can transparently tolerate database failures without needing to write complex failure handling code. RDS Proxy automatically routes traffic to a new database instance while preserving application connections. RDS Proxy also bypasses Domain Name System (DNS) caches to reduce failover times by up to 66% for RDS and Aurora Multi-AZ databases. During database failovers, the application may experience increased latencies and ongoing transactions may have to be retried.

Improved security and centralized credentials management: Amazon RDS Proxy aids you in building more secure applications by giving you a choice to enforce IAM based authentication with relational databases. RDS Proxy also enables you to centrally manage database credentials through AWS Secrets Manager.

Q: When should I connect to the database directly versus using Amazon RDS Proxy?

Depending on your workload, Amazon RDS Proxy can add an average of 5 milliseconds of network latency to query or transaction response time. If your application cannot tolerate 5 milliseconds of latency or does not need connection management and other features enabled by RDS Proxy, you may want your application to connect directly to the database endpoint.

Q: How will serverless applications benefit from Amazon RDS Proxy?

Amazon RDS Proxy transforms your approach to building modern serverless applications that leverage the power and simplicity of relational databases. First, RDS Proxy enables serverless applications to scale efficiently by pooling and reusing database connections. Second, with RDS Proxy, you no longer need to handle database credentials in your Lambda code. You can use the IAM execution role associated with your Lambda function to authenticate with RDS Proxy and your database. Third, you don't need to manage any new infrastructure or code to utilize the full potential of serverless applications backed by relational databases. RDS Proxy is fully managed and scales its capacity automatically based on your application demands.

Q: Which database engines does Amazon RDS Proxy support?

The Amazon RDS Proxy preview is available for Amazon Aurora with MySQL compatibility and Amazon RDS for MySQL. Support for additional database engines will come soon.

Q: How can I enable Amazon RDS Proxy?

You enable Amazon RDS Proxy for your RDS database with just a few clicks in the RDS console. While enabling RDS Proxy, you specify the VPC and subnets you want to access RDS Proxy from. As a Lambda user, you can enable RDS Proxy for your RDS database and set up a Lambda function to access it with just a few clicks in the Lambda console. You can learn more about getting started in the Amazon RDS User Guide.

Q: Can I access Amazon RDS Proxy using APIs?

- Yes. You can use Amazon RDS Proxy APIs to create a proxy and then define target groups to associate the proxy with specific database instances or clusters. For example:

```
aws rds create-db-proxy
        --db-proxy-name '…'
        --engine-family <mysql|postgresql>
        --auth [{}, {}]
        --role-arn '…'
        --subnet-ids {}
        --require-tls <true|false>
        --tags {}
aws rds register-db-proxy-targets
        --target-group-name '…'
        --db-cluster-identifier  '…'
        --db-instance-identifier '…'
```

Amazon RDS on VMware FAQs

General

Q: What is RDS on VMware?

A: RDS on VMware delivers AWS-managed relational databases in on-premises VMware environments. Managing relational databases is complex and time-consuming, and RDS on VMware makes it easy to setup, scale, and operate relational databases in your VMware vSphere clusters. RDS on VMware uses the same simple interface, AWS Management Console, to manage databases both on-premises and in AWS.

Q: Which relational database engines does RDS on VMware support?

A: RDS on VMware supports MySQL 5.7, PostgreSQL 10.9, and Microsoft SQL Server 2016 SP2 Enterprise Edition. For Microsoft SQL Server, you will need to provide your own media and license (on-premises customer-provided license).

Q: What are the monitoring capabilities of this solution?

A: You can use Amazon CloudWatch to monitor database metrics in RDS on VMware. (Normal charges for Amazon CloudWatch will apply).

Q: Does all my data residing in RDS on VMware stay on-premises?

A: Yes, data in your databases managed by RDS on VMware stays on-premises.

Q: Does RDS on VMware support AWS Direct Connect?

A: Yes, RDS on VMware supports connecting to the AWS Region over AWS Direct Connect.

Q: Does Amazon RDS on VMware support Amazon Aurora?

A: No. RDS on VMware does not support Amazon Aurora.

Q: Are there CLI/API available to access RDS on VMware?

A: Yes, you can use the same AWS CLI and RDS APIs to manage RDS databases running in AWS and in your on-premises VMware environment.

Q: What AWS Regions is RDS on VMware available in?

A: RDS on VMware is available in AWS US East (Northern Virginia) Region.

Getting Started

Q: What are the prerequisites to use this service?

- vSphere v6.5 or higher VMware vSphere Enterprise Plus Edition
- AWS Account
- AWS Business or Enterprise support
- For more details on prerequisites, please refer to RDS on VMware User's Guide

Q: Can I use my own Microsoft SQL Server license with RDS on VMware?

A: Yes, you need to bring your own media and your own Microsoft SQL Server license (on-premises customer provided license) to create RDS on VMware managed Microsoft SQL Server databases.

Q: How do I onboard RDS on VMware onto my vSphere cluster?

A: To onboard RDS on VMware, you create a Custom Availability Zone from the AWS Management Console. You then download the RDS on VMware installer from the AWS Management Console to the on-premises vSphere cluster where you want to use the service. When you run the installer, it deploys the local components for RDS on VMware on your vSphere cluster and connects your cluster to the Amazon RDS service running in the AWS Region. Once this process is complete, your newly created Custom Availability Zone will become

“Active”. You can then create a new database using the AWS Management Console, CLI, or APIs by choosing the appropriate database engine and instance size. Please note that for Microsoft SQL Server, customers have to provide their own media and on-premises license to create the database.

Q: Where do I go to create or manage a database?

A: You can create, modify, and manage your RDS on VMware databases using the same AWS Management Console, CLI, and APIs as you do with Amazon RDS in AWS.

How it Works

Q: How does RDS on VMware help manage my on-premises databases?

A: RDS on VMware is comprised of a set of VMs running on your vSphere infrastructure, connected through a dedicated VPN tunnel to the AWS Region. This service provides a single pane of glass experience via the AWS Management Console, CLI, and APIs to manage RDS databases running on-premises and in AWS. This service helps you to automatically perform the common database management tasks, including database provisioning, operating system and database patching, backup, point-in-time restore, compute scaling, instance health monitoring, and failover, freeing you to focus on your applications.

Q: Can RDS on VMware perform my backups and, if necessary, my restores?

A: Yes, RDS on VMware allows you to configure automated daily backups and to take user-initiated backups on-demand (DB snapshots). You can restore the database instance from a DB snapshot for all supported engines or to a specific point-in-time for MySQL and PostgreSQL.

RDS on VMware allows you to specify automated backup retention period of up to 35 days for each database in your fleet. It is important to note that the RDS on VMware restore mechanism does not perform in-place replacement of existing databases. RDS on VMware creates a new database instance and restores your data to new volumes, allowing you to decide the best path forward in your specific situation.

Q: Does RDS on VMware provide availability protection for my databases?

A: Yes, every RDS instance benefits from local, on-premises health monitoring. When RDS on VMware detects that the database instance is unhealthy, the service replaces the unhealthy database instance. Customers can continue to access their databases using the same FQDN.

Q: Does RDS on VMware patch my databases?

A: Yes. RDS on VMware takes care of both OS and database engine patching with minimal downtime. Patching happens during a configurable maintenance window for database instances.

Q: What happens if the network connection from my vSphere cluster to the AWS Region is disrupted?

A: Your database instance availability is not impacted due to disruption in the network connection to the AWS Region from your vSphere cluster. In case of VPN disconnectivity, you will not be able to initiate any new database management operations using the AWS Management Console, CLI, or APIs. If AWS services are not reachable, database monitoring metrics will not be sent to Amazon CloudWatch.

Q: Will Amazon RDS on VMware support VADP (VMware's vStorage API for Data Protection)?

A: No. RDS on VMware does not support VADP or any other external backup solution.

Q: Will RDS on VMware support customers using VMware NSX?

A: Yes, RDS on VMware works in a VMware NSX environment.

Q: Will RDS on VMware support vSAN?

A: Yes, RDS on VMware supports vSphere Cluster running vSAN.

Q: Can I migrate my existing database data to RDS on VMware database instances?

A: To migrate existing data from SQL Server databases, you can use the database migration tools provided by your database vendor to perform a migration.

Q: Does RDS on VMware support Amazon RDS resource tags?

A: Yes, RDS on VMware supports resource tags.

Help and Support

Q: Will AWS support my managed environment and databases?

A: Yes. You can engage with AWS Support for any issues with RDS on VMware, just as you would for any other AWS service.

Amazon Redshift FAQs

General

Learn about what's new with Amazon Redshift on the What's New page.

View more detailed information and usage guidance in the Documentation.

Q: What is Amazon Redshift?
Amazon Redshift is a fast, fully managed data warehouse that makes it simple and cost-effective to analyze all your data using standard SQL and your existing Business Intelligence (BI) tools. It allows you to run complex analytic queries against petabytes of structured data using sophisticated query optimization, columnar storage on high-performance storage, and massively parallel query execution. Most results come back in seconds. With Redshift, you can start small for just $0.25 per hour with no commitments and scale out to petabytes of data for $1,000 per terabyte per year, less than a tenth the cost of traditional solutions. Amazon Redshift also includes Amazon Redshift Spectrum, allowing you to directly run SQL queries against exabytes of unstructured data in Amazon S3 data lakes. No loading or transformation is required, and you can use open data formats, including Avro, CSV, Ion, JSON, ORC, Parquet, and more. Redshift Spectrum automatically scales query compute capacity based on the data being retrieved, so queries against Amazon S3 run fast, regardless of data set size.

Amazon Redshift gives you fast querying capabilities over structured data using familiar SQL-based clients and business intelligence (BI) tools using standard ODBC and JDBC connections. Queries are distributed and parallelized across multiple physical resources. You can easily scale an Amazon Redshift data warehouse up or down with a few clicks in the AWS Management Console or with a single API call. Amazon Redshift automatically patches and backs up your data warehouse, storing the backups for a user-defined retention period.

Amazon Redshift uses replication and continuous backups to enhance availability and improve data durability and can automatically recover from component and node failures. In addition, Amazon Redshift supports Amazon Virtual Private Cloud (Amazon VPC), SSL, AES-256 encryption, and Hardware Security Modules (HSMs) to protect your data in transit and at rest.

As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources you use. Amazon Redshift lets you pay as you go. You can even try Amazon Redshift for free.

For information about Amazon Redshift regional availability, see the AWS Region Table.

Q: Why would I use Amazon Redshift over an on-premises data warehouse?

On-premises data warehouses require significant time and resource to administer, especially for large datasets. In addition, the financial cost associated with building, maintaining, and growing self-managed, on-premise data warehouses is very high. As your data grows, you have to constantly trade-off what data to load into your data warehouse and what data to archive in storage so you can manage costs, keep ETL complexity low, and deliver good performance. Amazon Redshift not only significantly lowers the cost and operational overhead of a data warehouse, but with Redshift Spectrum, it also makes it easy to analyze large amounts of data in its native format without requiring you to load the data.

Q: How do I use Amazon Redshift features that are in preview?

When creating a Amazon Redshift cluster you can pick three tracks for maintenance: Current, Trailing, or Preview. Within the Preview track, PREVIEW_FEATURES should be selected to use Redshift features that are available in preview.

Q: What is AQUA (Advanced Query Accelerator) for Amazon Redshift?

AQUA is a new distributed and hardware-accelerated cache for Redshift. Learn more and sign up to be considered for the preview.

Q: What is Redshift Spectrum?

Redshift Spectrum is a feature of Amazon Redshift that enables you to run queries against exabytes of unstructured data in Amazon S3, with no loading or ETL required. When you issue a query, it goes to the Amazon Redshift SQL endpoint, which generates and optimizes a query plan. Amazon Redshift determines what data is local and what is in Amazon S3, generates a plan to minimize the amount of Amazon S3 data that needs to be read, requests Redshift Spectrum workers out of a shared resource pool to read and process data from Amazon S3.

Redshift Spectrum scales out to thousands of instances if needed, so queries run quickly regardless of data size. In addition, you can use the exact same SQL for Amazon S3 data as you do for your Amazon Redshift queries today and connect to the same Amazon Redshift endpoint using your same BI tools. Redshift Spectrum lets you separate storage and compute, allowing you to scale each independently. You can setup as many Amazon Redshift clusters as you need to query your Amazon S3 data lake, providing high availability and limitless concurrency. Redshift Spectrum gives you the freedom to store your data where you want, in the format you want, and have it available for processing when you need it.

For information about Redshift Spectrum regional availability, please visit the Amazon Redshift pricing page.

Q: What does Amazon Redshift manage on my behalf?

Amazon Redshift manages the work needed to set up, operate, and scale a data warehouse, from provisioning the infrastructure capacity to automating ongoing administrative tasks such as backups, and patching. Amazon Redshift automatically monitors your nodes and drives to help you recover from failures. For Redshift Spectrum, Amazon Redshift manages all the computing infrastructure, load balancing, planning, scheduling and execution of your queries on data stored in Amazon S3.

Q: How does the performance of Amazon Redshift compare to most traditional databases for data warehousing and analytics?

Amazon Redshift uses a variety of innovations to achieve up to ten times higher performance than traditional databases for data warehousing and analytics workloads:

- *Columnar Data Storage:* Instead of storing data as a series of rows, Amazon Redshift organizes the data by column. Unlike row-based systems, which are ideal for transaction processing, column-based systems are ideal for data warehousing and analytics, where queries often involve aggregates performed over large data sets. Since only the columns involved in the queries are processed and columnar data is stored sequentially on the storage media, column-based systems require far fewer I/Os, greatly improving query performance.
- *Advanced Compression:* Columnar data stores can be compressed much more than row-based data stores because similar data is stored sequentially on disk. Amazon Redshift employs multiple compression techniques and can often achieve significant compression relative to traditional relational data stores. When loading data into an empty table, Amazon Redshift automatically samples your data and selects the most appropriate compression scheme.
- *Massively Parallel Processing (MPP):* Amazon Redshift automatically distributes data and query load across all nodes. Amazon Redshift makes it easy to add nodes to your data warehouse and enables you to maintain fast query performance as your data warehouse grows.
- *Redshift Spectrum:* Redshift Spectrum enables you to run queries against exabytes of data in Amazon S3. There is no loading or ETL required. Even if you don’t store any of your data in Amazon Redshift, you can still use Redshift Spectrum to query datasets as large as an exabyte in Amazon S3. When you issue a query, it goes to the Amazon Redshift SQL endpoint, which generates the query plan. Amazon Redshift determines what data is local and what is in Amazon S3, generates a plan to minimize the amount of Amazon S3 data that needs to be read, requests Redshift Spectrum workers out of a shared resource pool to read and process data from Amazon S3, and pulls results back into your Amazon Redshift cluster for any remaining processing.

Q: How do I get started with Amazon Redshift?

You can sign up and get started within minutes from the Amazon Redshift detail page or via the AWS Management Console. If you don't already have an AWS account, you'll be prompted to create one.

To use the Redshift Spectrum feature, you need to first store your data in Amazon S3. You can then define the metadata about that data in your Amazon Redshift cluster or register the metadata you may already have in your Hive metastore with your cluster. You can issue a CREATE EXTERNAL SCHEMA SQL command in your Amazon Redshift cluster to define or register a database in your catalog as an external schema within Amazon Redshift. You can then issue queries against Amazon S3 using the same SQL you use for local tables and any BI tool that supports Amazon Redshift today. The external database definition you create using Amazon Redshift SQL is registered in the same data catalog that Amazon Athena uses. You can optionally manage the external database definition from the Amazon Athena Catalog as well.

Visit our Getting Started page to see how to try Amazon Redshift for free.

Q: How do I create and access an Amazon Redshift data warehouse cluster?

You can easily create an Amazon Redshift data warehouse cluster by using the AWS Management Console or the Amazon Redshift APIs. You can start with a single node, 160GB data warehouse and scale all the way to petabytes or more with a few clicks in the AWS Console or a single API call.

The single node configuration enables you to get started with Amazon Redshift quickly and cost-effectively and scale up to a multi-node configuration as your needs grow. A Redshift data warehouse cluster can contain from 1-128 compute nodes, depending on the node type. For details, please see our documentation.

The multi-node configuration requires a leader node that manages client connections and receives queries, and two compute nodes that store data and perform queries and computations. The leader node is provisioned for you automatically and you are not charged for it.

Simply specify your preferred Availability Zone (optional), the number of nodes, node types, a master name and password, security groups, your preferences for backup retention, and other system settings. Once you've chosen your desired configuration, Amazon Redshift will provision the required resources and set up your data warehouse cluster.

Once your data warehouse cluster is available, you can retrieve its endpoint and JDBC and ODBC connection string from the AWS Management Console or by using the Redshift APIs. You can then use this connection string with your favorite database tool, programming language, or Business Intelligence (BI) tool. You will need to authorize network requests to your running data warehouse cluster. For a detailed explanation, please refer to our Getting Started Guide.

Q: What is the maximum storage capacity per compute node? What is the recommended amount of data per compute node for optimal performance?

You can create a cluster using either RA3, DC, or DS node types. RA3 node types enable you to scale and pay for compute and storage independently. You choose the number of instances you need based on performance requirements, and only pay for the managed storage that you use.

RA3 is available now as RA3.16XL, which allows building a cluster with up to 8 petabytes in managed storage. With RA3, customers pay for the actual storage they use. The RA3 clusters run with minimum 2 nodes and the minimum sized RA3.16XL cluster can support up-to 128 TB. Redshift managed storage uses large, high-performance SSDs in each Amazon Redshift RA3 instance for fast local storage and Amazon S3 for longer-term durable storage. If the data in an instance grows beyond the size of the large local SSDs, Redshift managed storage automatically offloads that data to Amazon S3. Customers pay the same low rate for Redshift managed storage regardless of whether the data sits in high performance SSDs or in Amazon S3. For workloads that require a lot of storage, but not as much compute capacity, this lets customers automatically scale their data warehouse storage capacity without adding and paying for additional instances

DC node types are available in two sizes. The Large has 160GB of SSD storage, 2 Intel Xeon E5-2670v2 (Ivy Bridge) virtual cores and 15GiB of RAM. The Eight Extra Large is sixteen times bigger with 2.56TB of SSD storage, 32 Intel Xeon E5-2670v2 virtual cores, and 244GiB of RAM. You can get started with a single

DC2.Large node for $0.25 per hour and scale all the way up to 128 8XL nodes with 326TB of SSD storage, 3,200 virtual cores, and 24TiB of RAM.

DS node types are available in two sizes, Extra Large and Eight Extra Large. The Extra Large (XL) has 3 HDDs with a total of 2TB of magnetic storage, whereas Eight Extra Large (8XL) has 24 HDDs with a total of 16TB of magnetic storage. DS2.8XLarge has 36 Intel Xeon E5-2676 v3 (Haswell) virtual cores and 244GiB of RAM, and DS2.XL has 4 Intel Xeon E5-2676 v3 (Haswell) virtual cores, and 31GiB of RAM.

Please see our pricing page for more detail.

Q: When would I use Amazon Redshift vs. Amazon RDS?

Both Amazon Redshift and Amazon RDS enable you to run traditional relational databases in the cloud while offloading database administration. Customers use Amazon RDS databases primarily for online-transaction processing (OLTP) workload while Redshift is used primarily for reporting and analytics. Amazon Redshift harnesses the scale and resources of multiple nodes and uses a variety of optimizations to provide order of magnitude improvements over traditional databases for analytic and reporting workloads against very large data sets. Amazon Redshift provides an excellent scale-out option as your data and query complexity grows if you want to prevent your reporting and analytic processing from interfering with the performance of your OLTP workload. Now, with the new Federated Query feature (preview), you can easily query data across your Amazon RDS or Aurora database services with Redshift.

Q: When would I use Amazon Redshift or Redshift Spectrum vs. Amazon EMR?

You should use Amazon EMR if you use custom code to process and analyze extremely large datasets with big data processing frameworks such as Apache Spark, Hadoop, Presto, or Hbase. Amazon EMR gives you full control over the configuration of your clusters and the software you install on them.

Data warehouses like Amazon Redshift are designed for a different type of analytics altogether. Data warehouses are designed to pull together data from lots of different sources, like inventory, financial, and retail sales systems. In order to ensure that reporting is consistently accurate across the entire company, data warehouses store data in a highly structured fashion. This structure builds data consistency rules directly into the tables of the database. Amazon Redshift is the best service to use when you need to perform complex queries on massive collections of structured and semi-structured data and get super fast performance.

While the Redshift Spectrum feature is great for running queries against data in Amazon Redshift and S3, it really isn't a fit for the types of use cases that enterprises typically ask from processing frameworks like Amazon EMR. Amazon EMR goes far beyond just running SQL queries. Amazon EMR is a managed service that lets you process and analyze extremely large data sets using the latest versions of popular big data processing frameworks, such as Spark, Hadoop, and Presto, on fully customizable clusters. With Amazon EMR, you can run a wide variety of scale-out data processing tasks for applications such as machine learning, graph analytics, data transformation, streaming data, and virtually anything you can code.

You can use Redshift Spectrum together with EMR. Redshift Spectrum uses the same approach to store table definitions as Amazon EMR. Redshift Spectrum can support the same Apache Hive Metastore used by Amazon EMR to locate data and table definitions. If you're using Amazon EMR and have a Hive Metastore already, you just have to configure your Amazon Redshift cluster to use it. You can then start querying that data right away along with your Amazon EMR jobs. Therefore, if you're already using EMR to process a large data store, you can use Redshift Spectrum to query that data right at the same time without interfering with your Amazon EMR jobs.

Query services, data warehouses, and complex data processing frameworks all have their place, and they are used for different things. You just need to choose the right tool for the job.

Q: When should I use Amazon Athena vs. Redshift Spectrum?

Amazon Athena is the simplest way to give any employee the ability to run ad-hoc queries on data in Amazon S3. Athena is serverless, so there is no infrastructure to setup or manage, and you can start analyzing your data immediately.

If you have frequently accessed data, that needs to be stored in a consistent, highly structured format, then you should use a data warehouse like Amazon Redshift. This gives you the flexibility to store your structured, frequently accessed data in Amazon Redshift, and use Redshift Spectrum to extend your Amazon Redshift queries out to the entire universe of data in your Amazon S3 data lake. This gives you the freedom to store your data where you want, in the format you want, and have it available for processing when you need.

Q: Why should I use Amazon Redshift instead of running my own MPP data warehouse cluster on Amazon EC2?

Amazon Redshift automatically handles many of the time-consuming tasks associated with managing your own data warehouse including:

- *Setup:* With Amazon Redshift, you simply create a data warehouse cluster, define your schema, and begin loading and querying your data. Provisioning, configuration and patching are all managed for you.
- *Data Durability:* Amazon Redshift replicates your data within your data warehouse cluster and continuously backs up your data to Amazon S3, which is designed for eleven nines of durability. Amazon Redshift mirrors each drive's data to other nodes within your cluster. If a drive fails, your queries will continue with a slight latency increase while Redshift rebuilds your drive from replicas. In case of node failure(s), Amazon Redshift automatically provisions new node(s) and begins restoring data from other drives within the cluster or from Amazon S3. It prioritizes restoring your most frequently queried data so your most frequently executed queries will become performant quickly.
- *Scaling:* You can add or remove nodes from your Amazon Redshift data warehouse cluster with a single API call or via a few clicks in the AWS Management Console as your capacity and performance needs change. You can also schedule your scaling and resize operations by using the scheduler capability in Redshift.
- *Automatic Updates and Patching:* Amazon Redshift automatically applies upgrades and patches your data warehouse so you can focus on your application and not on its administration.
- *Exabyte Scale Query Capability:* Redshift Spectrum enables you to run queries against exabytes of data in Amazon S3. There is no loading or ETL

required. Even if you don’t store any of your data in Amazon Redshift, you can still use Redshift Spectrum to query datasets as large as an exabyte in Amazon S3.

Billing

Q: How will I be charged and billed for my use of Amazon Redshift?

You pay only for what you use, and there are no minimum or setup fees. Billing commences for a data warehouse cluster as soon as the data warehouse cluster is available. Billing continues until the data warehouse cluster terminates, which would occur upon deletion or in the event of instance failure. You are billed based on:

- *Compute node hours:* Compute node hours are the total number of hours you run across all your compute nodes for the billing period. Node usage hours are billed for each hour your data warehouse cluster is running in an available state. If you no longer wish to be charged for your data warehouse cluster, you must terminate it to avoid being billed for additional node hours. Partial node hours consumed are billed as full hours. You are billed for 1 unit per node per hour, so a 3-node data warehouse cluster running persistently for an entire month would incur 2,160 instance hours. You will not be charged for leader node hours; only compute nodes will incur charges.
- *Backup Storage:* Backup storage is the storage associated with your automated and manual snapshots for your data warehouse. Increasing your backup retention period or taking additional snapshots increases the backup storage consumed by your data warehouse. There is no additional charge for backup storage up to 100% of your provisioned storage for an active data warehouse cluster. For example, if you have an active Single Node XL data warehouse cluster with 2TB of local instance storage, we will provide up to 2TB-Month of backup storage at no additional charge. Backup storage beyond the provisioned storage size and backups stored after your cluster is terminated are billed at standard Amazon S3 rates.
- *Data transfer:* There is no data transfer charge for data transferred to or from Amazon Redshift and Amazon S3 within the same AWS Region. For all

other data transfers into and out of Amazon Redshift, you will be billed at standard AWS data transfer rates.

- *Data scanned:* With Redshift Spectrum, you are charged for the amount of Amazon S3 data scanned to execute your query. There are no charges for Redshift Spectrum when you're not running queries. If you store data in a columnar format, such as Parquet or RC, your charges will go down as Redshift Spectrum will only scan the columns needed by the query, rather than processing entire rows. Similarly, if you compress your data, using one of Redshift Spectrum's supported formats, your costs will also go down. You pay the standard Amazon S3 rates for data storage and Amazon Redshift instance rates for the cluster used.

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

For Amazon Redshift pricing information, please visit the Amazon Redshift pricing page.

Data Integration and Loading

Q: How do I load data into my Amazon Redshift data warehouse?

You can load data into Amazon Redshift from a range of data sources including Amazon S3, Amazon DynamoDB, Amazon EMR, AWS Glue, AWS Data Pipeline and or any SSH-enabled host on Amazon EC2 or on-premises. Amazon Redshift attempts to load your data in parallel into each compute node to maximize the rate at which you can ingest data into your data warehouse cluster. Clients can connect to Amazon Redshift using ODBC or JDBC and issue 'insert' SQL commands to insert the data. Please note this is slower than using S3 or DynamoDB since those methods load data in parallel to each compute node while SQL insert statements load via the single leader node. For more details on loading data into Amazon Redshift, please view our Getting Started Guide.

Q: How do I load data from my existing Amazon RDS, Amazon EMR, Amazon DynamoDB, and Amazon EC2 data sources to Amazon Redshift?

You can use our COPY command to load data in parallel directly to Amazon Redshift from Amazon EMR, Amazon DynamoDB, or any SSH-enabled host. Redshift Spectrum also enables you to load data from Amazon S3 into your cluster with a simple INSERT INTO command. This could enable you to load data from various formats such as Parquet and RC into your cluster. Note that if you use this approach, you will accrue Redshift Spectrum charges for the data scanned from Amazon S3. The Redshift Federated Query (Preview) feature enables you to combine data from your Amazon RDS and Aurora (PostgreSQL).

In addition, many ETL companies have certified Amazon Redshift for use with their tools, and a number are offering free trials to help you get started loading your data. AWS Data Pipeline provides a high performance, reliable, fault tolerant solution to load data from a variety of AWS data sources. You can use AWS Data Pipeline to specify the data source, desired data transformations, and then execute a pre-written import script to load your data into Amazon Redshift. In addition, AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy to prepare and load data for analytics. You can create and run an AWS Glue ETL job with a few clicks in the AWS Management Console.

Q: I have a lot of data for initial loading into Amazon Redshift. Transferring via the Internet would take a long time. How do I load this data?

You can use AWS Import/Export to transfer the data to Amazon S3 using portable storage devices. In addition, you can use AWS Direct Connect to establish a private network connection between your network or datacenter and AWS. You can choose 1Gbit/sec or 10Gbit/sec connection ports to transfer your data.

Security

Q: How does Amazon Redshift keep my data secure?

Amazon Redshift encrypts and keeps your data secure in transit and at rest using industry-standard encryption techniques. To keep data secure in transit, Amazon Redshift supports SSL-enabled connections between your client application and your Redshift data warehouse cluster. To keep your data secure at rest, Amazon Redshift encrypts each block using hardware-accelerated AES-256 as it is written to disk. This takes place at a low level in the I/O subsystem, which encrypts everything written to disk, including intermediate query results. The blocks are backed up as is, which means that backups are encrypted as well. By default, Amazon Redshift takes care of key management but you can choose to manage your keys using your own hardware security modules (HSMs) or manage your keys through AWS Key Management Service.

Redshift Spectrum supports Amazon S3's Server Side Encryption (SSE) using your account's default key managed used by the AWS Key Management Service (KMS).

Q: Can I use Amazon Redshift in Amazon Virtual Private Cloud (Amazon VPC)?

Yes, you can use Amazon Redshift as part of your VPC configuration. With Amazon VPC, you can define a virtual network topology that closely resembles a traditional network that you might operate in your own datacenter. This gives you complete control over who can access your Amazon Redshift data warehouse cluster. You can use Redshift Spectrum with an Amazon Redshift cluster that is part of your VPC.

Q: Can I access my Amazon Redshift compute nodes directly?

No. Your Amazon Redshift compute nodes are in a private network space and can only be accessed from your data warehouse cluster's leader node. This provides an additional layer of security for your data.

Availability and Durability

Q: What happens to my data warehouse cluster availability and data durability if a drive on one of my nodes fails?

Your Amazon Redshift data warehouse cluster will remain available in the event of a drive failure however you may see a slight decline in performance for certain queries. In the event of a drive failure, Amazon Redshift will transparently use a replica of the data on that drive which is stored on other drives within that node. In addition, Amazon Redshift will attempt to move your data to a healthy drive or will replace your node if it is unable to do so. Single node clusters do not support data replication. In the event of a drive failure, you will need to restore the cluster from snapshot on S3. We recommend using at least two nodes for production.

Q: What happens to my data warehouse cluster availability and data durability in the event of individual node failure?

Amazon Redshift will automatically detect and replace a failed node in your data warehouse cluster. The data warehouse cluster will be unavailable for queries and updates until a replacement node is provisioned and added to the DB. Amazon Redshift makes your replacement node available immediately and loads your most frequently accessed data from S3 first to allow you to resume querying your data as quickly as possible. Single node clusters do not support data replication. In the event of a drive failure, you will need to restore the cluster from snapshot on S3. We recommend using at least two nodes for production.

Q: What happens to my data warehouse cluster availability and data durability if my data warehouse cluster's Availability Zone (AZ) has an outage?

If your Amazon Redshift data warehouse cluster's Availability Zone becomes unavailable, you will not be able to use your cluster until power and network access to the AZ are restored. Your data warehouse cluster's data is preserved so you can start using your Amazon Redshift data warehouse as soon as the AZ becomes available again. In addition, you can also choose to restore any existing snapshots to a new AZ in the same Region. Amazon Redshift will restore your most frequently accessed data first so you can resume queries as quickly as possible.

Q: Does Amazon Redshift support Multi-AZ Deployments?

Currently, Amazon Redshift only supports Single-AZ deployments. You can run data warehouse clusters in multiple AZ's by loading data into two Amazon Redshift data warehouse clusters in separate AZs from the same set of Amazon S3 input files. With Redshift Spectrum, you can spin up multiple clusters across AZs and access data in Amazon S3 without having to load it into your cluster. In addition, you can also restore a data warehouse cluster to a different AZ from your data warehouse cluster snapshots.

Backup and Restore

Q: How does Amazon Redshift back up my data? How do I restore my cluster from a backup?

Amazon Redshift replicates all your data within your data warehouse cluster when it is loaded and also continuously backs up your data to S3. Amazon Redshift always attempts to maintain at least three copies of your data (the original and replica on the compute nodes and a backup in Amazon S3). Redshift can also asynchronously replicate your snapshots to S3 in another region for disaster recovery.

By default, Amazon Redshift enables automated backups of your data warehouse cluster with a 1-day retention period. You can configure this to be as long as 35 days.

Free backup storage is limited to the total size of storage on the nodes in the data warehouse cluster and only applies to active data warehouse clusters. For example, if you have total data warehouse storage of 8TB, we will provide at most 8TB of backup storage at no additional charge. If you would like to extend your backup retention period beyond one day, you can do so using the AWS Management Console or the Amazon Redshift APIs. For more information on automated snapshots, please refer to the Amazon Redshift Management Guide. Amazon Redshift only backs up data that has changed so most snapshots only use up a small amount of your free backup storage.

When you need to restore a backup, you have access to all the automated backups within your backup retention window. Once you choose a backup from which to restore, we will provision a new data warehouse cluster and restore your data to it.

Q: How do I manage the retention of my automated backups and snapshots?

You can use the AWS Management Console or ModifyCluster API to manage the period of time your automated backups are retained by modifying the RetentionPeriod parameter. If you wish to turn off automated backups altogether, you can set up the retention period to 0 (not recommended).

Q: What happens to my backups if I delete my data warehouse cluster?

When you delete a data warehouse cluster you have the ability to specify whether a final snapshot is created upon deletion. This enables a restore of the deleted data warehouse cluster at a later date. All previously created manual snapshots of your data warehouse cluster will be retained and billed at standard Amazon S3 rates, unless you choose to delete them.

Scalability

Q: How do I scale the size and performance of my Amazon Redshift data warehouse cluster?

If you would like to increase query performance or respond to CPU, memory or I/O over-utilization, you can increase the number of nodes within your data warehouse cluster using Elastic Resize via the AWS Management Console or the ModifyCluster API. When you modify your data warehouse cluster, your requested changes will be applied immediately. Metrics for compute utilization, storage utilization, and read/write traffic to your Amazon Redshift data warehouse cluster are available free of charge via the AWS Management Console or Amazon CloudWatch APIs. You can also add additional, user-defined metrics via Amazon Cloudwatch custom metric functionality.

With the Concurrency Scaling feature, you can support virtually unlimited concurrent users and concurrent queries, with consistently fast query performance. When concurrency scaling is enabled, Amazon Redshift automatically adds additional cluster capacity when you need it to process an increase in concurrent read queries.

With Redshift Spectrum, you can run multiple Amazon Redshift clusters accessing the same data in Amazon S3. You can use different clusters for different use cases. For example, you can use one cluster for standard reporting and another for data science queries. Your marketing team can use their own clusters different from your operations team. Depending on the type and number of nodes in your local cluster, and the number of files need to be processed for your query, Redshift Spectrum automatically distributes the execution of your query to several Redshift Spectrum workers out of a shared resource pool to read and process data from Amazon S3, and pulls results back into your Amazon Redshift cluster for any remaining processing.

Q: Will my data warehouse cluster remain available during scaling?

It depends. When you using the Concurrency Scaling feature, the cluster is fully available for read and write during concurrency scaling. With Elastic Resize, the cluster is unavailable for 4 to 8 minutes of the resize period. With the Redshift RA3 storage elasticity in managed storage, the cluster is fully available and data is automatically moved between managed storage and compute nodes.

Concurrency

Q: How do I manage resources to ensure that my Redshift cluster can provide consistently fast performance during periods of high concurrency?

A typical data warehouse has significant variance in concurrent query usage over the course of a day. It is more cost-effective to add resources just for the period during which they are required rather than provisioning to peak demand. Amazon Redshift handles this automatically on your behalf.

Concurrency Scaling is a feature in Amazon Redshift that provides consistently fast query performance, even with thousands of concurrent queries. With this feature, Amazon Redshift automatically adds transient capacity when needed to handle heavy demand. Amazon Redshift automatically routes queries to scaling clusters, which are provisioned in seconds and begin processing queries immediately.

This feature is free for most customers. Each Amazon Redshift cluster earns up to one hour of free Concurrency Scaling credits per day. This gives you predictability in your month-to-month cost, even during periods of fluctuating analytical demand.

Q: What is Elastic Resize and how is it different from Concurrency Scaling?

Elastic Resize adds or removes nodes from a single Redshift cluster within minutes to manage its query throughput. For example, an ETL workload for certain hours in a day or month-end reporting may need additional Redshift resources to complete on time. Concurrency Scaling adds additional cluster resources to increase the overall query concurrency.

Q: Can I access the Concurrency Scaling clusters directly?

No. Concurrency Scaling is a massively-scalable pool of Redshift resources, to which customers do not have direct access.

Querying and Analytics

Q: Are Amazon Redshift and Redshift Spectrum compatible with my preferred business intelligence software package and ETL tools?

Amazon Redshift uses industry-standard SQL and is accessed using standard JDBC and ODBC drivers. You can download Amazon Redshift custom JDBC and ODBC drivers from the Connect Client tab of the Redshift Console. We have validated integrations with popular BI and ETL vendors, a number of which are offering free trials to help you get started loading and analyzing your data. You can also go to the AWS Marketplace to deploy and configure solutions designed to work with Amazon Redshift in minutes.

Redshift Spectrum supports all Amazon Redshift client tools. The client tools can continue to connect to the Amazon Redshift cluster endpoint using ODBC or JDBC connections. No changes are required.

You use exactly the same query syntax and have the same query capabilities to access tables in Redshift Spectrum as you have for tables in the local storage of your Redshift cluster. External tables are referenced using the schema name defined in the CREATE EXTERNAL SCHEMA command where they were registered.

Q: What data formats and compression formats does Redshift Spectrum support?

Redshift Spectrum currently supports many open source data formats, including Avro, CSV, Grok, Ion, JSON, ORC, Parquet, RCFile, RegexSerDe, SequenceFile, TextFile, and TSV.

Redshift Spectrum currently supports Gzip and Snappy compression.

Q: What happens if a table in my local storage has the same name as an external table?

Just like with local tables, you can use the schema name to pick exactly which one you mean by using schema_name.table_name in your query.

Q: I use a Hive Metastore to store metadata about my S3 data lake. Can I use Redshift Spectrum?

Yes. The CREATE EXTERNAL SCHEMA command supports Hive Metastores. We do not currently support DDL against the Hive Metastore.

Q: How do I get a list of all external database tables created in my cluster?

You can query the system table SVV_EXTERNAL_TABLES to get that information.

Monitoring

Q: How do I monitor the performance of my Amazon Redshift data warehouse cluster?

Metrics for compute utilization, storage utilization, and read/write traffic to your Amazon Redshift data warehouse cluster are available free of charge via the AWS Management Console or Amazon CloudWatch APIs. You can also add additional, user-defined metrics via Amazon Cloudwatch's custom metric functionality. The AWS Management Console provides a monitoring dashboard that helps you monitor the health and performance of all your clusters. Amazon Redshift also provides information on query and cluster performance via the AWS Management Console. This information enables you to see which users and queries are consuming the most system resources and diagnose performance issues by viewing query plans and execution statistics. In addition, you can see the resource utilization on each of your compute nodes to ensure that you have data and queries that are well balanced across all nodes.

Maintenance

Q: What is a maintenance window? Will my data warehouse cluster be available during software maintenance?

Amazon Redshift periodically performs maintenance to apply fixes, enhancements and new features to your cluster. You can change the scheduled maintenance windows by modifying the cluster, either programmatically or by using the Redshift Console. During these maintenance windows, your Amazon Redshift cluster is not available for normal operations. For more information about maintenance windows and schedules by region, see Maintenance Windows in the Amazon Redshift Management Guide.

AWS Database Migration Service FAQs

Q: Will AWS Database Migration Service help me convert my Oracle PL/SQL and SQL Server T-SQL code to Amazon Aurora or MySQL and PostgreSQL stored procedures? >>

Q: How do I get started with AWS Database Migration Service? >>

Q: In addition to one-time data migration, can I use AWS Database Migration Service for continuous data replication? >>

Q: How are AWS Database Migration Service (DMS) and AWS Schema Conversion Tool (SCT) related? >>

Q: What sources and targets does AWS Database Migration Service support? >>

Q: What sources and targets does AWS Schema Conversion Tool support? >>

Q: Why should I use AWS Database Migration Service instead of my own self-managed replication solution? >>

Q: Can you summarize the database migration steps using AWS Database Migration Service for me? >>

Q: Are these steps different for continuous data replication? >>

Q: Can I monitor the progress of a database migration task? >>

Q: How do I integrate AWS Database Migration Service with other applications? >>

Q: Can I replicate data from encrypted data sources? >>

Q: Does AWS Database Migration Service migrate the database schema for me? >>

Q: Can I use DMS to perform bi-directional replication? >>

Q: How much does DMS cost? >>

Amazon Corretto FAQs

General

Q: What is Amazon Corretto?

A: Corretto is a build of the Open Java Development Kit (OpenJDK) with long-term support from Amazon. Corretto is certified using the Java Technical Compatibility Kit (TCK) to ensure it meets the Java SE standard and is available on Linux, Windows, and macOS. It includes patches from Amazon that have proven useful in running our own services.

Q: Why should I use Corretto?

A: Corretto is a reliable build of OpenJDK with the assurance of long-term support provided at no cost to you. Amazon runs Corretto internally on thousands of production services. Every modification we make to Corretto fixes or mitigates a problem we found running OpenJDK. Amazon also plans to apply urgent fixes (including security) when they are available and ready to use, outside of the regular quarterly cycle.

Q: How is Corretto different from OpenJDK?

A: Corretto is a distribution of Open JDK with patches included by Amazon that are not yet integrated in the corresponding OpenJDK update projects. We focus on patches that improve performance or stability in OpenJDK, chosen based on Amazon's observations running large services.

Q: What kinds of patches does Amazon intend to include in Corretto?

A: Patches will include security fixes, performance enhancements (e.g., speeding up frequently-used functions), garbage collection scheduling, and preventing out-of-memory situations, as well as improved monitoring, reporting, and thread management.

Q: Is there any cost associated with using Corretto?

A: Corretto is distributed by Amazon under an Open Source license at no cost to you. It is licensed under the terms of the GNU Public License version 2 with the Class Path Exception (GPLv2 with CPE). Amazon does not charge for its use or distribution.

Q. What does long-term support (LTS) mean for Corretto?

A: Amazon Corretto is a no-cost, multiplatform, production-ready distribution of the Open Java Development Kit (OpenJDK) that comes with long-term support (LTS). LTS includes Amazon's commitment to provide performance enhancements and security updates at no cost until at least the specified date for the relevant release version (e.g., June 2023 for Corretto 8). Updates are planned to be released quarterly. Amazon also plans to apply urgent fixes (including security) outside of the regular quarterly cycle when they are available and ready to use.

Q: What is included in Corretto's long-term support?

A: Long-term support (LTS) for Corretto includes performance enhancements and security updates for Corretto 8 until at least June 2023 at no cost. Updates are planned to be released quarterly. Amazon will provide LTS for Corretto 11 with quarterly updates until at least August 2024.

LTS for Corretto is unrelated to AWS Support Plans, which provide expert guidance and assistance for achieving your objectives on AWS. If you already have an AWS Support Plan, Corretto is covered on the same basis as all other supported AWS Services and software. For those who do not have a plan, it may or may not make sense for you to purchase a plan if your only intention is to receive assistance with Corretto. Please visit the website to determine if it is right for you. There are currently no plans to launch Corretto-specific assistance plans. As always, our roadmaps are a reflection of our customer feedback and we welcome your feature requests at the Corretto GitHub repository.

Q. What should I do if I need help with Corretto?

A: For general questions about installing or running Corretto, please see our documentation. If you have an issue related to OpenJDK, please open an issue with the upstream OpenJDK project. If you have a specific issue with Corretto or feature request that is not applicable to OpenJDK, please open an issue or a feature request in the Corretto GitHub repository. If you already have an AWS Support Plan you can reach out for assistance with Corretto through your plan.

Using Amazon Corretto

Q: Can I use Corretto as a drop-in replacement for other JDKs?

A: Corretto is designed as a drop-in replacement for all Java SE distributions unless you are using features (e.g., Java Flight Recorder) not available in OpenJDK. Once Corretto binaries are installed on a host and correctly invoked to run your Java applications (e.g., using the *alternatives* command on Linux), existing command-line options, tuning parameters, monitoring, and anything else in place will continue to work as before.

Q: What operating systems does Corretto 8 support?

A: Corretto 8 installation packages are distributed by Amazon for Linux, Windows, and macOS. In addition, an official Docker image is hosted on Docker Hub.

Windows builds are supported on versions 7, 8, 10, Server 2008, Server 2012, and Server 2016.

Mac OS X builds are supported on 10.10 (Yosemite) and later.

Linux builds are supported on Red Hat Enterprise Linux 6+, CentOS 6+, Ubuntu Linux 14+, Debian Linux 8+, Amazon Linux AMI, and SuSE 12+.

Q: What operating systems does Corretto 11 support?

A: Corretto 11 installation packages are distributed by Amazon for Linux, Windows, and macOS.

Windows builds are supported on versions 7, 8, 10, Server 2008, Server 2012, and Server 2016.

Mac OS X builds are supported on 10.10 (Yosemite) and later.

Linux builds are supported on Red Hat Enterprise Linux 6+, CentOS 6+, Ubuntu Linux 14+, Debian Linux 8+, Amazon Linux AMI, and SuSE 12+.

Licensing and Open Source

Q: What are Corretto's license terms?

A: Corretto is released under the same open source license as OpenJDK, which is licensed under the GNU Public License version 2 with the Class Path Exception (GPLv2 with CPE). You can use Corretto as you would use OpenJDK.

Q: How does Amazon contribute to OpenJDK?

A: Amazon started contributing to OpenJDK in 2017 and we plan to increase contributions in both number and complexity.

Q: How can I contribute to Corretto?

A: Amazon encourages contributions to the OpenJDK project as the way to get code into Corretto. This way the whole OpenJDK community benefits from your changes. If your contribution is specific to Corretto, such as to the build logic, the code is available on GitHub, where we will evaluate issues and pull requests.

AWS Cloud Development Kit FAQs

General

Q: What is AWS CDK?
The AWS Cloud Development Kit (AWS CDK) is an open-source software development framework for defining cloud infrastructure as code with modern programming languages and deploying it through AWS CloudFormation.

Q: What can I do with the AWS CDK CLI?
You can use the AWS CDK CLI to interact with your CDK applications. CDK CLI enables you to list the stacks defined in your CDK app, synthesize the stacks into CloudFormation templates, determine the differences between running stack instances and the stacks defined in your CDK code, and deploy stacks to any public AWS Region.

Q: How does AWS CDK work?
You use the AWS CDK framework to author AWS CDK projects which are executed to generate CloudFormation templates. AWS CDK projects can be executed using the AWS CDK command line or in a continuous delivery system.

Q: How do I get started with AWS CDK?
The best way to get started with AWS CDK is to work through the Getting Started section of the AWS CDK Developer Guide. Within a few minutes, you can define and deploy your first AWS CDK application. For a more in-depth guided tutorial, check out https://CDKworkshop.com.

Q: Are there sample AWS CDK applications that I can use as a starting point?
Yes, you can find a number of AWS CDK examples, in multiple programming languages, at https://github.com/aws-samples/aws-cdk-examples. The AWS

Developer Guide also includes examples at https://docs.aws.amazon.com/cdk/latest/guide/examples.html.

Q: Why do I need a JavaScript runtime installed to use AWS CDK?
AWS builds the business logic of AWS Construct Library packages in TypeScript, and provides mappings into each of the supported programming languages. This allows us to ensure that AWS CDK constructs behavior is consistent language to language, and allows us to offer a comprehensive set of construct packages that are available in all languages. The code you write in your AWS CDK project is all native in the programming language you prefer, and the JavaScript runtime is an implementation detail to your programming experience. You can reference to the jsii project at https://github.com/aws/jsii.

Q: Where can I find the AWS CDK source code?
The AWS CDK code is open source and is available through GitHub at https://github.com/awslabs/aws-cdk.

Q: How can I contribute?
We are developing AWS CDK in the open and welcome anyone who wants to contribute to our code base. You can learn more at https://github.com/awslabs/aws-cdk/blob/master/CONTRIBUTING.md.

Q: What programming languages does AWS CDK support?
AWS CDK is generally available in JavaScript, TypeScript, Python, Java, and C#. We are planning AWS CDK bindings for other languages in the future, so vote for your favorite on GitHub.

Q: In which regions can I use AWS CDK?
AWS CDK is available to define and deploy AWS resources in all public regions. Since AWS CDK leverages the CloudFormation service, refer to Regional Products and Services for details about specific resource availability per AWS Region.

Q: Are there any resource limits or restrictions applied to AWS CDK?
Since AWS CDK leverages CloudFormation, AWS CDK applications are subject to the same limits imposed by the CloudFormation service.

AWS Construct Library

Q: What is the AWS Construct Library?
AWS CDK constructs are delivered in the AWS Construct Library, and represent abstractions of cloud infrastructure logic that are packaged for reuse and expose a rich programmatic interface. Constructs can be defined locally or published to package managers such as npm, Maven, NuGet, or PyPI for sharing across organizations.

Q: Which services are available in the AWS Construct Library?
We have coverage for many common AWS services and features with rich, high-level constructs, and complete coverage of the lower-level CloudFormation resources, including new resources shortly after they are available. We keep the AWS Construct Library up to date by autogenerating the resource-level APIs every time the CloudFormation specification changes. On top of these autogenerated APIs, we provide handcrafted, higher-level abstractions that make it even easier to work with each service. We do not have this high-level, convenient abstraction for every AWS service yet, but are adding new features all the time with a focus on services most used in modern cloud applications. Visit the AWS Construct Library API reference to learn more.

Q: How long will I have to wait for AWS CDK to support new features that are added to CloudFormation?
We keep the AWS Construct Library in sync with the CloudFormation resource specification by adding new CloudFormation features to AWS CDK shortly after their publication.

Q: How do I share and manage AWS CDK constructs with my team?
AWS CDK construct libraries are just like any other library. They are consumed through the package manager of the programming language you use; keeping those libraries up to date is part of your normal workflow. All packages support semantic versioning, allowing you to make conscious choices about when to migrate to new infrastructure models.

CloudFormation and SAM

Q: What is the relationship between AWS CDK and CloudFormation?
You can think of the AWS CDK as a developer-centric toolkit that leverages the full power of modern programming languages to define your AWS infrastructure as code. When AWS CDK applications are run, they compile down to fully formed CloudFormation JSON/YAML templates that are then submitted to the CloudFormation service for provisioning. Because the AWS CDK leverages CloudFormation, you still enjoy all the benefits CloudFormation provides such as safe deployment, automatic rollback, and drift detection.

Q: What is the relationship between AWS CDK and AWS SAM?
AWS Serverless Application Model and AWS CDK both abstract AWS infrastructure as code making it easier for you to define your cloud infrastructure. AWS SAM is specifically focused on serverless use cases and architectures and allows you to define your infrastructure in compact, declarative JSON/YAML templates. AWS CDK offers broad coverage across all of AWS services and allows you to define cloud infrastructure in modern programming languages like TypeScript, Python, C#, and Java. Both AWS SAM and AWS CDK leverage CloudFormation as the provisioning engine for your infrastructure stacks.

If you prefer defining your serverless infrastructure in concise declarative templates, SAM is the better fit. If you want to define your AWS infrastructure in a familiar programming language, we encourage you to try out AWS CDK. In either case, you can rely on repeatable, safe infrastructure deployment through CloudFormation.

Q: Can I use the AWS SAM CLI to locally test Lambda functions defined with AWS CDK?
Yes, you can locally synthesize your AWS CDK application into a CloudFormation template and then reference the generated AWS Lambda handler ID in the SAM CLI. See the SAM CLI section in the CDK Developer Guide for more details.

Cost

Q: Under what license is AWS CDK distributed?
AWS CDK is distributed under the Apache License, Version 2.0. See LICENSE and

NOTICE for more information.

Q: How much does AWS CDK cost?

There is no additional charge for AWS CDK. You pay for AWS resources (such as Amazon EC2 instances, Elastic Load Balancing load balancers, etc.) created using AWS CDK in the same way as if you created them manually. You only pay for what you use, as you use it; there are no minimum fees and no required upfront commitments.

AWS Cloud9 FAQs

General

Q: What is AWS Cloud9?

AWS Cloud9 is a cloud-based integrated development environment (IDE) that lets you write, run, and debug your code with just a browser. It combines the rich code editing features of an IDE such as code completion, hinting, and step-through debugging, with access to a full Linux server for running and storing code. For more information see our AWS Cloud9 User Guide.

Q: Who should use AWS Cloud9?

Anybody who writes code can use AWS Cloud9. Those developing applications using Node.js (JavaScript), Python, PHP, Ruby, Go, and C++ can use Cloud9 and have immediate access to a fully configured development environment in their browsers with preinstalled runtimes, package managers, and debugging tools. With Cloud9, you are no longer tied to a single development machine and can access your development environment from any internet-connected computer.

AWS developers and those evaluating new AWS services can use AWS Cloud9 for easy access to their AWS resources through a preconfigured AWS Command Line Iinterface (AWS CLI), ready to run commands against AWS services. Those developing serverless applications on AWS Lambda using Node.js or Python can use the built-in tools in Cloud9 to create, edit, run, debug, and deploy their Lambda functions from within the IDE.

Q: Which programming languages are supported?

AWS Cloud9 supports over 40 programming languages, including Node.js (JavaScript), Python, PHP, Ruby, Go, and C++. It includes features such as syntax highlighting, outline view, code hinting, code completion, application runners, and step-through debugging for many popular programming languages. To learn more about the language features supported in Cloud9, please visit the Language Support topic of our user guide.

Q: What web browsers can I use to access AWS Cloud9?

AWS Cloud9 is fully supported on the recent versions of Google Chrome, Safari, Firefox, and Microsoft Edge.

Q: What is the pricing for AWS Cloud9?

There is no additional charge for AWS Cloud9. If you use an Amazon EC2 instance for your AWS Cloud9 development environment, you pay only for the compute and storage resources (i.e., an EC2 instance, an EBS volume) that are used to run and store your code. You can also connect your Cloud9 development environment to an existing Linux server (e.g., on-premises server) via SSH for no additional charge. See the AWS Cloud9 pricing page for more details.

Q: What are the other IDEs supported by AWS?

AWS offers a broad selection of IDE support to facilitate development of applications for AWS. To learn more about the IDE toolkits supported by AWS, visit the IDE Toolkits section on the AWS Tools page.

Q: What if I see an error when working with AWS Cloud9?

You can find some of the errors you might encounter and their possible solutions in the Troubleshooting topic of our user guide.

Using AWS Cloud9

Q: How do I get started with AWS Cloud9?

You can sign in to the AWS Management Console, and select AWS Cloud9. The console will guide you through the options to select the Linux server that you want to connect with Cloud9. You can either launch a new Amazon EC2 instance (AWS Cloud9 EC2 environment) or connect your existing Linux server (AWS

Cloud9 SSH environment) in a few simple steps. Once you've created a Cloud9 environment, you can access your IDE and write code in a fully configured development environment. For more information, see our documentation about setting up AWS Cloud9 and then complete a basic tutorial.

Q: What is an AWS Cloud9 development environment?

An AWS Cloud9 development environment is where the project code files are stored and the tools used to develop the application are run. Each environment has unique IDE settings stored with it. This enables you to easily create and switch between many different development environments, each one customized with the tools, runtimes, files, and IDE settings required for a specific project.

Q: What are the types of AWS Cloud9 development environments?

There are two types of AWS Cloud9 environments that you can use.

- **AWS Cloud9 EC2 environment –** Enables you to launch a new Amazon EC2 instance that Cloud9 connects to. By default, these instances stop 30 minutes after you close the IDE and start automatically when you open the IDE.
- **AWS Cloud9 SSH environment –** Enables you to connect an existing Linux server with Cloud9. There are certain dependencies that are required on the Linux server that you want to use with Cloud9 SSH environments. Please visit our documentation for more details.

Q: Can I use my existing Amazon EC2 or Amazon Lightsail instance with AWS Cloud9?

Yes. You can use SSH environments to connect an existing Linux-based EC2 or Lightsail instance with AWS Cloud9.

Q: How do I edit my code?

The AWS Cloud9 IDE has an advanced code editor with features such as auto-completion, code folding, hinting, syntax highlighting, and line manipulation. The code editor enables you to choose from over 30 color schemes that control syntax highlighting and the UI. You can also fully customize the Cloud9 UI by editing your stylesheet.

Q: What tools and packages are preinstalled on AWS Cloud9 EC2 environments?

AWS Cloud9 EC2 environments come preinstalled with commonly used development tools such as Git and Docker. They also include language runtimes and package managers for many popular programming languages such as Node.js and Python. To view the full list of tools and packages preinstalled on Cloud9 EC2 environments, please visit our documentation.

Q: How do I run my code?

The AWS Cloud9 IDE has a run button in the toolbar and built-in runners for over 10 different languages that will automatically start your application with the latest code changes. For full control over how you run your software, you can also customize existing runners, create your own runners, or run your code from the terminal.

Q: How do I run CLI commands?

The AWS Cloud9 IDE has a built-in terminal window that can interactively run CLI commands. You also have full administrative privileges on the instance (sudo rights), allowing you to install any additional tools required for development or to host your application.

Q: How do I connect to source control management systems?

You can open the terminal window within the IDE and access your source control system using the same command line tools that you would use on your local machine. AWS Cloud9 EC2 environments come preinstalled with Git to enable easy access to your source code.

Q: Which AWS Regions does AWS Cloud9 support?

See Regional Products and Services for details.

Q: Where does AWS Cloud9 store my code?

Any data that you store in your AWS Cloud9 environment such as code files, packages, or dependencies is always stored in your resources. If you use an EC2 environment, your data is stored in the associated Amazon Elastic Block Store (EBS) volume that exists in your AWS account. If you use an SSH environment, your data is stored in local storage on your Linux server.

AWS Cloud9 Associated Resources

Q: What are the resources created by AWS Cloud9 for Amazon EC2 environments?

When you create an Amazon EC2 environment, AWS Cloud9 creates the required compute and storage resources in your AWS account. These resources include an Amazon EC2 instance, an 8-GB Amazon Elastic Block Store (EBS) volume, an Amazon EC2 security group, and an AWS CloudFormation stack. You have access to these resources through the individual AWS service consoles. When you delete your environment, Cloud9 automatically deletes these resources for you.

Q: Does AWS Cloud9 manage resources created in AWS Cloud9 for Amazon EC2 environments?

In addition to creating and deleting your AWS Cloud9 EC2 environment resources on your behalf, Cloud9 can also automatically start and stop the EC2 instances to reduce your costs. You are responsible for all other administrative tasks on these resources, such as installing software patches on your EC2 instances and performing backup of your EBS volumes.

Q: Are my Amazon EC2 instances in AWS Cloud9 environments always running?

No. AWS Cloud9 provides a default auto-hibernation setting of 30 minutes for your Amazon EC2 instances created through Cloud9. With this setting, your EC2 instances automatically stop 30 minutes after you close the IDE and restart only when you reopen the IDE. As a result, you typically only incur EC2 instance charges for when you are actively working. When your instance requires a restart, you lose any active terminal sessions in the IDE and can experience some wait time while opening your IDE. Depending on your use case, you can configure the auto-hibernation setting and even elect to keep your EC2 instance “always on”.

Q: Can I change my Amazon EC2 instance type for an existing EC2 environment?

Yes. You can change the Amazon EC2 instance type that you initially selected with your AWS Cloud9 environment. To do this, you navigate to the instance in the EC2 console, locate your instance, and follow the instructions in Amazon EC2 documentation.

Environment Sharing

Q: How do I share my AWS Cloud9 environment with other people?

You can share your AWS Cloud9 environment by clicking the **Share** button in the top right of your IDE. You are prompted for the AWS Identity and Access Management (IAM) user name and the desired access levels for the person you want to collaborate with. Once you enter these details, the environment is available to both the participants for real-time collaboration on IDE features and command line sessions.

Q: Can I share an AWS Cloud9 environment with IAM users in a different AWS account?

No. AWS Cloud9 environments can currently be shared only with the IAM users within the same AWS account. If you want to invite a new user that doesn’t have an IAM user access, you can follow the link to create a new IAM user in the Share dialog box.

Using AWS Cloud9 with AWS Services

Q: How do I access AWS services from AWS Cloud9?

AWS Cloud9 EC2 environments come preinstalled with the AWS CLI, which is authenticated with the permissions of the logged-in AWS user automatically. This enables you to run interactive CLI commands against AWS services from the built-in terminal window in Cloud9 without any additional configuration.

Q: How can I develop serverless applications for AWS Lambda using AWS Cloud9?

You can access the built-in tools for AWS Lambda from the AWS Resources panel in the IDE. You can use these tools to import existing or create new Lambda functions in Node.js and Python. You can easily run, preview, debug, and deploy these functions directly from the IDE. AWS Cloud9 also provides support for the AWS Serverless Application Model (AWS SAM) framework. This enables you to easily manage multiple Lambda functions and serverless resources in your application. If you provisioned your project using AWS CodeStar, any changes committed to the application will be built and deployed directly to Lambda on git push.

Q: Can I locally test my AWS Lambda functions using AWS Cloud9?

Yes. AWS Cloud9 can simulate the AWS Lambda execution environment for Node.js and Python to run your functions locally in the IDE. This enables you to test your serverless applications with step-through debugging without uploading your application changes to Lambda. Once tested, you can also deploy your application changes to Lambda directly from the IDE.

Q: How do I use AWS Cloud9 with AWS CodeStar?

You can launch AWS Cloud9 environments directly from AWS CodeStar and immediately start editing and committing your CodeStar project code in the Cloud9 IDE. Any code changes that you commit to your project source repository from Cloud9 are automatically built and deployed using the tools provisioned by CodeStar. To learn more about using this integration, please visit the AWS CodeStar documentation.

AWS CodeBuild FAQs

General

Q: What is AWS CodeBuild? >>

Q: Why should I use CodeBuild? >>

Q: What is the pricing for CodeBuild? >>

Q: Can I use CodeBuild to automate my release process? >>

Using CodeBuild

Q: What is a build project? >>

Q: How do I configure a build project? >>

Q: Which source repositories does CodeBuild support? >>

Q: Which programming frameworks does CodeBuild support? >>

Q: Which preconfigured Windows build runtimes does CodeBuild provide? >>

Q: What happens when a build is run? >>

Q: How do I set up my first build? >>

Q: Can I use CodeBuild with Jenkins? >>

Q: How can I view past build results? >>

Q: How can I debug a past build failure? >>

Q: Why is build.general1.small not supported for .NET Core for Windows build environments? >>

Q: How do I receive notifications or alerts for any events in AWS CodeBuild? >>

Security

Q: Can I encrypt the build artifacts stored by CodeBuild? >>

Q: How does CodeBuild isolate builds that belong to other customers? >>

Q: Can I use AWS Identity and Access Management (IAM) to manage access to CodeBuild? >>

Regions

Q: Which regions does CodeBuild support? >>

AWS CodeCommit FAQs

General

Q: What is AWS CodeCommit? >>

Q: What is Git? >>

Q: Who should use AWS CodeCommit? >>

Q: How is AWS CodeCommit different from other Git-based source control systems? >>

Q: How does AWS CodeCommit compare to a versioned S3 bucket? >>

Using AWS CodeCommit

Q: How do I get started with AWS CodeCommit? >>

Q: How do I create a repository? >>

Q: How do I update files in my repository? >>

Q: How do I import my existing repository to AWS CodeCommit? >>

Q: What Git operations are currently supported by AWS CodeCommit? >>

Q: Does AWS CodeCommit support Git submodules? >>

Q: What are the service limits when using AWS CodeCommit? >>

Q: What is the maximum size for a single file that I can store in CodeCommit? >>

Q: How do I backup my repository? >>

Q: How do I restore a deleted AWS CodeCommit repository? >>

Q: How do I manage code reviews with AWS CodeCommit? >>

Q: How do I integrate my continuous integration system with AWS CodeCommit? >>

Q: How do I create webhooks using AWS CodeCommit? >>

Q: Can I get a history of AWS CodeCommit Git operations and API calls made in my account for security analysis and operational troubleshooting purposes? >>

Security

Q: Can I use AWS Identity and Access Management (IAM) to manage access to AWS CodeCommit? >>

Q: What communication protocols are supported by AWS CodeCommit? >>

Q: What ports should I open in my firewall for access to AWS CodeCommit? >>

Q: How do I encrypt my repository in AWS CodeCommit? >>

Q: Can I enable cross-account access to my repository? >>

Regions

Q: Which regions does AWS CodeCommit support? >>

Billing

Q: How much does AWS CodeCommit cost? >>

Q: What is the definition of an active user in AWS CodeCommit? >>

Q: Which Git requests are considered towards the monthly allowance? >>

AWS CodeDeploy FAQs

General

Q: What is AWS CodeDeploy?
AWS CodeDeploy is a service that automates code deployments to any instance, including Amazon EC2 instances and instances running on-premises. AWS CodeDeploy makes it easier for you to rapidly release new features, helps you avoid downtime during deployment, and handles the complexity of updating your applications. You can use AWS CodeDeploy to automate deployments, eliminating the need for error-prone manual operations, and the service scales with your infrastructure so you can easily deploy to one instance or thousands.

Q: Who should use AWS CodeDeploy?
AWS CodeDeploy is designed for developers and administrators who need to deploy applications to any instance, including Amazon EC2 instances and instances running on-premises. It is flexible and can also be used by anyone wanting to update software or run scripts on their instances.

Q: What types of applications can be deployed with AWS CodeDeploy?
AWS CodeDeploy can be used for deploying any type of application. To use AWS CodeDeploy, you specify the files to copy and the scripts to run on each instance during the deployment. AWS CodeDeploy is programming language and architecture agnostic, so you can use scripts for any custom deployment logic.

Q: What operating systems does AWS CodeDeploy support?
AWS CodeDeploy supports a wide variety of operating systems. AWS CodeDeploy provides agents that have been tested on Amazon Linux, Red Hat Enterprise Linux, Ubuntu Server, and Microsoft Windows Server. If you want to use other operating systems, the AWS CodeDeploy agent is available as open source software here. For more information on operating system support, see AWS CodeDeploy Documentation.

Q: Will AWS CodeDeploy work with my existing tool chain?
Yes. AWS CodeDeploy works with a variety of configuration management systems, continuous integration and deployment systems, and source control systems. For more information, see product integrations page.

Q: How is AWS CodeDeploy different from other AWS deployment and management services such as AWS Elastic Beanstalk and AWS OpsWorks?
AWS CodeDeploy is a building block service focused on helping developers deploy and update software on any instance, including Amazon EC2 instances and instances running on-premises. AWS Elastic Beanstalk and AWS OpsWorks are end-to-end application management solutions.

Q: Does AWS CodeDeploy support on-premises instances?
Yes. AWS CodeDeploy supports any instance that can install the CodeDeploy agent and connect to AWS public endpoints.

Concepts

Q: What is an application?
An application is a collection of software and configuration to be deployed to a group of instances. Typically, the instances in the group run the same software. For example, if you have a large distributed system, the web tier will likely constitute one application and the data tier another application.

Q: What is a revision?
A revision is a specific version of deployable content, such as source code, post-build artifacts, web pages, executable files, and deployment scripts, along with an AppSpec file. The AWS CodeDeploy Agent can access a revision from GitHub or an Amazon S3 bucket.

Q: What is a deployment group?
A deployment group is the AWS CodeDeploy entity for grouping EC2 instances or AWS Lambda functions in a CodeDeploy deployment. For EC2 deployments, it is a set of instances associated with an application that you target for a deployment. You can add instances to a deployment group by specifying a tag, an Auto Scaling group name, or both. In an AWS Lambda deployment, a deployment group defines a set of AWS CodeDeploy configurations for future serverless Lambda deployment to the group, like alarms and rollbacks.

You can define multiple deployment groups for an application such as staging and production. For information on tags, see Working with Amazon EC2 Tags in the Console. For more information on deploying to Auto Scaling groups, see Auto Scaling Integration.

Q: What is a deployment configuration?
A deployment configuration specifies how the behavior for how deployment should proceed, including how to handle deployment failure, through for a deployment group. You can use a deployment configuration to perform zero-downtime deployments to multi-instance deployment groups. For example, if your application needs at least 50% of the instances in a deployment group to be up and serving traffic, you can specify that in your deployment configuration so that a deployment does not cause downtime. If no deployment configuration is associated with either the deployment or the deployment group, then by default AWS CodeDeploy will deploy to one instance at a time. For more information on deployment configuration, see Instance Health.

Q: What are the parameters that I need to specify for a deployment?
There are three parameters you specify for a deployment:

1. Revision - Specifies what to deploy.

2. Deployment group - Specifies where to deploy.

3. Deployment configuration - An optional parameter that specifies how to deploy.

Q: What is an AppSpec file?
An AppSpec file is a configuration file that specifies the files to be copied and scripts to be executed. The AppSpec file uses the YAML format, and you include it in the root directory of your revision. The AppSpec file is used by the AWS CodeDeploy Agent and consists of two sections. The files section specifies the source files in your revision to be copied and the destination folder on each instance. The hooks section specifies the location (as relative paths starting from the root of the revision bundle) of the scripts to run during each phase of the deployment. Each phase of a deployment is called a deployment lifecycle event. The following is a sample AppSpec file. For more information on an AppSpec file, including all the options that can be specified, see AppSpec File Reference.

```
version: 0.0

os: linux

files:

# You can specify one or more mappings in the files section.

  - source: /

    destination: /var/www/html/WordPress

hooks:
 # The lifecycle hooks sections allows you to specify deployment script
ApplicationStop:
# Step 1: Stop Apache and MySQL if running.
    - location: helper_scripts/stop_server.sh
BeforeInstall:
# Step 2: Install Apache and MySQL.
# You can specify one or more scripts per deployment lifecycle event.
    - location: deploy_hooks/puppet-apply-apache.sh
    - location: deploy_hooks/puppet-apply-mysql.sh
 AfterInstall:
# Step 3: Set permissions.
    - location: deploy_hooks /change_permissions.sh
      timeout: 30
      runas: root
# Step 4: Start the server.
    - location: helper_scripts/start_server.sh
      timeout: 30
      runas: root
```

Q: What are deployment lifecycle events?
A deployment goes through a set of predefined phases called deployment lifecycle events. A deployment lifecycle event gives you an opportunity to run code as part of the deployment. The following table lists the different deployment lifecycle events currently supported, in their order of execution, along with examples of when you may want to use them.

Deployment Lifecycle Event	Description
ApplicationStop	This is the first deployment lifecycle event that occurs even before the revision gets downloaded. The AppSpec file and scripts used for this deployment lifecycle event are from the last successfully deployed revision. You can use the ApplicationStop deployment lifecycle event if you want to gracefully stop the application or remove currently installed packages in preparation of a deployment.
DownloadBundle	During this deployment lifecycle event, the agent copies the revision files to a temporary location on the instance. This deployment lifecycle event is reserved for the agent and cannot be used to run user scripts.
BeforeInstall	You can use the BeforeInstall deployment lifecycle event for preinstall tasks such as decrypting files and creating a backup of the current version.
Install	During this deployment lifecycle event, the agent copies the revision files from the temporary location to the final destination folder. This deployment lifecycle event is reserved for the agent and cannot be used to run user scripts.
AfterInstall	You can use the AfterInstall deployment lifecycle event for tasks such as configuring your application or changing file permissions.
ApplicationStart	You typically use the ApplicationStart deployment lifecycle event to restart services that were stopped during ApplicationStop.
ValidateService	ValidateService is the last deployment lifecycle event and is an opportunity to verify that the deployment completed successfully.

Getting started

Q: How do I get started with AWS CodeDeploy?

You can sign in to the AWS Management Console and start using AWS CodeDeploy. If you are looking for a quick overview of the service, see Getting Started, which includes a step-by-step tutorial.

Using AWS CodeDeploy

Q: Are there any prerequisites for using an existing Amazon EC2 instance with AWS CodeDeploy?
The Amazon EC2 instance must be associated with an IAM instance profile and should be running a supported operating system. For more information, see Use an Existing Amazon EC2 Instance.

Q: What are the typical steps to go through for deploying an application using AWS CodeDeploy?
The following diagram shows the typical steps during a deployment. Creating an application and deployment group (see the Concepts section for an explanation of these terms) are typically one-time setup tasks per application. The recurring actions are uploading a revision and deploying it. For a detailed explanation, including step-by-step instructions for each of these tasks, see Deployments.

Q: How can I access AWS CodeDeploy?
You can access AWS CodeDeploy using the AWS Management Console, the AWS Command Line Interface (AWS CLI), the AWS SDKs, and the AWS CodeDeploy APIs.

Q: What changes do I need to make to my code to deploy using AWS CodeDeploy?
You don’t need to make any changes to your code. You simply add a configuration file (called an AppSpec file) in the root directory of your revision bundle that specifies the files to be copied and scripts to be executed.

Q: How can I deploy an application from my source control system using AWS CodeDeploy?
If you are using GitHub, you can deploy a revision in a .zip, .tar, or .tar.gz format from your repository directly to instances. For other source control systems, you can bundle and upload the revision to an Amazon S3 bucket in a .zip, .tar, or .tar.gz format and specify the Amazon S3 location when doing a deployment. If your application needs a build step, make sure that the GitHub repository or the Amazon S3 bucket contains the post-build artifacts. For more information on using GitHub with AWS CodeDeploy, see our product integrations page. For more information on using Amazon S3 for storing revisions, see Push a Revision.

Q: How will AWS CodeDeploy work with my configuration management tool?
You can invoke your configuration management tool from any deployment lifecycle event

hook in the AppSpec file. For example, if you have a Chef recipe that you want to run as part of a deployment, you can do so by specifying it in the appropriate deployment lifecycle event hook in the AppSpec file. In addition, you can leverage your configuration management system to install the AWS CodeDeploy agent on instances. For samples that illustrate using AWS CodeDeploy with configuration management systems such as Chef, Puppet, Ansible, and Saltstack, see our product integrations page.

Q: Can I use AWS CodeDeploy with continuous integration and deployment systems?
Yes. You can integrate AWS CodeDeploy with your continuous integration and deployment systems by calling the public APIs using the AWS CLI or AWS SDKs. You can find prebuilt integrations and samples on our product integrations page.

Q: How do I get my application on the instances that I just added to the deployment group?
Deploy the latest revision to the deployment group for the newly added instances to get your application. Except for Amazon EC2 instances that are launched as part of an Auto Scaling group, AWS CodeDeploy doesn't automatically deploy the latest revision to newly added instances.

Q: How does AWS CodeDeploy work with Auto Scaling?
You can associate an Auto Scaling group with a deployment group to make sure that newly launched instances always get the latest version of your application. Every time a new Amazon EC2 instance is launched for that Auto Scaling group, it will be first put in a Pending state and a deployment of the last successful revision for that deployment group triggered on that Amazon EC2 instance. If the deployment completes successfully, the state of the Amazon EC2 instance is changed to InService. If that deployment fails, the Amazon EC2 instance is terminated, a new Amazon EC2 instance is launched in Pending state, and a deployment triggered for the newly launched EC2 instance. For more information on Auto Scaling group instance lifecycle events, see Auto Scaling Group Lifecycle.

Q: How do I track the status of a deployment?
You can track the status of a deployment using the AWS Management Console, the AWS Command Line Interface (AWS CLI), the AWS SDKs, and the AWS CodeDeploy APIs.You can see the overall status of a deployment and drill down further to see the status of each instance and the status of each deployment lifecycle event for the instance. You can also see the log entries corresponding to any failure, making it easy to debug deployment issues without having to log into the instance.

Q: Can I stop an in-flight deployment?
Yes. When you stop an in-flight deployment, the AWS CodeDeploy service will instruct the agent on each instance to stop executing additional scripts. To get your application back to a consistent state, you can either redeploy the revision, or deploy another revision.

Q: How do I roll back an application to the previous revision?
To roll back an application to a previous revision, you just need to deploy that revision. AWS CodeDeploy keeps track of the files that were copied for the current revision and removes them before starting a new deployment, so there is no difference between redeploy and roll back. However, you need to make sure that the previous revisions are available for roll back.

Q: Can I use a versioned Amazon S3 bucket to store revisions?
Yes. You can use a versioned Amazon S3 bucket and specify the version ID to uniquely identify a revision.

Q: What are the service limits when using AWS CodeDeploy?
For information on the service limits, see Limits. To increase your service limits, submit a request through the AWS Support Center.

Q: Can I get a history of AWS CodeDeploy API calls made on my account for security analysis and operational troubleshooting purposes?
Yes. To receive a history of AWS CodeDeploy API calls made on your account, you simply turn on AWS CloudTrail in the AWS Management Console.

Q: How do I receive notifications or alerts for any events in AWS CodeDeploy?
You can create notifications for events impacting your deployments. Notifications will come in the form of Amazon SNS notifications. Each notification will include a status message as well as a link to the resources whose event generated that notification. Notifications has no additional cost; but, you may be charged for other AWS services utilized by notifications, such as Amazon SNS. To learn how to get started with notifications, see the notifications user guide. Additionally, customers using AWS Chatbot can configure notifications to be sent to their Slack Channels or Amazon Chime chat rooms. For more details please check here.

Security

Q: Can I use AWS CodeDeploy to deploy an application to Amazon EC2 instances running within an Amazon Virtual Private Cloud (VPC)?
Yes, but the AWS CodeDeploy agent installed on the Amazon EC2 instances must be able to access the public AWS CodeDeploy and Amazon S3 service endpoints. For more information, see AWS CodeDeploy Endpoints and Amazon S3 Endpoints.

Q: Can I use AWS Identity and Access Management (IAM) to manage access to AWS CodeDeploy?
Yes. AWS CodeDeploy supports resource-level permissions. For each AWS CodeDeploy resource, you can specify which user has access and to which actions. For example, you can set an IAM policy to let a user deploy a particular application but only list revisions for other applications. You can therefore prevent users from inadvertently making changes to the wrong application. For more information on using IAM with AWS CodeDeploy, see Access Permissions Reference.

Regions

Q: Which regions does AWS CodeDeploy support?
Please refer to Regional Products and Services for details of CodeDeploy availability by region.

Q: How do I deploy an AWS CodeDeploy application to multiple regions?
AWS CodeDeploy performs deployments with AWS resources located in the same region. To deploy an application to multiple regions, define the application in your target regions, copy the application bundle to an Amazon S3 bucket in each region, and then start the deployments using either a serial or parallel rollout across the regions.

Billing

Q: How much does AWS CodeDeploy cost?
There is no additional charge for code deployments to Amazon EC2 instances through AWS CodeDeploy. You pay $0.02 per on-premises instance update using AWS CodeDeploy. Please see the Pricing page for more details.

AWS CodePipeline FAQs

General

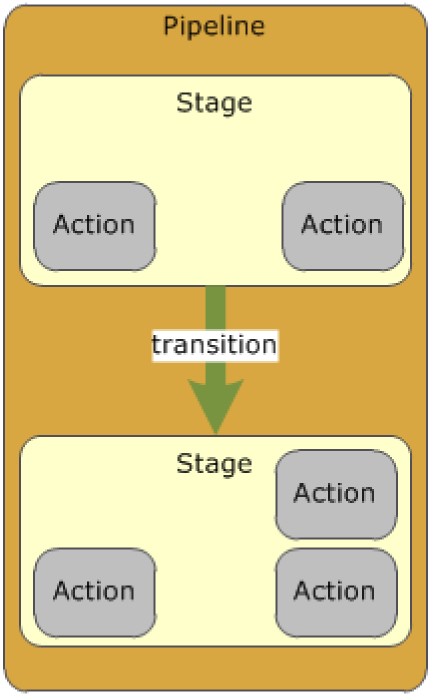
Q: What is AWS CodePipeline?
AWS CodePipeline is a continuous delivery service that enables you to model, visualize, and automate the steps required to release your software. With AWS CodePipeline, you model the full release process for building your code, deploying to pre-production environments, testing your application and releasing it to production. AWS CodePipeline then builds, tests, and deploys your application according to the defined workflow every time there is a code change. You can integrate partner tools and your own custom tools into any stage of the release process to form an end-to-end continuous delivery solution.

Q: Why should I use AWS CodePipeline?
By automating your build, test, and release processes, AWS CodePipeline enables you to increase the speed and quality of your software updates by running all new changes through a consistent set of quality checks.

Q: What is continuous delivery?
Continuous delivery is a software development practice where code changes are automatically built, tested, and prepared for a release to production. AWS CodePipeline is a service that helps you practice continuous delivery. Learn more about continuous delivery here.

Concepts

The diagram below represents the concepts discussed in this section.



Q: What is a pipeline?
A pipeline is a workflow construct that describes how software changes go through a release process. You define the workflow with a sequence of stages and actions.

Q: What is a revision?
A revision is a change made to the source location defined for your pipeline. It can include source code, build output, configuration, or data. A pipeline can have multiple revisions flowing through it at the same time.

Q: What is a stage?
A stage is a group of one or more actions. A pipeline can have two or more stages.

Q: What is an action?
An action is a task performed on a revision. Pipeline actions occur in a specified order, in serial or in parallel, as determined in the configuration of the stage. For more information, see Edit a Pipeline and Action Structure Requirements in AWS CodePipeline.

Q: What is an artifact?

When an action runs, it acts upon a file or set of files. These files are called artifacts. These artifacts can be worked upon by later actions in the pipeline. For example, a source action will output the latest version of the code as a source artifact, which the build action will read in. Following the compilation, the build action will upload the build output as another artifact, which will be read by the later deployment actions.

Q: What is a transition?

The stages in a pipeline are connected by transitions, and are represented by arrows in the AWS CodePipeline console. Revisions that successfully complete the actions in a stage will be automatically sent on to the next stage as indicated by the transition arrow. Transitions can be disabled or enabled between stages.

Using AWS CodePipeline

Q: How do I get started with AWS CodePipeline?

You can sign in to the AWS Management Console, create a pipeline, and start using the service. If you want an introduction to AWS CodePipeline, see Getting Started, which includes step-by-step tutorials. Or, see the Pipeline Starter Kit to quickly provision a preconfigured release pipeline with a Jenkins build server using an AWS CloudFormation template.

Q: How do I start a pipeline?

After you create a pipeline, it will automatically trigger a run to release the latest revision of your source code. From then on, every time you make a change to your source location, a new run is triggered. In addition, you can re-run the last revision through a pipeline using the Release Change button in the pipeline console.

Q: How do I stop a pipeline?

To stop a pipeline, you can disable a transition from one stage to another. Once disabled, your pipeline will continue to run revisions through the actions, but it will not promote revisions through the disabled transition to later stages. For more details, see Disable or Enable Transitions in AWS CodePipeline.

Q: Can I edit an existing pipeline?
Yes. You can use the AWS CodePipeline console or AWS CLI to add or remove stages in a pipeline as well as to add, edit, or remove actions in a stage.

Q: Can I create a copy of an existing pipeline?
Yes. You can use the get-pipeline AWS CLI command to get the JSON structure of your existing pipeline. You can then use that JSON and the create-pipeline AWS CLI command to create a new pipeline with the same structure as the existing one.

Q: Can actions run in parallel?
Yes. You can configure one or more actions to run in parallel for any given stage.

Q: How can I practice continuous delivery for my serverless applications and AWS Lambda functions?
You can release updates to your serverless application by including the AWS Serverless Application Model template and its corresponding files in your source code repository. You can use AWS CodeBuild in your pipeline to package your code for deployment. You can then use AWS CloudFormation actions to create a change set and deploy your serverless application. You have the option to extend your workflow with additional steps such as manual approvals or automated tests. Learn more here.

Q: How can I provision and manage my AWS resources through a release workflow process?
Using AWS CodePipeline and AWS CloudFormation, you can use continuous delivery to automatically build and test changes to your AWS CloudFormation stacks before promoting them to production stacks. This release process lets you rapidly and reliably make changes to your AWS infrastructure. You can extend your workflow with additional actions such as manual approvals, test actions, or invoke AWS Lambda actions. For more details, see Continuous Delivery with AWS CloudFormation page.

Q: What product integrations are available with AWS CodePipeline?
AWS CodePipeline integrates with AWS services such as AWS CodeCommit,

Amazon S3, AWS CodeBuild, AWS CodeDeploy, AWS Elastic Beanstalk, AWS CloudFormation, AWS OpsWorks, Amazon ECS, and AWS Lambda. In addition, AWS CodePipeline integrates with a number of partner tools. For details see the product integrations page. Finally, you can write your own custom actions and integrate any existing tool with CodePipeline. For more details on custom actions, see the Create and Add a Custom Action in AWS CodePipeline page.

Q: Can I get a history of AWS CodePipeline API calls?
Yes. To receive a history of AWS CodePipeline API calls made on your account for security analysis and operational troubleshooting purposes, you simply turn on AWS CloudTrail in the AWS Management Console. For more information, see Logging AWS CodePipeline API calls by Using AWS CloudTrail.

Q: What are the service limits when using AWS CodePipeline?
For information on the service limits, see Limits.

Q: How do I receive notifications or alerts for any events in AWS CodePipeline?
You can create notifications for events impacting your pipelines. Notifications will come in the form of Amazon SNS notifications. Each notification will include a status message as well as a link to the resources whose event generated that notification. Notifications has no additional cost; but, you may be charged for other AWS services utilized by notifications, such as Amazon SNS. To learn how to get started with notifications, see the notifications user guide. Additionally, customers using AWS Chatbot can configure notifications to be sent to their Slack Channels or Amazon Chime chat rooms. For more details please check here.

Partners

Q: What do I need to do to integrate with AWS CodePipeline?
If you're interested in becoming an AWS partner who integrates your developer service with AWS CodePipeline, please contact codepipeline-request@amazon.com.

Security

Q: Can I use AWS Identity and Access Management (IAM) to manage access to AWS CodePipeline?
Yes. AWS CodePipeline supports resource-level permissions. You can specify which user can perform what action on a pipeline. For example, you can provide a user read-only access to a pipeline, if you want them to see the pipeline status but not modify the pipeline. You can also set permissions for any stage or action within a pipeline. For more information on using IAM with AWS CodePipeline, see Access Permissions Reference.

Q: Can I enable the pipeline in one AWS account to be accessed by an IAM user in another AWS account?
Yes. You can create an IAM role in the AWS account that owns the pipeline to delegate access to the pipeline and any related resources to an IAM user in another account. For a walkthrough on enabling such a cross account access, see Walkthrough: Delegating Access Across AWS Accounts For Accounts You Own Using IAM Roles and Configure Cross-Account Access to a Pipeline.

Regions

Q: Which regions does AWS CodePipeline support?
Please refer to Regional Products and Services for details of CodePipeline availability by region.

Billing

Q: How much does AWS CodePipeline cost?
For details on AWS CodePipeline cost, see the pricing page.

AWS CodeStar FAQs

General

Q: What is AWS CodeStar?
AWS CodeStar is a cloud-based development service that provides the tools you need to quickly develop, build, and deploy applications on AWS. With AWS CodeStar, you can set up your entire continuous delivery toolchain in minutes, allowing you to start releasing code faster. AWS CodeStar makes it easy for your whole team to work together securely, with built-in role-based policies that allow you to easily manage access and add owners, contributors, and viewers to your projects. Each AWS CodeStar project comes with a unified project dashboard and integration with Atlassian JIRA software, a third-party issue tracking and project management tool. With the AWS CodeStar project dashboard, you can easily track your entire software development process, from a backlog work item to production code deployment.

Q: Why should I use AWS CodeStar?
You should use CodeStar whenever you want to quickly set up a software development project on AWS, whether you're starting with a full set of tools for a team-based project or only setting up a trial project with a source repository. AWS CodeStar can also be used by anyone interested in learning more about continuous delivery by starting with a full tool chain for a sample project. AWS CodeStar guides you through the setup experience with project templates that set up real applications and be modified at any point in the future to suit your needs.

Q: What can I do with AWS CodeStar?
Start developing on AWS in minutes. AWS CodeStar makes it easy for you to set up your entire development and continuous delivery toolchain for coding, building, testing, and deploying your application code. To start a project, you can choose from a variety of AWS CodeStar templates for Amazon EC2, AWS Lambda, and AWS Elastic Beanstalk. When you choose a project template, the underlying AWS services are provisioned in minutes, allowing you to quickly start coding and deploying your applications.
Work across your team securely. AWS CodeStar enables you to collaborate across your team in a secure manner. AWS CodeStar simplifies the process of setting up project access for teams because it provides built-in role-based policies that follow AWS security best practices. You can easily manage access for project owners, contributors, and viewers without needing to manually configure your own policy for each service.

Manage software delivery easily. AWS CodeStar provides an easy way to coordinate your day-to-day development activities through a unified project dashboard. This lets you monitor application activity, and track progress across all stages of your software development process, including code commits, builds, tests, and deployments, from a central place. AWS CodeStar integrates Atlassian JIRA, a third-party issue tracking and project management tool, allowing you to easily manage JIRA issues directly in the AWS CodeStar dashboard.
Choose from a variety of project templates. With AWS CodeStar project templates, you can easily develop a variety of applications including websites, web applications, web services, and Alexa skills. AWS CodeStar project templates include the code for getting started on supported programming languages including Java, JavaScript, PHP, Ruby, and Python.

Q: How much does AWS CodeStar cost?
There is no additional charge for AWS CodeStar. You pay for AWS resources (e.g. EC2 instances, Lambda executions or S3 buckets) used to in your CodeStar projects. You only pay for what you use, as you use it; there are no minimum fees and no upfront commitments.

Features and functions

Q: How do I get started with AWS CodeStar?
Getting started with AWS CodeStar can be done in a few minutes through the AWS CodeStar console. First, select one of the available CodeStar project templates, which will automatically provision all of the resources needed for your project. Once your project has been provisioned, you can see your running application from the “Application endpoints” tile. Use the steps in the CodeStar console to connect to the AWS CodeCommit source repository for your project and begin coding. You can use the project dashboard to track and manage changes in the release process and see the most recent project activity.

Q: What types of applications can I build with AWS CodeStar?
CodeStar can be used for building web applications, web services and more. The applications run on Amazon EC2, AWS Elastic Beanstalk or AWS Lambda. Project templates are available in several different programming languages including Java, Node.js (Javascript), PHP, Python and Ruby.

Q: How do I add, remove, or change users for my AWS CodeStar projects?
You can add, change or remove users for your CodeStar project through the “Team” section of the CodeStar console. You can choose to grant the users Owner, Contributor or Viewer permissions. You can also remove users or change their roles at any time.

Q: How do AWS CodeStar users relate to IAM users?
CodeStar users are IAM users that are managed by CodeStar to provide pre-built, role-based access policies across your development environment; Because CodeStar users are built on IAM, you still get the administrative benefits of IAM. For example, if you add an existing IAM user to a CodeStar project, the existing global account policies in IAM are still enforced.

Q: Can I work on my AWS CodeStar projects directly from an IDE?
Yes. By installing the AWS Toolkit for Eclipse or Visual Studio you gain the ability to easily configure your local development environment to work with CodeStar Projects; Once installed, developers can then select from a list of available CodeStar projects and have their development tooling automatically configured to clone and checkout their project's source code, all from within their IDE.

Q: How do I configure my project dashboard?
Project dashboards can be configured to show the tiles you want, where you want them; To add or remove tiles, click on the "Tiles" drop-down on your project dashboard. To change the layout of your project dashboard, drag the tile to your desired position.

Q: Are there third party integrations that I can use with AWS CodeStar?
AWS CodeStar works with Atlassian JIRA to integrate issue management with your projects; In addition, you can add partner actions to your project's AWS CodePipeline. To see a list of the available CodePipeline actions, see the AWS CodePipeline integrations page.

Q: I am a third party tools vendors. Can I integrate with AWS CodeStar?
We are starting to build out an integration program for AWS Partner Network (APN) members. If you are already an APN member and interested in learning more, please contact aws-codestar-request@amazon.com.

Q: Can I use AWS CodeStar to help manage my existing AWS applications?
No. AWS CodeStar helps customers quickly start new software projects on AWS. Each CodeStar project includes development tools, including AWS CodePipeline, AWS CodeCommit, AWS CodeBuild and AWS CodeDeploy, that can be used on their own and with existing AWS applications. Customers who are interested in how these tools can help them with their existing AWS applications can visit the respective service pages to learn more.

Regions

Q: In what regions is AWS CodeStar available?
See Regional Products and Services for details. The CodeStar console displays all of your development projects across all regions in a single, centralized view; Your CodeStar project will be saved to the region your console is set to.

Q: Can I use AWS CodeStar to launch applications in other regions?
No. CodeStar configures and manages Code services resources, like a CodeCommit repository, in the regions that you specify in your CodeStar project configuration.

AWS Device Farm FAQs

Testing on real mobile devices

General

Q: What is AWS Device Farm? >>

Q: Who should use AWS Device Farm and why? >>

Q: What types of apps does AWS Device Farm support? >>

Q: Does AWS Device Farm use simulators or emulators? >>

Setting up tests & remote access sessions

Q: How do I get started with AWS Device Farm? >>

Q: Which browsers does the AWS Device Farm console support? >>

Q: Which browsers are supported for testing web applications? >>

Q: What is the maximum file size for apps and tests? >>

Q: Do I need to instrument my app or supply source code? >>

Q: Do you store my app, tests, and other files on your servers? For how long? >>

Q: How do you clean up devices after my testing is completed? >>

Q: Do you modify my app? >>

Selecting devices

Q: Which devices are available in AWS Device Farm? How do you select the devices in your fleet? >>

Q: Does AWS Device Farm have international devices from markets like Europe, China, and India? >>

Q: How do I select devices? Can I retest on the same device? >>

Q: Are any apps pre-installed on AWS Device Farm test devices? >>

Q: Are devices able to communicate with other services or systems that are available on the Internet? >>

Q: Can I test different carrier connections and conditions? >>

Q: Can I make phone calls or send SMS from the devices? >>

Q: Can I use the device camera? >>

Testing your app

Q: I don’t have any automated test scripts yet. What do the built-in tests do? >>

Q: What does Fuzz do? >>

Q: I test using an automation framework. Which frameworks do you support? >>

Q: Which test frameworks do you support for web applications? >>

Q: Can you add support for a modified framework or one I designed myself? >>

Q: How does AWS Device Farm decide when to take a screenshot during a test? >>

Q: Android: Is Google Play Services available on your devices? Which version is installed? >>

Q: Android: Is there a default Google account on the devices? >>

Q: Does AWS Device Farm support record and playback automation or do I have to write my scripts? >>

Q: iOS: Do I need to add your UDIDs to my provisioning profile? >>

Q: iOS: My app does not contain debug symbols. Can I supply a dSYM file to AWS Device Farm? >>

Q: Android: My app is obfuscated. Can I still test my app on AWS Device Farm? >>

Q: My app serves ads. Will they be displayed on your devices? Will my ad provider flag this as abuse and ban my account? >>

Q: Can I access the machine hosting the device or access its shell as part of my tests? Can I reach the Internet from it? >>

Q: I'd like to supply media or other data for my app to consume. How do I do that? >>

Q: My app requires dependencies to test all functionality. Can I install other apps? >>

Q: Can I test upgrade flows for my app? How do I install an old version of my app? >>

Q: My app makes use of location services. Can I specify the physical location of the device? >>

Q: Can I run localization tests? How do I change the language of the device? >>

Q: How long does it take before my test starts? >>

Q: What is the maximum test time allowed? >>

Q: Does AWS Device Farm provide a way to run tests and get results through an API? >>

Reviewing results

Q: What's in an AWS Device Farm test report? >>

Q: Which device logs are included in an AWS Device Farm report? >>

Q: My tests generate and save additional log files. Will I see them in my AWS Device Farm reports? >>

Pricing

Q: How much does AWS Device Farm cost? >>

Q: How does the free trial work? >>

Q: What is a device minute? >>

Q: How does the free trial work? >>

Q: What is the unmetered plan and how do device slots work? >>

Q: What if my testing needs change and I need to add or remove device slots? >>

Q: If I'm on an unmetered plan, can I still make use of metered billing? >>

Q: What is a private device? >>

Q: How do private device subscriptions work and how are they priced? >>

Q: Can I use both private devices and public devices? >>

Testing on desktop browsers

Q: What is Selenium? >>

Q: What is Desktop Browser Testing on AWS Device Farm? >>

Q: How do I get started with Desktop Browser Testing on AWS Device Farm? >>

Q: What operating system are the browsers hosted on? >>

Q: What desktop browsers does AWS Device Farm support? >>

Q: What desired capabilities does AWS Device Farm support? >>

Q: What artifacts are available for troubleshooting test failures? >>

Q: Can I use AWS Device Farm to test my web app on real mobile devices? >>

Q: What are limits of Desktop Browser Testing on AWS Device Farm? >>

Q: How much does Desktop Browser Testing on AWS Device Farm cost? >>

Q: What is instance minute? >>

AWS X-Ray FAQs

General

Q: What is AWS X-Ray?
AWS X-Ray helps developers analyze and debug production, distributed applications, such as those built using a microservices architecture. With X-Ray, you can understand how your application and its underlying services are performing to identify and troubleshoot the root cause of performance issues and errors. X-Ray provides an end-to-end view of requests as they travel through your application, and shows a map of your application's underlying components. You can use X-Ray to analyze both applications in development and in production, from simple three-tier applications to complex microservices applications consisting of thousands of services.

Q: Why should I use X-Ray?
Currently, if you build and run distributed applications, you have to rely on a per-service or per-resource process to track requests for your application as it travels across various components that make up your application. This problem is further complicated by the varying log formats and storage mediums across frameworks, services, and resources your application runs on or uses. This makes it difficult to correlate the various pieces of data and create an end-to-end picture of a request from the time it originates at the end-user or service to when a response is returned by your application. X-Ray provides a user-centric model, instead of service-centric or resource-centric model, for collecting data related to requests made to your application. This model enables you to create a user-centric picture of requests as they travel across services and resources. By correlating and aggregating data on your behalf, X-Ray enables you to focus on improving the experience for end-users of your application.

Q: What can I do with X-Ray?
X-Ray makes it easy for you to:

- **Create a service map –** By tracking requests made to your applications, X-Ray can create a map of services used by your application. This provides you with a view of connections among services in your application, and enables you to create a dependency tree, detect latency or errors when working across AWS Availability Zones or Regions, zero in on services not operating as expected, and so on.
- **Identify errors and bugs –** X-Ray can automatically highlight bugs or errors in your application code by analyzing the response code for each request made to your application. This enables easy debugging of application code without requiring you to reproduce the bug or error.
- **Build your own analysis and visualization apps –** X-Ray provides a set of query APIs you can use to build your own analysis and visualizations apps that use the data that X-Ray records.

Core concepts

Q: What is a trace?
An X-Ray trace is a set of data points that share the same trace ID. For example, when a client makes a request to your application, it is assigned a unique trace ID. As the request makes its way through services in your application, the services relay information regarding the request back to X-Ray using this unique trace ID. The piece of information relayed by each service in your application to X-Ray is a segment, and a trace is a collection of segments.

Q: What is a segment?
An X-Ray segment encapsulates all the data points for a single component (for example, authorization service) of the distributed application. Segments include system-defined and user-defined data in the form of annotations and are composed of one or more sub-segments that represent remote calls made from the service. For example, when your application makes a call to a database in response to a request, it creates a segment for that request with a sub-segment representing the database call and its result. The sub-segment can contain data such as the query, table used, timestamp, and error status.

Q: What is an annotation?
An X-Ray annotation is system-defined or user-defined data associated with a segment. A segment can contain multiple annotations. System-defined annotations include data added to the segment by AWS services, whereas user-defined annotations are metadata added to a segment by a developer. For example, a segment created by your application can automatically be injected with region data for AWS service calls, whereas you might choose to add region data yourself for calls made to non-AWS services.

Q: What are errors?
X-Ray errors are system annotations associated with a segment for a call that results in an error response. The error includes the error message, stack trace, and any additional information (for example, version or commit ID) to associate the error with a source file.

Q: What is sampling?
To provide a performant and cost-effective experience, X-Ray does not collect data for every request that is sent to an application. Instead, it collects data for a statistically significant number of requests. X-Ray should not be used as an audit or compliance tool because it does not guarantee data completeness.

Q: What is the X-Ray agent?
The X-Ray agent collects data from log files and sends them to the X-Ray service for aggregation, analysis, and storage. The agent makes it easier for you to send data to the X-Ray service, instead of using the APIs directly, and is available for Amazon Linux AMI, Red Hat Enterprise Linux (RHEL), and Windows Server 2012 R2 or later operating systems.

Using AWS X-Ray

Q: How do I get started with X-Ray?
You can get started with X-Ray by including the X-Ray language SDK in your application and installing the X-Ray agent. For more information see the X-Ray user guide.

Q: What types of applications can I use with X-Ray?
X-Ray can be used with distributed applications of any size to trace and debug both synchronous requests and asynchronous events. For example, X-Ray can be used to trace web requests made to a web application or asynchronous events that utilize Amazon SQS queues.

Q: Which AWS services can I use with X-Ray?
You can use X-Ray with applications running on EC2, ECS, Lambda, Amazon SQS, Amazon SNS and Elastic Beanstalk. In addition, the X-Ray SDK automatically captures metadata for API calls made to AWS services using the AWS SDK. In addition, the X-Ray SDK provides add-ons for MySQL and PostgreSQL drivers.

Q: What code changes do I need to make to my application to use X-Ray?
If you're using Elastic Beanstalk, you will need to include the language-specific X-Ray libraries in your application code. For applications running on other AWS services, such as EC2 or ECS, you will need to install the X-Ray agent and instrument your application code.

Q: Does X-Ray provide an API?
Yes, X-Ray provides a set of APIs for ingesting request data, querying traces, and configuring the service. You can use the X-Ray API to build analysis and visualization applications in addition to those provided by X-Ray.

Regions

Q: In which regions is X-Ray available?
See Regional Products and Services for details.

Q: Can I use X-Ray to track requests from applications or services spread across multiple regions?
Yes, you can use X-Ray to track requests flowing through applications or services across multiple regions. X-Ray data is stored locally to the processed region but with enough information to enable client applications to combine the data and provide a global view of traces. Region annotation for AWS services will be added automatically, however, customers will need to

instrument custom services to add the regional annotation to make use of the cross-region support.

Data handling

Q: How long does it take for trace data to be available in X-Ray?
Trace data sent to X-Ray is generally available for retrieval and filtering within 30 seconds of it being received by the service.

Q: How far back can I query the trace data? How long does X-Ray store trace data for?
X-Ray stores trace data for the last 30 days. This enables you to query trace data going back 30 days.

Q: Why do I sometimes see partial traces?
X-Ray makes the best effort to present complete trace information. However, in some situations (connectivity issues, delay in receiving segments, and so on) it is possible that trace information provided by the X-Ray APIs will be partial. In those situations, X-Ray tags traces as incomplete or partial.

Q: My application components run in their own AWS accounts. Can I use X-Ray to collect data across AWS accounts?
Yes, the X-Ray agent can assume a role to publish data into an account different from the one in which it is running. This enables you publish data from various components of your application into a central account.

Amazon WorkSpaces FAQs

General

Q: What is Amazon WorkSpaces?

Amazon WorkSpaces is a managed, secure cloud desktop service. You can use Amazon WorkSpaces to provision either Windows or Linux desktops in just a few minutes and quickly scale to provide thousands of desktops to workers across the globe. You can pay either monthly or hourly, just for the WorkSpaces you launch, which helps you save money when compared to traditional desktops and on-premises VDI solutions. Amazon WorkSpaces help you eliminate the complexity in managing inventory, OS versions and patches, and Virtual Desktop Infrastructure (VDI), which helps simplify your desktop delivery strategy. With Amazon WorkSpaces, your users get a fast, responsive desktop of their choice that they can access anywhere, anytime, from any supported device.

Q: What is an Amazon WorkSpace?

An Amazon WorkSpace is a cloud-based virtual desktop that can act as a replacement for a traditional desktop. A WorkSpace is available as a bundle of operating system, compute resources, storage space, and software applications that allow a user to perform day-to-day tasks just like using a traditional desktop.

Q: How do I connect to my Amazon WorkSpace?

A user can connect to a WorkSpace from any supported device using the free Amazon WorkSpaces client application on supported devices including Windows and Mac computers, Chromebooks, iPads, Fire tablets, Android tablets, or using Chrome or Firefox web browsers. Users will connect using credentials set up by an administrator or using their existing Active Directory credentials if you've chosen to integrate your Amazon WorkSpaces with an existing Active Directory domain. Once the user is connected to a WorkSpace they can perform all the usual tasks they would do on a desktop computer.

Q: How can I get started with Amazon WorkSpaces?

To get started with Amazon WorkSpaces, you will need an AWS account. You can use this account to sign into the AWS Management Console and you can then quickly provision

Amazon WorkSpaces for yourself and any other users in your organization who might require one. To provision an Amazon WorkSpace, first select a user from your directory. Next, select an Amazon WorkSpaces bundle for the user. The Amazon WorkSpaces bundle specifies the resources you need, which desktop operating system you want to run, how much storage you want to use and the software applications you want prepackaged. Finally, choose a running mode for their Amazon WorkSpace – pick AlwaysOn if you want to use monthly billing, or AutoStop if you want to use hourly billing. Once your WorkSpace is provisioned, the user will receive an email with instructions for connecting to their WorkSpace. You can use this same process to provision multiple WorkSpaces at the same time.

Q: Which Amazon WorkSpaces bundles are available?

You can find the latest information on Amazon WorkSpaces bundles here.

Q: Which operating systems are available for use with Amazon WorkSpaces?

Amazon WorkSpaces offers Amazon Linux WorkSpaces built on Amazon Linux 2 LTS, or Windows 10 desktop experiences. The Windows 10 desktop experiences is powered by Windows Server 2016. If your organization is eligible to bring their own Windows Desktop licenses, you can run the Windows 10 Enterprise operating system on your Amazon WorkSpaces.

Q: What are the root and user volumes mapped to for Amazon Linux WorkSpaces and Amazon WorkSpaces with Windows?

For Amazon Linux WorkSpaces, root volume is mapped to /, and user volume is mapped to /home

For Windows, root volume is mapped to C: drive, and user volume is mapped to D: drive

Q: Can I migrate users from an Amazon WorkSpaces Windows 7 bundle to a Windows 10 bundle?

Yes. WorkSpaces migrate enables WorkSpaces migration to a new bundle or compute type with the user volume data preserved. You could perform migrate operations to move your users to the Windows 10 Desktop experience. To get started, go to the Amazon WorkSpaces console, select the WorkSpace, click “Action > Migrate WorkSpaces”, then select a target bundle with the Windows 10 desktop experience.

Q: How does a user get started with their Amazon WorkSpace once it has been provisioned?

When Amazon WorkSpaces are provisioned, users receive an email providing instructions on where to download the WorkSpaces clients they need, and how to connect to their WorkSpace. If you are not integrating with an existing Active Directory, the user will have the ability to set a password the first time they attempt to connect to their WorkSpace. If the AWS Directory Services AD Connector has been used to integrate with an existing Active Directory domain, users will use their regular Active Directory credentials to log in.

Q: What does a user need to use an Amazon Workspace?

A user needs to have an Amazon WorkSpace provisioned for them, and a broadband Internet connection. To use an Amazon WorkSpaces client application to access their WorkSpace, they will need a supported client device (PC, Mac, Linux, iPad, Kindle Fire, or Android tablet), and an Internet connection with TCP ports 443 & 4172, and UDP port 4172 open.

Q: Once users connect to their Amazon WorkSpace can they personalize it with their favorite settings?

An administrator can control what a user can personalize in their WorkSpace. By default, users can personalize their WorkSpaces with their favorite settings for items such as wallpaper, icons, shortcuts, etc. These settings will be saved and persist until a user changes them. If an administrator wishes to lock down a WorkSpace using tools like Group Policy for Windows, this will restrict a user’s ability to personalize their WorkSpaces.

Q: Can users install applications on their Amazon WorkSpace?

By default, users are configured as local administrators of their WorkSpaces. Administrators can change this setting and can restrict users’ ability to install applications with a technology such as Group Policy.

Q: Are Amazon WorkSpaces persistent?

Yes. Each WorkSpace runs on an individual instance for the user it is assigned to. Applications and users’ documents and settings are persistent.

Q: How is a user’s data backed up?

The user volume on a WorkSpace is backed up every 12 hours. In case of a WorkSpace failure, AWS can restore this volume from the last backup. If Amazon WorkDocs Sync is enabled on a WorkSpace, the folder a user chooses to sync will be continuously backed up and stored in Amazon WorkDocs.

Q: Do users need an AWS account?

No. An AWS account is only needed to provision WorkSpaces. To connect to WorkSpaces, users will require only the information provided in the invitation email they will receive when their WorkSpace is provisioned.

Q: If I am located a significant distance from the region where my Amazon WorkSpace is located, will I have a good user experience?

If you are located more than 2000 miles from the regions where Amazon WorkSpaces is currently available, you can still use the service, but your experience may be less responsive. The easiest way to check performance is to use the Amazon WorkSpaces Connection Health Check Website. You can also refer to the Regional Products and Services page for details of Amazon WorkSpaces service availability by region.

Q: Does Amazon WorkSpaces offer a set of public APIs?

Yes, public APIs are available for creating and managing Amazon WorkSpaces programmatically. APIs are available via the AWS CLI and SDK; you can learn more about the APIs in the documentation.

Q: Do the Amazon WorkSpaces APIs log actions in AWS CloudTrail?

Yes. Actions on Amazon WorkSpaces performed via the WorkSpaces APIs will be included in your CloudTrail audit logs.

Q: Is there Resource Permission support with the Amazon WorkSpaces APIs?

Yes. You can specify which Amazon WorkSpaces resources users can perform actions on. For details see the documentation.

Q: Do I need to use the AWS Management Console to get started with Amazon WorkSpaces?

To get started with Amazon WorkSpaces, you will need to register a directory with the WorkSpaces service. You can use AWS Management Console or Amazon WorkSpaces APIs to register a directory with the WorkSpaces service and then create and manage WorkSpaces.

Q: Can I deploy my WorkSpaces in the AWS GovCloud (US) Regions?

Yes. You can deploy WorkSpaces in the AWS GovCloud (US West) region to meet US federal, state, and local government requirements. Go here for details on the AWS GovCloud (US) Regions.

Bundles and Custom Images

Q: What applications are available with Amazon Linux WorkSpaces?

Amazon Linux WorkSpaces come with a curated selection of applications at no additional cost that include LibreOffice, Firefox Web Browser, Evolution mail, Pidgin IM, GIMP, and other desktop utilities and tools. You can always add more software from the Amazon Linux repositories using yum. To install an available package from the Amazon Linux repositories, simply type “yum install [package-name]”. You can also add software from RPM based public and private Linux repositories at any time.

Q: What applications are available with Amazon WorkSpaces with Windows 10 Experience?

Amazon WorkSpaces come with a default set of applications at no additional cost that include Internet Explorer 11, and Firefox. You can choose to add “Plus” application bundles to your Amazon WorkSpaces with Windows 10 which include Microsoft Office Professional 2016, and Trend Micro Worry-Free Business Security, for an additional monthly fee.

Q: Can I create custom images for Amazon WorkSpaces?

Yes, as an administrator you can create a custom image from a running WorkSpace. Once you have customized your WorkSpace with your applications and settings, select the WorkSpace in the console and select “Create Image.” This creates an image with your applications and settings. Custom images created from Amazon WorkSpaces with GPU-enabled bundles (Graphics and GraphicsPro) can only be used with Graphics bundles. Custom images created from Value, Standard, Performance, Power, or PowerPro bundles can only be used with those bundles. Most WorkSpace images are available within 45 minutes. See the custom image documentation for more detail.

Q: How do I launch an Amazon WorkSpace from a custom image?

To launch an Amazon WorkSpace from a custom image, you will first need to pair the custom image with a hardware type you want that WorkSpace to use, which results in a bundle. You can then publish this bundle through the console, then select the bundle when launching new WorkSpaces.

Q: What is the difference between a bundle and an image?

An image contains only the OS, software and settings. A bundle is a combination of both that image and the hardware from which a WorkSpace can be launched.

Q: How many custom images can I create?

As an administrator, you can create as many custom images as you need. Amazon WorkSpaces sets default limits, but you can request an increase in these limits here. To see the default limits for Amazon WorkSpaces, please visit our documentation.

Q: Can I update the image in an existing bundle?

Yes. You can update an existing bundle with a new image that contains the same tier of software (for example containing the Plus software) as the original image.

Q: Can I copy my Amazon WorkSpaces Images to other AWS Regions?

Yes, you can use the WorkSpaces console, APIs, or CLI to copy your WorkSpaces Images to other AWS Regions where WorkSpaces is available. Log on to the WorkSpaces console and navigate to the "**Images**" section from the left hand navigation menu. Simply select the image you would like to copy, click on the "**Actions**" button and select the "**Copy Image**" option to get started.

Q: How can I tell if the Image I copied is available for me to use?

As soon as you initiate a copy operation, you will be provided a unique identifier for the new Image being created as a copy of the original one. You can use that identifier to look up the status of that Image in the destination Region through the WorkSpaces console, APIs, or CLI.

Q: Can I cancel a pending Image copy operation?

Once initiated, you cannot cancel a pending Image copy operation. You can delete the Image in the destination Region if the Image is not required.

Q: Are there any data transfer fees for copying Images?

No. There are no additional fees for copying Images across Regions. Maximum Image limits for your account in destination AWS Region will still apply. Once you reach this limit you will not be able to copy more Images.

Q: Can I bulk copy multiple Images to another Region?

You can copy Images one by one to another AWS Region. You can use CopyWorkspaceImage API to programmatically copy Images.

Q: Can I copy a BYOL Image to another AWS Region?

Yes. You can copy a BYOL WorkSpace Image to another AWS Region if the destination AWS Region is enabled for BYOL.

Q: Can I copy an Image to the same Region?

Yes. You can use the copy Image operation to make a copy of the WorkSpaces Image in the same region.

Q: What type of Amazon Elastic Block Store (EBS) volumes does Amazon WorkSpaces offer?

All Amazon WorkSpaces launched after 31st January 2017 are built on general purpose solid-state drives (SSD) EBS volumes for both root and user volumes. Amazon WorkSpaces launched prior to 31st January 2017 are configured with EBS magnetic volumes. You can switch your Amazon WorkSpaces using magnetic EBS volumes to SSD EBS volumes by rebuilding them (more information can be found here). You can learn more about SSD EBS volumes here, and magnetic EBS volumes here.

Q: Can I use custom images to launch WorkSpaces with SSD volumes, even if they were created using WorkSpaces with magnetic EBS volumes?

Yes. You can use your custom images to launch WorkSpaces with SSD EBS volumes, even if they were created using WorkSpaces with magnetic EBS volumes.

Q: Do I need to provide an AMI build using WorkSpaces with SSD EBS volumes when using my own Windows desktop licenses (BYOL)?

No. You can use the AMIs you built as part of the BYOL process without any additional changes.

Q: How do I deploy applications to my users?

You have flexibility in how you deploy the right set of applications to users. First, you choose which image type to build from, either basic or Plus, which determines the default applications that will be in the WorkSpaces. Second, you can install additional software on a WorkSpace and create a custom image which can be used to launch more WorkSpaces. For more detail see the bundle documentation.

Q: Which software can I install on an Amazon WorkSpace?

For Amazon Linux, any application available in the Amazon Linux repositories is compatible and can be installed using yum install [package-name].

For Windows, any applications that are compatible with the Windows 10 experience provided by Windows Server 2016, should run on your WorkSpaces. We recommend testing any software you would like to deploy on a ‘test’ WorkSpace before delivering it to more users. You are responsible for ensuring that you remain compliant with any licensing restrictions associated with any software you intend to install on a WorkSpace.

Storage and Hardware Bundles

Q: Can I increase the size of my Amazon WorkSpaces storage volumes?

Yes. You can increase the size of the root and user volumes attached to your WorkSpaces at any time. When you launch new WorkSpaces, you can select bundled storage configurations for root and user volumes, or choose your preferred storage size greater than the provided storage configurations. For storage configurations with 80 GB Root volume, you can choose 10 GB, 50 GB or 100 GB for User volume. You can use storage configurations with 175 GB to 2000 GB Root volume along with 100 GB to 2000 GB User volume. Please note that you need to set the Root volume to 175 GB in order to expand the User volume in the range of 100GB to 1000GB. After your WorkSpaces have been launched, you can only increase the size of the volumes using the above configurations to up to 2000 GB for each Root and User volume.

Q: Can I decrease the size of storage volumes?

No. To ensure that your data is preserved, the volume sizes of either volume cannot be reduced after a WorkSpace is launched. You can launch a Value, Standard, Performance, Power or PowerPro WorkSpace with a minimum of 80 GB for the root volume and 10 GB for the user volume. You can launch a Graphics or GraphicsPro WorkSpace with a minimum of 100 GB for the root volume and 100 GB for the user volume. For more information about configurable storage, see Modifying WorkSpaces.

Q: How do I change the size of my Amazon WorkSpaces storage volumes?

You can change the size of your storage volumes via the Amazon WorkSpaces management console, or through the Amazon WorkSpaces API.

WorkSpaces users can also increase the size of their storage volume directly in the WorkSpaces client if this self-service management capability is enabled by the WorkSpaces administrator.

Q: Is the storage configuration for a WorkSpace preserved when I rebuild it?

Yes, each rebuild preserves the storage allocation size for WorkSpaces when using default bundles. If a WorkSpace has its volumes extended, and is rebuilt, the larger volume sizes will be preserved, even if the bundle's drive sizes are smaller.

Q: Is the storage configuration for a WorkSpace preserved when I restore it?

Yes, each restore preserves your existing storage allocation size when using WorkSpaces default bundles. For example, restoring a WorkSpace with 80GB Root and 100GB User volumes will result in a rebuilt WorkSpace with 80GB Root and 100GB User.

If the storage allocation of a Custom bundle is increased and a linked WorkSpace is restored, the WorkSpace volumes will be increased to match the bundle’s new volume sizes.

Q: What data can I retain after a WorkSpaces migrate?

All data in the latest snapshot of the original user volume will be retained. For a Windows WorkSpace, the D drive data captured by the latest snapshot will be retained after migration and the C drive will be newly created from the target bundle image. In addition, migrate attempts to move data from the old user profile to the new one. Data that cannot be moved to the new profile will be preserved in a .notMigrated folder. For more information, please refer to the documentation.

Q: Can I move an existing WorkSpace from a public bundle to a custom bundle?

Yes. The WorkSpaces migrate function allows you to replace your WorkSpace’s root volume with a base image from another bundle. Migrate will recreate the WorkSpace using a new root volume from the target bundle image, and the user volume from the latest original user volume snapshot. For detailed information about migrate, please refer to the documentation.

Q: What’s the difference between migrate and rebuild?

WorkSpaces Migrate allows you to switch to a new bundle and have your user profile regenerated. Rebuild just refreshes your WorkSpace with a root volume generated from the base image of the original bundle.

Q: What happens if I rebuild my WorkSpace after migrate?

Migrate associates your WorkSpace with a new bundle. And a rebuild after migration will uses the newly associated bundle to generate the root volume.

Q: Can I expand Amazon WorkSpaces magnetic storage volumes?

No, configurable storage volumes are only available when using solid state drives (SSD). Any WorkSpaces launched before February 2017 might still use magnetic storage volumes. To switch from magnetic to SSD drives, rebuild your WorkSpaces.

Q: How do custom images affect my root volume size?

The root volume size of WorkSpaces launched from a custom image is, by default, the same size as the custom image. For example, if your custom image has a root volume of 100 GB, all WorkSpaces launched from that image also have a root volume size of 100 GB. You can increase your root volume size when you launch your WorkSpace, or any time after that.

Q: Can I change my Amazon WorkSpaces bundle without performing WorkSpaces migrate?

Yes. You can switch between Value, Standard, Performance, Power, or PowerPro bundles by using the Amazon WorkSpaces management console or the WorkSpaces API. When you switch hardware bundles, your WorkSpaces restart immediately. When they resume, your operating system, applications, data, and allocated storage on both the root and user volumes are all preserved.

For example, you can launch a Standard bundle (2vCPU, 4 GiB), and later expand the volume size on both volumes to 500 GB. You can then switch to the Performance bundle (2vCPU, 7.5 GiB) while preserving your operating system, applications, and data in the expanded volume.

Users can also change their WorkSpaces bundle directly from the WorkSpaces client if this self-service management capability is enabled by their WorkSpaces administrator.

Q: How can I track my storage and bundle switch requests?

You can use AWS CloudTrail to track the changes that you have requested.

Q: I currently bring my own Windows licenses. Can I expand my storage volumes and switch my WorkSpaces bundles?

Yes. You can take advantage of both these features even if you bring your own Windows desktop licenses. By default, you can switch WorkSpaces bundles for up to 20% of the total number of your WorkSpaces in a week. To switch more than 20% of your WorkSpaces, contact us.

Q: Does a WorkSpace running in AutoStop mode need to be running to apply a change to the bundle type?

No. When you make a change, we start a WorkSpace that isn’t running, apply the bundle change, restart it so that the changes take effect, and then stop it again.
For example, you change the bundle type on a stopped Standard (2vCPU, 4 GiB) WorkSpace to Performance. We start your Standard WorkSpace, apply the bundle change, and restart it. Following the restart, your WorkSpace has Performance hardware (2vCPU, 7.5 GiB).

Q: How do I get charged if I change storage size or hardware bundle during a month?

For either change, you get charged the monthly price for AlwaysOn or the monthly fee for AutoStop WorkSpaces prorated on a per day basis.

For example, if you increase the volume on the 10th of a month on an AlwaysOn Power WorkSpace with 175 GB, and 100 GB for root and user volumes respectively, you are charged $78 for the Power WorkSpace and $11.6 for 20 days of additional 175 GB at $0.1/GB-month (in US-East-1). Similarly, switching a bundle—for example, from Value to Standard—on the 15th of a month results in 15 days of Value WorkSpaces charge ($12.5 in US-East-1) and 15 days of Standard WorkSpaces charge ($17.5 in US-East-1).

Q: How often can I increase volume sizes or change hardware bundle of a WorkSpace?

You can increase volume sizes or change a WorkSpace to a larger hardware bundle once in a 6-hour period. You can also change to a smaller hardware bundle once in a 30-day period. For a newly launched WorkSpace, you must wait 6 hours before requesting a larger bundle.

For example, if you increase the root and user volume of a Standard WorkSpace on 5th Dec at 11:00 AM and change it to Performance WorkSpace at the same time, on 5th Dec at 4:00 PM, you can again increase the root and user volume, and change the hardware bundle. If you change the Performance WorkSpace to a Standard WorkSpace on 6th Dec at 12:00 and want to go to a further smaller bundle (Value), you would be able to make this change on 6th Jan at 12:00.

GPU-Enabled Bundles

Q: Does Amazon WorkSpaces offer GPU-enabled cloud desktops?

Yes. Amazon WorkSpaces offers a Graphics bundle for general purpose graphics applications such as CAD/CAM software, commercial and industrial modeling, prototyping, and mainstream graphics development. Amazon WorkSpaces also offers a GraphicsPro bundle for performance intensive graphics applications such as 3D visualizations, graphics rendering, video encoding, and high end gaming. Graphics and GraphicsPro bundles are available in English and Japanese.

Q: What are GPU-enabled bundles from Amazon WorkSpaces?

GPU-enabled bundles from Amazon WorkSpaces are cloud desktops optimized for workloads that benefit from graphics hardware acceleration. You can choose the Graphics or the GraphicsPro bundle, depending on the performance requirements of your graphics workload.

The Graphics bundles are well-suited good for general-purpose graphics workloads such as computer-aided design, manufacturing, and engineering software. Each Graphics bundle comes with a high-performance NVIDIA GPU with 1,536 CUDA cores and 4 GB of video memory. Each Graphics bundle includes 8 vCPUs, 15 GiB of RAM, 4 GB of video memory, and 100 GB of storage on the user volume, and 100 GB of general-purpose persistent storage on the root volume. Graphics bundles provide a Windows 10 desktop experience powered by Windows Server 2016.

The GraphicsPro bundles are ideal for complex graphics applications such as 3D visualizations, 3D rendering, image processing, video encoding, media encoding, seismic visualization, and data mining. GraphicsPro bundles come with a dedicated high-performance NVIDIA Tesla M60 GPU with 2048 parallel processing cores, and a hardware encoder capable of supporting up to 10 H.265 (HEVC) 1080p30 streams and up to 18 H.264 1080p30 streams. Each GraphicsPro bundle contains 16 vCPUs, 122 GiB of RAM, 8 GB of video memory, and a minimum of 100 GB for the root volume and 100 GB for the user volume. GraphicsPro bundles provide a Windows 10 desktop experience powered by Windows Server 2016.

Q: In which AWS Regions can I launch GPU-enabled Amazon WorkSpaces bundles?

You can launch Graphics or GraphicsPro bundles in the following AWS Regions: US East (N. Virginia), US West (Oregon), EU (Ireland), EU (Frankfurt), Asia Pacific (Sydney), Asia Pacific (Tokyo), and Asia Pacific (Singapore).

Q: Can I create a custom image for my GPU-enabled bundles?

Yes. Custom images created from a GPU-enabled Amazon WorkSpaces bundle can only be used with the same type of bundle. For instance, you cannot use an image made from a Graphics bundle to launch a GraphicsPro WorkSpace.

Q: How do I get started with GPU-enabled Amazon WorkSpaces bundles?

You can launch Graphics or GraphicsPro bundles using the Amazon WorkSpaces Management Console, or the Amazon WorkSpaces API. When launching a new WorkSpace, simply select the Graphics or GraphicsPro bundle.

Q: How much bandwidth do GPU-enabled Amazon WorkSpaces consume?

Bandwidth used by GPU-enabled Amazon WorkSpaces bundles depends on the tasks being performed. If there aren’t many changes taking place on the screen, the bandwidth used is generally less than 300 kbps. If there is context switching between multiple windows, or if 3D models are being manipulated, bandwidth use can increase to several megabits per second.

Windows BYOL

Q: Can I bring my Windows Desktop licenses to Amazon WorkSpaces?

Yes, you can bring your own Windows 10 desktop licenses to WorkSpaces if they meet Microsoft’s licensing requirements. WorkSpaces gives you an option to run Windows 10 desktop images on physically dedicated hardware, which lets you maintain license compliance for your Windows desktops when you bring your own licenses to WorkSpaces.

Q: Can I bring my own Windows Desktop licenses for Amazon WorkSpaces Graphics bundles?

Yes, you can. Please contact us if this is something you’d like to do.

Q: What versions of Windows desktop licenses can I bring to Amazon WorkSpaces?

If your organization meets the licensing requirements set by Microsoft, you can bring your Windows 10 Enterprise or Professional licenses to Amazon WorkSpaces. You cannot use Windows OEM licenses for your Amazon WorkSpaces. Please consult with Microsoft if you have any questions about your eligibility to bring your own Windows Desktop licenses.

Q: What benefits are there in bringing my own Windows desktop licenses to Amazon WorkSpaces?

By bringing your own Windows Desktop licenses to Amazon WorkSpaces, you will save $4 per Amazon WorkSpace per month when being billed monthly, and you will save money on the hourly usage fee when being billed hourly (see the Amazon WorkSpaces pricing page for more information). Additionally, you can now use a single golden image to manage your physical and virtual desktop deployments.

Q: What are the requirements for bringing my Windows desktop Licenses to Amazon WorkSpaces?

You need an active and eligible Microsoft Volume Licensing (VL) agreement with Software Assurance contracts to bring your Windows 10 Desktop licenses to Amazon WorkSpaces. Please consult with your Microsoft representative to confirm your eligibility in bringing your Windows Desktop licenses to Amazon WorkSpaces.

Q: How do I get started with bringing my Windows desktop licenses to Amazon WorkSpaces?

In order to ensure that you have adequate dedicated capacity allocated to your account, please reach out to your AWS account manager or sales representative to enable your account for BYOL. alternatively, you can create a Technical Support case with Amazon WorkSpaces to get started with BYOL.

Once enabled for your account, it's easy to bring your existing Windows 10 Desktop OS to Amazon WorkSpaces. First, import your existing Windows desktop OS using the VM Import API. Then create your new WorkSpaces image, based on the imported VM, using the Create Image action on the Images page in the WorkSpaces Admin Console. Finally, create a custom WorkSpaces bundle using the Bundles tabs in the WorkSpaces Admin Console. You can then launch your newly created custom WorkSpaces bundle as new WorkSpaces for your users through the WorkSpaces Management console.

You can see more information on the BYOL process in our documentation.

Q: How will I activate my Windows 10 Desktop operating system on Amazon WorkSpaces?

You can activate your Windows 10 Desktop operating system using existing Microsoft activation servers that are hosted in your VPC, or ones that can be reached from the VPC in which Amazon WorkSpaces are launched.

Q: Can I create a new custom image of the Windows 10 Desktop image uploaded to Amazon WorkSpaces?

Yes. You can use the standard WorkSpaces image management functionality to further customize the Windows 10 Desktop image and save it as a new Amazon Workspaces image in your account.

Q: Can I launch new Amazon WorkSpaces using one of the pre-configured public bundles in the same directory with custom Windows bundles I brought to WorkSpaces?

No. Your custom WorkSpaces that support BYOL for Windows 10 Desktops are launched on physically dedicated hardware to meet license compliance requirements with Microsoft. WorkSpaces launched in a directory marked for dedicated hardware can only be from the custom bundle you created that has your Windows 10 Desktop image.
If you wish to launch WorkSpaces from public bundles to users in the same domain, you can create a new AWS AD Connector directory that points to the same Microsoft Active Directory as your Windows 10 Desktop WorkSpaces, and launch WorkSpaces in that directory as you normally would through the AWS Management Console or the WorkSpaces SDK and CLI.

Q: Would I need to commit to a certain number of Amazon WorkSpaces if I want to bring my own Windows desktop license?

Yes, you need to commit to running 200 Amazon WorkSpaces in a region per month on hardware that is dedicated to you. If you want to bring your own Windows desktop licenses for graphics use cases, you need to commit to at least 4 monthly or 20 hourly GPU-enabled WorkSpaces.

Q: How long will it take before I can launch Amazon WorkSpaces using my own Windows desktop licenses and image?

It can take a few hours after you perform the “Create Image” operation for your custom Windows desktop image to be available to use. You can check the status of your custom image in the WorkSpaces Console, API, or CLI.

Q: Will all of my dedicated Amazon WorkSpaces launch in a single AZ?

No. Amazon WorkSpaces launched on dedicated hardware will be balanced across two AZs. You select the AZs for Amazon WorkSpaces when you create the directory in which your Amazon WorkSpaces will be launched, and subsequent launches of Amazon WorkSpaces are automatically load balanced across the AZs selected when you created the directory.

Q: What happens when I terminate Amazon WorkSpaces that are launched on physically dedicated hardware?

You can terminate Amazon WorkSpaces when you no longer need them. You will only be billed for the Amazon WorkSpaces that are running.

Q: What happens to Amazon WorkSpaces that are rebuilt, restored, or restarted on physically dedicated hardware?

Amazon WorkSpaces that are rebuilt, restored, or restarted can be placed on any available physical server allocated to your account. A restart, restore, or rebuild of an Amazon WorkSpace can result in that instance being placed on a different physical server that has been allocated to your account.

Amazon Linux WorkSpaces

Q: What is Amazon Linux WorkSpaces?

Amazon Linux WorkSpaces are enterprise ready cloud desktops that organizations can provide to developers, engineers, students or office workers to get their work done.

Q: What can I do with Amazon Linux WorkSpaces?

Developers can develop software with their favorite applications like AWS CLI, AWS SDK tools, Visual Studio Code, Eclipse and Atom. Analysts can run simulations using MATLAB and Simulink. Office workers can use pre-installed applications like Libre Office for editing documents, spreadsheets, and presentations, Evolution for email, Firefox for web browsing, GIMP for image editing, Pidgin for instant messaging, and many others. You can always install more applications from the Amazon Linux repositories or other RPM based Linux repositories.

Q: Which applications and tools come with Amazon Linux WorkSpaces?

Amazon Linux WorkSpaces include a selection of desktop utilities and tools, development tools, and general productivity applications. Developers can quickly get started using packages like OpenJDK 8, Python, C/C++, AWS CLI, and AWS SDK. General office workers can use Libre Office for document editing, spread sheets, and presentations, Firefox for web browsing, GIMP for photo editing, Pidgin for IM, Evolution for mails, Atril for PDF documents and more for day to day productivity tasks. You can always install more applications from the Amazon Linux repositories or from other RPM based Linux repositories.

Q: How do I get started with Amazon Linux WorkSpaces?

To get started, simply create or select users from your configured directory, select Amazon Linux WorkSpaces bundles, and launch. Your users will receive instructions via email for connecting to their WorkSpaces. Please see here for the list of available hardware bundles.

Q: How much does it cost to use Amazon Linux WorkSpaces?

Amazon Linux WorkSpaces are available with both the hourly and monthly billing options. Detailed pricing is available here.

Q: Which package manager does Amazon Linux supports?

Amazon Linux is RPM based and uses yum package manager.

Q: Which repositories are available with Amazon Linux WorkSpaces?

Amazon Linux WorkSpaces are connected to the Amazon Linux core and extras repositories. You can always add other RPM based Linux repositories.

Q: How can I request new packages for the Amazon Linux repositories?

You can request new packages for the Amazon Linux repositories using the AWS developer forums here. Packages will be added at the sole discretion of Amazon Web Services.

Q: How will I receive package updates for the Amazon Linux WorkSpaces?

Amazon Linux WorkSpaces are regularly patched and updated from the Amazon Linux repositories.

Q: What directory types are supported for Amazon Linux WorkSpaces?

Amazon Linux WorkSpaces currently support Active Directory, an on-premises directory available via AD Connector and Microsoft Active Directory on AWS.

Q: What hardware bundles are available for Amazon Linux WorkSpaces?

Amazon Linux WorkSpaces are available with different hardware bundle in all regions where the Amazon WorkSpaces service operates. For a complete list, please see here.

Q: Can I customize my Amazon Linux WorkSpaces?

Yes. You can customize settings and install additional software on Amazon Linux WorkSpaces. You can also create custom images using the Amazon WorkSpaces console or API and use those images to launch WorkSpaces with your customizations for other users in your organization.

Q: Is sudo access enabled by default on Amazon Linux WorkSpaces?

By default, Amazon Linux WorkSpaces users get sudo access while root user is disabled for them. You can always modify permissions by editing /etc/sudoers file.

Compliance and Security

Q: Is Amazon WorkSpaces HIPAA eligible?

Yes. If you have an executed Business Associate Agreement (BAA) with AWS, you can use Amazon WorkSpaces with the AWS accounts associated with your BAA. If you don’t have an executed BAA with AWS, contact us and we will put you in touch with a representative from our AWS sales team. For more information, see, HIPAA Compliance.

Q: Is Amazon WorkSpaces PCI compliant?

Yes. Amazon WorkSpaces is PCI compliant and conforms to the Payment Card Industry Data Security Standard (PCI DSS). PCI DSS is a proprietary information security standard administered by the PCI Security Standards Council, which was founded by American Express, Discover Financial Services, JCB International, MasterCard Worldwide and Visa Inc. PCI DSS applies to all entities that store, process or transmit cardholder data (CHD) and/or sensitive authentication data (SAD) including merchants, processors, acquirers, issuers, and service providers. The PCI DSS is mandated by the card brands and administered by the Payment Card Industry Security Standards Council. For more information, see PCI DSS Compliance.

Q: Which credentials should be used to sign in to Amazon WorkSpaces?

Users sign into their WorkSpace using their own unique credentials, which they can create after a WorkSpace has been provisioned for them. If you have integrated the Amazon WorkSpaces service with an existing Active Directory domain, users will sign in with their regular Active Directory credentials. Amazon WorkSpaces also integrates with your existing RADIUS server to enable multi-factor authentication (MFA).

Q: Can I control the client devices that access my Amazon WorkSpaces?

Yes. You can restrict access to Amazon WorkSpaces based on the client OS type, and using digital certificates. You can choose to block or allow macOS, Microsoft Windows, Linux, iOS, Android, Chrome OS, zero client, and the WorkSpaces web access client.

Q: What is a digital certificate?

A digital certificate is a digital form of identity that is valid for a specified period of time, which is used as a credential that provides information about the identity of an entity, as well as other supporting information. A digital certificate is issued by a certificate authority (CA), and the CA guarantees the validity of the information in the certificate.

Q: What devices use digital certificates to control access to Amazon WorkSpaces?

Digital certificates can be used to block or allow WorkSpaces access from macOS and Microsoft Windows client devices.

Q: How do I use digital certificates to control access to Amazon WorkSpaces?

To use digital certificates to block or allow access to Amazon WorkSpaces, you upload your root certificates to the WorkSpaces management console and distribute your client certificates to the macOS and Windows devices you want to trust. To distribute your client certificates, use your preferred solution such as Microsoft System Center Configuration Manager (SCCM), or Mobile-Device Management (MDM) software. For more information, see Restrict WorkSpaces Access to Trusted Devices.

Q: How many root certificates can be imported to an Amazon WorkSpaces directory?

For each Amazon WorkSpaces directory, you can import up to two root certificates each for macOS and Microsoft Windows devices. If two root certificates are imported, WorkSpaces will present both root certificates to the client device, and the client device will use the first certificate that chains up to either root certificate.

Q: Can I control client device access to Amazon WorkSpaces without using digital certificates?

Yes. You can control access to Amazon WorkSpaces using the device type only.

Q: Can I use digital certificates to control Amazon WorkSpaces access from iOS, Android, Chrome OS, or zero clients?

At this time Amazon WorkSpaces can use digital certificates only with macOS and Microsoft Windows client devices.

Q: What is Multi-Factor Authentication (MFA)?

Multi-Factor Authentication adds an additional layer of security during the authentication process. Users must validate their identity by providing something they know (e.g. password), as well as something they have (e.g. hardware or software generated one-time password (OTP).

Q: What delivery methods are supported for MFA?

Amazon supports one time passwords that are delivered via hardware and software tokens. Out of band tokens, such as SMS tokens are not currently supported.

Q: Is there support for Google Authenticator and other virtual MFA solutions?

Google Authenticator can be used in conjunction with RADIUS. If you are running a Linux-based RADIUS server, you can configure your RADIUS fleet to use Google Authenticator through a PAM (Pluggable Authentication Module) library.

Q: Which Amazon WorkSpaces client applications support Multi-Factor Authentication (MFA)?

MFA is available for Amazon WorkSpaces client applications on the following platforms - Windows, Mac, Linux, Chromebooks, iOS, Fire, Android, and PCoIP Zero Clients. MFA is also supported when using web access to access Amazon WorkSpaces through Chrome or Firefox web browsers.

Q: What happens if a user forgets the password to access their Amazon WorkSpace?

If either AD Connector or AWS Microsoft AD is used to integrate with an existing Active Directory domain, the user would follow your existing lost password process for your domain, such as contacting an internal helpdesk. If the user is using credentials stored in a directory managed by the WorkSpaces service, they can reset their password by clicking on the “Forgot Password” link in the Amazon WorkSpaces client application.

Q: How will Amazon WorkSpaces be protected from malware and viruses?

You can install your choice of anti-virus software on your users’ WorkSpaces. The Plus bundle options offer users access to anti-virus software, and you can find more details on this here. If you choose to install your own anti-virus software, please ensure that it does not block UDP port 4172, as this will prevent users connecting to their WorkSpaces.

Q: How do I remove a user’s access to their Amazon WorkSpace?

To remove a user’s access to their WorkSpace, you can disable their account either in the directory managed by the WorkSpaces service, or in an existing Active Directory that you have integrated the WorkSpaces service with.

Q: Does WorkSpaces work with AWS Identity and Access Management (IAM)?

Yes. Please see our documentation.

Q: Can I select the Organizational Unit (OU) where computer accounts for my WorkSpaces will be created in my Active Directory?

Yes. You can set a default Organizational Unit (OU) in which computer accounts for your WorkSpaces are created in your Active Directory. This OU can be part of the domain to which your users belong, or part of a domain that has a trust relationship with the domain to which your users belong, or part of a child domain in your directory. Please see our documentation for more details.

Q: Can I use Amazon VPC Security groups to limit access to resources (applications, databases) in my network or on the Internet from my WorkSpaces?

Yes. You can use Amazon VPC Security groups to limit access to resources in your network or the Internet from your WorkSpaces. You can select a default Amazon VPC Security Group for the WorkSpaces network interfaces in your VPC as part of the directory details on the WorkSpaces console. Please see our documentation for more details.

Q: What is an IP Access Control Group?

An IP Access Control Group is a feature that lets you specify trusted IP addresses that are permitted to access your WorkSpaces. An Access Control group is made up of a set of rules, each rule specifies a specific permitted IP address or range of addresses. you can create up to 25 IP Access Control groups with up to 10 rules per group specifying the IP addresses or IP ranges accessible to your Amazon WorkSpaces.

Q: Can I implement IP address-based access controls for WorkSpaces?

Yes. With this feature you can create up to 25 IP Access Control groups with up to 10 rules per group specifying the IP addresses or IP ranges accessible to your Amazon WorkSpaces.

Q: How can I implement IP address-based access controls?

You will have two ways to do this:

1. From the WorkSpaces management console on the ‘IP Access Controls’ page, you can create access control groups by selecting ‘Create IP Access Control Groups’ and entering a group name and description. You can then add rules to your IP Access Control Group by selecting the group and going to the Rules tab, selecting ‘Edit’ and adding up to 10 rules, entering the IP addresses to allow along with a description of each rule. You can apply your IP Access Control Groups to WorkSpaces Directories on the ‘Update Directory Details’ page. See IP Access Control Groups for details.

2. From the AWS Command Line Interface you can call Amazon WorkSpaces APIs to create, delete, and describe groups, create and delete rules from each group, and to add and remove groups from directories. See Amazon WorkSpaces API Reference for details.

Q: Can IP address-based access controls be used with all WorkSpaces clients?

Yes. This feature can be used with the macOS, iPad, Windows desktop, Android tablet, Chromebook clients, and web access. This feature also supports zero clients using MFA.

Q: Which Zero Client configurations are compatible with the IP Based Access Controls feature?

Zero Clients using MFA can be used with IP Based Access Controls, along with any compatible Zero Clients which do not use PCoIP Connection Manager to connect to WorkSpaces. Any connections through PCoIP Connection Manager will not be able to access WorkSpaces if IP Based Access Controls are enabled.

Q: Are there any scenarios where a non-whitelisted IP address could access a WorkSpace?

Yes. If web access is enabled, when accessing WorkSpaces through the web access client, if the IP address changes from a whitelisted IP to a non-whitelisted IP address after the user's credentials are validated and before the WorkSpace session begins to launch, the non-whitelisted IP address would be allowed. The initial connection would require a whitelisted IP address.

Q: How are IP addresses whitelisted if users are accessing the WorkSpaces through a Network address translation (NAT)?

You will need to whitelist your public IPs with this feature, so if you have a NAT, you will need to allow access from the IPs coming from it. In this case you will be allowing access any time a user accesses WorkSpaces through a NAT.

Q: How should IP addresses be whitelisted for VPNs?

If you want to allow access from VPNs, you will need to add the public IPs of the VPN. In this case you will be allowing access any time a user accesses WorkSpaces through the VPN with public IPs whitelisted.

Q: Can I customize the login workflow for my end users login experience?

WorkSpaces supports the use of the URI (uniform resource identifier) WorkSpaces:// to open the WorkSpaces client and optionally enter the registration code, user name, and/or multi-factor authentication (MFA) code (if MFA is used by your organization).

Q: How do I enable URI?

You can create your unique URI links by following the WorkSpaces URI formatting documented in Customize How Users Log in to their WorkSpaces in the Amazon WorkSpaces Administration Guide. By providing these links to users, you enable them to use the URI on any device that has the WorkSpaces client installed. URI links can contain human-readable sensitive information if you choose to include the registration code, user name, and/or MFA information, so take precautions with how and whom you share URI information.

Q. How can I securely implement a URI for my WorkSpaces users?

WorkSpaces URIs are ideal for organizations that have an existing secure internal landing page or service portal where users access their applications. You can customize and then share the WorkSpaces URI to the portal, which gives users an easy way to access their WorkSpaces without requiring them to enter their user name and MFA multiple times or label their URI to remember which WorkSpaces registration code belongs which directory.

Encryption

Q: Does Amazon WorkSpaces support encryption?

Yes. Amazon WorkSpaces supports root volume and user volume encryption. Amazon WorkSpaces uses EBS volumes that can be encrypted on creation of a WorkSpace, providing encryption for data stored at rest, disk I/O to the volume, and snapshots created from the volume. Amazon WorkSpaces integrates with the AWS KMS service to allow you to specify the keys you want to use to encrypt the volumes.

Q: Which Amazon WorkSpaces bundle types support encryption?

Encryption is supported on all Amazon WorkSpaces hardware and software bundle types. This includes the Windows 10 desktop experiences, and the Value, Standard, Performance, Power, PowerPro, Graphics, and GraphicsPro bundles. It also includes all Plus application bundles. Additionally, any custom bundles also support encryption.

Q: How can I encrypt a new Amazon WorkSpace?

When creating a new Amazon WorkSpace from the console or the Amazon WorkSpaces APIs, you will have the option to specify which volume(s) you want encrypted along with a key ARN from your KMS keys for encryption. Note that during the launch of a WorkSpace, you can specify whether you want encryption for the user volume, root volume or both volumes, and the key provided will be used to encrypt the volumes specified.

Q: Can I use different keys to encrypt the root and user volumes of a WorkSpace?

The root and user volumes are encrypted using a single key.

Q: Do I need to provide a new KMS key for each WorkSpace that I want to encrypt?

You can use the same KMS key to encrypt the volumes of up to 500 Amazon WorkSpaces.

Q: Can Amazon WorkSpaces create a KMS key on my behalf?

Amazon WorkSpaces creates a default master key upon your first attempt to launch a WorkSpace through the AWS Management Console. You cannot manage the lifecycle of default master keys. To control the full lifecycle of a key, configure WorkSpaces to use a KMS custom customer master key (CMK). To create a KMS custom CMK, visit the KMS console or use KMS APIs to create your own keys. Note that you can use a default key generated by KMS for your WorkSpaces which will be made available to you on your first attempt to launch Amazon WorkSpaces with encryption through the AWS Management Console.

Q: What are the prerequisites for using KMS keys to encrypt Amazon WorkSpaces?

In order to use KMS keys to encrypt Amazon WorkSpaces, the key must not be disabled, and should not have exceeded its limits (learn more about limits here). You also need to have the correct permissions and policies associated with the key to use it for encryption. To learn more about the correct permissions and policies needed on the keys, please refer to our documentation.

Q: How will I be notified if my KMS key does not meet the pre-requisites outlined above?

When you launch a new WorkSpace with the key specified, the WorkSpaces service will verify if the key is valid and eligible to be used for encryption. If the key is not valid, the launch process will fail quickly and notify you of the error associated with the key. Please note that if you change the key settings while the WorkSpace is being created, there is a chance that provisioning will fail and you will be notified of this failure through the AWS Management Console or through the DescribeWorkSpaces API call.

Q: How will I be able to tell which Amazon WorkSpaces are encrypted and which ones are not?

You will be able to see if a WorkSpace is encrypted or not from the AWS Management Console or using the Amazon WorkSpaces API. In addition to that, you will also be able to tell which volume(s) on the WorkSpace were encrypted, and the key ARN that was used to encrypt the WorkSpace. For example, the DescribeWorkSpaces API call will return information about which volumes (user and/or root) are encrypted and the key ARN that was used to encrypt the WorkSpace.

Q: Can I enable encryption of volumes on a running Amazon WorkSpace?

Encryption of WorkSpaces is only supported during the creation and launch of a WorkSpace.

Q: What happens to a running Amazon WorkSpace when I disable the key in the KMS console?

A running WorkSpace will not be impacted if you disable the KMS key that was used to encrypt the user volume of the WorkSpace. Users will be able to login and use the WorkSpace without interruption. However, restarts, rebuilds, and restores of WorkSpaces that were encrypted using a KMS key that has been disabled (or the permissions/policies on the key have been modified) will fail. If the key is re-enabled and/or the correct permissions/policies are restored, restarts, rebuilds, and restores of the WorkSpace will work again.

Q: Is it possible to disable encryption for a running Amazon WorkSpace?

Amazon WorkSpaces does not support disabling encryption for a running WorkSpace. Once a WorkSpace is launched with encryption enabled, it will always remain encrypted.

Q: Will snapshots of an encrypted user volume also be encrypted?

Yes. All snapshots of the user volume will be encrypted using the same key that was used to encrypt the user volume of the WorkSpace when it was created. The user volume once encrypted stays encrypted throughout its lifecycle. Please note that Amazon WorkSpaces does not take snapshots of the root volume of a running WorkSpace.

Q: Can I rebuild an Amazon WorkSpace that has been encrypted?

Yes. Rebuilds of a WorkSpace will work as long as the key that was used to encrypt the WorkSpace is still valid. The WorkSpace volume(s) stay encrypted using the original key after it has been rebuilt.

Q: Can I restore an Amazon WorkSpace that has been encrypted?

Yes. A WorkSpace restore will work as long as the key that was used to encrypt the WorkSpace is still valid. The WorkSpace volume(s) stay encrypted using the original key after it has been restored.

Q: Can I create a custom image from a WorkSpace that has been encrypted?

Creating a custom image from a WorkSpace that is encrypted is not supported.

Q: Will the performance of my WorkSpace be impacted because the volume(s) are encrypted?

You can expect a minimum increase in latency on IOPS on encrypted volumes.

Q: Will encryption impact the launch time of an Amazon WorkSpace?

The launch time of a WorkSpace that only requires user volume encryption are similar to those of an unencrypted WorkSpace. The launch time of a WorkSpace that requires root volume encrypt will take several more minutes.

Q: Will encryption be supported for BYOL WorkSpaces?

Yes. Amazon WorkSpaces will support encryption for BYOL WorkSpaces.

Q: Will I be able to use the same KMS key to encrypt Amazon WorkSpaces in a different region?

No. Encrypted resources in one region cannot be used in a different region, because a KMS key belongs to the region in which it was created.

Q: Is there a charge for encrypting volumes on Amazon WorkSpaces?

There is no additional charge for encrypting volumes on WorkSpaces, however you will have to pay standard AWS KMS charges for KMS API requests and any custom CMKs that are used to encrypt WorkSpaces. Please see AWS KMS pricing here. Please note that the Amazon WorkSpaces services makes a maximum of five API calls to the KMS service upon launching, restarting or rebuilding a single WorkSpace.

Q: Can I rotate my KMS keys?

Yes. You can use KMS to rotate your custom CMKs. You can configure a custom CMK that you create to be automatically rotated by KMS on an annual basis. There is no impact to WorkSpaces encrypted before the CMK rotation, they will work as expected.

WorkDocs Sync Client

Q: What is the Amazon WorkDocs sync client?

The Amazon WorkDocs sync client is a client application that you can install on your Amazon WorkSpaces with Windows, which continuously, automatically, and securely syncs documents from your Amazon WorkSpace to your Amazon WorkDocs location. You can also install the Amazon WorkDocs sync client on a Mac or Windows to sync documents across all desktops they may be using. When an Amazon WorkSpace is launched, users will have a link on their desktop so that they can install the Amazon WorkDocs sync client. The client can be downloaded here.

Q: Can I enable or disable Amazon WorkDocs sync for a user's Amazon WorkSpace?

When you create a directory, or use AD Connector to integrate with an existing Active Directory, you can choose to enable or disable Amazon WorkDocs sync for that directory. Currently you cannot enable or disable Amazon WorkDocs sync on a per-user basis.

Q: How do I synchronize documents between an Amazon WorkSpace and a Mac or Windows PC?

To enable synchronization, all you need to do is install the Amazon WorkDocs sync client on your Amazon WorkSpace with WIndows and PCs you would like to synchronize with. Once you've done this, simply select the folders you want to sync.

Q: Is Single Sign-On (SSO) supported?

Yes. Single Sign-On (SSO) can be enabled so that when users are signed in to their Amazon WorkSpace they will be automatically signed in to their Amazon WorkDocs sync client, and will not be required to provide credentials when they access the web client from their Amazon WorkSpace. You can enable SSO by visiting the AWS Directory Service area of the AWS Management Console, clicking the directory ID link for your directory and selecting the Apps & Services tab. For more information and detailed setup see our documentation.

Amazon WorkSpaces Application Manager (WAM)

Q: What is Amazon WorkSpaces Application Manager?

Amazon WorkSpaces Application Manager (Amazon WAM) offers a fast, flexible, and secure way for you to deploy and manage applications for Amazon WorkSpaces with Windows. Amazon WAM accelerates software deployment, upgrades, patching, and retirement by packaging Microsoft Windows desktop applications into virtualized application containers that run as though they are natively installed.

Q: How are Amazon WAM applications delivered to users?

Amazon WAM delivers desktop apps to users' WorkSpaces with Windows as virtualized app containers using a unique cloud delivery technology. The applications execute on a WorkSpace from within the virtualized container and provide performance similar to natively-installed applications.

Q: How can I get started with Amazon WAM?

To get started with Amazon WAM, select your level of subscription (Lite or Standard,) build an application catalog in the AWS Management Console and assign applications to your users running Amazon WorkSpaces with Windows. You can build an application catalog using applications for which you own licenses, proprietary applications built in-house, and applications from the AWS Marketplace for Desktop Apps.

After your catalog is available, you can use the AWS Management Console to assign applications from the catalog to your Amazon WorkSpaces users. Applications from the catalog can be made required or optional. Required applications are automatically installed on the appropriate WorkSpaces; optional applications are made available to users for on-demand installation.

Q: How do I upload my applications to Amazon WAM?

You can package your applications using the Amazon WAM Studio, validate using the Amazon WAM Player, and then upload your applications to Amazon WAM. For more information, see the Amazon WAM User Guide on packaging and validating.

Q: What type of applications can be delivered using Amazon WAM?

Any application compatible with Microsoft Windows 8, Windows 10, Windows Server 2012, and Windows Server 2016 can be delivered to WorkSpaces using Amazon WAM. Both 32-bit and 64-bit applications are supported.

Q: Can I track application use with Amazon WAM?

You can track usage for any applications assigned to users.

Q: In which AWS regions is Amazon WAM available?

To see a list of AWS regions where Amazon WAM is currently available, please visit Region Table.

Q: Which Amazon WorkSpaces experiences work with Amazon WAM?

You can use Amazon WAM to deploy and manage applications for Amazon WorkSpaces running the Windows 10 desktop experience.

Q: Which AWS Directory Service directories does Amazon WAM support?

Amazon WAM can be used with AWS Directory Services AD Connector and Simple AD, or AWS Managed Microsoft AD. NOTE: If you have set up a trust relationship between AWS Managed Microsoft AD in the AWS Cloud and your existing on-premises Microsoft Active Directory, you can't assign applications to users in your on-premises Microsoft Active Directory.

Q: Do Amazon WorkSpaces need Internet access to use Amazon WAM?

Yes, Amazon WorkSpaces need an Internet connection to receive applications via Amazon WAM.

Q: How do I get Amazon WAM on my users' Amazon WorkSpaces?

Your users can install the Amazon WAM desktop app on their Amazon WorkSpaces via a shortcut located on the desktop by default.

Q: How do end users access applications that are assigned using Amazon WAM?

Users can open the Amazon WAM desktop app and see all the applications available to them. You can set up applications to be required or optional. Required applications are automatically installed on user's WorkSpace, and optional applications can be installed via the Amazon WAM desktop app. For more information about the Amazon WAM desktop app, see the Amazon WAM User Guide.

Q: How many applications can I add to my Amazon WAM catalog?

There is no limit to the number of applications you can add to your Amazon WAM catalog. However, storage charges apply to applications that you upload to Amazon WAM, after the first 100 GB of storage used for your applications.

Q: How many applications can I deliver to each Amazon WorkSpaces user via Amazon WAM?

You can assign up to 50 applications to each Amazon WorkSpaces user.

Q: Can I use tags to categorize applications in my Amazon WAM catalogs?

Yes, you can assign tags to applications and service-related charges for WAM by simply tagging your Amazon WorkSpaces. To learn more about assigning tags to your Amazon WorkSpaces, follow the steps listed on this web page: Tagging WorkSpaces.

Q: How will I be billed for Amazon WAM?

The Lite plan is available at no cost, and the Standard plan costs $5/user/month for each user enrolled in the WAM Standard plan with one or more applications assigned. There may be a cost for applications from AWS Marketplace for Desktop Applications that users activate.

Q: Can I have users on both the Lite and the Standard plans?

No. You can subscribe to either the Lite or Standard plan, and all users will be on the same plan.

Q: Can I change my subscription plan during the billing period?

Yes. On the "Subscription plan" page" of the WAM console you can upgrade or downgrade your plan and view the feature details for the two subscription plans. You have the opportunity to view the current usage before confirming the upgrade.

Q: What will happen to my applications if I downgrade from the Standard to the Lite plan?

Users will be moved to the most up to date version of applications from AWS Marketplace for Desktop Apps, and will lose access to any applications that you packaged and uploaded to Amazon WAM.

Q: Is there a limit for storage of my app packages?

Both the Lite and Standard plans include 100GB of storage for the apps, and S3 charges will apply for additional storage.

Q: Can I share an Amazon WAM package with another AWS account?

Yes. Packages created and approved by you within your AWS account can be shared with other AWS accounts in the same region. You can set up package sharing via the Packages tab on the Amazon WAM console by adding package permissions to the AWS account to which you wish to share the package.

Q: Can I set limits on the packages that I share with other AWS accounts?

No. At this time, you cannot place any restrictions on packages that are shared.

Q: How do I use an Amazon WAM package that is shared with me?

You can use an Amazon WAM package shared with you by creating an application and assigning the application to your users.

Q: Can I make any changes to a package that has been shared with my account?

No. A package made available to you by another AWS account cannot be modified.

Q: How do I know if I can trust a package that has been shared with my account?

Always verify that your package is shared from a trusted source. Verify the source by validating the AWS account ID and check if it is an account that you trust.

Q: Can I delete an Amazon WAM package?

Yes. You can delete an Amazon WAM package that belongs to your account within an AWS region by launching Amazon WAM Studio in your packaging instance. Once you delete a package, all versions of the package will be deleted. Also, you can only delete packages that don’t have apps assigned or have not been shared with another AWS account. If you have an application created, you will first need to delete the application before you can delete the package. If you have shared a package with another AWS account, you will first need to remove sharing of the package before deleting the package.

Q: What happens to an Amazon WAM package once it is deleted?

Once an Amazon WAM package is deleted, it will no longer be available from within your account. The package will be fully deleted once any accounts you shared the package with have deleted applications using the package.

AWS Marketplace for Desktop Apps

Q: What is AWS Marketplace for Desktop Apps?

AWS Marketplace for Desktop Apps is a new category in the AWS Marketplace that can deploy applications to Amazon WorkSpaces with Windows through Amazon WAM. The AWS Marketplace for Desktop Apps includes both applications you can purchase on a monthly basis and free apps. You can find applications from developers such as Microsoft, Corel and Foxit and popular open source titles.

Q: How do I use desktop applications from AWS Marketplace?

You can subscribe to applications from the AWS Marketplace for Desktop Apps via Amazon WorkSpaces console. Start by selecting the Application Catalog in Amazon WorkSpaces console, browse and add applications from the AWS Marketplace to your application catalog. Once the applications are in your catalog you can assign the applications to your WorkSpaces users. The applications can then be accessed by users via the Amazon WorkSpaces Application Manager (Amazon WAM) desktop app.

Q: How will I be charged for applications from the AWS Marketplace for Desktop Apps?

You will be charged the price listed on AWS Marketplace for Desktop Apps for each application on a monthly subscription basis. Software subscriptions are billed monthly, even if they are used on Amazon WorkSpaces set to bill hourly. A subscription is activated and charged the first time a user launches an application and will renew monthly until access to the application is removed for that user. Charges for an application are prorated for the remainder of the first month in which a user launches them. Subsequent months are billed for the entire month. Subscriptions that are removed in the middle of a month will not receive a refund for the remainder of the month.

Q: How do I unsubscribe from an application?

To unsubscribe from an application, simply remove the users and groups assigned to use the application. Once this is completed, the application will immediately not be available to your users and there will be no new charges for the application in the following month.

Q: Can Amazon WorkSpaces end users access the AWS Marketplace for Desktop Apps directly?

No, only the administrator of the WorkSpaces account will see the entire AWS Marketplace in the WorkSpaces console. End users will only see the applications you provisioned for them.

Q: Where can I view charges for my application subscriptions from AWS Marketplace for Desktop Apps?

You can view the charges for application subscriptions from AWS Marketplace for Desktop Apps by signing in to the AWS billing console and viewing the AWS Marketplace section in the estimate bill. You can view the applications subscribed, monthly price, and total charge for each application.

Q: How do I get support for the applications I use from AWS Marketplace for Desktop Apps?

After subscribing to the application on AWS Marketplace for Desktop Apps, you can select the application details to view support information. Expand the support information to view details on how to obtain support.

Client Access, Web Access, and User Experience

Q: Where can I download the Amazon WorkSpaces client application?

You can download the Amazon WorkSpaces client application for free on the client download website.

Q: Can I use any other client (e.g., an RDP client) with Amazon WorkSpaces?

No. You can use any of the free clients provided by AWS, which includes client applications for Windows, macOS, Chromebooks, iOS, Fire tablets, and Android tablets, or Chrome or Firefox web browsers, to access your Amazon WorkSpaces.

Q: Which operating systems are supported by the Amazon WorkSpaces client applications?

Amazon WorkSpaces clients are available for the following operating systems:
- Microsoft Windows 7, Windows 8, and Windows 10
- Apple macOS (10.8.1 and above)
- Linux (Ubuntu Linux 18.04 and above)
- Google Chrome OS (45 and above)
- Apple iOS (8.0 and above)
- Google Android (4.4 and above)
- Amazon Fire OS 4 and Fire OS 5

Q: Which tablet devices are supported by the Amazon WorkSpaces client application?

Amazon WorkSpaces clients are available for the following devices:
- Apple iPad Pro 12.9-inch and 9.7-inch models

- Apple iPad Mini 2, 3 and 4
- Apple iPad Air and iPad Air 2
- Amazon Fire tablets released after 2012: Fire 7", Fire HD 6/7/8/10, Fire HDX 8.9", Kindle Fire 7", and Kindle Fire HDX 7/8.9
- Samsung and Nexus tablets

While we expect other popular Android tablets running Android version 4.4 to work correctly with the Amazon WorkSpaces client, there may be some that are not compatible. If you are interested in support for a particular device, please let us know via the Amazon WorkSpaces forum.

Q: Which smartphones are supported by the Amazon WorkSpaces client application?

Amazon WorkSpaces clients are available for the following devices:
Samsung Galaxy S8 and S8+ with Samsung DeX Station
If you are interested in support for a particular device, please let us know via the Amazon WorkSpaces forum.

Q: What is a PCoIP Zero Client?

A PC-over-IP (PCoIP) Zero Client is a single-purpose hardware device that can enable access to Amazon WorkSpaces. Zero Clients include hardware optimization specifically for the PCoIP protocol, and are designed to require very little administration.

Q: Can I use PCoIP Zero Clients with Amazon WorkSpaces?

Yes, Amazon WorkSpaces supports PCoIP Zero Client devices that have the Teradici Tera2 chipset. For a complete list of Zero Clients that are compatible with Amazon WorkSpaces please reference Teradici's website.

Q: Will my Amazon WorkSpace running in AutoStop running mode preserve the state of applications and data when it stops?

Amazon WorkSpaces preserve the data and state of your applications when stopped. On reconnect, your Amazon WorkSpace will resume with all open documents and running programs intact. Graphics enabled WorkSpaces with Windows preserve your data when stopped, and any running application sessions will be closed. On reconnect, your Amazon WorkSpace will retain your files and folders in your last saved location.

Q: How do I resume my Amazon WorkSpace after it stops?

By logging into your Amazon WorkSpace from the Amazon WorkSpaces client application, the service will automatically restart your Amazon WorkSpace. When you first attempt to log in, the client application will notify you that your Amazon WorkSpace was previously stopped, and that your new session will start once your WorkSpace has resumed.

Q: How long does it take for my Amazon WorkSpace to be available once I attempt to log in?

If your Amazon WorkSpace has not yet stopped, your connection is almost instantaneous. If you Amazon WorkSpace has already stopped, in most cases it will be available within sixty to ninety seconds.

Q: Which peripherals can be used with the Amazon WorkSpaces client applications?

Amazon WorkSpaces clients support:
• Keyboard, mouse, and touch input (touch input is only supported on tablet clients). Amazon WorkSpaces do not currently support a 3D mouse.
• Audio output to client device
• Analog and USB headsets

Q: What kind of headsets can be used for audio conversations?

Most analog and USB headsets will work for audio conversations through WorkSpaces running Windows. For USB headsets, you should ensure they show up as a playback device locally on your client computer.

Q: Can I use the built in microphone and speakers for making audio calls?

Yes. For the best experience, we recommend using a headset for audio calls. However, you may experience an echo when using the built in microphone and speakers with certain communication applications.

Q: Does Audio-in work with mobile clients such as Android, iOS, and Chromebooks?

Audio-in is supported on the Windows, OSX and iOS clients.

Q: How do I enable Audio-in for my WorkSpaces?

Audio-in is enabled for all new WorkSpaces.

For WorkSpaces with Windows, enabling the WorkSpaces Audio-in capability requires local logon access inside your WorkSpace. If you have a Group Policy restricting user local logon in your WorkSpace, we will detect it and not apply the Audio-in update to the WorkSpace. You can remove the Group Policy and the Audio-in capability will be enabled after the next reboot.

For WorkSpaces with Linux, admins can disable/enable the feature by setting the directive "pcoip.enable_audio=0" in PCoIP configuration file "/etc/pcoip-agent/pcoip-agent.conf".

Q: Should I update my custom images to take advantage of Audio-in?

Yes. We always recommend you refresh your custom images on a regular basis to take advantage of the latest features. WorkSpaces launching from custom images that have not been recently updated may take longer to be available to users. Once a WorkSpace is updated for Audio-In you can use it to create an updated custom image which will include Audio-in support by default.

Q: Does WorkSpaces support devices with high DPI screens?

Yes. The Amazon WorkSpaces desktop client application will automatically scale the in-session display to match the DPI settings of the local device.

Q: How many monitors does Amazon WorkSpaces support? What monitor resolution is supported?

Amazon WorkSpaces Value, Standard, Performance, Power, PowerPro and GraphicsPro bundles support a maximum of four displays. The maximum supported resolution depends on the number of displays, as shown in the following table:

Displays	Maximum Supported Resolution
1	3840 x 2160 pixels
2	3840 x 2160 pixels
4	1920 x 1200 pixels

Graphics bundles support only a single monitor configuration with a maximum resolution of 2560x1600.

Q: Will my bandwidth usage be higher when I use four monitors, or I use 4k Ultra HD resolution?

Yes. The bandwidth requirements for WorkSpaces depends on two factors (a) the number of screens it has to stream to and (b) the amount of pixel changes taking place in each screen.

Q: Can each monitor have different orientation?

Yes. You can have some of your monitors in landscape mode and others in portrait mode to suit your desktop productivity needs.

Q: Will Amazon WorkSpaces remember my monitor settings between sessions?

The fullscreen mode setting will be preserved. If you quit a WorkSpaces session in the fullscreen mode, you will be able to log into the fullscreen mode next time. However, display configurations will not be saved. Every time you initiate a WorkSpaces session, the client application extracts the EDID of uses your local setup configuration and sends that to the WorkSpaces host to deliver an optimal display experience.

Q: What happens to my display settings when I connect to my WorkSpace from a different desktop?

When you connect from a different desktop computer, the display settings of that computer will take precedence to deliver an optimal display experience.

Q: Will the iPad and Android applications support Keyboard/Mouse input?

The Android client supports both keyboard and mouse input. The iPad client supports keyboard and mouse (SwiftPoint GT mouse) inputs. While we expect most popular keyboard and mouse devices to work correctly, there may be devices that may not be compatible. If you are interested in support for a particular device, please let us know via the Amazon WorkSpaces forum.

Q: Can I access my Amazon WorkSpaces through a web browser?

Yes, you can use Amazon WorkSpaces web access to log in to your Amazon WorkSpace with Windows through Chrome or Firefox web browsers. You do not need to install any software, and you can connect from any network that can access the public Internet. To enable WorkSpaces web access, first, if you have an exisiting WorkSpace with Windows 10, you should reboot it. Then your WorkSpaces admin needs to enable web access from the AWS Console in the WorkSpaces Directory Details – Access Control Options section. Once these steps are complete, to access your WorkSpace through a browser, simply visit the Amazon WorkSpaces web access page using a supported browser and enter your WorkSpaces registration code and then login to the WorkSpace with your username and password.

Q: What is Amazon WorkSpaces web access?

Amazon WorkSpaces web access allows you to access your Amazon WorkSpace with Windows from Chrome or Firefox running on a computer connected to any network that can access the public Internet. web access does not exclude users from using native Amazon

WorkSpaces client applications to connect to their WorkSpaces with Windows; users can choose between web access and native client applications. Web access is available here.

Q: Which web browsers can I use to access Amazon WorkSpaces web access?

Amazon WorkSpaces web access works with the latest Google Chrome and Firefox versions running on Windows, Mac, or Linux. Mobile versions of Chrome and Firefox are not currently supported.

Q: Can I enable web access for Non-English based Amazon WorkSpaces?

Yes. Web access support is currently available on WorkSpaces with English (US), Japanese, Korean, and French (Canadian) based versions Windows desktops.

Q: Do I need to install any additional software in order to access my Amazon WorkSpaces through a web browser?

No, you do not need to install any programs, add-ins, or plugins in order to access your Amazon WorkSpaces through a supported web browser.

Q: How do I get started using web access to log in to my Amazon WorkSpaces?

First, your Amazon WorkSpace needs to be enabled for web access. This can be done through the AWS Management Console by your IT administrator. Once this is complete, you can log in using web access, available here. The first time you log in, you will be asked to enter the registration code that was provided in your welcome email.

Q: How will I know if my Amazon WorkSpace has been enabled for web access?

If your Amazon WorkSpace has been set to block web access, you will receive an error message when you attempt to log in, informing you to contact your system administrator to enable web access.

Q: Can I use Web Access to access my Amazon WorkSpaces on any network?

Yes. You can use web access on any network that can access the public Internet. If you can browse the web, then you can connect to your Amazon WorkSpace.

Q: Which Amazon WorkSpaces bundles support web access?

You can use web access to connect to the Value, Standard, Performance, Power, and PowerPro Amazon WorkSpaces with Windows 10 or Windows Server 2016 operating systems. Graphics, GraphicsPro bundles and Amazon Linux WorkSpaces currently do not support web access.

Q: What local devices can I use when connecting to my Amazon WorkSpace through Chrome or Firefox?

You will be able to use your mouse and keyboard as input devices. Local peripheral devices —including printers, USB drives, webcams, and microphones—will not be available. Though clipboard redirection will not work across your local operating system and your Amazon WorkSpace, copy and paste operations within your WorkSpace will work.

Q: In which regions is web access available?

Amazon WorkSpaces web access is available in all regions where Amazon WorkSpaces is available.

Q: Do I need to enter a registration code to use web access?

The first time you log in using web access, you will be asked to enter the registration code that was provided in your welcome email. At the moment, web access does not offer the ability to store multiple different registration codes.

Q: When using a web browser to access my Amazon WorkSpace, how can I control my session?

You can use the connection bar along the top of your browser window to control your session. The connection bar allows you to disconnect, enter and exit full screen mode, and send a “Ctrl-Alt-Del” key sequence to the Amazon WorkSpace. It can be pinned in place, or set to hide automatically.

Q: How do I disconnect from my Amazon WorkSpace when accessing it through a web browser?

You can disconnect using the “Disconnect” command in the connection bar, by closing the browser tab, or by quitting the browser program. Web access does not support reconnecting to your Amazon WorkSpace - you must log in again to reconnect.

Q: Will Amazon WorkSpaces support additional client devices and virtual desktop operating systems?

We continually review our roadmap to see what features we can add to address our customers' requirements. If there is a client device or virtual desktop operating system that you'd like Amazon WorkSpaces to support, please email us with details of your request.

Q: What is the end user experience when Multi-Factor Authentication (MFA) is enabled?

Users will be prompted for their Active Directory username and password, followed by their OTP. Once a user passes both Active Directory and RADIUS validation, they will be logged in to their Amazon WorkSpace. To learn more, visit our documentation.

Q: How can I determine the best region to run my Amazon WorkSpaces?

The Amazon WorkSpaces Connection Health Check Website compares your connection speed to each Amazon WorkSpaces region and recommends the fastest one.

Q: Which languages are supported by Amazon WorkSpaces?

Amazon WorkSpaces bundles that provide the Windows 10 desktop experience currently support English (US), French (Canadian), Korean, and Japanese. You can also download and install language packs for Windows directly from Microsoft. For more information, visit this page. Amazon WorkSpaces client applications currently support English (US), German, Chinese (Simplified), Japanese, French (Canadian), Korean, and Portuguese.

Maintenance and Setup

Q: Does the Amazon WorkSpaces service have maintenance windows?

Yes. Amazon WorkSpaces enables maintenance windows for both AlwaysOn and AutoStop WorkSpaces by default.

For AlwaysOn (monthly) WorkSpaces, the maintenance schedule is controlled by the OS settings on the WorkSpace. The default maintenance window is a four-hour period from 00h00 – 04h00 (this time window is based on the time zone settings you have set for your Amazon WorkSpaces) each Sunday morning. During this time your WorkSpaces may not be available.

For AutoStop (hourly) WorkSpaces, the default maintenance window is typically from 00h00 to 05h00 everyday starting on the 3rd Monday of the month in the time zone of the WorkSpaces's AWS region. The Maintenance window might take up to two weeks. WorkSpaces can be maintained on any day in the maintenance window. You can set the Maintenance mode for AutoStop WorkSpaces in the WorkSpaces management console. For more information see Manage the WorkSpace Running Mode. The maintenance window for AutoStop WorkSpaces is currently not configurable.

Q: Can I opt out of maintenance windows for my WorkSpaces?

It is highly recommended to keep your WorkSpaces maintained regularly. If you want to run your own WorkSpaces maintenance schedule, it is possible to opt out of the service default maintenance windows for Windows WorkSpaces.

For AutoStop (hourly) WorkSpaces, you can disable the Maintenance mode on the console. For AlwaysOn Windows WorkSpaces, the maintenance window is controlled by the system settings and can be configured via Automatic Updates GPO settings. Currently, you cannot opt out of the maintenance windows for AlwaysOn Amazon Linux WorkSpaces.

Q: Will my Amazon WorkSpaces require software updates?

Your Amazon WorkSpaces provide users with the Amazon Linux cloud desktops, Windows 10 experience, provided by Windows Server 2016. The underlying OS, and any applications installed in the WorkSpace may need updates.

Q: How will my Amazon WorkSpaces be patched with software updates?

By default, your Amazon WorkSpaces are configured to install software updates. Amazon Linux WorkSpaces will be updated to install the latest security and software patches, and Amazon WorkSpaces with Windows have Windows Updates turned on. You can customize these settings, or use an alternative patch management approach. Updates are installed at 2am each Sunday.

Q: What action is needed to receive updates for the Amazon WorkSpaces service?

No action is needed on your part. Updates are delivered automatically to your Amazon WorkSpaces during the maintenance window. During the maintenance window, your WorkSpaces may not be available.

Q: Can I turn off the software updates for the Amazon WorkSpaces service?

No. The Amazon WorkSpaces service requires these updates to be provided to ensure normal operation of your users' WorkSpaces.

Q: I don't want to have Windows Update automatically update my Amazon WorkSpaces. How can I control updates and ensure they are tested in advance?

You have full control over the Windows Update configuration in your WorkSpaces, and can use Active Directory Group Policy to configure this to meet your exact requirements. If you would like to have advance notice of patches so you can plan appropriately we recommend you refer to Microsoft Security Bulletin Advance Notification for more information.

Q: How are updates for applications installed in my WorkSpaces provided?

Amazon WorkSpaces running Amazon Linux are updated via pre-configured Amazon Linux yum repositories hosted in each WorkSpaces region and the updates are automatically installed. Patches and updates requiring a reboot are installed during our weekly maintenance window.

For all other applications, updates can be delivered via the automatic update service for each application if one is available. For applications without an automatic update service, you will need to evaluate the software vendor’s recommended updating approach and follow that if necessary.

Q: How do I manage my WorkSpaces?

The WorkSpaces Management console lets you provision, restart, rebuild, restore, and delete WorkSpaces. To manage the underlying OS for the WorkSpaces, you can use standard Microsoft Active Directory tools such as Group Policy or your choice of Linux orchestration tools to manage the WorkSpaces. In the case when you have integrated WorkSpaces with an existing Active Directory domain, you can manage your WorkSpaces using the same tools and techniques you are using for your existing on-premises desktops. If you have not integrated with an existing Active Directory, you can set up a Directory Administration WorkSpace to perform management tasks. Please see the documentation for more information.

You can also give WorkSpaces users the ability to perform common tasks on their own by enabling self-service management. Once enabled, WorkSpaces users can do things like restart, rebuild, restore, increase volume size, change compute type, and change running mode directly from the WorkSpaces client with no IT or helpdesk intervention.

Q: Can I use tags to categorize my Amazon WorkSpaces resources?

Yes, you can assign tags to existing Amazon WorkSpaces resources including WorkSpaces, directories registered with WorkSpaces, images, custom bundles, and IP Access Control Groups. You can also assign tags during the creation of new Amazon WorkSpaces and new IP Access Control Groups. You can assign up to 50 tags (key/value pairs) to each Amazon WorkSpaces resource using the AWS Management Console, the AWS Command Line Interface, or the Amazon WorkSpaces API. These tags automatically get applied to all Amazon WorkSpaces Application Manager (WAM) applications and WAM-related service charges associated with a WorkSpace. To learn more about assigning tags to your Amazon WorkSpaces resources, follow the steps listed on this web page: Tag WorkSpaces Resources.

Q: Can I control whether my users can access Amazon WorkSpaces web access?

Yes. You can use the AWS Management Console to control whether Amazon WorkSpaces in your directory can be accessed using web access, by visit the directory details page. Note: this setting can only be applied to all Amazon WorkSpaces in a directory, not at an individual Amazon WorkSpace level.

Q: What is the difference between restarting and rebuilding a WorkSpace?

A restart is just the same as a regular operating system (OS) reboot. A rebuild will retain the user volume on the WorkSpace but will return the WorkSpace to its original state (any changes made to the system drive will not be retained).

Q: What is the difference between WorkSpaces Rebuild and Restore?

A rebuild will retain the user volume on the WorkSpace but will return the WorkSpace to its original state (any changes made to the system drive will not be retained). A restore will retain both the root and user volumes on the WorkSpace but will return the WorkSpace to the last healthy state as detected by the service.

Q: How do I remove an Amazon WorkSpace I no longer require?

To remove a WorkSpace you no longer require, you can “delete” the Workspace. This will remove the underlying instance supporting the WorkSpace and the WorkSpace will no longer exist. Deleting a WorkSpace will also remove any data stored on the volumes attached to the WorkSpace, so please confirm you have saved any data you must keep prior to deleting a WorkSpace.

Q: Can I provide more than one Amazon Workspace per user?

No. You can currently only provide one WorkSpace for each user.

Q: How many Amazon WorkSpaces can I launch?

You can launch as many Amazon WorkSpaces as you need. Amazon WorkSpaces sets default limits, but you can request an increase in these limits here. To see the default limits for Amazon WorkSpaces, please visit our documentation.

Q: What is the network bandwidth that I need to use my Amazon WorkSpace?

The bandwidth needed to use your WorkSpace depends on what you're doing on your WorkSpace. For general office productivity use, we recommend that a bandwidth download speed of between 300Kbps up and 1Mbps. For graphics intensive work we recommend bandwidth download speeds of 3Mbps.

Q: What is the maximum network latency recommended while accessing a Workspace?

While the remoting protocol has a maximum round trip latency recommendation of 250 ms, the best user experience will be achieved at less than 100 ms.

Q: Is there a recommended power plan or power settings for my WorkSpaces?

Yes. For WorkSpaces running Windows, we recommend selecting the "High Performance" power plan in Windows. For WorkSpaces running Linux, you should select a power plan that optimizes for performance.

Q: Does WorkSpaces need any Quality of Service configurations to be updated on my network?

If you wish to implement Quality of Service on your network for WorkSpaces traffic, you should prioritize the WorkSpaces interactive video stream which is comprised of real time traffic on UDP port 4172. If possible, this traffic should be prioritized just after VoIP to provide the best user experience.

Q: Is MFA on Amazon WorkSpaces available in my region?

Support for MFA is available in all AWS Regions where Amazon WorkSpaces is offered.

Q: What are the prerequisites for setting up a PCoIP Zero Client?

Zero Clients should be updated to firmware version 4.6.0 (or newer). You will need to run the PCoIP Connection Manager to enable the clients to successfully connect to Amazon WorkSpaces. Please consult the Amazon WorkSpaces documentation for a step by step guide on how to properly setup the PCoIP Connection Manager, and for help on how to find and install the necessary firmware required for your Zero Clients

Q: How do I get support with Amazon WorkSpaces?

You can get help from AWS Support, and you can also post in the Amazon WorkSpaces Forum.

Billing and Pricing

Q: How does billing work for Amazon WorkSpaces?

You can pay for your Amazon WorkSpaces either by the hour, or by the month. You only pay for the WorkSpaces you launch, and there are no upfront fees and no term commitments. The fees for using Amazon WorkSpaces include use of both the infrastructure (compute, storage, and bandwidth for streaming the desktop experience to the user) and the software applications listed in the bundle.

Q: How much does an Amazon WorkSpace cost?

Please see our pricing page for the latest information.

Q: Can I pay for my Amazon WorkSpaces by the hour?

Yes, you can pay for your Amazon WorkSpaces by the hour. Hourly pricing is available for all WorkSpaces bundles, and in all AWS regions where Amazon WorkSpaces is offered.

Q: How does hourly pricing work for Amazon WorkSpaces?

Hourly pricing has two components: an hourly usage fee, and a low monthly fee for fixed infrastructure costs. Hourly usage fees are incurred only while your Amazon WorkSpaces are actively being used, or undergoing routine maintenance. When your Amazon WorkSpaces are not being used, they will automatically stop after a specified period of inactivity, and hourly metering is suspended. When your Amazon WorkSpaces resume, hourly charges begin to accrue again.

Q: How do I get started with hourly billing for my Amazon WorkSpaces?

To launch an Amazon WorkSpace to be billed hourly, simply select a user, choose an Amazon WorkSpaces bundle (a configuration of compute resources and storage space), and specify the AutoStop running mode. When your Amazon WorkSpace is created, it will be billed hourly.

Q: What is the difference between monthly pricing and hourly pricing for Amazon WorkSpaces?

With monthly billing, you pay a fixed monthly fee for unlimited usage and instant access to a running Amazon WorkSpace at all times. Hourly pricing allows you to pay for your Amazon WorkSpaces by the hour and save money on your AWS bill when your users only need part-time access to their Amazon WorkSpaces. When your Amazon WorkSpaces being billed hourly are not being used, they automatically stop after a specified period of inactivity, and hourly usage metering is suspended.

Q: How do I select hourly billing or monthly billing for my Amazon WorkSpaces?

To make hourly billing possible, Amazon WorkSpaces now operates in two running modes – AutoStop and AlwaysOn. The AutoStop running mode allows you to pay for your Amazon WorkSpaces by the hour. The AlwaysOn running mode is used when paying a fixed monthly fee for unlimited usage of your Amazon WorkSpaces. You can easily choose between monthly and hourly billing by selecting the running mode when you launch Amazon WorkSpaces through the AWS Management Console, the Amazon WorkSpaces APIs, or the Amazon WorkSpaces Command Line Interface. You can also switch between running modes for your Amazon WorkSpaces at any time.

Q: When do I incur charges for my Amazon WorkSpace when paying by the hour?

Hourly usage fees start accruing as soon as your Amazon WorkSpace is running. Your Amazon WorkSpace may resume in response to a login request from a user, or to perform routine maintenance.

Q: When do I stop incurring charges for my Amazon WorkSpaces when paying by the hour?

Hourly usage charges are suspended when your Amazon WorkSpaces stop. AutoStop automatically stops your WorkSpaces a specified period of time after users disconnect, or when scheduled maintenance is completed. The specified time period is configurable and is set to 60 minutes by default. Note that partial hours are billed as a full hour, and the monthly portion of hourly pricing does not suspend when your Amazon WorkSpaces stop.

Q: Can I force hourly charges to suspend sooner?

You can manually stop Amazon WorkSpaces from the AWS Management Console, or by using the Amazon WorkSpaces APIs. To stop the monthly fee associated with your hourly Amazon WorkSpaces, you need to remove the Amazon WorkSpaces from your account (note: this also deletes all data stored in those Amazon WorkSpaces).

Q: Can I switch between hourly and monthly billing?

Yes, you can switch from hourly to monthly billing for your Amazon WorkSpaces at any time by switching the running mode to AlwaysOn in the AWS Management Console, or through the Amazon WorkSpaces APIs. When you switch, billing immediately changes from hourly to monthly, and you are charged a prorated amount at the monthly rate for the remainder of the month, along with the monthly and hourly usage fees already billed for the month. Your Amazon WorkSpaces will continue to be charged monthly unless you switch the running mode back to AutoStop.

You can switch from monthly to hourly billing by setting the running mode to AutoStop in the AWS Management Console or through the Amazon WorkSpaces APIs. Switching from monthly to hourly billing will take effect the following month as you will have already paid for your Amazon WorkSpaces for that month. Your Amazon WorkSpaces will continue to be charged hourly unless you switch the running mode back to AlwaysOn. Your Amazon WorkSpaces will continue to be charged hourly unless you switch the running mode back to AlwaysOn. Please note that billing renewals happen at 00:00 Pacific Time on the first of each month.

WorkSpaces users can also switch between monthly and hourly billing directly from the WorkSpaces client if this self-service management capability is enabled by their WorkSpaces administrator.

Q: If I don’t use my Amazon WorkSpace for the full month, are the fees prorated?

If you’re paying for your Amazon WorkSpaces monthly, your Amazon WorkSpaces are charged for the full month’s usage. If you’re paying hourly (AutoStop running mode), you are charged for the hours during which your Amazon WorkSpaces are running or undergoing maintenance, plus a monthly fee for fixed infrastructure costs. In both cases, the monthly fee is prorated in the first month only.

Q: Will I be charged the low monthly fee associated with hourly billing if I don’t use my Amazon WorkSpaces in a given month?

Yes, you will be charged a small monthly fee for the Amazon WorkSpaces bundle you selected. If you’ve chosen an Amazon WorkSpaces Plus bundle, you will be charged for the software subscription as well. You can find the monthly fees for all Amazon WorkSpaces on the pricing page here.

Q: How are the Plus software bundles charged when I pay hourly for my Amazon WorkSpaces?

Plus bundles are always charged monthly, even if you’re paying for your Amazon WorkSpaces by the hour. If you selected a Plus bundle when you launched your WorkSpaces, you will incur the listed fee for the Plus software bundle even if you do not use those Amazon WorkSpaces in a particular month.

Q: Will I be able to monitor how many hours my Amazon WorkSpaces have been running?

Yes, you will be able to monitor the total number of hours your Amazon WorkSpaces have been running in a given period of time through the Amazon CloudWatch “UserConnected” metric.

Q: Does Amazon WorkSpaces pricing include bandwidth costs?

Amazon WorkSpaces pricing includes network traffic between the user's client and their WorkSpace. Web traffic from WorkSpaces (for example, accessing the public Internet, or downloading files) will be charged separately based on current AWS EC2 data transfer rates listed here.

Q: How will I be charged for Amazon WorkSpaces that I launch that are based on a custom image?

There is no additional charge for Amazon WorkSpaces created from custom images. You will be charged the same as the underlying bundles on which the customized images are based.

Q: Can I use custom images for Amazon WorkSpaces that are billed hourly?

Yes. You can launch Amazon WorkSpaces billed hourly from images that you create and upload. There is no additional charge for Amazon WorkSpaces launched from custom images. You will be charged the same as the underlying bundles on which the customized images are based.

Q: Is there a charge to use Amazon WorkSpaces client applications?

The Amazon WorkSpaces client applications are provided at no additional cost, and you can install the clients on as many devices as you need to. You can access these here.

Q: Is there an additional charge to access Amazon WorkSpaces using web access?

There is no additional charge to access Amazon WorkSpaces using web access. For Amazon WorkSpaces set to bill hourly, you will keep getting billed for the time you leave a browser tab open with an actively running Amazon WorkSpace.

Q: Can I use tags to obtain usage and cost details for Amazon WorkSpaces, Amazon WorkSpaces Application Manager (WAM), and WAM applications on my AWS monthly billing report?

Yes. By setting tags to appear on your monthly Cost Allocation Report, your AWS monthly bill will also include those tags. You can then easily track costs according to your needs. To do this, first assign tags to your Amazon WorkSpaces by following the steps listed on this web page: Tagging WorkSpaces. Next, select the tag keys to include in your cost allocation report by following the steps listed on this web page: Setting Up Your Monthly Cost Allocation Report.

Q: Are there any costs associated with tagging Amazon WorkSpaces?

There are no additional costs when using tags with your Amazon WorkSpaces.

Q: What does the Amazon WorkSpaces Application Manager (Amazon WAM) cost?

Amazon WAM is available in two versions - lite or standard. The Amazon WAM lite subscription is available at no charge, and the Amazon WAM standard subscription costs $5/user/month. You can learn more about Amazon WAM here.

Q: Can I pay for Amazon WAM on an hourly basis?

Amazon WAM is not available for hourly billing. You will still be charged monthly for Amazon WAM usage, even if you're using Amazon WAM to deliver applications to an Amazon WorkSpace being billed hourly.

Q: Do I have to pay to use the Amazon WAM Studio or Amazon WAM Player?

No. There is no additional charge for using the Studio or Player. You will be charged for AWS resources such as the Amazon EC2 instance hours, EBS storage, and bandwidth when using the Studio to package your applications for Amazon WAM.

Q: What are the requirements for schools, universities, and public institutions to reduce their WorkSpaces licensing?

Schools, universities, and public institutions may qualify for reduced WorkSpaces licensing fees. Please reference the Microsoft Licensing Terms and Documents for qualification requirements. If you think you may qualify, please create a case with the AWS support center here. Select Regarding:<Account and Billing Support>, Service:<Billing>, Category:<Qualify as Educational institution>, and enter the required info. We will review your information and work with you to reduce your fees and costs.

Q: What do I need to provide to qualify as a school, university, or public institution?

You will need to provide AWS your institution's full legal name, principle office address, and public website URL. AWS will use this information to qualify you for reduced user fees for qualified educational institutions with your WorkSpaces. Please note: The use of Microsoft software is subject to Microsoft's terms. You are responsible for complying with Microsoft licensing. If you have questions about your licensing or rights to Microsoft software, please consult your legal team, Microsoft, or your Microsoft reseller. You agree that we may provide the information to Microsoft in order to apply educational pricing to your Amazon WorkSpaces usage.

Q: Does qualification for Amazon WorkSpaces reduced user fees affect other AWS cloud services?

No, your user fees are specific to Amazon WorkSpaces, and do not affect any other AWS cloud services or licenses you have.

Q: Is there a charge for streaming data between my WorkSpaces and End Users' devices?

The charges for the Service include the cost of streaming data between your WorkSpaces and End Users' devices unless you stream via VPN, in which case you will be charged VPN data transfer rates in addition to any applicable Internet data transfer changes. Other WorkSpace data transfer will be charged using Amazon EC2 data transfer pricing.

Free Tier

Q: Am I eligible to take advantage of the Amazon WorkSpaces Free Tier offer?

The Amazon WorkSpaces Free Tier offer is available to new or existing AWS customers that have not previously used WorkSpaces. The Free Tier allows you to gain hands-on experience with Amazon WorkSpaces, at no cost, so that you can evaluate the service.

Q: What Amazon WorkSpaces bundles are available as part of the Free Tier?

The Amazon WorkSpaces Free Tier allows you to provision two Standard bundle WorkSpaces with 80 GB Root and 50 GB User volumes. The Standard bundle WorkSpace offers a cloud desktop with 2 vCPUs, 4 GB of memory, 80 GB Root and 50 GB User volume of SSD-based storage, and you can choose between Amazon Linux WorkSpaces, Amazon WorkSpaces with Windows 10 desktop experiences powered by Windows Server. As with all WorkSpaces, your WorkSpace comes with the pre-installed applications, and access to Amazon WorkDocs with 50 GB included storage.

Q: What is included with the Amazon WorkSpaces Free Tier?

The WorkSpaces Free Tier includes two Standard bundle WorkSpaces with 80 GB Root and 50 GB User volumes, for 40 hours of combined use per month, for two calendar months. As with all bundles, your WorkSpace comes with the pre-installed applications, and access to Amazon WorkDocs with 50 GB included storage.

Q: Can I use any other Amazon WorkSpaces bundles as part of the Free Tier?

The Amazon WorkSpaces Free Tier includes the Standard bundle only.

Q: What is the duration of the Amazon WorkSpaces Free Tier?

The Free Tier offer starts when you launch your first Amazon WorkSpace, and expires at the end of the second calendar month. For example, if you launched your first WorkSpace on the 15th of the month, the Free Tier offer extends to the end of the next month.

Q: If I use less than 40 hours in my first month of Free Tier use, do the remaining hours roll over to the next month?

The Amazon WorkSpaces Free Tier allows you to use a combined total of 40 hours per month. Unused hours expire when the new calendar month starts.

Q: What happens if I use my WorkSpaces for more than 40 hours in a calendar month during the Free Tier period?

In the event you exceed 40 hours of use in a month during the Free Tier period, you are billed at the current hourly rate for Amazon WorkSpaces.

Q: What happens if I convert my Amazon WorkSpaces from AutoStop (hourly billing) to AlwaysOn (monthly billing) before my Free Tier period expires?

To qualify for the Free Tier, your Amazon WorkSpaces need to run in the AutoStop running mode. You can change the running mode of your WorkSpaces to AlwaysOn, but this action converts your WorkSpaces to monthly billing, and your Free Tier period will end. To learn more about how billing works when switching running modes, see the Q:

Q: Hourly billing for Amazon WorkSpaces includes a fee for hours used, and a monthly infrastructure cost. Is the monthly infrastructure cost waived during the Amazon WorkSpaces Free Tier?

The monthly infrastructure fee for Amazon WorkSpaces is waived for Free Tier use, even if you use more than 40 hours in a month. If you do exceed 40 hours in a month, you are billed for your additional usage at the current hourly rate, which is available at Amazon WorkSpaces Pricing.

Q: What happens when my Amazon WorkSpaces Free Tier period ends?

When your Free Tier period ends, your Amazon WorkSpaces convert to Standard bundle WorkSpaces billed at the current hourly rate. In addition, the monthly infrastructure fee will start to apply. For current rates, see Amazon WorkSpaces Pricing.

Q: How can I track my Amazon WorkSpaces Free Tier usage?

To track your Amazon WorkSpaces usage, go to the My Account page in the AWS management console and see your current and past activity by service, and region. You can also download usage reports. For more information, see Understanding Your Usage with Billing Reports.

Connectivity

Q: Can I use an HTTPS proxy to connect to my Amazon WorkSpaces?

Yes, you can configure a WorkSpaces Client app to use an HTTPS proxy. Please see our documentation for more information.

Q: Can I connect Amazon WorkSpaces to my VPC?

Yes. The first time you connect to the WorkSpaces Management Console, you can choose an easy ‘getting started’ link that will create a new VPC and two associated subnets for you as well as an Internet Gateway and a directory to contain your users. If you choose to access the console directly, you can choose which of your VPCs your WorkSpaces will connect to. If you have a VPC with a VPN connection back to your on-premises network, then your WorkSpaces will be able to communicate with your on-premises network (you retain the usual control you have over network access within your VPC using all of the normal configuration options such as security groups, network ACLS, and routing tables).

Q: Can I connect to my existing Active Directory with my Amazon WorkSpaces?

Yes. You can use AD Connector or AWS Microsoft AD to integrate with your existing on-premises Active Directory.

Q: Will my Amazon WorkSpaces be able to connect to the Internet to browse websites and download applications?

Yes. You have full control over how your Amazon WorkSpaces connect to the Internet based on regular VPC configuration. Depending on what your requirements are you can either deploy a NAT instance for Internet access, assign an Elastic IP Address (EIP) to the Elastic Network Interface (ENI) associated with the WorkSpace, or your WorkSpaces can access the Internet by utilizing the connection back to your on-premises network.

Q: Can I use IPv6 addresses for my Amazon WorkSpaces bundles?

Yes. You can use IPv6 addresses for Value, Standard, Performance, Power, and PowerPro bundles. At this time, IPv6 addresses are not supported in Graphics, or GraphicsPro bundles.

Q: Can my Amazon WorkSpaces connect to my applications that are running in Amazon EC2 such as a file server?

Yes. Your WorkSpaces can connect to applications such as a fileserver running in Amazon EC2 (both "Classic" and VPC networking environments). All you need to do is ensure appropriate route table entries, security groups and network ACLs are configured so that the WorkSpaces can reach the EC2 resources you would like them to be able to connect to.

Q: What are the pre-requisites for using my digital certificates on Amazon WorkSpaces?

To use your certificates to manage which client devices can access Amazon WorkSpaces, you need to distribute your client certificates using your preferred solution such as Microsoft System Center Configuration Manager (SCCM), or a Mobile-Device Management (MDM) software solution to the devices you want to trust. Your root certificates are imported into the WorkSpaces management console. For more information, please see Restrict WorkSpaces Access to Trusted Devices.

Q: What are the pre-requisites for enabling MFA on Amazon WorkSpaces?

To enable MFA on WorkSpaces, you will need to configure AD Connector, and have an on-premises RADIUS server(s). Your on-premises network must allow inbound traffic over the default RADIUS server port (1812) from the AD Connector server(s). Additionally, you must ensure that usernames match between Active Directory and your RADIUS server. To learn more, visit our documentation.

Directories

Q: Do I need to set up a directory to use the Amazon WorkSpaces service?

Each user you provision a WorkSpace for needs to exist in a directory, but you do not have to provision a directory yourself. You can either have the WorkSpaces service create and manage a directory for you and have users in that directory created when you provision a WorkSpace. Alternatively, you can integrate WorkSpaces with an existing, on-premises

Active Directory so that users can continue to use their existing credentials meaning that they can get seamless applications to existing applications.

Q: If I use a directory that the Amazon WorkSpaces service creates for me, can I configure or customize it?

Yes. Please see our documentation for more details.

Q: Can I integrate Amazon WorkSpaces with my existing on-premises Active Directory?

Yes. You can use AD Connector or AWS Microsoft AD to integrate with your existing on-premises Active Directory.

Q: How do I integrate Amazon WorkSpaces with my on-premises Microsoft Active Directory?

There are two ways you can integrate Amazon WorkSpaces with your on-premises Microsoft Active Directory (AD): you can set up an interforest trust relationship with your AWS Microsoft AD domain controller, or you can use AD Connector to proxy AD authentication requests.

To configure an interforest trust relationship between your on-premises Microsoft AD and your AWS Microsoft AD please see the documentation here. To configure AD Connector, please see the documentation here.

Once a trust is established, you can select the domain where your user accounts reside directly in the Amazon WorkSpaces console, and proceed to provisioning WorkSpaces for your users. Please note that usernames across domains need to be unique per instance of AWS Microsoft AD.

Q: There are two options for integrating Amazon WorkSpaces with my on-premises Microsoft Active Directory. Which one should I use?

You can integrate Amazon WorkSpaces with your on-premises Microsoft Active Directory (AD) either by setting up an interforest trust relationship with your AWS Microsoft AD domain controller, or by using AD Connector to proxy AD authentication requests.

When using interforest trust, you only need a single trust relationship between your on-premises AD and your AWS Microsoft AD domain controller. You can assign Amazon WorkSpaces to users in any of your on-premises domains, and AWS Microsoft AD automatically discovers and routes authentication requests to the correct domain controller. This option works well when your environment consists of multiple on-premises Microsoft AD domains.

When using AD Connector, a separate AD Connector is required for each of your on-premises Microsoft AD domains with users that will need WorkSpaces assigned to them. Using AD Connector works well for environments with a single on-premises domain, or for proof-of-concept projects.

For more information, please visit this page.

Q: Can I use the Amazon WorkSpaces APIs to create new WorkSpaces for users across domains when I have an interforest trust relationship established with AWS Microsoft AD?

Yes. When using the Amazon WorkSpaces API to launch WorkSpaces, you will need to specify the domain name as part of the username, in this format: “NETBIOS\username” or “corp.example.com\username”. For more information, please visit this page.

Q: Can I apply the same Group Policy object settings from my on-premises Microsoft Active Directory to Amazon WorkSpaces?

Yes. If you’re using an interforest trust relationship between your on-premises Microsoft AD and your AWS Microsoft AD domain controller, you will need to ensure that your Group Policy object (GPO) settings are replicated across domains before they can be applied to Amazon WorkSpaces. If you are using AD Connector, your GPO settings will be applied to your WorkSpaces much like any other computer in your domain.

Q: Can I apply Active Directory policies to my Amazon WorkSpaces using the directory that the WorkSpaces service creates for me?

Yes. Please see our documentation for more details.

Q: What happens to my directory when I remove all of my Amazon WorkSpaces?

You may keep your AWS directory in the cloud and use it to domain join EC2 instances or provide directory users access to the AWS Management Console. You may also delete your directory.

If there are no WorkSpaces being used with your Simple AD or AD Connector for 30 consecutive days, you will be charged for this directory as per the AWS Directory Service pricing terms. If you delete your Simple AD or AD Connector you can always create a new one when you want to start using WorkSpaces again.

Q: Which AWS Directory Services support the use of PCoIP Zero Clients?

PCoIP Zero Clients can be used with the AD Connector and Simple AD directory services from AWS. Currently, Zero Clients cannot be used with the AWS Directory Service for Microsoft Active Directory.

CloudWatch Monitoring

Q: What does Amazon CloudWatch monitor for Amazon WorkSpaces?

Amazon WorkSpaces is integrated with both CloudWatch Metrics and CloudWatch Events.

You can use Amazon CloudWatch Metrics to review health and connection metrics for individual WorkSpaces and all WorkSpaces belonging to a directory. You can set up CloudWatch Alarms on these metrics to be alerted about changes to WorkSpaces health, or about issues your users may have connecting to their WorkSpaces.

You can use CloudWatch Events to view, search, download, archive, analyze, and respond to successful WorkSpace logins. Amazon WorkSpaces client applications send WorkSpaces Access events to CloudWatch Events when a user successfully logs in to a WorkSpace. All Amazon WorkSpaces client applications send these events.

Q: Will I be able to monitor how many hours my Amazon WorkSpaces have been running?

Yes, you will be able to monitor the total number hours your Amazon WorkSpaces has been running in a given period of time through Amazon CloudWatch “UserConnected” metric.

Q: In what regions can I use Amazon WorkSpaces with CloudWatch Metrics?

CloudWatch Metrics are available with Amazon WorkSpaces in all AWS regions where WorkSpaces is available.

Q: What does CloudWatch Metrics cost?

There is no additional cost for using CloudWatch Metrics with WorkSpaces via the CloudWatch console. There may be additional charges for setting up CloudWatch Alarms and retrieving CloudWatch Metrics via APIs. Please see CloudWatch pricing for more information.

Q: How do I get started with CloudWatch Metrics for my Amazon WorkSpaces?

CloudWatch Metrics are enabled by default for all your WorkSpaces. Visit the AWS Management Console to review the metrics and set up alarms.

Q: What metrics are supported for the Amazon WorkSpaces client application and PCOIP Zero Clients?

Please see the documentation for more information on Amazon CloudWatch metrics with Amazon WorkSpaces.

Q: What metrics are supported for Amazon WorkSpaces web access usage?

The following metrics are currently supported for reporting on Amazon WorkSpaces web access usage:

- Available
- Unhealthy
- UserConnected
- Maintenance

Please see the documentation for more information on Amazon CloudWatch Metrics with Amazon WorkSpaces.

Q: What CloudWatch Events are generated by Amazon WorkSpaces?

Successful WorkSpace logins. Amazon WorkSpaces sends access event information to CloudWatch Events when a user successfully logs in to a WorkSpace from any WorkSpaces client application.

Q: How can I utilize CloudWatch Events with WorkSpaces?

You can use CloudWatch Events to view, search, download, archive, analyze, and respond based on rules that you configure. You can either use the AWS Console under CloudWatch to view and interact with CloudWatch Events or use services such as Lambda, ElasticSearch, Splunk and other partner solutions using Kinesis Streams or Firehose to take actions based on your event data. For storage, CloudWatch Events recommends using Kinesis to push data to S3. For more information on how to use CloudWatch Events, see the Amazon CloudWatch Events User Guide.

Q: What information is included in WorkSpaces Access Events?

Events are represented as JSON objects which include WAN IP address, WorkSpaces ID, Directory ID, Action Type (ex. Login), OS platform, Timestamp and a Success/Failure indicator for each successful login to WorkSpaces. See our documentation for more details here.

Q: What does CloudWatch Events cost?

There is no additional cost for using CloudWatch Events with Amazon WorkSpaces. You will be charged for any other services you use that take action based on CloudWatch Events, such as Amazon ElasticSearch, and AWS Lambda. This also includes other CloudWatch services such as CloudWatch Metrics, CloudWatch Logs, and CloudWatch Alarms if your usage surpasses the CloudWatch Free Tier limits. All of these services are integrated with and can be triggered from CloudWatch Events.

Printing

Q: Can I print from my Amazon WorkSpace?

Yes, Amazon WorkSpaces with Windows support local printers, network printers, and cloud printing services. Amazon WorkSpaces with Amazon Linux support network printers, and cloud printing services.

Q: How do I enable printer auto-redirection for my Amazon WorkSpace?

By default, local printer auto-redirection is disabled. You can use the Group Policy settings to enable this feature. This will ensure that your local printer is set as the default every time you connect to your WorkSpace.

Q: How do I print to my local printer?

If you have a local printer configured, it will show up in your WorkSpaces printer menu the next time you connect to your WorkSpace. If not, you will need to configure a local printer outside of your WorkSpace. Once this is done, select your local printer from the print menu, and select print.

Q: Why can’t I see my local printer from the printing menu?

Most printers are already supported by Amazon WorkSpaces. If your printer is not recognized, you may need to install the appropriate device driver on your WorkSpace.

Q: How do I print to a network printer?

Any printer which is on the same network as your Amazon WorkSpace and is supported by Windows Server 2016 can be added as a network printer. Once a network printer is added, it can be selected for printing from within an application.

Q: Can I use my Amazon WorkSpace with a cloud printing service?

You can use cloud printing services with your WorkSpace including, but not limited to, Cortado ThinPrint®.

Q: Can I print from my tablet or Chromebook?

The Amazon WorkSpaces clients for tablets and Chromebook support cloud printing services including, but not limited to, Cortado ThinPrint®. Local and network printing are not currently supported.

User Self Service Management

Q: What self-service management capabilities are available for Amazon WorkSpaces?

You can choose to let users accomplish typical management tasks for their own WorkSpace, including restart, rebuild, change compute type, and change disk size. You can also let users switch from monthly to hourly billing (and back). You can choose to enable specific self-service management capabilities that suit your needs directly in the WorkSpaces Admin Console.

Q: How do I get started with self-service management capabilities for my WorkSpaces users?

Self-service management capabilities are enabled by default when you register a directory with WorkSpaces. You can choose to not enable them when you register a directory.

You can modify specific self-service management capabilities from the WorkSpaces console. On the Directories page, select the directory you want to modify for self-service management. Next, select “Update Details” under the “Actions” menu. You can find all self-service management capabilities options under the “User Self Service Permissions” section. You can also use WorkSpaces APIs to modify self-service management capabilities.

Q: How do end users access self-service management capabilities?

Self-service management capabilities are available to users through the WorkSpaces client on Windows and Mac devices.

Q: Do I need to log into WorkSpaces to use self-service management capabilities?

Yes, you must authenticate to use any self-service management capabilities.

Q: Can I continue to use my WorkSpace while a self-service management actions is being performed?

You can continue to use your WorkSpace while disk size or running mode is being changed. Restarting, rebuilding, restoring, and changing compute type requires disconnecting from your WorkSpaces session.

Q: How much does it cost to use self-service management capabilities?

Self-service management capabilities are available at no additional cost. You can enable self service management for tasks such as changing the WorkSpace bundle type, or increasing the volume size. When end users perform these tasks, the billing rate for those WorkSpaces may change.

Amazon AppStream 2.0 FAQs

General

Q: What is Amazon AppStream 2.0?

Amazon AppStream 2.0 is a fully managed application streaming service that provides users instant access to their desktop applications from anywhere. Amazon AppStream 2.0 simplifies application management, improves security, and reduces costs by moving a company's applications from their users' physical devices to the AWS Cloud. The Amazon AppStream 2.0 streaming protocol provides users a responsive, fluid performance that is almost indistinguishable from a natively installed application. With Amazon AppStream 2.0, organizations can realize increased flexibility, improved scalability, and the agility to support a broad range of compute and storage requirements for their applications.

Q: What's the difference between the original Amazon AppStream and Amazon AppStream 2.0?

Amazon AppStream 2.0 is the next-generation desktop application streaming service from AWS. Amazon AppStream was an SDK-based service that customers could use to set up their own streaming service with DIY engineering. AppStream 2.0 provides a fully managed streaming service with no DIY effort. AppStream 2.0 offers a greater range of instance types; streams desktop applications to HTML5-compatible web browsers with no plugins required; provides dual-monitor support on web browsers and 4-monitor, 4K monitor, and USB peripheral support through the AppStream 2.0 client for Windows. In addition, AppStream 2.0 simplifies application lifecycle management and lets your applications access services in your VPC.

Q: Can I continue to use the original Amazon AppStream service?

No. You cannot use the original Amazon AppStream service. Amazon AppStream 2.0 offers a greater range of instance types, streams desktop applications with no rewrite, simplifies application lifecycle management, and allows your apps to access services in your VPC.

Q: What are the benefits of streaming over rendering content locally?

Interactively streaming your application from the cloud provides several benefits:

Instant-on: Streaming your application with Amazon AppStream 2.0 lets your users start using your application immediately, when using an image builder or Always-On fleet, without the delays associated with large file downloads and time-consuming installations.

Remove device constraints: You can leverage the compute power of AWS to deliver experiences that wouldn’t normally be possible due to the GPU, CPU, memory, or physical storage constraints of local devices.

Multi-platform support: You can take your existing applications and start streaming them to a computer without any modifications.

Easy updates: Because your application is centrally managed by Amazon AppStream 2.0, updating your application is as simple as providing a new version of your application to Amazon AppStream 2.0. That's all you need to do to immediately upgrade all your users, without any action on their part.

Improved security: Unlike traditional boxed software and digital downloads, where your application is available for theft or reverse engineering, Amazon AppStream 2.0 stores and executes your application securely in AWS data centers, and only provides an interactive pixel stream to users.

Q: Do some applications work better with Amazon AppStream 2.0 than others?

Many types of applications work well as streaming applications, including CAD, CAM, CAE, 3D modeling, simulation, games, video and photo-editing software, medical imaging, and life sciences applications. These applications benefit most from streaming because the application runs on the vast computational resources of AWS, yet your users can interact with the application using low-powered devices, with very little noticeable change in application performance.

Q: Does Amazon AppStream 2.0 support microphones?

Yes. Amazon AppStream 2.0 supports most analog and USB microphones, including built-in microphones.

Q: Does Amazon AppStream 2.0 support USB devices such as 3D mice?

Yes. Amazon AppStream 2.0 supports most USB devices such as 3D mice through the Windows Client. All USB devices are disabled by default. Administrators can enable USB devices for their users.

Q: How do users enable audio input in an Amazon AppStream 2.0 streaming session?

Users enable audio input from the Amazon AppStream 2.0 toolbar by selecting the Settings icon and selecting Enable Microphone.

Q: What browser support audio-input in an Amazon AppStream 2.0 session?

Most popular HTML5 compatible browsers support audio-input in Amazon AppStream 2.0 session, including Chrome, Edge, and Firefox. Microsoft Internet Explorer 11 (IE11) does not support audio-input, and the microphone option will not appear on the Amazon AppStream 2.0 toolbar in streaming sessions running in IE11.

Q: What does a user need to access applications streamed from Amazon AppStream 2.0?

A user needs to have applications set up by an administrator, a modern web browser that can support HTML5, a broadband internet connection with at least 2 Mbps capability, and outbound access to the internet via HTTPS (443). For web-based AppStream 2.0 streaming sessions, up to two monitors are supported. To use up to four monitors, 4K monitors and USB peripherals such as 3D mice, users can download and use the AppStream 2.0 client for Windows.

Q: What is the AppStream 2.0 Windows Client?

The AppStream 2.0 client for Windows is a native application that is designed for users who require additional functionality not available from web browsers during their AppStream 2.0 streaming sessions. The AppStream 2.0 client lets users use multiple monitors and USB peripherals such as 3D mice with their applications. The client also supports keyboard shortcuts, such as Alt + Tab, clipboard shortcuts, and function keys. The AppStream 2.0 client is supported on the following versions of Windows: Windows 7, Windows 8, Windows 8.1, and Windows 10. Both 32-bit and 64-bit versions of Windows are supported.

Q: What are the system requirements for using the AppStream 2.0 Windows Client?

The minimum system requirements are 2 GB of ram and 150 MB of disk space.

Q: What monitor configurations are supported by the AppStream 2.0 Windows Client?

For browser-based streaming sessions, AppStream 2.0 supports the use of up to two monitors with a maximum display resolution of 2560x1440 pixels per monitor. The AppStream 2.0 client for Windows supports up to 4 monitors with a maximum display resolution of 2560x1440 pixels per monitor. For streaming sessions that are supported by the Graphics Design and Graphics Pro instance families, the AppStream 2.0 client also supports the use of up to 2 monitors with a maximum display resolution of 4096x2160 pixels per monitor.

Q: How can I deploy the AppStream 2.0 Windows Client to my users?

Users can download and install the Windows Client. To use USB peripherials, a users need local administrator rights to install the AppStream 2.0 USB driver. You can remotely install the Windows Client using remote deployment tools like Microsoft System Center Configuration Manager (SCCM). Learn more in our documentation.

Q: Can users configure location and language settings for their applications?

Yes. Users can set the time zone, locale, and input method to be used in their streaming sessions to match their location and language preferences.

Q: Can users copy and paste between their local device and their Amazon AppStream 2.0 streaming applications?

Yes. Users can use the Windows Client and Google Chrome to access their streaming applications can copy and paste text between their local device and their streaming applications in the same way they copy and paste between applications on their local device - for example, using keyboard shortcuts. For other browsers, users can use the Amazon AppStream 2.0 web clipboard tool.

Q: Can my Amazon AppStream 2.0 applications run offline?

No. Amazon AppStream 2.0 requires a sustained internet connection or network route to an AppStream 2.0 streaming VPC endpoint to access your applications.

Q: What does Amazon AppStream 2.0 manage on my behalf?

Streaming resources: Amazon AppStream 2.0 launches and manages AWS resources to host your application, deploys your application on those resources, and scales your application to meet client demand.

Simplified app management: Amazon AppStream 2.0 delivers the latest version of an application instantly to users, and eliminates the pain of patching and updating applications on every end-user device. Because your application is centrally managed by Amazon AppStream 2.0, updating your application is as simple as providing a new version of your application to Amazon AppStream 2.0. Applications can be assigned to users dynamically and removed instantly at any time, improving business flexibility and reducing costs.

Q: Can I use tags to categorize AppStream 2.0 resources?

Yes. you can assign tags to manage and track the following Amazon AppStream 2.0 resources: Image builders, images, fleets, and stacks. AWS enables you to assign metadata to your AWS resources in the form of tags. Tags let you categorize your AppStream 2.0 resources so you can easily identify their purpose and track costs accordingly. For example, you can use tags to identify all resources used by a particular department, project, application, vendor, or use case. Then, you can use AWS Cost Explorer to identify trends, pinpoint cost drivers, and detect anomalies in your account.

You can assign or remove tags using the AppStream 2.0 management console, command line interface, or API. Tags have a key and a corresponding value, and you can assign up to 50 tags per AppStream 2.0 resource.

Q. What resources can I create with AWS CloudFormation?

With CloudFormation, you can automate creating fleets, deploying stacks, adding and managing user pool users, launching image builders, and creating directory configurations alongside your other AWS resources.

Q: How do I use my AWS Direct Connect, AWS VPN, or other VPN tunnel to stream my applications?

First, create an Amazon Virtual Private Cloud (Amazon VPC) endpoint in the same Amazon VPC as your AWS Direct Connect, AWS VPN, or other VPN tunnel. Then, specify the VPC endpoint when creating a new stack, modifying an existing one, or creating a new image builder. Your users will then use the VPC endpoint when they stream their applications. To learn more about the AppStream 2.0 streaming VPC endpoints, see Creating and Streaming From VPC Interface Endpoints in the AppStream 2.0 Administration Guide.

Try sample applications

Q: Can I try sample applications?

Yes. Visit Try Sample Applications low-friction, setup-free trial experience for Amazon AppStream 2.0 service.

Q: What do I need to start using Try It Now?

You need an AWS account and a broadband Internet connection with at least 1 Mbps bandwidth to use Try It Now. You also need a browser capable of supporting HTML5.

Q: Will I be charged for using Try It Now?

You won't be charged any AWS fees for using Try It Now. However, you may incur other fees such as Internet or broadband charges to connect to the Try It Now experience.

Q: What applications can I use with Try It Now?

Try It Now includes popular productivity, design, engineering, and software development applications running on Amazon AppStream 2.0 for you to try. To see the full list of available applications, go to the Try It Now catalog page after signing in with your AWS account.

Q: How long can I stream applications via Try It Now?

You can stream the applications included in Try It Now for up to 30 minutes. At the end of 30 minutes, your streaming session is automatically terminated and any unsaved data will be deleted.

Q: Can I save files within Try It Now?

You can save files to your Amazon AppStream 2.0 session storage and download them to your client device before your streaming session ends. Your files are not saved when you disconnect from your Try It Now session, or when your session ends, and any unsaved data will be deleted.

Q: Can I submit an application to be included in Try It Now?

Yes. You can submit a request to include your application in Try It Now. After your request is received, AWS usually reviews the request and responds within 10 business days.

Getting started

Q: How do I get started with Amazon AppStream 2.0?

You can begin using Amazon AppStream 2.0 by visiting the AWS Management Console, or by using the AWS SDK. Visit Stream Desktop Applications for a 10 step tutorial.

Q: What resources do I need to set up to stream my applications using Amazon AppStream 2.0?

You need to create an Amazon AppStream 2.0 stack in your AWS account to start streaming applications to your users. A stack includes a fleet of Amazon AppStream 2.0 instances that executes and streams applications to end users. Each instance is launched using an Amazon

AppStream 2.0 image containing your applications, and uses an instance type that you select for your fleet. To learn more about Amazon AppStream 2.0 resources, please visit this page.

Q: How do I create an Amazon AppStream 2.0 image to import my applications?

You can create an Amazon AppStream 2.0 image using Image Builder via the AWS Management Console. Image Builder allows you to install and test your applications just as you would with any Windows desktop, and then create an image. You can complete all the install, test, and creation steps for the image without leaving the console.

Q: What instance types are available to use with my Amazon AppStream 2.0 fleet?

Amazon AppStream 2.0 provides a menu of instance types for configuring a fleet or an image builder. You can select the instance type that best matches your applications and end-user requirements. You can choose from General Purpose, Compute Optimized, Memory Optimized, Graphics Design, Graphics Desktop, or Graphics Pro instance families.

Q: Can I change an instance type after creating a fleet?

Yes. You can change your instance type after you have created a fleet. To change the instance type, you will need to stop the fleet, edit the instance type, and then start the fleet again. For more information, see Set up AppStream 2.0 Stacks and Fleets.

Q: Can I connect Amazon AppStream 2.0 instances to my VPC?

Yes. You can choose the VPCs to which your Amazon AppStream 2.0 instances (fleet and image builders) connect. When you create your fleet, or launch Image Builder, you can specify one or more subnets in your VPC. If you have a VPC with a VPN connection to your on-premises network, then Amazon AppStream 2.0 instances in your fleet can communicate with your on-premises network. You retain the usual control you have over network access within your VPC, using all the normal configuration options such as security groups, network access control lists, and routing tables. For more information about creating a VPC and working with subnets, see Working with VPCs and Subnets.

Q: Can I use custom branding with Amazon AppStream 2.0?

Yes. You can customize your users' Amazon AppStream 2.0 experience with your logo, color, text, and help links in the application catalog page. To replace AppStream 2.0's default branding and help links, log in to the AppStream 2.0 console, navigate to Stacks, and select a your application stack. Then, click Branding, choose Custom, select your options, and click Save. Your custom branding will apply to every new application catalog launched using SAML 2.0 single-sign-on (SSO) or the CreateStreamingURL API. You can revert to the default AppStream 2.0 branding and help links at any time. To learn more, visit Add Your Custom Branding to Amazon AppStream 2.0.

Q: Can I define default application settings for my users?

Yes, you can set default application settings for your users. This includes application connection profiles, browser settings, and installing plugins.

Q: Can users save their application settings?

Yes. You can enable persistent application and Windows settings for your users on AppStream 2.0. Your users' plugins, toolbar settings, browser favorites, application connection profiles, and other settings will be saved and applied each time they start a streaming session. Your users' settings are stored in an S3 bucket you control in your AWS account.

To learn more about persistent application settings, see Enable Application Settings Persistence for Your AppStream 2.0 Users.

Q: Am I charged for persistent user application settings?

There is no additional AppStream 2.0 charge to use this feature. However, persistent user application settings are stored in an Amazon S3 bucket in your account, and you will be billed for the S3 storage used for your user’s settings data. See Amazon S3 pricing or Enable Application Settings Persistence for Your AppStream 2.0 Users for more information.

Q: Is there a limit to the file size of my users' persistent application settings?
By default, the maximum user profile file size is 1 GB. See Enable Application Settings Persistence for Your AppStream 2.0 Users to increase this.

Q: Will my users' application settings persist across stacks?
Yes. Your users' application settings persist across stacks.

Q: How are my users' application settings secured?
Your users' application settings are encrypted in transit to the S3 bucket in your account using Amazon S3's SSL endpoints. Your users' application settings are encrypted at rest using S3-managed encryption keys.

Q: Can I dynamically entitle users to apps?
Yes, you can use the dynamic app framework APIs to build a dynamic app provider that specifies what apps uers can launch at run-time. The apps provided can be virtualized apps

that are delivered from a Windows file share or other storage technology. To learn more, see Manage App Entitlement with the Dynamic App Framework.

Images

Q: How can I create images with my own applications?

You can use Amazon AppStream 2.0 Image Builder to create images with your own applications. To learn more, please visit the tutorial found on this page.

Q: With which operating system do my apps need to be compatible?

Amazon AppStream 2.0 streams applications that can run on the following 64-bit OS versions - Windows Server 2012 R2, Windows Server 2016 and Windows Server 2019. You can add support for 32-bit applications by using the WoW64 extensions. If your application has other dependencies, such as the .NET framework, include those dependencies in your application installer.

Q: Can I install anti-virus software on my Amazon AppStream 2.0 image to secure my applications?

You can install any tools, including anti-virus programs on your AppStream 2.0 image. However, you need to ensure that these applications do not block access to the AppStream 2.0 service. We recommend testing your applications before publishing them to your users.

Q: Can I customize the operating system using group policies?

Any changes that are made to the image using Image Builder through local group policies will be reflected in your AppStream 2.0 images. Any customizations made with domain based group policies can only be applied to domain joined fleets.

Q: How will my Amazon AppStream 2.0 images be updated with updates from the AppStream 2.0 service?

AppStream 2.0 regularly releases base images that include Microsoft Windows operating system updates and AppStream 2.0 agent updates. The AppStream 2.0 agent software runs on your streaming instances and enables your users to stream applications. When you create a new image, the Always use latest agent version option is selected by default. When this option is selected, any new image builder or fleet instance that is launched from your image will always use the latest AppStream 2.0 agent version. If you deselect this option, your image will use the agent version you selected when you launched the image builder.

Windows operating system updates are released only through base images. To keep your operating system updated in your images, you need to rebuild your images using the latest AWS base image.

Q: How will my Amazon AppStream 2.0 images be updated with Windows updates from Microsoft?

You will need to create new AppStream 2.0 images to apply Windows updates. To do this, you can create a new image builder instance from an existing image, apply Microsoft updates, and create a new image. Existing streaming instances will be replaced with instances launched from the new image within 16 hours or immediately after users have disconnected from them, whichever is earlier. You can immediately replace all the instances in the fleet with instances launched from the latest image by stopping the fleet, changing the image used, and starting it again.

Q: How do I update my applications in an existing image?

To update applications on the image, or to add new applications, launch Image Builder using an existing image, update your applications and create a new image. Existing streaming instances will be replaced with instances launched from the new image within 16 hours or immediately after users have disconnected from them, whichever is earlier. You can immediately replace all the instances in the fleet with instances launched from the latest image by stopping the fleet, changing the image used, and starting it again.

Q: Can I connect my Amazon AppStream 2.0 applications to my existing resources, such as a licensing server?

Yes. Amazon AppStream 2.0 allows you to launch streaming instances (fleets and image builders) in your VPC, which means you can control access to your existing resources from your AppStream 2.0 applications. For more information, see Network Settings for Fleet and Image Builder Instances.

Q. Can I copy my Amazon AppStream 2.0 images?

Yes. You can copy your Amazon AppStream 2.0 application images across AWS Regions. To copy an image, launch the AppStream 2.0 console and select the region that contains your existing image. In the navigation pane, choose Images, select your existing image, click Actions, select Copy, and pick your target AWS Region. You can also use the CopyImage API to programmatically copy images. Visit Tag and Copy an Image for more information.

Q: Can I share application images with other AWS Accounts?

Yes. You can share your AppStream 2.0 application images with other AWS accounts within the same AWS Region. You control the shared image and can remove it from another AWS account at any time. To learn more, visit Administer Your Amazon AppStream 2.0 Image

Q: What permissions can I give other AWS accounts when I share my application image(s) with them?

You maintain full privileges to the application image. You can share the image with other AWS accounts, granting them permission to either create image builders, use for fleets, or both. These permissions can later be revoked. However, if you granted the destination AWS account permission to create image builders, you will not be able to revoke access to the image builders or images they create from your image.

Q: If I share an application image with another AWS account, can I delete it or remove permissions?

Yes. You control the image. In order to delete the image, you will first have to stop sharing the image from all AWS accounts you shared it with. The AWS accounts you shared the image with will no longer see the image in their Image Registry, and will be unable to select it for new or existing fleets. Existing streaming instances in the fleets will continue to stream applications, but the fleet will terminate existing unused instances. If you originally granted permissions for creating image builders, they will be unable to create new image builders from it, but existing ones will continue to work. Images in the destination account created from image builders from the shared image will continue to work.

Graphics instances

Q: Does Amazon AppStream 2.0 offer GPU-accelerated instances?

Yes. Amazon AppStream 2.0 offers Graphics Design, Graphics Desktop and Graphics Pro instance families.

Graphics Design instances are ideal for delivering applications such as Adobe Premiere Pro, Autodesk Revit, and Siemens NX that rely on hardware acceleration of DirectX, OpenGL, or OpenCL. Powered by AMD FirePro S7150x2 Server GPUs and equipped with AMD Multiuser GPU technology, instances start from 2 vCPU, 7.5 GiB system memory, and 1 GiB graphics memory, to 16 vCPUs, 61 GiB system memory, and 8 GiB graphics memory.

The Graphics Desktop instance family offers a single instance type with an NVIDIA GPU based on K520 with 1,536 CUDA cores, 8 vCPUs, 15 GiB system memory, and 4 GiB graphics memory. This instance type is ideal for running desktop graphics applications such as Siemens NX, SolidWorks, ESRI ArcGIS, and other applications that use DirectX, OpenGL, OpenCL, and CUDA. The Graphics Desktop family is a powerful yet economical choice, with pricing that starts at tens-of-cents per hour.

The Graphics Pro instance family offers three different instance types to support the most demanding graphics applications. Powered by NVIDIA Tesla M60 GPUs with 2048 parallel processing cores, there are three Graphics Pro instances types starting from 16 vCPUs, 122 GiB system memory, and 8 GiB graphics memory, to 64 vCPUs, 488 GiB system memory, and 32 GiB graphics memory. These instance types are ideal for graphic workloads that need a massive amount of parallel processing power for 3D rendering, visualization, and video encoding, including applications such as Petrel from Schlumberger Software, Landmark's DecisionSpace, or MotionDSP's Ikena. For more information on available instance types and pricing, see Amazon AppStream 2.0 Pricing.

Fleets

Q: What types of fleets are available with Amazon AppStream 2.0?

Amazon AppStream 2.0 offers two fleet types: Always-On and On-Demand. Always-On fleet instances are in a running state, even if no users are connected. This is best when your users need high availability and instant access to their applications. On-Demand fleets instances don’t start until a user connects to an instance within the fleet. This fleet type is best when your users can wait up to 2 minutes to start their applications, and for streaming applications that have sporadic use.

Q: Can I switch my Amazon AppStream 2.0 Always-On fleet to On-Demand or vice versa?

You can only specify the fleet type when you create a new fleet, and you cannot change the fleet type once the fleet has been created.

Q: What are the benefits to Always-On and On-Demand fleets for Amazon AppStream 2.0?

Always-On fleets are best for when your users need high availability and instant access to their applications. On-Demand fleets instances don’t start until a user connects to an instance within the fleet, and is best for when your users can wait up to 2 minutes to start their applications, and for streaming applications that have sporadic use.

	On-Demand	Always-On

Instances	Stopped	Running
User session start	Up to 2 minutes	Instant on
Optimized for	Up to 2 minutes	Instant availability of applications
Use cases	Use cases where cost savings are critical such as education	Businesses that need instant availability of applications

Platform support

Q: What client operating systems and browsers are supported?

Amazon AppStream 2.0 can stream your applications to HTML5-compatible browsers, including the latest versions of Google Chrome, Mozilla Firefox, Microsoft Internet Explorer, and Microsoft Edge, on desktop devices, including Windows, Mac, Chromebooks, and Linux PCs. The AppStream 2.0 client for Windows lets your users use 4 monitors, 4K monitors and and USB peripherals such as 3D mouse with your applications on AppStream 2.0. The AppStream 2.0 client for Windows is supported on the following versions of Windows: Windows 7, Windows 8, Windows 8.1, and Windows 10. Both 32-bit and 64 bit versions of Windows are supported.

Q: What server operating system is supported?

Amazon AppStream 2.0 streams applications that can run on the following 64-bit OS versions - Windows Server 2012 R2, Windows Server 2016 and Windows Server 2019. You can add support for 32-bit applications by using the WoW64 extensions. If your application has other dependencies, such as the .NET framework, include those dependencies in your application installer.

Q: Which AWS regions does Amazon AppStream 2.0 support?

Please refer to the AWS Regional Products and Services page for details of Amazon AppStream 2.0 service availability by region

Q: What instance types are available to use with my Amazon AppStream 2.0 fleet?

Amazon AppStream 2.0 provides a menu of instance types for configuring a fleet. You can select the instance type that best matches your applications and end-user requirements. You can choose from General Purpose, Compute Optimized, Memory Optimized, Graphics Design, Graphics Desktop, or Graphics Pro instance families.

Auto scaling

Q: How does Amazon AppStream 2.0 scale?

Amazon AppStream 2.0 uses Fleet Auto Scaling to launch Amazon AppStream 2.0 instances running your application and to adjust the number of servers to match the demand for end-user sessions. Each end-user session runs on a separate instance, and all the apps that are streamed within a session run on the same instance. An instance is used to stream applications for only one user, and is replaced with a new instance at the end of the session.

Q: What scaling policy does Amazon AppStream 2.0 support?

Amazon AppStream 2.0 supports fixed and dynamic scaling policies. Use a fixed scaling policy to keep a constant number of Amazon AppStream 2.0 instances and users who can start a streaming session. Use a dynamic scaling policy to scale based on the use of Amazon AppStream 2.0 instances in your environment.

Q: What is an Amazon AppStream 2.0 Fleet Auto Scaling policy?

A Fleet Auto Scaling policy is a dynamic scaling policy that allows you to scale the size of your fleet to match the supply of available instances to user demand. You can define scaling policies that adjust the size of your fleet automatically based on a variety of utilization metrics, and optimize the number of running instances to match user demand.

Q: How can I create auto scaling policies for my Amazon AppStream 2.0 fleet?

You can create automatic scaling policies from the Fleets tab in the AppStream 2.0 console, or by using the AWS SDK.

Q: Which Amazon AppStream 2.0 CloudWatch metrics can I use to build Fleet Auto Scaling polices?

You can use the following metrics to build your Fleet Auto Scaling policies:

• Capacity utilization: you can scale your fleet based on the percentage of instances in your fleet that are being used

• Available capacity: you can scale your fleet based on the number of available instances in your fleet
• Insufficient capacity error: you can provision new instances when users can’t start streaming sessions due to lack of capacity

For more information, please see Fleet Auto Scaling for Amazon AppStream 2.0.

Q: Can my Amazon AppStream 2.0 fleet have more than one associated Fleet Auto Scaling policy?

Yes. You can have up to 50 Fleet Auto Scaling policies associated with a single fleet. Each policy allows you to set a single criteria and action for resizing your fleet.

Q: What is the minimum size I can set for my Amazon AppStream 2.0 fleet when using Fleet Auto Scaling policies?

You can set your Fleet Auto Scaling policies to scale in to zero instances. Scaling policies associated with your fleet decrease fleet capacity until it reaches your defined minimum, or the default setting of one if you haven’t set a minimum. For more information, please see Fleet Auto Scaling for Amazon AppStream 2.0.

Q: What is the maximum size I can set for my Amazon AppStream 2.0 fleet when using Fleet Auto Scaling policies?

Fleet Auto Scaling policies increase fleet capacity until it reaches your defined maximum size or until service limits apply. For more information, please see Fleet Auto Scaling for Amazon AppStream 2.0. For service limit information, please see Amazon AppStream 2.0 Service Limits.

Q: Are there additional costs for using Fleet Auto Scaling policies with Amazon AppStream 2.0 fleets?

There are no charges for using Fleet Auto Scaling policies. However, each CloudWatch alarm that you create and use to trigger scaling policies for your AppStream 2.0 fleets may incur additional CloudWatch charges. For more information, see Amazon CloudWatch Pricing.

Persistent storage

Q: Does Amazon AppStream 2.0 offer persistent storage so that I can save and access files between sessions?

Yes. Amazon AppStream 2.0 offers multiple options for persistent file storage to allow users to store and retrieve files between their application streaming sessions. You can use a home folder backed by Amazon S3, Google Drive for G Suite, or Microsoft OneDrive for Business. Each of these are accessed from the my files tab within an active AppStream 2.0 streaming session, and content can be saved or opened directly from the File menu in most apps.

Home folders are AppStream 2.0's native persistent storage option. Users can access a home folder on their streaming instance and save content in their folder. Files are stored in an S3 bucket which is automatically created in your AWS account. To learn more, visit Enable and Administer Home Folders for Your AppStream 2.0 Users.

You can enable **Google Drive for G Suite**, and users can link their G Suite account to access files on Google Drive. Changes are automatically synced with Google Drive. To learn more, visit Enable and Administer Google Drive for Your AppStream 2.0 Users.

You can enable **Microsoft OneDrive for Business**, and users can link their OneDrive for Business account to access their files on OneDrive. Changes are automatically synced with OneDrive for Business. To learn more, visit Enable and Administer OneDrive for Your AppStream 2.0 Users.

Q: How do users access persistent storage from their Amazon AppStream 2.0 sessions?

Users can access a home folder during their application streaming session. Any file they save to their home folder will be available for use in the future. They can also connect their G Suite account to access Google Drive and connect their Microsoft OneDrive for Business account to access OneDrive within AppStream 2.0. New files added or changes made to existing files within a streaming session are automatically synced between AppStream 2.0 and their persistent storage options.

Q. Can I enable multiple persistent storage options for an Amazon AppStream 2.0 stack?

Yes. You can enable Home Folders, Google Drive for G Suite, and Microsoft OneDrive for Business. To optimize your internet bandwidth, create a VPC endpoint for Amazon S3 and authorize AppStream 2.0 to access your VPC endpoint. This routes Home Folders data through your VPC and Google Drive or OneDrive data through the public Internet.

Q. How do I enable Google Drive for G Suite for Amazon AppStream 2.0?

When creating an Amazon AppStream 2.0 stack, select the option to enable Google Drive for the stack, provide your G Suite domain names, and create the stack. To learn more, visit Enable and Administer Google Drive for Your AppStream 2.0 Users.

Q. Can a user remove their Google Drive for G Suite account?

Yes. Users can remove permissions that AppStream 2.0 has to their Google account from their Google account permissions page.

Q. Can I control which Google Drive for G Suite accounts integrate with AppStream 2.0?

Yes. Only user accounts with your G Suite organization's domain name can use their Google Drive account. Users cannot link any other accounts. To learn more, visit Enable and Administer Google Drive for Your Users.

Q. What kind of data can users store in Google Drive during a streaming session?

Any file type that is supported by Google Drive can be stored during the streaming session. For more details on the file types supported by Google Drive, refer to Google Drive FAQs.

Q. Can users transfer files from their device to Google Drive during their streaming session?

Yes. Users can transfer files to and from from their device and Google Drive using the MyFiles feature in the streaming session toolbar. Visit Enable Persistent Storage for Your AppStream 2.0 Users to learn more.

Q. How do I enable Microsoft OneDrive for Business for Amazon AppStream 2.0?

When creating an Amazon AppStream 2.0 stack, select the option to enable OneDrive for Business for the stack, provide your OneDrive for Business domain names, and create the stack. To learn more, visit Enable and Administer OneDrive for Your AppStream 2.0 Users.

Q. Can I control which Microsoft OneDrive for Business accounts integrate with AppStream 2.0?

Yes. Only user accounts with your OneDrive for Business domain names can use their accounts. Users cannot link any other accounts. To learn more, visit Enable and Administer OneDrive for Your AppStream 2.0 Users.

Q. Can a user remove Microsoft OneDrive for Business?

Yes. Users can remove permissions that AppStream 2.0 has to their OneDrive for Business online account.

Q. What kind of data can users store in Microsoft OneDrive for Business during a streaming session?

Any file type that is supported by OneDrive for Business can be stored during the streaming session. For more details on the file types supported by OneDrive for Business, refer to OneDrive for Business documentation.

Q. Can users transfer files from their device to Microsoft OneDrive for Business during their streaming session?

Yes. Users can transfer files to and from from their device and OneDrive for Business using the MyFiles feature in the streaming session toolbar. To learn more, visit Enable and Administer OneDrive for Your AppStream 2.0 Users.

Monitoring

Q: How do I monitor usage of my Amazon AppStream 2.0 fleet resources?

There are two ways you can monitor your Amazon AppStream 2.0 fleet. First, the AppStream 2.0 console provides a lightweight, real-time view of the state of your AppStream 2.0 fleet, and offers up to two weeks of historical usage data. Metrics are displayed automatically, and don't require any setup.

Second, you can access AppStream 2.0 metrics using CloudWatch. The CloudWatch console allows you to specify reporting intervals, create custom dashboards and graphs, and set alarms.

To learn more, see Monitoring Amazon AppStream 2.0 Resources.

Q: What information can I get from the Amazon AppStream 2.0 usage metrics?

You can see the size of your Amazon AppStream 2.0 fleet, the number of running instances, the number of instances available to accept new connections, and the utilization of your fleet. You can track these metrics over time so that you can optimize your fleet settings to suit your needs.

Using Amazon CloudWatch, you can also set alarms to notify you of changes to your fleet, or when there is insufficient capacity to support your users.

For the complete list of available metrics, see Monitoring Amazon AppStream 2.0 Resources.

Q: Can I create custom Amazon CloudWatch metrics for Amazon AppStream 2.0?

Yes, you can create custom metrics for Amazon AppStream 2.0. For more information, see Publish Custom Metrics.

Q: How frequently are Amazon AppStream 2.0 metrics published to Amazon CloudWatch?

Amazon AppStream 2.0 sends metrics to Amazon CloudWatch every 1 minute. The metrics are stored in CloudWatch using the standard retention policy. For more information, see Amazon CloudWatch FAQs.

Q: How do I create CloudWatch alarms for Amazon AppStream 2.0?

You can create Amazon CloudWatch alarms for Amazon AppStream 2.0 using the CloudWatch console or the CloudWatch APIs.

Q: Are there additional costs for using CloudWatch metrics with Amazon AppStream 2.0?

There is no additional charge for viewing CloudWatch metrics for AppStream 2.0. You may incur additional charges for setting up CloudWatch alarms and retrieving metrics via the CloudWatch APIs. For more information, see Amazon CloudWatch Pricing.

Q: Does Amazon AppStream 2.0 offer a set of public APIs?

Yes, Amazon AppStream 2.0 includes APIs that you can use to easily integrate and extend the service. The APIs enable you to create, update, and delete Amazon AppStream 2.0 resources, and provide detailed information about resource states. You can create URLs for administrators to connect to their image builders to install applications, and create URLs for users to access their AppStream 2.0 applications. See our API reference for more information.

Streaming

Q: What streaming protocol does Amazon AppStream 2.0 use?

Amazon AppStream 2.0 uses NICE DCV to stream your applications to your users. NICE DCV is a proprietary protocol used to stream high-quality, application video over varying network conditions. It streams video and audio encoded using standard H.264 over HTTPS. The protocol also captures user input and sends it over HTTPS back to the applications being streamed from the cloud. Network conditions are constantly measured during this process and information is sent back to the encoder on the server. The server dynamically responds by altering the video and audio encoding in real time to produce a high-quality stream for a wide variety of applications and network conditions.

Q: What is the maximum network latency recommended while accessing Amazon AppStream 2.0?

While the remoting protocol has a maximum round-trip latency recommendation of 250 ms, the best user experience is achieved at less than 100 ms. If you are located more than 2000 miles from the AWS Regions where Amazon AppStream 2.0 is currently available, you can still use the service, but your experience may be less responsive.

Security

Q: How do I restrict network access from fleets and image builders launched in my VPC?

Security groups enable you to specify network traffic that is allowed between your streaming instances and resources in your VPC. You can restrict network access by assigning an image builder or fleet to the security groups in your VPC. For more information, refer to Security Group for Your VPC.

Q: Can I use existing VPC security groups to secure AppStream 2.0 fleets and image builders?

Yes. You can assign an image builder or fleet to existing security groups in your VPC.

Q: How many security groups can I apply to a fleet or image builder?

You can assign an image builder or fleet to up to five security groups.

Q: Can I change the security groups to which my fleets are assigned after they have been created?

Yes. You can change the security groups to which your fleets are assigned, so long as they are in the stopped status.

You can also change the rules of a security group in your VPC at any time using the Amazon EC2 console. Note that the new rules will apply to all resources assigned to that security group. For more information, refer to Security Groups for your VPC.

Q: Can I change the security groups to which my image builders are assigned after they have been created?

No. You cannot change the security groups to which your fleets are assigned after they have been created. To assign an image builder to a different security groups, you will need to create a new image builder.

You can also change the rules of a security group in your VPC at any time using the Amazon EC2 console. Note that the new rules will apply to all resources assigned to that security group. For more information, refer to Security Groups for your VPC.

Q: How is the data stored in my user's home folders secured?

Files and folders in your users' home folders are encrypted in transit using Amazon S3's SSL endpoints. Files and folders are encrypted at rest using Amazon S3-managed encryption keys.

Q: How is the data from my streamed application encrypted to the client?

The streamed video and user inputs are sent over HTTPS and are SSL-encrypted between the Amazon AppStream 2.0 instance executing your applications, and your end users.

Q. Can I control data transfer between AppStream 2.0 and my users' devices?

Yes. You can choose whether to allow users to transfer data between their streaming applications and their local device through copy or paste, file upload or download, or print actions. To learn move, visit Create Fleets and Stacks.

Identity

Q: How do I authenticate users with Amazon AppStream 2.0 applications?

There are three options to authenticate users with Amazon AppStream 2.0: you can use built-in user management, you can build a custom identity, or you can set up federated access using SAML 2.0.

When using built-in user management, you can set up and manage your users in the AppStream 2.0 management console from the User Pool tab. To add a new user, all you need is their first and last name, and an e-mail address. To learn more about user management within AppStream 2.0, see Using the AppStream 2.0 User Pool.

When using federated sign-in to authenticate users, you will set up identity federation using SAML 2.0, which allows you to use your existing user directory to control access to applications available via AppStream 2.0. For details on setting up SAML integration, see the steps outlined here.

When building an entitlement service, you should authenticate users either with a custom identity or by using a service such as Login with Amazon. After your custom identity has authenticated a user, it should call into Amazon AppStream 2.0 to create a new streaming URL. AppStream 2.0 returns a URL for the session that can be opened in a browser to start the streaming session.

Q: Can I use Amazon AppStream 2.0 with my existing user directory, including Microsoft Active Directory?

Yes. Amazon AppStream 2.0 supports identity federation using SAML 2.0, which allows you to use your existing user directory to manage end user access to your AppStream 2.0 apps. For details on setting up SAML integration, see the steps outlined here.

Q: What type of identity federation does Amazon AppStream 2.0 support?

Amazon AppStream 2.0 supports federation using SAML 2.0 (Identity Provider initiated). This type of federated access allows a user to sign in by first authenticating with an identity federation provider, after which they can access their AppStream 2.0 apps.

Q: What are the requirements for setting up identity federation with Amazon AppStream 2.0?

To configure identity federation with Amazon AppStream 2.0, you need a SAML 2.0 Identity Provider that links to an existing LDAP-compatible directory, such as Microsoft Active Directory. Microsoft Active Directory Federation Services (ADFS), Ping Identity, Okta, and Shibboleth, are all examples of SAML 2.0 Identity Providers that will work with AppStream 2.0.

Q: Can I control which users access my Amazon AppStream 2.0?

Yes. When using built-in user management, you can control which users have access to your Amazon AppStream 2.0 stacks in the User Pool tab of the AppStream 2.0 management console. To learn more about user management within AppStream 2.0, see Using the AppStream 2.0 User Pool.

When you use SAML 2.0, you can control which users have access to your Amazon AppStream 2.0 stacks by mapping the users in your federation service to the IAM role that has access permissions to the stack. Please refer to the AppStream 2.0 documentation for detailed information and step-by-step guidelines for popular federation services.

Q: Can I enable multi-factor authentication for my users?

Yes. You can enable Multi-Factor Authentication when using federation with SAML 2.0 or when using your own entitlement service.

Q: Can users choose which Amazon AppStream 2.0 stack they want to access during signing-in?

Yes. You can setup every Amazon AppStream 2.0 stack as an entity or a package in your federation service. This allows your users to select which stack they want to access while signing in from your application portal.

Q: Who can access the management console for my Amazon AppStream 2.0 application?

You can use AWS Identity and Access Management (IAM) to add users to your AWS account and grant them access to view and manage your Amazon AppStream 2.0 application. For more information, see “What is IAM?” in the IAM User Guide.

Microsoft Active Directory domain support

Q: Can I join Amazon AppStream 2.0 image builders to Microsoft Active Directory domains?

Yes, Amazon AppStream 2.0 images can be joined to your Microsoft Active Directory domains. This allows you to apply your existing AD policies to your streaming instances, and provides your users with single sign on access to Intranet sites, file shares, and network printers from within their applications. Your users are authenticated using a SAML 2.0 provider of your choice, and can access applications that require a connection to your AD domain.

Q: What Microsoft Active Directory versions are supported?

Microsoft Active Directory Domain Functional Level Windows Server 2008 R2 and newer are supported by Amazon AppStream 2.0.

Q: Which AWS Directory Services directory options are supported by Amazon AppStream 2.0?

Amazon AppStream 2.0 supports AWS Directory Services Microsoft AD. Other options such as AD Connector and Simple AD are not supported. To learn more about AWS Microsoft AD see What Is AWS Directory Service.

Q: How do I join my Amazon AppStream 2.0 instances to my Microsoft Active Directory domain?

To get started you will need a Microsoft Active Directory domain that is accessible from an Amazon VPC, the credentials of a user with authority to join the domain, and the domain Organizational Unit (OU) you want to join to your fleet. For more information, see Using Active Directory Domains with AppStream 2.0.

Q: Can I use my existing Organization Units (OU) structure with Amazon AppStream 2.0?

Yes, you can use your existing Organizational Unit (OU) structure with Amazon AppStream 2.0. To learn more, see Using Active Directory Domains with AppStream 2.0.

Q: What gets joined to my Microsoft Active Directory domain by Amazon AppStream 2.0?

Amazon AppStream 2.0 will automatically create a unique computer object for every image builder and fleet instance you configure to be joined to your Microsoft Active Directory domain.

Q: How can I identify Amazon AppStream 2.0 computer objects in my Microsoft Active Directory domain?

Amazon AppStream 2.0 computer objects are only be created in the Microsoft Active Directory Organization Unit (OU) you specify. The description field indicates that the object is an AppStream 2.0 instance, and to which fleet the object belongs. To learn more, see Using Active Directory Domains with AppStream 2.0.

Q: How are computer objects that are created by Amazon AppStream 2.0 deleted from my Microsoft Active Directory domain?

Computer objects created by Amazon AppStream 2.0 that are no longer used will remain in your Active Directory (AD) if the AppStream 2.0 fleet or image builder is deleted, you update a fleet or image builder to a new OU, or select a different AD. To remove unused objects you will have to delete them manually from your AD domain. To learn more, see Using Active Directory Domains with AppStream 2.0.

Q: How do I provide users with access to Amazon AppStream 2.0 streaming instances that are joined to a Microsoft Active Directory domain?

To enable user access, you will need to set up federated access using a SAML 2.0 provider of your choice. This allows you to use your existing user directory to control access to streaming applications available via Amazon AppStream 2.0. For details on setting up SAML 2.0 integration, see the steps outlined at Setting Up SAML.

Q: Can I connect my users that are managed through User Pools to my Active Directory domain?

No. At this time we do not support User Pools users connecting to domain joined resources. To learn more about User Pools see, Using the AppStream 2.0 User Pool.

Pricing and billing

Q: How much does Amazon AppStream 2.0 cost?

You are charged for the streaming resources in your Amazon AppStream 2.0 environment, and monthly user fees per unique authorized user accessing applications via Amazon AppStream 2.0. You pay for these on-demand, and never have to make any long-term commitments.

The streaming resources consist of Amazon AppStream 2.0 instances in your Amazon AppStream 2.0 fleet as well as image builder instances. You have the option to have Always-On and On-Demand fleets. For Always-On fleets you pay for instances in your fleet that are running, even if users are not connected. These instances are billed per hour, and the price per hour is based on the instance type you select. For On-Demand fleets you pay for the instances in your fleet that are running only when a user is connected. These instances are billed per hour, and the price per hour is based on the instance type you select. In an On-Demand fleet If an instance is running but not connected to a user, you pay a nominal hourly On-Demand Stopped Instance fee, which is the same for all instance types within a region. Image builder instances are only available as always on, and you pay for instances that are running, even if users are not connected. The charge for Always-On and On-Demand fleet instances as well as image builder instances includes the cost of the storage volumes used by the Amazon AppStream 2.0 image, and outbound bandwidth used by the streaming protocol.

You can control the number of running instances using fixed or dynamic scaling policies.

The monthly user fee is used to pay for the Microsoft Remote Desktop Services Subscriber Access License (RDS SAL). This fee is charged per unique authorized user, and is charged in full (not pro-rated), regardless of when a user first accesses Amazon AppStream 2.0 in that month. Schools, universities, and public institutions may qualify for reduced user fees. Please reference the Microsoft Licensing Terms and Documents for qualification requirements. If you think you may qualify, please contact us. We will review your information and work with you to reduce your Microsoft RDS SAL fee. There is no user fee incurred when using image builder instances.

Q: Can I bring my own licenses and waive the user fees?

Yes. If you have Microsoft License Mobility, you may be eligible to bring your own Microsoft RDS CAL licenses and use them with Amazon AppStream 2.0. For users covered with your own licenses, you won’t incur the monthly user fees. For more information about using your existing Microsoft RDS SAL licenses with Amazon AppStream 2.0, please visit this page, or consult with your Microsoft representative.

Q: What are the requirements for schools, universities, and public institutions to reduce their user fee?

Schools, universities, and public institutions may qualify for reduced user fees. Please reference the Microsoft Licensing Terms and Documents for qualification requirements. If you think you may qualify, please contact us. We will review your information and work with you to reduce your Microsoft RDS SAL fee. There is no user fee incurred when using image builder instances.

Q: What do I need to provide to qualify as a school, university, or public institution?

You will need to provide AWS your institution's full legal name, principal office address, and public website URL. AWS will use this information to qualify you for AppStream 2.0's reduced user fees for qualified educational institutions. Please note: The use of Microsoft software is subject to Microsoft’s terms. You are responsible for complying with Microsoft licensing. If you have questions about your licensing or rights to Microsoft software, please consult your legal team, Microsoft, or your Microsoft reseller. You agree that we may provide the information to Microsoft in order to apply educational pricing to your Amazon AppStream 2.0 usage.

Q. Does qualification for Amazon AppStream 2.0's reduced RDS SAL user fees affect other AWS cloud services?

No, your user fees are specific to Amazon AppStream 2.0, and do not affect any other AWS cloud services or licenses you have.

Q: Can I use tags to obtain usage and cost details for Amazon AppStream 2.0 on my AWS monthly billing report?

Yes. When you set tags to appear on your monthly Cost Allocation Report, your AWS monthly bill will also include those tags. You can then easily track costs according to your needs. To do this, first assign tags to your Amazon AppStream 2.0 resources by following the steps in Tagging Your AppStream 2.0 Resources. Next, select the tag keys to include in your cost allocation report by following the steps in Setting Up Your Monthly Cost Allocation Report.

Q: Are there any costs associated with tagging Amazon AppStream 2.0 resources?

There are no additional costs when using tags with Amazon AppStream 2.0.

Compliance

Q: Is Amazon AppStream 2.0 HIPAA eligible?

Yes. If you have an executed Business Associate Addendum (BAA) with AWS, you can use Amazon AppStream 2.0 with the AWS accounts associated with your BAA to stream desktop applications with data containing protected health information (PHI). If you don’t have an executed BAA with AWS, contact us and we will put you in touch with a representative from our AWS sales team. For more information, see HIPAA Compliance.

Q: Is AppStream 2.0 PCI Compliant?

Yes. Amazon AppStream 2.0 is PCI compliant and conforms to the Payment Card Industry Data Security Standard (PCI DSS). PCI DSS is a proprietary information security standard administered by the PCI Security Standards Council, which was founded by American Express, Discover Financial Services, JCB International, MasterCard Worldwide and Visa Inc. PCI DSS applies to all entities that store, process or transmit cardholder data (CHD) and/or sensitive authentication data (SAD) including merchants, processors, acquirers, issuers, and service providers. The PCI DSS is mandated by the card brands and administered by the Payment Card Industry Security Standards Council. For more information, see PCI DSS Compliance.

Q: Is Amazon AppStream 2.0 included in the System and Organizational Controls (SOC) reports?

Yes. Amazon AppStream 2.0 is included in the AWS System and Organizational Controls (SOC) reports. AWS System and Organization Controls Reports are independent third-party examination reports that demonstrate how AWS achieves key compliance controls and objectives. The purpose of these reports is to help you and your auditors understand the AWS controls established to support operations and compliance. You can learn more about the AWS Compliance programs by visiting AWS Compliance Programs or by visiting the Services in Scope by Compliance Program.

Amazon WorkLink FAQs

General

Q: What is Amazon WorkLink?

Amazon WorkLink is a fully managed, cloud-based service that enables secure one-click access to internal websites and web apps from mobile devices. With Amazon WorkLink, employees can access internal websites as seamlessly as they access any public website. Employees can simply type a URL in Chrome or Safari, or click on a link in their email. Amazon WorkLink works behind the scenes to isolate and render corporate web content using the compute and networking infrastructure of AWS, and send that content securely to employee devices. Amazon WorkLink reduces the risk of information loss or theft because data is never stored or cached on devices, and IT administrators can enforce their own security and access policies.

Q: Why should I use Amazon WorkLink?

Amazon WorkLink enables your employees to access corporate websites in Chrome or Safari as seamlessly as they access any other public website. Amazon WorkLink does not store or cache data on user devices because your web content is rendered in AWS and sent to user devices as encrypted Scalable Vector Graphics (SVG). Because user devices do not connect directly to your network, Amazon WorkLink eliminates the path for device-based malware to reach resources in your network. Amazon WorkLink does not require content migration and can be configured to access your web content regardless of where it is hosted. Amazon WorkLink honors your existing security policies so you don’t have to manage access. As a fully managed service, Amazon WorkLink automatically handles the deployment, capacity provisioning, scaling, and updates to browsers and resources in the cloud.

Q: How is Amazon WorkLink related to other EUC services from AWS?

Amazon WorkLink is the latest addition to the AWS End-User Computing category, which today consists of WorkSpaces, AppStream 2.0, and Workdocs. Each service is designed to provide secure access to a different environment – WorkSpaces for full desktops, AppStream for applications, WorkLink for web content, and WorkDocs for documents.

Getting Started

Q: How do I get started with Amazon WorkLink?

You can get started with Amazon WorkLink from the AWS Management Console. First, select the region that will serve as your home region (this is where Amazon WorkLink will be deployed, your websites rendered, and user analytics generated). Then, link your existing SAML-based identity provider with Amazon WorkLink. Next, associate web domains that your employees will access through Amazon WorkLink . Lastly, you need a VPC to connect Amazon WorkLink with your corporate content.

Q: How does Amazon WorkLink communicate with my corporate network?

Amazon WorkLink containers execute on customer-specific EC2 instances provisioned on-demand. You can create a VPC in your account, identify subnets for Amazon WorkLink traffic, and give Amazon WorkLink permission to create Elastic Network Interfaces (X-ENI) that will be linked to EC2 rendering hosts allocated to you. Your content can exist within that VPC (for example, applications hosted on an EC2 instance), in another VPC that is peered with it, or on-premises. Resources running in a VPC can use AWS PrivateLink. Resources hosted on-premises must be accessible via an IPsec tunnel, AWS Direct Connect, or the new AWS Transit Gateway.

Q: How do my end users get started with Amazon WorkLink?

One you have completed setup in the AWS Management Console, you can use the provided email template to invite employees to download the WorkLink app. Your end-users can simply download the WorkLink app from their device app store, log in with their corporate credentials, and start accessing internal websites using their browser.

Compatibility

Q: What devices are supported at launch?

Amazon WorkLink supports devices running iOS 12+ and Android 6+. The Amazon WorkLink app is available for download in the iOS App Store and the Google Play Store.

Q: Do some applications work better with Amazon WorkLink than others?

Amazon WorkLink uses Chromium as part of its core rendering engine, so a good rule of thumb is that if it works in Chrome, it will work in Amazon WorkLink. That said, Amazon WorkLink will work best for sites that are not graphically intense (such as sites that deliver video with more than 60 frames per second, or sites that include Flash, Silverlight, or other plugins).

Q: Does Amazon WorkLink work with SaaS applications?

Amazon WorkLink supports SaaS web apps served via domains that you own. For example, if you use JIRA or SAP with a custom domain for your enterprise.

Q: Does Amazon WorkLink work with e-mail?

Amazon WorkLink supports web interfaces to email. For example, you can enable end users to access email via Outlook Web Access served via a custom domain. However, Amazon WorkLink does not today support email in native email clients.

Security

Q: How does Amazon WorkLink manage user access and authentication?

Amazon WorkLink is designed to work with your existing systems and not add extra layers of user management. Amazon WorkLink supports user authentication and federated sign-in using any SAML 2.0 compliant identity providers, such as AWS SSO, OneLogin, Okta, or Ping Identity.

Q: How is my data protected?

During an Amazon WorkLink session, customer data is isolated in the cloud and securely rendered. Once the session ends, that data is deleted. Throughout this process, data in transit is protected by enterprise-grade encryption.

Q: What are the key security differentiators enabled by Amazon WorkLink?

Amazon WorkLink is an AWS service, so your content is always handled in a secure environment consistent with AWS standards. As a user of Amazon WorkLink, a part of the cloud is dedicated to you and only handles your data. Amazon WorkLink only handles pages that you associate with your fleet. Your content is delivered in a format that enables seamless interactions for end users while enhancing the controls on how that data is secured on the user device.

Q: Does Amazon WorkLink prevent web browsers from caching corporate data?

Yes, Amazon WorkLink streams a representation (image) of your content only to the browsers that enforce web standards on content caching. The streaming process prevents corporate content from being cached on user devices.

Q: Is corporate data stored on end-user devices?

Content from your websites is never stored on the end-user web browsers. Session information (i.e., cookies) is encrypted in the cloud and stored on end-user devices in an encrypted format that cannot be decrypted on the device. Because the encrypted data is unusable on the device, there is no need to wipe devices if a device is lost or stolen or if a user leaves the company.

Q: How do I associate a corporate website with Amazon WorkLink ?

In your AWS Management Console, you can associate your domains to Amazon WorkLink by providing the domain name and the domain certificate via Amazon

Certificate Manager.

Q: Does Amazon WorkLink have visibility of personal browsing activity?

No, Amazon WorkLink only handles domains you associate with the service and requires administrators to provide proof of ownership (via Amazon Certificate Manager). It is not possible for administrators to associate public domains they do not own (such as news or social media websites) with the service.

Monitoring

Q: What information can I get from Amazon WorkLink usage metrics?

Amazon WorkLink usage metrics provide the following information:

- BrowsingSessionId: An unique identifier representing a user's browsing session
- UserId: Registered user email.
- DeviceId: Unique identifier representing the device registered with Amazon WorkLink.
- DomainName: domain name provided by you during set up.
- ErrorMessage: Message describing error, if any, encountered during the browsing session.
- HTTPStatusCode: Status code of the HTTP request directed via Amazon WorkLink.
- URI: Path that follows the domain name when the request was sent by the user browser.

Q: How can I create custom usage metrics for Amazon WorkLink?

Amazon WorkLink delivers usage data logs to you via an Amazon Kinesis stream. Amazon Kinesis makes it easy to collect, process, and analyze data so you can get timely insights and react quickly to new information. You can store, process, and analyze the logs with familiar tools or data store of your choice. For example, you can stream these logs to Amazon S3 and use tools such as Splunk to analyze the information. Similarly, you can direct this data to Amazon

Redshift via Kinesis Data Firehose and use Amazon QuickSight to generate reports and dashboards.

Q: Do the Amazon WorkLink APIs log actions in AWS CloudTrail?

Yes. To receive a history of Amazon WorkLink API calls made on your account, you can simply turn on CloudTrail in the AWS Management Console.

Pricing and Availability

Q: How much does Amazon WorkLink cost?

Amazon WorkLink is a pay-as-you-go service with no minimum fees, upfront commitments, or long-term contracts. Users have unlimited access to the content you enable with Amazon WorkLink, and you are charged monthly based on the number of users that connect to the service. Please see our pricing page for the latest information.

Q: What are all the regions where Amazon WorkLink is available?

Amazon WorkLink is generally available in the following AWS regions: US East (N. Virginia), US East (Ohio), US West (Oregon), and Europe (Ireland).

Learn about features

Learn more about Amazon WorkLink features.

Learn more »

Sign up for a free account

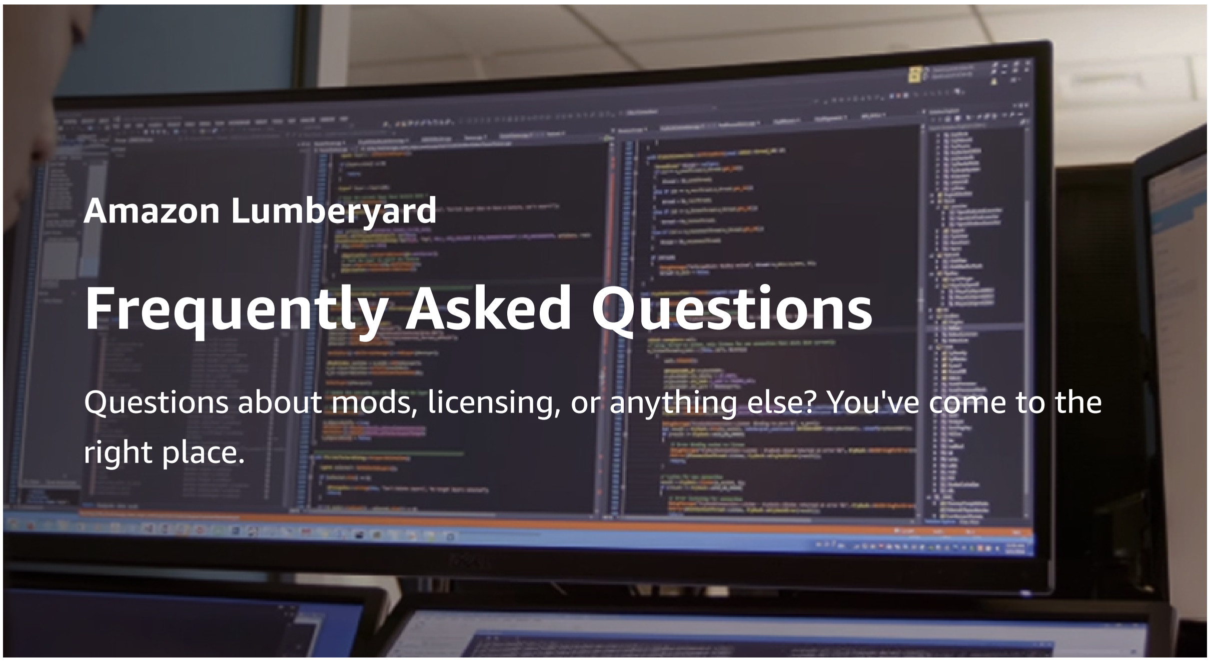
Instantly get access to the AWS Free Tier.

Sign up »

Start building in the console

Get started building with Amazon WorkLink in the AWS Console.

Sign in »



Amazon Lumberyard

Frequently Asked Questions

Questions about mods, licensing, or anything else? You've come to the right place.

General

Q. What is Amazon Lumberyard? ⊕

Q. What do you mean by "free"? Do I owe you 5% of my revenues? ⊕

Q. Is Amazon Game Studios using Lumberyard to build games? ⊕

Q. Is Lumberyard based on other technologies? ⊕

Q. Do I really get source code access to Lumberyard? ⊕

Q. What kind of support is available for Lumberyard? ⊕

Q. Can I use Lumberyard for non-game purposes, such as architecture, simulations, and animated movies? ⊕

Q. What are the system requirements for building a game with the Lumberyard Editor and tools? ⊕

Q. What device platforms does Lumberyard support? ⊕

Q. Does Lumberyard support VR? ⊕

Q. How do I get started with Xbox and PlayStation game development? ⊕

Lumberyard and AWS

Q. If I build a single-player game that uses no cloud connectivity, do I have to pay to use the engine? ⊕

Q. Do I need an AWS account to use Lumberyard?

Q. Do I have to run my game on AWS?

Q. Is there a surcharge or other additional fee over and above AWS service rates for Lumberyard customers?

Q. How do I authorize my team of developers to use Cloud Canvas and AWS via the Lumberyard Editor?

Q. Can I grant certain team members permissions or restrictions to access specific AWS services in Cloud Canvas?

Q. Which AWS services are available in Cloud Canvas?

Licensing and Mods

Q. What are the license terms for Lumberyard?

Q. Do I have to sell my Lumberyard game on Amazon? ⊕

Q. Can I take Lumberyard and make my own game engine and distribute it? ⊕

Q. Is Lumberyard “open source”? ⊕

Q. Can I make plugins or tools for Lumberyard? ⊕

Q. Can I redistribute source code modifications to Lumberyard? ⊕

Q. Do I get to set my own terms when I post a fork on GitHub, or share improvements on a forum? ⊕

Q. Can I submit code that adds features or fixes bugs that I find? ⊕

Q. Can my Lumberyard game connect to services like Steamworks, Xbox Live, PSN, Apple Game Center, Google Play Games, or console social services? ⊕

Q. Can my game use an alternate web service instead of AWS? ⊕

Q. Is it okay for me to use my own servers? ⊕

Q. Can I use the game assets that are included with Lumberyard in my game? ⊕

Q. Can I redistribute assets from Lumberyard or Lumberyard sample projects? ⊕

Q. Can I use Lumberyard in a way not permitted by the Service Terms? ⊕

Q. Does Lumberyard support integrations with third-party middleware? ⊕

Q. Can I include Lumberyard’s tools so my players can build mods for my game? ⊕

Registration

Q. Where can I tell you about my Lumberyard game? ⊕

Other

Q. How do I submit feedback or suggestions? ⊕

Q. I’d love to join your team. Are you hiring? ⊕

AWS IoT Core FAQs

Q. What is AWS IoT Core?

AWS IoT Core is a managed cloud platform that lets connected devices easily and securely interact with cloud applications and other devices. AWS IoT Core can support billions of devices and trillions of messages, and can process and route those messages to AWS endpoints and to other devices reliably and securely. With AWS IoT Core, your applications can keep track of and communicate with all your devices, all the time, even when they aren't connected.

Q. What does AWS IoT Core offer?

Connectivity between devices and the AWS cloud. First, with AWS IoT Core you can communicate with connected devices securely, with low latency and with low overhead. The communication can scale to as many devices as you want. AWS IoT Core supports standard communication protocols (HTTP, MQTT, and WebSockets are supported currently). Communication is secured using TLS.

Connectivity between devices and the AWS cloud. First, with AWS IoT Core you can communicate with connected devices securely, with low latency and with low overhead. The communication can scale to as many devices as you want. AWS IoT Core supports standard communication protocols (HTTP, MQTT, and WebSockets are supported currently). Communication is secured using TLS.

Processing data sent from connected devices. Secondly, with AWS IoT Core you can continuously ingest, filter, transform, and route the data streamed from connected devices. You can take actions based on the data and route it for further processing and analytics.

Application interaction with connected devices. Finally, AWS IoT Core accelerates IoT application development. It serves as an easy to use interface for applications running in the cloud and on mobile devices to access data sent from connected devices, and send data and commands back to the devices.

Q. How does AWS IoT Core work?

Connected devices, such as sensors, actuators, embedded devices, smart appliances, and wearable devices, connect to AWS IoT Core over HTTPS, WebSockets, or secure MQTT.

Included in AWS IoT Core is a **Device Gateway** that allows secure, low-latency, low-overhead, bi-directional communication between connected devices and your cloud and mobile applications.

AWS IoT Core also contains a **Rules Engine** which enables continuous processing of data sent by connected devices. You can configure rules to filter and transform the data. You also configure rules to route the data to other AWS services such as DynamoDB, Kinesis, Lambda, SNS, SQS, CloudWatch, Elasticsearch Service with built-in Kibana integration, as well as to non-AWS services, via Lambda for further processing, storage, or analytics.

There is also a **Registry** where you can register and keep track of devices connected to AWS IoT Core, or devices that may connect in the future. The **Device Shadow** in AWS IoT Core enables cloud and mobile applications to query data sent from devices and send commands to devices, using a simple REST API, while letting AWS IoT Core handle the underlying communication with the devices. The Device Shadow accelerates application development by providing a uniform interface to devices, even when they use one of the several IoT communication and security protocols with which the applications may not be compatible. The Device Shadow also accelerates application development by providing an always available interface to devices even when the connected devices are constrained by intermittent connectivity, limited bandwidth, limited computing ability or limited power.

Communication with AWS IoT Core is secure. The service requires all of its clients (connected devices, server applications, mobile applications, or human users) to use strong authentication (X.509 certificates, AWS IAM credentials, or 3rd party authentication via AWS Cognito). All communication is encrypted. AWS IoT Core also offers fine-grained authorization to isolate and secure communication among authenticated clients.

Q. What is 2lemetry and how does it relate to AWS IoT?

2lemetry was acquired by AWS in 2015, and their capabilities provided foundational elements such as the MQTT Message Broker and the Rules Engine for AWS IoT Core.

Q. In which regions is AWS IoT Core available?

See the AWS Region Table for the current list of regions for AWS IoT Core.

You can use AWS IoT Core regardless of your geographic location, as long as you have access to one of the above AWS regions.

Q. How do I get started with using AWS IoT Core?

Use the AWS IoT Console or refer to the Quickstart section of our developer guide to test drive the AWS IoT Core in minutes.

Also, take a look at the AWS-powered Starter Kits provided by our partners.

Refer to the AWS IoT Core documentation for further details.

Q. Which languages does the AWS IoT Console support?

The AWS IoT Console supports English, French, Japanese, Korean, Simplified Chinese, German, Portuguese, Spanish, Italian and Traditional Chinese.

Q. How can I switch the console's language?

Click on the language at the bottom left corner of the console to pick the language. The language selection will persist throughout the consoles of different AWS services.

Q. What are the ways for accessing AWS IoT Core?

You can use the AWS Management Console, the AWS SDKs, the AWS CLI, and the AWS IoT Core APIs. Connected devices can use the AWS IoT Device SDKs to simplify the communication with AWS IoT Core.

The AWS IoT Core APIs and commands are largely divided into control plane operations and data plane operations. The control plane operations enable you to do tasks such as configuring security, registering devices, configuring rules for routing data, and setting up logging. The data plane operations enable you to ingest data from connected devices into AWS IoT Core with low latency and high throughput rate at a large scale.

Q. What communication and authentication protocols does AWS IoT Core support?

For control plane operations, AWS IoT Core supports HTTPS. For data plane operations, AWS IoT Core supports HTTPS, WebSockets, and secure MQTT – a protocol often used in IoT scenarios.

HTTPS and WebSockets requests sent to AWS IoT Core are authenticated using AWS IAM or AWS Cognito, both of which support the AWS SigV4 authentication. If you are using the AWS SDKs or the AWS CLI, the SigV4 authentication is taken care of for you under the hood. HTTPS requests can also be authenticated using X.509 certificates. MQTT messages to AWS IoT Core are authenticated using X.509 certificates.

With AWS IoT Core you can use AWS IoT Core generated certificates, as well as those signed by your preferred Certificate Authority (CA).

Q. Can devices that are NOT directly connected to the Internet access AWS IoT Core?

Yes, via a physical hub. Devices connected to a private IP network and devices using non-IP radio protocols such as ZigBee or Bluetooth LE can access AWS IoT Core as long as they have a physical hub as an intermediary between them and AWS IoT Core for communication and security.

Q. How should applications access AWS IoT Core?

Applications connecting to AWS IoT Core largely fall in two categories: 1. companion apps and 2. server applications. Companion apps are mobile or client-side browser applications that interact with connected devices via the cloud. A mobile app that lets a consumer remotely unlock a smart lock in the consumer’s house is an example of a companion app. Server applications are designed to monitor and control a large number of connected devices at once. An example of a server application would be a fleet management website that plots thousands of trucks on a map in real-time.

AWS IoT Core enables both companion apps and server applications to access connected devices via uniform, RESTful APIs. Applications also have the option to use pub/sub to communicate directly with the connected devices.

Typically the companion apps would authenticate using end-user identities which are managed either by your own identity store or a third party identity provider such as Facebook and Login with Amazon. For companion apps, use Amazon Cognito, which integrates with several identity providers. Cognito identities can be authorized to access AWS IoT Core, and their access can be restricted only to the resources relevant to them. For example, as a connected washing machine manufacturer, you can authorize your consumers to access your AWS IoT Core information pertaining only to their individual washing machines.

Server applications (such as a mapping application running on Amazon EC2) can use IAM roles to access AWS IoT Core.

Q. Can I get a history of AWS IoT Core API calls made on my account for security analysis and operational troubleshooting purposes?

Yes, to receive a history of AWS IoT Core API calls made on your account, you simply turn on CloudTrail in the AWS Management Console.

Q. How do I send feedback?

To send feedback, click on the “Feedback” link in the footer bar of the console.

Q: What is the Device Gateway?

The Device Gateway forms the backbone of communication between connected devices and the cloud capabilities such as the Rules Engine, Device Shadow, and other AWS and 3rd-party services.

The Device Gateway supports the pub/sub messaging pattern, which enables scalable, low-latency, and low-overhead communication. It is particularly useful for IoT scenarios where billions of devices are expected to communicate frequently and with minimal delay. Pub/sub involves clients publishing messages on logical communication channels called ‘topics’ and clients subscribing to topics to receive messages. The device gateway enables the communication between publishers and subscribers. Traditionally, organizations have had to provision, operate, scale, and maintain their own servers as device gateways to take advantage of pub/sub. AWS IoT Core has eliminated this barrier by providing the Device Gateway.

The Device Gateway scales automatically with your usage, without any operational overhead for you. AWS IoT Core supports secure communication with the device gateway, AWS-account level isolation, as well as fine-grained authorization within an AWS account. The device gateway currently supports publish and subscribe over secure MQTT and WebSockets, as well as publish over HTTPS.

Q. What is MQTT?

MQTT is a lightweight pub/sub protocol, designed to minimize network bandwidth and device resource requirements. MQTT also supports secure communication using TLS. MQTT is often used in IoT use cases. MQTT v3.1.1 is an OASIS standard, and the Device Gateway supports most of the MQTT specification.

Q. What is the Rules Engine?

The Rules Engine enables continuous processing of inbound data from devices connected to AWS IoT Core. You can configure rules in the Rules Engine in an intuitive, SQL-like syntax to automatically filter and transform inbound data. You can further configure rules to route data from AWS IoT Core to several other AWS services as well as your own or 3rd party services.

Here are just a few example use cases of rules:

- Filtering and transforming incoming messages and storing them as time series data in DynamoDB.
- Sending a push notification via SNS when the data from a sensor crosses a certain threshold.
- Saving a firmware file to S3
- Processing messages simultaneously from a multitude of devices using Kinesis
- Invoke Lambda to do custom processing on incoming data
- Sending a command to a group of devices with an automated republish

Q. How are the rules defined and triggered?

An AWS IoT Core rule consists of two main parts:

A SQL statement that specifies the pub/sub topics to apply the rule on, data transformation to perform, if any, and the condition under which the rule should be executed. The rule is applied on every message published on the specified topics.

An actions list that defines the actions to take when the rule is executed, that is, when an incoming message matches the condition specified in the rule.

Rule definitions use a JSON-based schema. You can directly edit the JSON or use the rules editor in the AWS Management Console.

As an example, here is a rule for saving temperature data from a sensor to DynamoDB whenever the temperature is above 50:

JSON

```
{
    "sql": "SELECT * from 'iot/tempSensors/#' WHERE temp > 50",
    "description": "Rule to save sensor data when temperature is a
    "actions": [
     {
            "dynamoDB": {
            "tableName": "HighTempTable",
            "roleArn": "arn:aws:iam::your-aws-account-id:role/dyna
            "hashKeyField": "key",
            "hashKeyValue": "${topic(3)}",
            "rangeKeyField": "timestamp",
            "rangeKeyValue": "${timestamp()}"
            }
        }
    ]
}
```

Sensors in this example are publishing on their topics under “iot/tempSensors/”. The first line of the rule defines the SQL SELECT statement used to query on the “iot/tempSensors/#” topic. It contains a WHERE clause that extracts the value of a ‘temp’ field in the message’s payload and checks if it passes the condition ‘greater than 50’. If the condition is met, the data is stored in the specified DynamoDB table. The example uses built-in functions for tasks such as traversing the message payload and getting current time.

Q. Where can I learn more about rules?

You can learn more about rule here Core Rules documentation

Q. What is the Registry and what should I use it for?

IoT scenarios can range from a small number of mission-critical devices to large fleets of devices. The Registry allows you to organize and track those devices. You can maintain a logical handle in the Registry for every device you are connecting to AWS IoT Core. Each device in the Registry can be uniquely identified and can have metadata such as model numbers, support contact, and certificates associated with it. You can search for connected devices in the Registry based on the metadata.

Q. What is a Thing Type?

Thing Types allow you to effectively manage your catalogue of devices by defining common characteristics for devices that belong to the same device category. In addition, a Thing associated with a Thing Type can now have up to 50 attributes including 3 searchable attributes.

Q. What is Simplified Permission Management?

This feature allows you to easily manage permission policies for a large number of devices by using variables that reference Registry or X.509 certificate properties. The integration of Registry and Certificate properties with device policies offers the benefits listed below:

- You can now reference Registry properties in device permission policies. Referencing device properties defined in the Registry allows your policies to reflect any changes made in the Registry. For example, by referencing the Thing Attribute named “building-address” as a variable in the policy, devices will automatically inherit a new set of permissions when they move buildings.
- You can share a single generic policy for multiple devices. A generic policy can be shared among the same category of devices instead of creating a unique policy per device. For example, a policy that references the “serial-number” as a variable, can be attached to all the devices of the same model. When devices of the same serial number connect, policy variables will be automatically substituted by their serial-number.

Q. What is the Device Shadow?

The Device Shadow enables cloud and mobile applications to easily interact with the connected devices registered in AWS IoT Core. The Device Shadow in AWS IoT Core contains properties of a connected device. You can define any set of properties applicable to your use case. For example, for a smart light bulb, you might define ‘on-or-off’, ‘color’, and ‘brightness’ as the properties. The connected device is expected to report the actual values of those properties, which are stored in the Device Shadow. Applications get and update the properties simply by using a RESTful API provided by AWS IoT Core. AWS IoT Core and the Device SDKs take care of synchronizing property values between the connected device and its Device Shadow in AWS IoT Core.

Q. Do I have to use the Registry and the Device Shadow?

You can have applications communicate directly to the connected devices using the Device Gateway and/or the Rules Engine in AWS IoT Core. However, we recommend using the Registry and the Device Shadow since they offer richer and more structured development and management experience that lets you focus on the unique value you want to create for your customers rather than having to focus on the underlying communication and synchronization between the connected devices and the cloud.

Q. What is the lifecycle of a device and its Device Shadow in AWS IoT Core?

- You register a device (such as a light bulb) in the Registry.
- You program connected device to publish a set of its property values or ‘state (“I am ON and my color is RED”) to the AWS IoT Core service.
- The last reported state is stored in the Device Shadow in AWS IoT Core.
- An application (such as a mobile app controlling the light bulb) uses a RESTful API to query AWS IoT Core for the last reported state of the light bulb, without the complexity of communicating directly with the light bulb.

- When a user wants to change the state (such as turning the light bulb from ON to OFF), the application uses a RESTful API to request an update, i.e. sets a ‘desired’ state for the device in AWS IoT Core. AWS IoT Core takes care of synchronizing the desired state to the device.
- The application gets notified when the connected device updates its state to the desired state.

Q. Where can I learn more about the Registry and the Device Shadow?

For more information on the Registry, see the Registry documentation. For more information on the Device Shadow, see the Device Shadow documentation.

Q. Can I configure fine-grained authorization in AWS IoT Core?

Yes. Similar to other AWS services, in AWS IoT Core you have fine-grained control over the set of API actions each identity is authorized to invoke. In addition, you have fine-grained control over the pub/sub topics that an identity can publish or subscribe to, as well as over the devices and the Device Shadow in the Registry that an identity can access.

Q. Where can I learn more about Security and Access Control in AWS IoT Core?

For more information, see AWS IoT Core Security and Identity.

Q. What is Just-in-time registration of certificates?

Just-in-time registration (JITR) of device certificates expands on the "Use Your Own Certificate" feature launched in April 2016 by simplifying the process of enrolling devices with AWS IoT Core. Prior to support for JITR, the device enrollment process required two steps: first, registering the Certificate Authority (CA) certificate to AWS IoT Core, then individually registering the device certificates that were signed by the CA. Now, with JITR you can complete the second step by auto-registering device certificates when devices connect to AWS IoT Core for the first time. This saves time spent on registering device certificates and allows devices to remain off-line during the manufacturing process. To further automate IoT device provisioning, you can create an AWS IoT Core rule with a Lambda action that activates the certificates and attaches policies. For more information, visit the Internet of Things Blog on AWS or Developer Documentation.

Q. What is the AWS IoT Device SDK?

The AWS IoT Device SDKs simplify and accelerate the development of code running on connected devices (micro-controllers, sensors, actuators, smart appliances, wearable devices, etc.). First, devices can optimize the memory, power, and network bandwidth consumption by using the Device SDKs. At the same time, Device SDKs enable highly secure, low-latency, and low-overhead communication with built-in TLS, WebSockets, and MQTT support. The Device SDKs also accelerate IoT application development by supporting higher level abstractions such as synchronizing the state of a device with its Device Shadow in AWS IoT Core.

AWS IoT Device SDKs are freely available as open-source projects. For more details visit our Device SDK page.

Q. Which programming languages and hardware platforms does the AWS IoT Device SDK support?

AWS currently offers the AWS IoT Device SDKs for C and Node.js languages, as well as for the Arduino Yún platform.

In addition, several hardware manufacturers have partnered with AWS to make the AWS IoT Device SDKs available on their respective platforms. You can find out more about the hardware platforms on our Getting Started page.

Lastly, AWS IoT Device SDKs are open-source. You can port them to the languages and hardware platforms of your choice if they are not supported already.

Q: Should I use AWS IoT Device SDK or the AWS SDKs?

The AWS IoT Device SDK complements the AWS SDKs. IoT projects often involve code running on micro-controllers and other resource-constrained devices. However, IoT projects often include application running in the cloud and on mobile devices that interact with the micro-controllers/resource-constrained devices. AWS IoT Device SDKs are designed to be used on the micro-controllers/resource-constrained devices, while the AWS SDKs are designed for cloud and mobile applications.

For more information on the AWS IoT Device SDKs, see AWS Device SDKs.

Q. Is AWS IoT Core available in AWS Free Tier?

Yes. Please visit our pricing page for more information.

Q. How much does AWS IoT Core cost?

Please visit our pricing page for information.

Q. What is the AWS IoT Core SLA?

The AWS IoT Core SLA stipulates that you may be eligible for a credit towards a portion of your monthly service fees if AWS IoT Core fails to achieve a Monthly Uptime Percentage of at least 99.9% for AWS IoT Core.

For full details on all of the terms and conditions of the SLA, as well as details on how to submit a claim, please see the AWS IoT Core SLA details page.

Q. Why should I use the AVS Integration for AWS IoT?

Until now, producing an Alexa Built-in device required on-device memory and compute to be at least 50MB RAM and ARM Cortex 'A' class microprocessors, increasing the engineering bill of materials (eBOM) and MSRP. Additionally, retrieving, buffering, decoding, and mixing audio on devices can be complex and time consuming. The high production cost and complexity makes it difficult for device makers to quickly go to market with differentiated, voice-enabled experiences on resource-constrained IoT devices.

The AVS Integration lowers the Alexa Built-in cost by up to 50% by offloading compute & memory intensive workloads to the cloud. This reduces the hardware requirements of integrating AVS on a device from 50MB to 1MB of RAM and from ARM Cortex ‘A’ class microprocessors to ARM Cortex ‘M’ class microcontrollers and brings Alexa to ubiquitous products such as light switches, thermostats, and small appliances. In addition to the simplified device side AVS integration, device makers using the AVS Integration also have access to simple and cost-effective ongoing device maintenance and management, enhanced device security, and analytics services via the scalable AWS IoT Core infrastructure.

With new categories of low-price Alexa Built-in devices available on the market, end users can now experience Alexa in new parts of their home, office, or hotel rooms for a truly ambient experience, where they talk directly to their surroundings rather than to an Alexa Family of Devices.

Q. How do I use the Alexa Voice Service (AVS) Integration?

Learn how to create low-cost Alexa Built-in devices with the AVS Integration for AWS IoT Core Getting Started Guide.

Q. How is AVS for IoT different from traditional AVS?

AVS Integration supports Device arbitration, Dialog, Multi-turn dialog, Timers, Alarms, Reminders, Flash Briefing, Routines, Alexa Announce, eBooks, and Skills. It does not support high-quality music playback, Whole Home Audio, Alexa Calling, Spotify, Bluetooth, and rich multi-modal displays.

Q. What types of devices can I build with AVS?

The AVS Integration for AWS IoT is a great solution for device makers producing low-cost, resource-constrained devices (including light switches, light bulbs, home hubs, home appliances and more) that want to allow their customers to talk to these products directly with the wake word “Alexa,” and receive voice responses and content instantly. These devices will have Built-in microphones and speakers that are capable of playing back dialog, alerts, and the news but are not adequate to support high-quality music playback. Device makers who want full music playback with richer Alexa Music Playback capabilities such as high fidelity (>128kbps) music streaming, Spotify, synchronized music streaming over multiple speakers, should continue to build these devices using the existing Alexa Built-in solutions.

Q. Do I get the Alexa Built-in badge using the AVS Integration for AWS IoT Core?

Similar to other Alexa Built in products, products built with the AVS Integration will need to pass the Alexa Voice Service product certification process comprising of Amazon-managed testing of security, acoustic performance, user experience, and functional testing to earn the Amazon Certified Alexa Built-in badge.

Q. What AWS regions will AVS Integration for IoT Core be available in at launch?

The AVS Integration for AWS IoT Core is available in all AWS regions where AWS IoT Core is available other than China (Beijing and Ningxia), Asia Pacific (Hong Kong) and Middle East (Bahrain). See the AWS Region Table for the current list of regions for AWS IoT Core.

AWS IoT 1-Click FAQs

Q. What is AWS IoT 1-Click?

AWS IoT 1-Click is a new service that makes it easy for simple devices to trigger AWS Lambda functions for required actions. With AWS IoT 1-Click, simple devices are ready to securely connect to AWS IoT Core right out of the box. This makes it easy for developers to deploy these devices into their applications without the need for creating, installing and managing certificates. AWS IoT 1-Click provides device manufacturers with the tools and APIs needed to provision certificates at the time of manufacturing and register the devices with AWS IoT.

Q. What are simple devices?

Simple devices are Internet connected devices that are connected through different methods including Wi-Fi, LTE-M, NB-IoT, SigFox etc. These devices have a single purpose that can be associated with actions such as AWS Lambda functions. This action can be triggered in a simple and scalable manner with AWS IoT 1-Click. Examples of simple devices include but are not limited to:

Button-like devices: These devices can be clicked to trigger particular actions.

Asset Trackers: Trackers on containers in warehouses or on trucks transporting materials enable location tracking for instant traceability and notification.

Temperature Sensors: Actions such as temperature control could be triggered based on temperatures reaching predefined thresholds, based on customer requirements.

Card readers: Devices to track entry/exit of authorized personnel into offices, laboratories, or factories as part of workplace procedures.

Q. Which simple devices are supported with AWS IoT 1-Click?

During preview, AWS IoT 1-Click will support the following simple devices:

- AWS IoT Enterprise button with Wi-Fi connectivity
- AT&T LTE-M button with AT&T LTE-M network connectivity

Both these devices will be pre-provisioned with the appropriate credentials at the time of manufacturing and can securely connect to AWS IoT Core right out of the box.

Q. Can I use my existing AWS IoT buttons for the AWS IoT 1-Click?

No, your existing AWS IoT buttons (both the 1st generation and the 2nd generation) will not be supported with AWS IoT 1-Click.

Q. In which regions will the AWS IoT 1-Click Preview be available?

AWS IoT 1-Click is available in the following AWS regions:

US West (Oregon), US East (N. Virginia), US East (Ohio), EU (Ireland), EU (Frankfurt), EU (London), Asia Pacific (Tokyo).

Q. How do I get started with AWS IoT 1-Click?

Visit our documentation to see the AWS IoT 1-Click getting started guide. This document has the details on how you can get started with AWS IoT 1-Click.

Q. What are the features within AWS IoT 1-Click?

AWS IoT 1-Click enables the following:

Choice of simple devices: AWS IoT 1-Click supports ready-to-use simple devices that will be ready to use right out of the box. Currently, we support the AWS IoT Enterprise Button and the AT&T LTE-M Button. Going forward, we are targeting support for different types of devices that connect over different methods.

Secure connectivity: Device manufacturers will be able to build simple devices with appropriate credentials to connect to AWS IoT Core at the time of manufacturing, thus enabling secure connectivity to the AWS cloud.

Grouping with contextual data: With AWS IoT 1-Click, customers will be able to group their devices into projects based on their requirements including usage and location. Individual device placements within their projects can be defined, with user-specified contextual data.

Choice of actions: AWS IoT 1-Click will enable customers to choose from a set of pre-defined Lambda actions including sending email and SMS and associate them to devices.Alternatively, customers can choose to use their own custom-defined Lambda functions, from their AWS accounts.

Usage and Status reports: Customers will be able to derive reports detailing usage and status of their deployed devices, using AWS IoT 1-Click. Reports include pre-defined reports for device activity such as number of clicks and device health such as remaining life on battery-operated devices or custom reports that can be generated using Amazon CloudWatch.

AWS IoT 1-Click will be available through the AWS Management Console and the AWS IoT 1-Click mobile app on the iOS and Android platforms. In addition, customers will be able to use AWS IoT 1-Click features through a number of public APIs.

Q. I am a device manufacturer and would like my devices to be supported with AWS IoT 1-Click. What needs to be done for these devices to be supported?

Please contact us at iot1click@amazon.com with details of your devices and we can work with you on your request.

Q. I was part of the AWS IoT 1-Click Preview Program. Will my buttons work with the General Availability of AWS IoT 1-Click?

Yes. However, you would need to recreate your projects and re-define your contextual data for devices in your account. You will also need to upgrade to the launch version of the AWS IoT 1-Click mobile app. There is no need to whitelist your account either for the console or the mobile app.

Q. I was part of the AWS IoT Button Enterprise Program. Will my buttons work on with AWS IoT 1-Click?

No, the AWS IoT Button Enterprise Program has been retired. The buttons that were supported with that program will not work with AWS IoT 1-Click.

Q. Will I be able to get branded custom labels on the AWS IoT Enterprise Buttons with AWS IoT 1-Click?

No, the AWS IoT 1-Click service does not include branded custom labels. The buttons come with standard AWS IoT Enterprise Button labels.

Q. Where can I find the list of all supported devices on AWS IoT 1-Click?

Please visit https://aws.amazon.com/iot-1-click/devices to view all supported devices on AWS IoT 1-Click.

Q. My button does not work in spite of being connected to the Internet?

Please visit our documentation to troubleshoot the problem. The LED colors on the button should help you to determine if the problem is with the physical device or the network. If the problem persists, please contact AWS support.

AWS IoT Analytics FAQs

Q. What is AWS IoT Analytics? >>

Q. How does AWS IoT Analytics work? >>

Q. Can I execute my custom analysis code on AWS IoT Analytics? >>

Q. How is a SQL data set different than a container data set? >>

Q. What are DeltaTime Windows? >>

Q. How do I execute my custom code container on AWS IoT Analytics at my preferred schedule? >>

Q. What retention policies do I've on my Data Stores and Channels? >>

Q. What type of message formats are supported with AWS IoT Analytics? >>

Q. Can I re-process my data from Channel to a Pipeline? >>

Q. How do I get the data into AWS IoT Analytics using the Ingestion API? >>

Q. Can I preview my messages in the Channel? >>

Q. Can I simulate my pipeline activity? >>

Q. What are the differences between AWS IoT Analytics and Amazon Kinesis Analytics? >>

Q. When do I use AWS IoT Analytics and when do I use Kinesis Analytics? >>

Q. When do I use AWS IoT Analytics and Amazon Kinesis together? >>

Q. When working with IoT data, when should I use AWS IoT Analytics vs. Amazon Kinesis Streams, Amazon Kinesis Analytics, and Amazon Kinesis Firehose? >>

Q. When do I use AWS IoT Analytics and when do I use Amazon Kinesis Video Streams? >>

AWS IoT Button FAQs

Getting started

Q. Are the buttons available outside of the US?

Yes, in addition to the United States, the buttons are available in the United Kingdom, Germany, France, Italy and Spain.

Q. I would like to learn about AWS IoT using the AWS IoT button. Is there a step- by-step tutorial?

Download the AWS IoT Button Dev app from the iOS or Google Play app stores to learn more.

Q. What more can I do with the AWS IoT Button?

You can use the button to count items, track usage, call or text and alert someone, start and stop a process or Internet-connected device. To count and track usage, you can store the clicks in Amazon DynamoDB. Follow this tutorial to create a DynamoDB rule using the AWS IoT rules engine. You can alert someone using Amazon SNS. Follow this tutorial to create an SNS rule using the AWS IoT rules engine. You can start and stop a process, call an external API endpoint, or connect to an Internet-connected device using AWS Lambda. Follow this tutorial to create a Lambda rule using the AWS IoT rules engine.

Q. How do I configure the button to use at home or office Wi-Fi?

Press and hold the button for five seconds until the LED starts flashing blue.

Use your phone or computer to connect to the Button ConfigureMe - XXXX Wi-Fi network. Use the last 8 digits of the serial number of your device as the WPA2-PSK password.

In a browser, navigate to http://192.168.0.1/index.html.

Select the network (SSID) you want the button to connect to and enter its password.

Please use the AWS IoT Button Dev app for iOS or Android

Q. Will the AWS IoT Button work anywhere?

The button will work wherever there is Wi-Fi (2.4 Ghz). It is designed to work in a home or office environment, but is not ideal for spaces with a congested Wi-Fi spectrum like expo halls, lunch rooms, keynote rooms, or hallways. If it can connect to the Internet, it will work.

What's in the payload sent from the device when you press it?

The payload contains the device serial number, the measured battery voltage, and a click type.

The following JSON template shows what is sent in the payload.

JSON

```json
{
"serialNumber": "GXXXXXXXXXXXXXXXXX",
"batteryVoltage": "mV",
"clickType": "SINGLE | DOUBLE | LONG"
}
```

A LONG clickType is sent if the first press lasts longer than 1.5 seconds. SINGLE and DOUBLE clickType payloads are sent for short clicks.

Q. What do the LED color patterns mean?

Color	Status
Blinking White	Connecting to Wi-Fi, getting IP address, connecting to AWS IoT.
Solid Green	Successfully connected to Wi-Fi and published a message to AWS IoT.
Blinking Blue	Soft AP mode.
Solid Orange	No Wi-Fi configured.

Red	Error. (See the Troubleshooting Guide for more information.)

Q. How long will the battery last?

The battery on the 2nd Gen AWS IoT Button will last for approximately 2,000 clicks. When the device battery runs out of charge, there is no way to recharge or replace the battery.

Q. Do you have examples of Lambda functions I can use?

You can use any of the blueprints in the AWS Lambda console.

Q. Do I need an AWS account as well as an Amazon.com account to use this device?

Yes. You need an AWS account to set up this device. An AWS account is different from an Amazon.com account. You need an Amazon.com account to purchase the button.

Q. How much is it going to cost?

With AWS, you pay only for the services you use, for as long as you use them, and with no long-term contracts. The button will use the AWS IoT service and any other service you consume. If you qualify for the free tier, there will be no additional charge. Otherwise, each service has separate pricing. You will be charged separately based on your usage.

Q. I have built a unique IoT project. Where can I post it?

We have worked with Hackster and they have created an AWS IoT Button page on Hackster.io, hackster.io/amazonwebservices/products/aws-iot-button. Feel free to post your project with details on how others can build what you have built.

Q. What's the difference between an AWS IoT Button and an AWS IoT 1-Click compatible device?

The AWS IoT Button connects to AWS IoT Core and can be configured using the “AWS IoT Button Dev” mobile app. AWS IoT 1-Click is a new service that supports multiple devices including the AWS IoT Enterprise Button which can be configured using the AWS IoT 1-Click Mobile app. For information about the AWS IoT 1-Click service and devices it supports: AWS IoT 1-Click

Troubleshooting

Q. I am having trouble provisioning the certificate and private key on my button. What should I do?

In the AWS Lambda Event Source wizard, if you are using Firefox, Chrome, or Opera, click the link to download the certificate and private key. If you are using Safari or Internet Explorer, right-click and choose Save As to download the certificate and the private key files to your computer.

For security reasons, AWS does not store the AWS IoT certificate and private key. If you do not download them to your computer, they will be lost. You can create a new certificate and private key or follow the steps in the AWS Lambda Event Source wizard if the current certificate does not work.

Q. I followed the blueprint wizard, but when I press the button, I get a green light. I don't get an email.

It's possible that your Lambda function does not have the right permissions to create resources in SNS. To confirm, go to the CloudWatch logs of the Lambda function you just created with the wizard. Add the required SNS permissions (sns:createTopic, sns:Publish, sns:subscribe, sns:listSubscriptions) to your execution role of the function.

In other cases, when you are not using the wizard, whenever you get a green light and nothing happens, follow these steps for debugging:

Test using the MQTT Client in the AWS IoT console. Subscribe to the "iotbutton/+" topic and see whether you receive a message when clicked.

In the AWS IoT console, under Settings, enable CloudWatch Logs, and then examine "AWSIoTLogs" in the Amazon CloudWatch console to see if the rule was fired.

For other services like AWS Lambda, examine the logs specific to the function in the CloudWatch logs to find the error.

Q. My button cannot connect and blinks a red pattern.

This indicates an error has occurred. Use this blinking sequence guide to troubleshoot the issue:

Blinking Pattern	Error
Short short short	There was an error connecting to the configured wireless network.
Short short long	There was an error obtaining an IP address from the network.
Short long short	There was an error performing the host name lookup. This can be the result of not being able to reach the DNS server or an incorrectly configured AWS IoT endpoint subdomain.
Short long long	Cannot connect to AWS IoT. This can be an issue with the network, but is most likely not an issue with the certificates.
Long short short	Cannot establish a secure connection with the server. This error is most likely due to an unknown or inactive certificate.
Long short long	Received HTTP 403 Forbidden. This can happen if your button's certificate is deactivated or expired.
Long long short	There is a problem sending to or receiving from AWS IoT. This is most likely just a networking error.
Long long long	Received an unknown HTTP response from AWS IoT.
Solid red	A fatal internal error occurred. Your only option is to retry.

Q. I am having trouble getting the device into setup mode. I have pressed the button for 5+ seconds but don’t see flashing blue.

If this happens, release the button, wait a few seconds and then try again. If you still have no luck, short press the button and verify you see an orange light. If you don’t see an orange light, get in touch with customer services by creating a support case here.

Q. I am an IAM user with restricted access.

You must provide iot:* permissions in your IAM user policy. I am having problems not addressed in this guide. Don't worry. We are here for you. If you cannot connect the button to AWS, click here to create a support case. The AWS Customer Service team will contact you. Feel free to post your technical service-related questions on the AWS IoT forum or AWS Lambda forum.

AWS IoT Core FAQs

Q. What is AWS IoT Core?

AWS IoT Core is a managed cloud platform that lets connected devices easily and securely interact with cloud applications and other devices. AWS IoT Core can support billions of devices and trillions of messages, and can process and route those messages to AWS endpoints and to other devices reliably and securely. With AWS IoT Core, your applications can keep track of and communicate with all your devices, all the time, even when they aren't connected.

Q. What does AWS IoT Core offer?

Connectivity between devices and the AWS cloud. First, with AWS IoT Core you can communicate with connected devices securely, with low latency and with low overhead. The communication can scale to as many devices as you want. AWS IoT Core supports standard communication protocols (HTTP, MQTT, and WebSockets are supported currently). Communication is secured using TLS.

Connectivity between devices and the AWS cloud. First, with AWS IoT Core you can communicate with connected devices securely, with low latency and with low overhead. The communication can scale to as many devices as you want. AWS IoT Core supports standard communication protocols (HTTP, MQTT, and WebSockets are supported currently). Communication is secured using TLS.

Processing data sent from connected devices. Secondly, with AWS IoT Core you can continuously ingest, filter, transform, and route the data streamed from connected devices. You can take actions based on the data and route it for further processing and analytics.

Application interaction with connected devices. Finally, AWS IoT Core accelerates IoT application development. It serves as an easy to use interface for applications running in the cloud and on mobile devices to access data sent from connected devices, and send data and commands back to the devices.

Q. How does AWS IoT Core work?

Connected devices, such as sensors, actuators, embedded devices, smart appliances, and wearable devices, connect to AWS IoT Core over HTTPS, WebSockets, or secure MQTT.

Included in AWS IoT Core is a **Device Gateway** that allows secure, low-latency, low-overhead, bi-directional communication between connected devices and your cloud and mobile applications.

AWS IoT Core also contains a **Rules Engine** which enables continuous processing of data sent by connected devices. You can configure rules to filter and transform the data. You also configure rules to route the data to other AWS services such as DynamoDB, Kinesis, Lambda, SNS, SQS, CloudWatch, Elasticsearch Service with built-in Kibana integration, as well as to non-AWS services, via Lambda for further processing, storage, or analytics.

There is also a **Registry** where you can register and keep track of devices connected to AWS IoT Core, or devices that may connect in the future. The **Device Shadow** in AWS IoT Core enables cloud and mobile applications to query data sent from devices and send commands to devices, using a simple REST API, while letting AWS IoT Core handle the underlying communication with the devices. The Device Shadow accelerates application development by providing a uniform interface to devices, even when they use one of the several IoT communication and security protocols with which the applications may not be compatible. The Device Shadow also accelerates application development by providing an always available interface to devices even when the connected devices are constrained by intermittent connectivity, limited bandwidth, limited computing ability or limited power.

Communication with AWS IoT Core is secure. The service requires all of its clients (connected devices, server applications, mobile applications, or human users) to use strong authentication (X.509 certificates, AWS IAM credentials, or 3rd party authentication via AWS Cognito). All communication is encrypted. AWS IoT Core also offers fine-grained authorization to isolate and secure communication among authenticated clients.

Q. What is 2lemetry and how does it relate to AWS IoT?

2lemetry was acquired by AWS in 2015, and their capabilities provided foundational elements such as the MQTT Message Broker and the Rules Engine for AWS IoT Core.

Q. In which regions is AWS IoT Core available?

See the AWS Region Table for the current list of regions for AWS IoT Core.

You can use AWS IoT Core regardless of your geographic location, as long as you have access to one of the above AWS regions.

Q. How do I get started with using AWS IoT Core?

Use the AWS IoT Console or refer to the Quickstart section of our developer guide to test drive the AWS IoT Core in minutes.

Also, take a look at the AWS-powered Starter Kits provided by our partners.

Refer to the AWS IoT Core documentation for further details.

Q. Which languages does the AWS IoT Console support?

The AWS IoT Console supports English, French, Japanese, Korean, Simplified Chinese, German, Portuguese, Spanish, Italian and Traditional Chinese.

Q. How can I switch the console's language?

Click on the language at the bottom left corner of the console to pick the language. The language selection will persist throughout the consoles of different AWS services.

Q. What are the ways for accessing AWS IoT Core?

You can use the AWS Management Console, the AWS SDKs, the AWS CLI, and the AWS IoT Core APIs. Connected devices can use the AWS IoT Device SDKs to simplify the communication with AWS IoT Core.

The AWS IoT Core APIs and commands are largely divided into control plane operations and data plane operations. The control plane operations enable you to do tasks such as configuring security, registering devices, configuring rules for routing data, and setting up logging. The data plane operations enable you to ingest data from connected devices into AWS IoT Core with low latency and high throughput rate at a large scale.

Q. What communication and authentication protocols does AWS IoT Core support?

For control plane operations, AWS IoT Core supports HTTPS. For data plane operations, AWS IoT Core supports HTTPS, WebSockets, and secure MQTT – a protocol often used in IoT scenarios.

HTTPS and WebSockets requests sent to AWS IoT Core are authenticated using AWS IAM or AWS Cognito, both of which support the AWS SigV4 authentication. If you are using the AWS SDKs or the AWS CLI, the SigV4 authentication is taken care of for you under the hood. HTTPS requests can also be authenticated using X.509 certificates. MQTT messages to AWS IoT Core are authenticated using X.509 certificates.

With AWS IoT Core you can use AWS IoT Core generated certificates, as well as those signed by your preferred Certificate Authority (CA).

Q. Can devices that are NOT directly connected to the Internet access AWS IoT Core?

Yes, via a physical hub. Devices connected to a private IP network and devices using non-IP radio protocols such as ZigBee or Bluetooth LE can access AWS IoT Core as long as they have a physical hub as an intermediary between them and AWS IoT Core for communication and security.

Q. How should applications access AWS IoT Core?

Applications connecting to AWS IoT Core largely fall in two categories: 1. companion apps and 2. server applications. Companion apps are mobile or client-side browser applications that interact with connected devices via the cloud. A mobile app that lets a consumer remotely unlock a smart lock in the consumer's house is an example of a companion app. Server applications are designed to monitor and control a large number of connected devices at once. An example of a server application would be a fleet management website that plots thousands of trucks on a map in real-time.

AWS IoT Core enables both companion apps and server applications to access connected devices via uniform, RESTful APIs. Applications also have the option to use pub/sub to communicate directly with the connected devices.

Typically the companion apps would authenticate using end-user identities which are managed either by your own identity store or a third party identity provider such as Facebook and Login with Amazon. For companion apps, use Amazon Cognito, which integrates with several identity providers. Cognito identities can be authorized to access AWS IoT Core, and their access can be restricted only to the resources relevant to them. For example, as a connected washing machine manufacturer, you can authorize your consumers to access your AWS IoT Core information pertaining only to their individual washing machines.

Server applications (such as a mapping application running on Amazon EC2) can use IAM roles to access AWS IoT Core.

Q. Can I get a history of AWS IoT Core API calls made on my account for security analysis and operational troubleshooting purposes?

Yes, to receive a history of AWS IoT Core API calls made on your account, you simply turn on CloudTrail in the AWS Management Console.

Q. How do I send feedback?

To send feedback, click on the “Feedback” link in the footer bar of the console.

Q: What is the Device Gateway?

The Device Gateway forms the backbone of communication between connected devices and the cloud capabilities such as the Rules Engine, Device Shadow, and other AWS and 3rd-party services.

The Device Gateway supports the pub/sub messaging pattern, which enables scalable, low-latency, and low-overhead communication. It is particularly useful for IoT scenarios where billions of devices are expected to communicate frequently and with minimal delay. Pub/sub involves clients publishing messages on logical communication channels called ‘topics’ and clients subscribing to topics to receive messages. The device gateway enables the communication between publishers and subscribers. Traditionally, organizations have had to provision, operate, scale, and maintain their own servers as device gateways to take advantage of pub/sub. AWS IoT Core has eliminated this barrier by providing the Device Gateway.

The Device Gateway scales automatically with your usage, without any operational overhead for you. AWS IoT Core supports secure communication with the device gateway, AWS-account level isolation, as well as fine-grained authorization within an AWS account. The device gateway currently supports publish and subscribe over secure MQTT and WebSockets, as well as publish over HTTPS.

Q. What is MQTT?

MQTT is a lightweight pub/sub protocol, designed to minimize network bandwidth and device resource requirements. MQTT also supports secure communication using TLS. MQTT is often used in IoT use cases. MQTT v3.1.1 is an OASIS standard, and the Device Gateway supports most of the MQTT specification.

Q. What is the Rules Engine?

The Rules Engine enables continuous processing of inbound data from devices connected to AWS IoT Core. You can configure rules in the Rules Engine in an intuitive, SQL-like syntax to automatically filter and transform inbound data. You can further configure rules to route data from AWS IoT Core to several other AWS services as well as your own or 3rd party services.

Here are just a few example use cases of rules:

- Filtering and transforming incoming messages and storing them as time series data in DynamoDB.
- Sending a push notification via SNS when the data from a sensor crosses a certain threshold.
- Saving a firmware file to S3
- Processing messages simultaneously from a multitude of devices using Kinesis
- Invoke Lambda to do custom processing on incoming data
- Sending a command to a group of devices with an automated republish

Q. How are the rules defined and triggered?

An AWS IoT Core rule consists of two main parts:

A SQL statement that specifies the pub/sub topics to apply the rule on, data transformation to perform, if any, and the condition under which the rule should be executed. The rule is applied on every message published on the specified topics.

An actions list that defines the actions to take when the rule is executed, that is, when an incoming message matches the condition specified in the rule.

Rule definitions use a JSON-based schema. You can directly edit the JSON or use the rules editor in the AWS Management Console.

As an example, here is a rule for saving temperature data from a sensor to DynamoDB whenever the temperature is above 50:

JSON

```json
{
    "sql": "SELECT * from 'iot/tempSensors/#' WHERE temp > 50",
    "description": "Rule to save sensor data when temperature is a
    "actions": [
     {
            "dynamoDB": {
            "tableName": "HighTempTable",
            "roleArn": "arn:aws:iam::your-aws-account-id:role/dyna
            "hashKeyField": "key",
            "hashKeyValue": "${topic(3)}",
            "rangeKeyField": "timestamp",
            "rangeKeyValue": "${timestamp()}"
            }
        }
    ]
}
```

Sensors in this example are publishing on their topics under “iot/tempSensors/”. The first line of the rule defines the SQL SELECT statement used to query on the “iot/tempSensors/#” topic. It contains a WHERE clause that extracts the value of a ‘temp’ field in the message’s payload and checks if it passes the condition ‘greater than 50’. If the condition is met, the data is stored in the specified DynamoDB table. The example uses built-in functions for tasks such as traversing the message payload and getting current time.

Q. Where can I learn more about rules?

You can learn more about rule here Core Rules documentation

Q. What is the Registry and what should I use it for?

IoT scenarios can range from a small number of mission-critical devices to large fleets of devices. The Registry allows you to organize and track those devices. You can maintain a logical handle in the Registry for every device you are connecting to AWS IoT Core. Each device in the Registry can be uniquely identified and can have metadata such as model numbers, support contact, and certificates associated with it. You can search for connected devices in the Registry based on the metadata.

Q. What is a Thing Type?

Thing Types allow you to effectively manage your catalogue of devices by defining common characteristics for devices that belong to the same device category. In addition, a Thing associated with a Thing Type can now have up to 50 attributes including 3 searchable attributes.

Q. What is Simplified Permission Management?

This feature allows you to easily manage permission policies for a large number of devices by using variables that reference Registry or X.509 certificate properties. The integration of Registry and Certificate properties with device policies offers the benefits listed below:

- You can now reference Registry properties in device permission policies. Referencing device properties defined in the Registry allows your policies to reflect any changes made in the Registry. For example, by referencing the Thing Attribute named “building-address” as a variable in the policy, devices will automatically inherit a new set of permissions when they move buildings.
- You can share a single generic policy for multiple devices. A generic policy can be shared among the same category of devices instead of creating a unique policy per device. For example, a policy that references the “serial-number” as a variable, can be attached to all the devices of the same model. When devices of the same serial number connect, policy variables will be automatically substituted by their serial-number.

Q. What is the Device Shadow?

The Device Shadow enables cloud and mobile applications to easily interact with the connected devices registered in AWS IoT Core. The Device Shadow in AWS IoT Core contains properties of a connected device. You can define any set of properties applicable to your use case. For example, for a smart light bulb, you might define ‘on-or-off’, ‘color’, and ‘brightness’ as the properties. The connected device is expected to report the actual values of those properties, which are stored in the Device Shadow. Applications get and update the properties simply by using a RESTful API provided by AWS IoT Core. AWS IoT Core and the Device SDKs take care of synchronizing property values between the connected device and its Device Shadow in AWS IoT Core.

Q. Do I have to use the Registry and the Device Shadow?

You can have applications communicate directly to the connected devices using the Device Gateway and/or the Rules Engine in AWS IoT Core. However, we recommend using the Registry and the Device Shadow since they offer richer and more structured development and management experience that lets you focus on the unique value you want to create for your customers rather than having to focus on the underlying communication and synchronization between the connected devices and the cloud.

Q. What is the lifecycle of a device and its Device Shadow in AWS IoT Core?

- You register a device (such as a light bulb) in the Registry.
- You program connected device to publish a set of its property values or ‘state (“I am ON and my color is RED”) to the AWS IoT Core service.
- The last reported state is stored in the Device Shadow in AWS IoT Core.
- An application (such as a mobile app controlling the light bulb) uses a RESTful API to query AWS IoT Core for the last reported state of the light bulb, without the complexity of communicating directly with the light bulb.

- When a user wants to change the state (such as turning the light bulb from ON to OFF), the application uses a RESTful API to request an update, i.e. sets a ‘desired’ state for the device in AWS IoT Core. AWS IoT Core takes care of synchronizing the desired state to the device.
- The application gets notified when the connected device updates its state to the desired state.

Q. Where can I learn more about the Registry and the Device Shadow?

For more information on the Registry, see the Registry documentation. For more information on the Device Shadow, see the Device Shadow documentation.

Q. Can I configure fine-grained authorization in AWS IoT Core?

Yes. Similar to other AWS services, in AWS IoT Core you have fine-grained control over the set of API actions each identity is authorized to invoke. In addition, you have fine-grained control over the pub/sub topics that an identity can publish or subscribe to, as well as over the devices and the Device Shadow in the Registry that an identity can access.

Q. Where can I learn more about Security and Access Control in AWS IoT Core?

For more information, see AWS IoT Core Security and Identity.

Q. What is Just-in-time registration of certificates?

Just-in-time registration (JITR) of device certificates expands on the "Use Your Own Certificate" feature launched in April 2016 by simplifying the process of enrolling devices with AWS IoT Core. Prior to support for JITR, the device enrollment process required two steps: first, registering the Certificate Authority (CA) certificate to AWS IoT Core, then individually registering the device certificates that were signed by the CA. Now, with JITR you can complete the second step by auto-registering device certificates when devices connect to AWS IoT Core for the first time. This saves time spent on registering device certificates and allows devices to remain off-line during the manufacturing process. To further automate IoT device provisioning, you can create an AWS IoT Core rule with a Lambda action that activates the certificates and attaches policies. For more information, visit the Internet of Things Blog on AWS or Developer Documentation.

Q. What is the AWS IoT Device SDK?

The AWS IoT Device SDKs simplify and accelerate the development of code running on connected devices (micro-controllers, sensors, actuators, smart appliances, wearable devices, etc.). First, devices can optimize the memory, power, and network bandwidth consumption by using the Device SDKs. At the same time, Device SDKs enable highly secure, low-latency, and low-overhead communication with built-in TLS, WebSockets, and MQTT support. The Device SDKs also accelerate IoT application development by supporting higher level abstractions such as synchronizing the state of a device with its Device Shadow in AWS IoT Core.

AWS IoT Device SDKs are freely available as open-source projects. For more details visit our Device SDK page.

Q. Which programming languages and hardware platforms does the AWS IoT Device SDK support?

AWS currently offers the AWS IoT Device SDKs for C and Node.js languages, as well as for the Arduino Yún platform.

In addition, several hardware manufacturers have partnered with AWS to make the AWS IoT Device SDKs available on their respective platforms. You can find out more about the hardware platforms on our Getting Started page.

Lastly, AWS IoT Device SDKs are open-source. You can port them to the languages and hardware platforms of your choice if they are not supported already.

Q: Should I use AWS IoT Device SDK or the AWS SDKs?

The AWS IoT Device SDK complements the AWS SDKs. IoT projects often involve code running on micro-controllers and other resource-constrained devices. However, IoT projects often include application running in the cloud and on mobile devices that interact with the micro-controllers/resource-constrained devices. AWS IoT Device SDKs are designed to be used on the micro-controllers/resource-constrained devices, while the AWS SDKs are designed for cloud and mobile applications.

For more information on the AWS IoT Device SDKs, see AWS Device SDKs.

Q. Is AWS IoT Core available in AWS Free Tier?

Yes. Please visit our pricing page for more information.

Q. How much does AWS IoT Core cost?

Please visit our pricing page for information.

Q. What is the AWS IoT Core SLA?

The AWS IoT Core SLA stipulates that you may be eligible for a credit towards a portion of your monthly service fees if AWS IoT Core fails to achieve a Monthly Uptime Percentage of at least 99.9% for AWS IoT Core.

For full details on all of the terms and conditions of the SLA, as well as details on how to submit a claim, please see the AWS IoT Core SLA details page.

Q. Why should I use the AVS Integration for AWS IoT?

Until now, producing an Alexa Built-in device required on-device memory and compute to be at least 50MB RAM and ARM Cortex 'A' class microprocessors, increasing the engineering bill of materials (eBOM) and MSRP. Additionally, retrieving, buffering, decoding, and mixing audio on devices can be complex and time consuming. The high production cost and complexity makes it difficult for device makers to quickly go to market with differentiated, voice-enabled experiences on resource-constrained IoT devices.

The AVS Integration lowers the Alexa Built-in cost by up to 50% by offloading compute & memory intensive workloads to the cloud. This reduces the hardware requirements of integrating AVS on a device from 50MB to 1MB of RAM and from ARM Cortex ‘A’ class microprocessors to ARM Cortex ‘M’ class microcontrollers and brings Alexa to ubiquitous products such as light switches, thermostats, and small appliances. In addition to the simplified device side AVS integration, device makers using the AVS Integration also have access to simple and cost-effective ongoing device maintenance and management, enhanced device security, and analytics services via the scalable AWS IoT Core infrastructure.

With new categories of low-price Alexa Built-in devices available on the market, end users can now experience Alexa in new parts of their home, office, or hotel rooms for a truly ambient experience, where they talk directly to their surroundings rather than to an Alexa Family of Devices.

Q. How do I use the Alexa Voice Service (AVS) Integration?

Learn how to create low-cost Alexa Built-in devices with the AVS Integration for AWS IoT Core Getting Started Guide.

Q. How is AVS for IoT different from traditional AVS?

AVS Integration supports Device arbitration, Dialog, Multi-turn dialog, Timers, Alarms, Reminders, Flash Briefing, Routines, Alexa Announce, eBooks, and Skills. It does not support high-quality music playback, Whole Home Audio, Alexa Calling, Spotify, Bluetooth, and rich multi-modal displays.

Q. What types of devices can I build with AVS?

The AVS Integration for AWS IoT is a great solution for device makers producing low-cost, resource-constrained devices (including light switches, light bulbs, home hubs, home appliances and more) that want to allow their customers to talk to these products directly with the wake word "Alexa," and receive voice responses and content instantly. These devices will have Built-in microphones and speakers that are capable of playing back dialog, alerts, and the news but are not adequate to support high-quality music playback. Device makers who want full music playback with richer Alexa Music Playback capabilities such as high fidelity (>128kbps) music streaming, Spotify, synchronized music streaming over multiple speakers, should continue to build these devices using the existing Alexa Built-in solutions.

Q. Do I get the Alexa Built-in badge using the AVS Integration for AWS IoT Core?

Similar to other Alexa Built in products, products built with the AVS Integration will need to pass the Alexa Voice Service product certification process comprising of Amazon-managed testing of security, acoustic performance, user experience, and functional testing to earn the Amazon Certified Alexa Built-in badge.

Q. What AWS regions will AVS Integration for IoT Core be available in at launch?

The AVS Integration for AWS IoT Core is available in all AWS regions where AWS IoT Core is available other than China (Beijing and Ningxia), Asia Pacific (Hong Kong) and Middle East (Bahrain). See the AWS Region Table for the current list of regions for AWS IoT Core.

AWS IoT Device Defender FAQs

Q. What is AWS IoT Device Defender?

AWS IoT Device Defender is a fully managed IoT security service that enables you to secure your IoT configurations on an ongoing basis. With AWS IoT Device Defender, you get tools to identify and respond to security issues. AWS IoT Device Defender audits your fleet to ensure it adheres to security best practices, continuously monitors your device fleets to detect any abnormal device behavior, alerts you about security issues as they arise, and recommends mitigation actions for these security issues.

Q. What are the key capabilities of AWS IoT Device Defender?

Audit AWS IoT Device Defender audits your device-related resources (such as X.509 certificates, IoT policies, and Client IDs) against AWS IoT security best practices (e.g., the principle of least privilege or unique identity per device). AWS IoT Device Defender reports configurations that are out of compliance with security best practices, such as multiple devices using the same identity, or overly permissive policies that can allow one device to read and update data for many other devices.

Detect AWS IoT Device Defender detects unusual device behavior that may be indicative of a compromise by continuously monitoring high-value security metrics from the device and AWS IoT Core (e.g., the number of listening TCP ports on your devices or authorization failure counts). You can specify normal device behavior for a group of devices by setting up behaviors (rules) for these metrics. AWS IoT Device Defender monitors and evaluates each datapoint reported for these metrics against user-defined behaviors (rules) and alerts you if an anomaly is detected.

Alerting AWS IoT Device Defender publishes alerts to the AWS IoT Console, Amazon CloudWatch, and Amazon SNS.

Mitigation AWS IoT Device Defender enables you to investigate issues by providing contextual and historical information about the device such as device metadata, device statistics, and historical alerts for the device. You can also use AWS IoT Device Management tools to perform mitigation steps such as revoking permissions, rebooting a device, resetting factory defaults, or pushing security fixes.

Q. How do customers secure devices today using AWS IoT and how does AWS IoT Device Defender help?

AWS IoT Core provides the security building blocks for you to securely connect devices to the cloud and to other devices. The building blocks allow enforcing security controls such as authentication, authorization, audit logging and end-to-end encryption at various levels of strictness based on your configurations. Following the AWS shared responsibility model, you own baselining security configurations regularly according to business requirements. However, human or systemic errors and authorized actors with bad intentions can introduce configurations with negative security impacts.

AWS IoT Device Defender helps you to continuously audit security configurations for compliance with security best practices and your own organizational security policies. The continuous audit is essential as misconfigurations can happen at any point of time. Additionally, security configurations can be impacted by the passage of time and new threats are constantly emerging. For example, cryptographic algorithms once known to provide secure digital signatures for device certificates can be weakened by advances in the computing and cryptanalysis methods.

AWS IoT Device Defender identifies opportunities to use AWS IoT security controls effectively. However, if security misconfigurations are not remediated or new attack vectors are disclosed publicly before devices are patched, the security of connected devices may be compromised. AWS IoT Device Defender complements preventative security controls in AWS IoT by helping you identify devices already compromised and initiating containment and corrective actions.

Q. Do I need to change device level code to use AWS IoT Device Defender?

No. You can audit your IoT configurations as well as monitor all cloud-side metrics with just a few clicks in the console. If you also want to monitor device-side metrics, you need to make some changes to your device code to publish device-side metrics to AWS IoT Device Defender. Reference implementation for a sample agent can be found here. AWS IoT Greengrass and FreeRTOS are fully integrated with AWS IoT Device Defender for both device-side and cloud-side metrics.

If your device platform has available specialized hardware that enables a trusted execution environment, we highly recommend implementing your device agent to run in a trusted environment. Consult your hardware security solution vendor for specific guidance on how to implement this type of design.

Q. How does AWS IoT Device Defender work?

AWS IoT Device Defender allows you to schedule audit tasks, monitor device activities, and receive notifications for audit violations and abnormal device behavior.

Audit tasks conduct assessments of your AWS IoT configurations. You can launch audit tasks on-demand or on a scheduled basis. To increase the accuracy of audit checks and minimize false positives, AWS IoT Device Defender incorporates the context of device interactions with AWS IoT Core.

AWS IoT Device Defender ingests and analyzes high-value security metrics collected from connected devices and their interactions with AWS IoT Core to continuously monitor device activities and detect abnormal device behavior. The metric data is continuously compared against user-provided security profiles. The collection and emittance of device metrics is optional. However, it is highly recommended. AWS IoT Device Defender provides reference implementation and documentation for device agents responsible to collect and emit the device-side metrics.

The results from scheduled audit tasks and any detected device activity abnormalities are published to the AWS IoT Console and are accessible through the AWS IoT Device Defender API. Additionally, you can configure AWS IoT Device Defender to send results to Amazon SNS topics for integration with security dashboards or triggering automated remediation workflows.

Q. Which AWS regions is AWS IoT Device Defender available in?

See the AWS Region Table for the current list of regions supported by AWS IoT Device Defender.

You can use AWS IoT Device Defender regardless of your geographic location, as long as you have access to one of the above AWS regions.

Q. Is AWS IoT Defender available in AWS Free Tier?

Yes. Visit the AWS IoT Device Defender pricing page for more information.

Q. How much does AWS IoT Device Defender cost?

You have the flexibility to use Audit and Detect independently, since they are both charged separately. Please visit the AWS IoT Device Defender pricing page for more information.

Q. When working with AWS IoT Device Defender, will I need to pay for AWS IoT Core Messages to report Detect metrics?

No, you will not need to pay for messages used to report device-side Detect metrics to AWS IoT Device Defender.

Q. When working with AWS IoT Device Defender, will I need to pay for AWS IoT Core Connectivity to report Detect metrics?

Yes, you will need to pay for connectivity if you connect with AWS IoT Core solely to report device-side Detect metrics to AWS IoT Device Defender. Please visit the AWS IoT Core pricing page for more information.

Q. How do I know the right values to set for the expected behavior of my devices in AWS IoT Device Defender?

Start by creating a security profile with restrictive behavior (e.g., low thresholds) and attach it to a ThingGroup for a representative set of devices. AWS IoT Device Defender will alert you with the metric datapoints emitted by the device for the behaviors that are violated. You can fine-tune the device behavior thresholds to match your use case.

AWS IoT Device Management FAQs

Q. What is AWS IoT Device Management? >>

Q. In which regions is AWS IoT Device Management available? >>

Q. What are the components of IoT Device Management? >>

Q. How does IoT Device Management help with device organization? >>

Q. How does IoT Device Management help with device updates? >>

Q. What is Fleet Indexing and Search? >>

Q. What type of devices does IoT Device Management support? >>

Q. How does Secure Tunneling help with troubleshooting devices? >>

Q. Are there any pre-requisites before using Secure Tunneling? >>

Q. In which regions is Secure Tunneling available? >>

Q. How do I get started? >>

Q. How much does AWS IoT Device Management cost? >>

Q. Is AWS IoT Device Management available in AWS Free Tier? >>

Q. What is the AWS IoT Device Management SLA? >>

AWS IoT Events FAQs

General

Q: What is AWS IoT Events?

A: AWS IoT Events is a new IoT service that helps companies continuously monitor their equipment and fleets of devices for failure or changes in operation and trigger alerts to respond when events occur. AWS IoT Events recognizes events across multiple sensors to identify operational issues, such as equipment slowdowns, and generates alerts such as notifying support teams of an issue. AWS IoT Events offers a managed complex event detection service on the AWS cloud, accessible through the AWS IoT Events console, a browser-based GUI where you can define and manage your event detectors, or direct ingest application program interfaces (APIs), code that allows two applications to communicate with each other. Understanding equipment or a process based on telemetry from a single sensor is often not possible; a complex event detection service will combine multiple sources of telemetry to gain full insight into equipment and processes. You define conditional logic and states inside AWS IoT Events to evaluate incoming telemetry data to detect events in equipment or a process. When AWS IoT Events detects an event, it can trigger pre-defined actions in another AWS service, such as sending alerts through Amazon Simple Notification Service (SNS).

Q: Why should I use AWS IoT Events?

A: AWS IoT Events makes it easy to detect and respond to events that happen across multiple IoT devices, equipment subsystems, and applications. For the vast majority of IoT customers, IoT deployments consist of multiple pieces of equipment with many independent sensors. For those customers, detecting when a critical event has occurred is hard and requires the creation of custom logic, which in turn requires the engagement of specialized system integrators. AWS IoT Events makes it possible to easily and cost effectively detect events system-wide and respond with appropriate actions to drive results such as optimizing manufacturing efficiency or improving production quality.

Getting Started

Q: How do I get started with AWS IoT Events?

A: To get started, sign up for an account. Log in to the console and create an endpoint to receive telemetry data you want to monitor, such as belt speed, motor voltage, amperage, and noise levels. Then, you can define your events to detect by writing simple ‘if-then-else’ statements, and selecting the alert or custom action to trigger when the event occurs. AWS IoT Events is stateful and reacts to the same input differently depending on the current state of equipment, such as ‘running’, ‘stuck’, or ‘off’. You can define states of equipment and combine this with incoming telemetry data and conditional logic, to make the right decision at the right time. Then, you can select a pre-built action to trigger, such as sending a message to a motor to shut down, whenever an event is detected.

Q: How do I receive inputs in AWS IoT Events?

A: AWS IoT Events accepts inputs from many IoT telemetry data sources, including sensor devices, management applications, and AWS IoT services. Any telemetry data input can be pushed to AWS IoT Events using a standard API interface (“Put_Signals” API). To directly send telemetry data from a device using AWS IoT Core, you can write a rule in AWS IoT Rules Engine to forward your IoT data into AWS IoT Events - identifying the detector via its Amazon Resource Name (ARN).

Q: How do I use AWS IoT Events to detect events?

A: AWS IoT Events uses custom, pre-defined conditional logic, such as ‘if-then-else’ statements, to understand events, such as when a motor might be stuck. An event detector definition includes inputs for evaluation, states that have been defined, conditional logic to evaluate, and actions to trigger. Within IoT Events, you can define an event detection in one of two ways. The first option is to use the AWS IoT Events console to define the conditions under which an event occurs and trigger actions when the conditions evaluate to “true”. The second option is to programmatically create an event detection by calling the “Create_Detector” API.

Q: How do I take an action when AWS IoT Events detects an event?

A: AWS IoT Events triggers actions when events are detected. Many of the actions that trigger are predefined in the AWS IoT Events actions library which makes it easier for you to reuse actions. Common actions used in AWS IoT Events are sending notifications using Amazon Simple Notification Service (SNS), triggering a function in AWS Lambda, writing a record to DynamoDB, republishing a message via AWS IoT Core, and starting workflows in AWS Step Functions. When configuring IoT Events, after defining the logic that will recognize a pattern in a number of inputs, you select the type of action to trigger.

Q: Can I create custom actions?

A: When using AWS IoT Events, you will have the option of triggering a function in AWS Lambda; this allows the execution of code without provisioning or managing servers and the ability to create custom actions.

Q: If my event is stateful, how do I use AWS IoT Events and the states of my equipment?

A: States are operational modes of equipment and processes. AWS IoT Events is stateful and reacts to the same input differently depending on the current state of equipment, such as ‘running’, ‘stuck’, or ‘off’. AWS IoT Events ties together inputs, states, and conditional logic to trigger actions. A software developer can use a graphical console builder to define states and transition between states. Alternatively, they may also define the same using the available APIs.

AWS IoT Greengrass FAQs

General Questions

What is AWS IoT Greengrass?

AWS IoT Greengrass is software that lets you run local compute, messaging, management, sync, and ML inference capabilities on connected devices in a secure way. With AWS IoT Greengrass, connected devices can run AWS Lambda functions, Docker containers, or both, execute predictions based on machine learning models, keep device data in sync, and communicate with other devices securely – even when not connected to the Internet.

AWS IoT Greengrass seamlessly extends AWS to devices so they can act locally on the data they generate, while still using the cloud for management, analytics, and durable storage. With AWS IoT Greengrass, you can use familiar languages and programming models to create your device software in the cloud, and then deploy it to your devices. AWS IoT Greengrass can be programmed to filter device data and only transmit necessary information back to the cloud.

How do I get started using AWS IoT Greengrass?

Click here to see the AWS IoT Greengrass getting started guide. You can click here to purchase a Raspberry Pi or review the list of qualified devices in the AWS IoT Partner Device Catalog.

Which AWS regions is AWS IoT Greengrass service available in?

Please refer to the AWS Regions Table for the most up to date information regarding region availability of AWS IoT Greengrass.

What are the major components of AWS IoT Greengrass? What does each component do?

AWS IoT Greengrass consists of a cloud service and three software distributions for IoT devices: AWS IoT Greengrass Core, AWS IoT Device SDK, and the AWS IoT Greengrass SDK.

The chart below compares those three software distributions to one another. AWS IoT Greengrass also works together with FreeRTOS. For more information about AWS IoT Greengrass and FreeRTOS, see the FAQ section called "Connecting FreeRTOS and other Devices to AWS IoT Greengrass."

	Purpose	Where it Runs
AWS IoT Greengrass Core	Provides local services (compute, messaging, state, security), and communicates locally with devices that run the AWS IoT Device SDK	64-bit CPU-based devices (x86 or Arm) that run a general-purpose OS such as Linux.
AWS IoT Device SDK	Allows devices to interact locally with AWS IoT Greengrass Cores	Almost any device that supports C++ or Python 2.7 and 3.7. Also included in FreeRTOS
AWS IoT Greengrass SDK	Allows Lambda functions to interact with local services inside an AWS IoT Greengrass Core	Inside a Lambda function deployed to AWS IoT Greengrass Core

What are AWS IoT Greengrass Core devices? What minimum hardware specifications are required?

The AWS IoT Greengrass Core software runs on a hub, gateway, or other device to automatically sync and interact with the cloud. AWS IoT Greengrass Core is designed to run on devices with a general-purpose processor that are powerful enough to run a general-purpose operating system, such as Linux. AWS IoT Greengrass requires at least 1GHz of compute (either Arm or x86), 128MB of RAM, plus additional resources to accommodate the desired OS, message throughput, and AWS Lambda execution depending on the use case. AWS IoT Greengrass Core can run on devices that range from a Raspberry Pi to a server-level appliance.

What AWS Lambda development languages are supported by AWS IoT Greengrass?

AWS IoT Greengrass supports Lambda functions authored in the following languages:

- Python 2.7 and 3.7
- Node v8.10 and v12.x
- Java 8
- C
- C++
- Any language that supports importing C libraries

Which Lambdas can be deployed to AWS IoT Greengrass?

Any Lambda that uses the Python 2.7 or 3.7, Node v8.10 or v12.x, or Java 8 Lambda Runtime can be deployed to AWS IoT Greengrass Core. Lambdas that get deployed to AWS IoT Greengrass must be packaged together with the AWS IoT Greengrass Core SDK. In addition, you can choose to also add the AWS SDK to your Lambda's package in order to easily interact with AWS services such as Amazon DynamoDB.

Please note: Some cloud services that your Lambda relies upon (e.g. DynamoDB) will not be available to your Lambda functions when AWS IoT Greengrass Core is in offline mode, and API calls to those services will fail in offline mode. In addition, your Lambda functions need to use the appropriate namespace for each AWS IoT Greengrass Core SDK and AWS SDK, if you include both in the same package.

Can I use AWS IoT Greengrass with a Docker container?

Yes, you can run Docker containers on an AWS IoT Greengrass device or run AWS IoT Greengrass in a Docker container environment.

You can deploy, run, and manage Docker containers with AWS IoT Greengrass. You can use any third-party tool to build Docker/Open Container Initiative (OCI) images, and your Docker images can be stored in Docker container registries, such as Amazon Elastic Container Registry (Amazon ECR), Docker Hub, or private Docker Trusted Registries (DTRs).

You can run AWS IoT Greengrass in a Docker container by configuring your AWS IoT Greengrass group to run with no Lambda containerization. To get started, you can access an AWS IoT Greengrass Docker file here and you can find documentation about how you can pull the AWS IoT Greengrass Docker image from Amazon ECR here. You can also deploy AWS IoT Greengrass as a snap, a containerized software package that can run on a variety of Linux distributions. To get started, you can access the AWS IoT Greengrass snap here and get started here.

Can I run AWS IoT Greengrass on Mac OS or Windows?

Yes, by running AWS IoT Greengrass with no Greengrass Lambda containerization at the group level in a Docker container, you'll be able to run AWS IoT Greengrass on Mac OS or Windows. You can learn more about this capability in our documentation.

What is the AWS IoT Greengrass SLA?

The AWS IoT Greengrass SLA for cloud management stipulates that you may be eligible for a credit towards a portion of your monthly service fees if AWS IoT Greengrass fails to achieve a Monthly Uptime Percentage of at least 99.9% for AWS IoT Greengrass cloud service.

For full details on all of the terms and conditions of the SLA, as well as details on how to submit a claim, please see the AWS IoT Greengrass SLA details page.

Local Resource Access

What is an AWS IoT Greengrass local resource?

"Local resource" refers to buses and peripherals that are physically present on the AWS IoT Greengrass host, or a file system volume on the AWS IoT

Greengrass host OS. For example, to communicate with devices connected via Modbus / CANBUS, an AWS IoT Greengrass Lambda function would need to access the serial port on the device. A local resource is defined at AWS IoT Greengrass group scope, and all Lambdas in the AWS IoT Greengrass group can use the defined local resources.

When would I access a local resource?

AWS IoT Greengrass local resource allows your Lambda functions to securely interact with hardware such as sensors and actuators. For example, your Lambda function can read video streams from the camera on the device or send command and control to GPIO.

Security

What is a hardware root of trust and why might I want one?

Hardware roots of trust provide tamper-protected trusted execution environments where a true random number generator can produce the private keys used for encryption functions. These hardware “secure elements” are resistant to malware tampering and are physically tied to a given IoT device, establishing a strong root of trust upon which software can be deployed safely.

How do I introduce hardware root of trust security to my AWS IoT Greengrass architecture?

First, you must run your AWS IoT Greengrass Core software on an edge device with a secure element. Following the hardware vendor’s directions, generate a private key on that secure element. Next, follow our documentation to update the config.json file settings to use the secure element private key.

Which partners offer hardware with a secure element?

For a current list of integrated hardware, visit the AWS Partner Device Catalog.

How are secure elements qualified to work with the Hardware Security Integration feature?

Secure element vendors have configured their secure elements to use a set of PKCS#11 standard APIs to integrate with AWS IoT Greengrass. Vendors use a set of testing tools to qualify that their hardware is configured correctly.

ML Inference

How can I use a ML model compiled with Amazon SageMaker Neo?

On AWS IoT Greengrass devices, you can perform ML inference on locally-generated data using models optimized with Amazon SageMaker Neo. To prepare your device for inference, you can follow the instructions on installing the Neo DLR runtime on your device. For more information, see Installing DLR.

You can compile a model in Amazon SageMaker Neo for your target hardware platform and store it in an Amazon S3 bucket. Then you can configure AWS IoT Greengrass to use the S3 bucket to deploy the Neo optimized model for local inference on the device.

How can I use a ML model not trained in Amazon SageMaker?

You can bring your ML model trained elsewhere by placing it in .tar.gz and .zip format in Amazon S3. You will then let AWS IoT Greengrass know the S3 URI and AWS IoT Greengrass will deploy to target devices.

Which AWS regions is AWS IoT Greengrass ML Inference available in?

AWS IoT Greengrass ML Inference is currently available in all the regions AWS IoT Greengrass is available in. Please refer to the AWS Regions Table for the most up to date information regarding region availability of AWS IoT Greengrass.

You can use AWS IoT Greengrass ML Inference regardless of your geographic location, as long as you have access to one of these AWS regions.

AWS IoT Greengrass Connectors

What are AWS IoT Greengrass Connectors?

AWS IoT Greengrass Connectors allow you to easily build complex workflows on AWS IoT Greengrass without having to worry about understanding device protocols, managing credentials, or interacting with external APIs. AWS IoT Greengrass Connectors allow you to interact with third-party applications, on-premises software, and AWS services without writing code. You can re-use common business logic from one AWS IoT Greengrass device to another through the ability to discover, import, configure, and deploy applications at the edge. You can also use AWS IoT Greengrass Secrets Manager to protect your keys and credentials in the cloud and at the edge. If an AWS IoT Greengrass Connector needs a secret to authenticate with an application or service, you can select and deploy a secret to the AWS IoT Greengrass Core as part of the AWS IoT Greengrass Connector configuration.

How can I add an AWS IoT Greengrass Connector to an AWS IoT Greengrass group?

AWS IoT Greengrass Connectors can be added via the “Connectors” section for each group in the AWS IoT Greengrass console. Once added, you configure the AWS IoT Greengrass Connector’s parameters and deploy the group to add them to your AWS IoT Greengrass Core device.

Who can use AWS IoT Greengrass Connectors?

Any AWS IoT Greengrass customer can use AWS IoT Greengrass Connectors in the AWS Management Console.

What AWS IoT Greengrass Connectors are available?

You can find available AWS IoT Greengrass Connectors in our documentation.

Protocol Adapters

How can I use AWS IoT Greengrass to implement alternative protocols?

Since Lambda functions running on AWS IoT Greengrass Cores have access to network resources, you can use Lambda to implement support for any protocol that is implemented on top of TCP/IP. In addition, you can also take advantage of AWS IoT Greengrass Local Resource Access to implement support for protocols that need access to hardware adapter/drivers.

AWS IoT Greengrass also provides a Modbus-RTU Protocol Adapter Connector that can help you to connect Modbus RTU devices. For more information, refer to the connector documentation here.

How can I use the OPC-UA Protocol with AWS IoT Greengrass?

You can use the IoT SiteWise connector to send industrial device data from OPC-UA servers to asset properties in AWS IoT SiteWise. Or, you can create a custom implementation that uses locally deployed Lambda functions to ingest and process OPC-UA data and then deliver the data to local or cloud targets.

Over the Air (OTA) Updates

What are AWS IoT Greengrass Over the Air (OTA) Updates?

From time to time, AWS will publish updated versions of the AWS IoT Greengrass Core software to provide the following benefits:

- Introduce new or improved features
- Bug fixes
- Security improvements

With AWS IoT Greengrass Over the Air Updates (OTA), customers can get all these benefits without having to manually download and reinstall the AWS IoT Greengrass Core software.

Do I have to use AWS IoT Greengrass OTA Updates?

No. You can always choose to download and install updates manually or follow a different software deployment process.

How will I be notified that new versions of AWS IoT Greengrass Core are available?

When new versions of AWS IoT Greengrass Core become available, we will announce it on the AWS IoT Greengrass software developer forum. You can find a link to that forum here.

AWS IoT Device Tester for AWS IoT Greengrass

What is AWS IoT Device Tester for AWS IoT Greengrass?

AWS IoT Device Tester for AWS IoT Greengrass is a test automation tool that lets you self-test and qualify AWS IoT Greengrass on your Linux-based devices. AWS IoT Device Tester provides a collection of automated tests that enable you to verify whether devices can run AWS IoT Greengrass and be authenticated by and interoperate with AWS IoT services.

Where do I get AWS IoT Device Tester for AWS IoT Greengrass?

You can get AWS IoT Device Tester for AWS IoT Greengrass here.

What does AWS IoT Device Tester for AWS IoT Greengrass test?

AWS IoT Device Tester for AWS IoT Greengrass verifies that the combination of a device's CPU architecture, Linux kernel configuration, and drivers work with AWS IoT Greengrass by testing the following:

- Required software packages have been installed
- Linux kernel containing AWS IoT Greengrass required kernel configuration (e.g. kernel configured for cgroups)
- Over the air updates
- Device can connect with IoT services and is able to run AWS Lambda functions
- Local resource access functionality
- Device shadow functionality

How do I get technical support for AWS IoT Device Tester for AWS IoT Greengrass?

Use any of the following channels to get support:

AWS Forum for AWS IoT Greengrass

Premium Support

Customer Support

How do I get my device listed in the AWS Partner Device Catalog?

If you are an AWS partner, the AWS Device Qualification Program defines the process to get your device listed in the catalog. A high level overview of the process is as follows:

1. Pass the AWS IoT Device Tester for AWS IoT Greengrass test
2. Log into the AWS Partner Network Portal
3. Upload the AWS IoT Device Tester report. Once the report is verified by AWS, and other device related artifacts such as picture and data sheet have been submitted, the device is listed in the AWS Partner Device Catalog.

In which regions is AWS IoT Device Tester for AWS IoT Greengrass available?

AWS IoT Device Tester for AWS IoT Greengrass is available in all the regions where AWS IoT Greengrass is available.

How much does AWS IoT Device Tester for AWS IoT Greengrass cost?

AWS IoT Device Tester for AWS IoT Greengrass is free to use. However, you will be responsible for any costs associated with AWS usage as part of testing. A single run of AWS IoT Device Tester that tests on a single AWS IoT Greengrass device will cost less than 20 cents.

Which CPU architectures and operating systems is AWS IoT Greengrass compatible with?

Operating systems and CPU architectures supported by AWS IoT Greengrass Core and tested for compatibility by AWS are listed here. Other Linux variants may also successfully run IoT Greengrass, but may not have been validated by the AWS IoT Greengrass team. You can validate other Linux variants for compatibility using the IoT Greengrass dependency checker on GitHub. Alternatively, you can run IoT Greengrass in "process mode" which lowers the compatibility threshold, but removes support for Linux containers.

What devices are compatible with AWS IoT Greengrass Core, and how can I get started quickly?

You can run AWS IoT Greengrass Core on a device that meets the minimum hardware and software requirements. You can also self-test your devices to see if they will run optimally with AWS IoT Greengrass and other AWS services using AWS IoT Device Tester. You can also discover and evaluate devices that are compatible with AWS IoT Greengrass in the AWS Partner Device Catalog.

How can I validate my device will run AWS IoT Greengrass Core?

To ensure your devices work with AWS IoT Greengrass Core, test it using the AWS IoT Device Tester for AWS IoT Greengrass. Download the tool and read the documentation.

Connecting FreeRTOS and Other Devices to AWS IoT Greengrass

How can I connect devices locally to AWS IoT Greengrass Core?

You can connect devices locally to AWS IoT Greengrass Core using FreeRTOS or the AWS IoT Device SDK. AWS IoT Greengrass discovery is available on the AWS IoT Device SDK via C++ and Python 2.7 and 3.7. For more information, refer to the AWS IoT Greengrass user guide. You can use the AWS IoT Greengrass discovery library in your FreeRTOS source code to find and connect to an AWS IoT Greengrass Core device.

What languages support AWS IoT Greengrass via the AWS IoT Device SDK?

AWS IoT Greengrass discovery is available on the AWS IoT Device SDK via C++ and Python 2.7 and 3.7. For more information, refer to the AWS IoT Greengrass developer guide.

Does FreeRTOS work with AWS IoT Greengrass?

Yes. FreeRTOS devices can connect directly to the cloud or connect to AWS IoT Greengrass. FreeRTOS runs on IoT endpoints and is often responsible for the ‘sensing’ and ‘acting’ in an IoT topology.

What is the difference between AWS IoT Greengrass and FreeRTOS?

AWS IoT Greengrass is software that lets you run local compute, messaging, data caching, sync, and ML inference capabilities for connected devices in a secure way. With AWS IoT Greengrass, connected devices can run AWS Lambda functions, Docker containers, or both, keep device data in sync, and communicate with other devices securely – even when not connected to the Internet. Using AWS Lambda, AWS IoT Greengrass ensures your IoT devices can respond quickly to local events, use Lambda functions running on AWS IoT Greengrass Core to interact with local resources, operate with intermittent connections, stay updated with over the air updates, and minimize the cost of transmitting IoT data to the cloud.

FreeRTOS is an open source, real-time operating system for microcontrollers that operates on the edge and does not generally support chipsets that could run AWS IoT Greengrass. These microcontroller devices are found on a variety of IoT endpoints such as fitness trackers, pacemakers, electricity meters, automotive transmissions, and sensor networks. FreeRTOS devices cannot run AWS IoT Greengrass Core but can connect, send, and receive messages to and from an AWS IoT Greengrass Core device for local processing at the edge.

The hardware requirements and operating systems are different on both devices.

	FreeRTOS	AWS IoT Greengrass
Software	Real-time operating system, runs on a microcontroller	Runtime for Linux devices and SDK for AWS IoT Greengrass aware devices
Hardware Requirements	>64**KB** RAM	>128**MB** of RAM
Category	Embedded systems, IoT endpoints	Edge devices, local gateways
Use Cases	Microcontroller-based devices	Industrial automation systems, wireless routers, video cameras, gateways

AWS IoT SiteWise FAQs

General Questions

Q: What is AWS IoT SiteWise?

A: AWS IoT SiteWise is a managed service that enables industrial enterprises to collect, store, organize and visualize thousands of sensor data streams across multiple industrial facilities. AWS IoT SiteWise includes software that runs on a gateway device that sits onsite in a facility, continuously collects the data from its historian, and sends it to the AWS Cloud. Data can also be ingested into AWS IoT SiteWise through AWS IoT Core via the MQTT protocol or by using a PUT API. With AWS IoT SiteWise, you can skip months of developing undifferentiated data collection and cataloging solutions. Instead, you can focus on using your data to detect and fix equipment issues, spot inefficiencies and improve production output.

Q: Why should I use AWS IoT SiteWise?

A: With AWS IoT SiteWise you can easily collect and gain insights into equipment data to reduce waste that commonly occurs in industrial operations. You can structure your sensor data by equipment, so you can easily retrieve it. AWS IoT SiteWise also computes performance metrics that you specify for your equipment and processes. These metrics can help identify various types of wastes such as equipment issues, production gaps, and quality defects.

Specifically, AWS IoT SiteWise enables you to:

Understand and improve processes across multiple facilities. Reducing waste often requires calculating equipment data metrics to track various business goals. With AWS IoT SiteWise, industrial engineers can group sensor data streams by production line, and group production lines into facilities. Analysts in the corporate HQ can then easily query sensor data across all facilities. With

AWS IoT SiteWise, you can create an authoritative source of data that is organized by equipment and processes for use across your entire organization.

Understand and resolve equipment issues efficiently. Industrial engineers need performance metrics to pinpoint issues with physical equipment. With AWS IoT SiteWise, a technician can understand the condition of each wind turbine and solar powered generator remotely and dispatch the right resources to fix an issue much faster. This leaves the engineers more time to focus on their core job of understanding and designing better systems, instead of coordinating operational issues in the field.

Visualize operational data of devices and equipment with SiteWise Monitor. Create a fully managed web application for visualizing and interacting with operational data from devices and equipment connected to AWS IoT, without writing code. Automatically discover and visualize data from assets that have already been ingested and modeled with AWS IoT SiteWise. You can view the current data values and view live trend charts of asset data, as well as view historical time series asset data plots across a user defined time period.

Q: What are the pre-requisites for using SiteWise Monitor?

A: Before using a SiteWise Monitor web application, your edge devices and equipment must be configured to send data to AWS IoT SiteWise. If edge data is stored in on-premises historians, then AWS IoT SiteWise software can be run on an edge gateway to transmit data to AWS. If edge data is transmitted to AWS IoT Core using the MQTT protocol, then AWS IoT SiteWise must be selected as the destination endpoint in the Rules Engine. Once data is flowing to AWS, digital models of edge devices and equipment must be created in AWS IoT SiteWise. Finally, you will have to link the equipment data ingested into AWS IoT SiteWise to the asset models created and instantiate the models that represent the actual edge devices and equipment.

Q: How do I setup SiteWise Monitor?

A: To create a web application, an Administrator user will login to the AWS Management Console and then open the AWS IoT SiteWise console. They will then navigate to the Getting Started page under the AWS IoT SiteWise Monitor menu. The Getting Started page guides the Administrator through a simple step-by-step workflow to 1/ create a web portal, 2/ configure an active directory for SSO login, 3/ select users as administrators of the web portal and 4/ add users that will have access to the web portal

Q: How is operational data secured within SiteWise Monitor?

A: SiteWise Monitor web application users are authenticated based on their enterprise identity credentials or built-in user credentials set up in AWS SSO. An Administrator user can setup which assets and asset data is accessible in a web application, and which users are authorized to use that web application.

AWS IoT Things Graph FAQs

AWS IoT Things Graph Overview

Q: What is AWS IoT Things Graph?

A: AWS IoT Things Graph is a fully managed service that makes it easy to rapidly build IoT applications by connecting devices (such as sensors and actuators) and web services (such as Amazon Rekognition) from different vendors that speak different protocols and don't work with each other out of the box. AWS IoT Things Graph allows you to represent things (devices and web services) as models that abstract capabilities from the underlying implementation and make them accessible via APIs, transform messages between various models so they can understand each other, and coordinate interactions between models to create workflows.

Q: Why should I use AWS IoT Things Graph?

A: AWS IoT Things Graph simplifies IoT application development and enables you to bring your applications to market faster. A model-based approach allows you to focus on the logic and functionality of your applications instead of dealing with the low-level details of integrating devices and services. IoT applications built using AWS IoT Things Graph are more maintainable, allowing you to swap in similar devices from different manufacturers without redoing your entire application or retrofitting existing installations. As your needs and available devices change, you can adapt without throwing away your existing investments. The simple drag-and-drop interface allows you to build and iterate on applications easily.

Q: What applications are best suited for AWS IoT Things Graph?

A: AWS IoT Things Graph is best suited for applications that encompass a wide diversity of devices and services from different manufacturers, that when combined in a specific order, can automate business operations. Such applications are abundant in verticals such as home automation and access management, building energy management, smart agriculture, smart cities, manufacturing, logistics, and oil and gas extraction.

Q: How does AWS IoT Things Graph work with other AWS IoT services?

A: AWS IoT Things Graph provides a higher-level abstraction on AWS IoT by allowing you to define things and their interactions, and making sure interactions work across things without having to worry about device specific details and managing sequencing of events. AWS IoT Things Graph builds on the capabilities of AWS IoT Core and AWS IoT Greengrass.

Q: In which regions is AWS IoT Things Graph available?

A: AWS IoT Things Graph is now generally available in the US East (Northern Virginia), US West (Oregon), EU (Ireland), APAC (Sydney), APAC (Seoul), and APAC (Tokyo) regions. Please see the AWS Region table for all the regions where AWS IoT Things Graph is available.

Q: How much does AWS IoT Things Graph cost?

A: Please see the AWS IoT Things Graph pricing page for more details.

Getting Started

Q: How do I get started with AWS IoT Things Graph?

A: AWS IoT Things Graph is generally available. You can start building workflows by logging into AWS IoT Things Graph Console (click here). The AWS IoT Things Graph console allows you to build models, create flows, configure your deployment, and manage and monitor your flows, all from a single location. If you haven’t already, sign up for an AWS account here.

Q: Where can I deploy my AWS IoT Things Graph application?

A: Once your application is built, you can deploy it to the AWS Cloud or to edge gateways running AWS IoT Greengrass. Deployment is easy and can be initiated with one click from the AWS IoT Things Graph console. Then, AWS IoT Things Graph asks you to select devices registered in the AWS IoT registry that are represented as models in your application. IoT Things Graph bundles your application details and the related models along with the AWS IoT Things Graph run-time, and pushes it to the AWS Cloud or your AWS IoT Greengrass-enabled device where it starts listening to messages and coordinating interactions.

Q: Which protocols does AWS IoT Things Graph handle?

A: AWS IoT Things Graph supports MQTT, HTTP, and Modbus.

404

Sorry, this page does not exist or is no longer available.

FreeRTOS FAQs

FreeRTOS

Q. What is FreeRTOS?

FreeRTOS is an open source, real-time operating system for microcontrollers that makes small, low-power edge devices easy to program, deploy, secure, connect, and manage. Distributed freely under the MIT open source license, FreeRTOS includes a kernel and a growing set of software libraries suitable for use across industry sectors and applications. To support a growing number of use cases, AWS provides software libraries that offer enhanced functionality including connectivity, security, and over-the-air updates. For example, you can use FreeRTOS to securely connect your small, low-powered devices to AWS cloud services like AWS IoT Core or to more powerful edge devices running AWS IoT Greengrass.

Q. What is the relationship between Amazon FreeRTOS and FreeRTOS?

Since 2017, Amazon FreeRTOS has been an extension of the FreeRTOS project, so we have unified the two names to reduce customer confusion. The FreeRTOS project now includes the additional connectivity libraries, security libraries, and IoT reference integrations.

Q. Which AWS region is FreeRTOS available in?

Please see the AWS Region Table for a complete list of regions where FreeRTOS is available. You can download FreeRTOS code from GitHub irrespective of your geographic location and AWS region availability.

Q. What are some use cases for FreeRTOS?

FreeRTOS can be used in embedded systems spanning industrial, commercial, and consumer applications. For example, smart meters, oil pump sensors, appliances, commercial security systems, fitness trackers, and sensor networks can all benefit from FreeRTOS. Smart meters are used in homes to monitor electricity usage in real-time. Fitness trackers send health data via the user's mobile device to the cloud for real-time monitoring or analytics. Utilities benefit from this data by enabling more efficient load balancing and power output from their generating stations. Oil pump sensors are used on oil rigs to monitor the output on wells that might be buried deep underwater. An oil rig might deploy FreeRTOS on those sensors and use an AWS IoT Greengrass Core to locally process data from pumps and valves in real-time. The AWS IoT Greengrass Core could then send batches of preprocessed pump sensor data to the cloud for analytics and data warehousing. To learn more about AWS IoT Greengrass, click here.

Q. How can a microcontroller developer get access to FreeRTOS?

FreeRTOS developers can download the FreeRTOS microcontroller device software from the FreeRTOS console, GitHub, or FreeRTOS.org.

Q. Who can benefit from FreeRTOS?

Semiconductor vendors manufacture microcontrollers and modules like connectivity sensors, security peripherals, and Ethernet controllers. These microcontrollers and modules are used by OEMs to build IoT devices.

OEMs include industrial companies, commercial enterprises, and consumer brands. Microcontroller developers can use FreeRTOS to easily design and develop a connected device and IoT applications.

Enterprises can use IoT connected devices that are powered by FreeRTOS to gain business and operational efficiency.

Q. What are the major components of FreeRTOS software?

FreeRTOS includes the FreeRTOS kernel, a real-time operating system kernel for microcontrollers, and libraries that support connectivity, security, and over-the-air updates. The connectivity stack includes MQTT, HTTP, TCP/IP, Wi-Fi, and Bluetooth Low Energy for cloud and local connectivity. Security libraries include a standard-based Berkeley socket interface for TLS and a PKCS#11 standard interface for crypto offload.

- FreeRTOS kernel: An MIT-licensed real-time operating system (RTOS) kernel for embedded microcontroller devices.
- MQTT: MQTT client library that you can use to create applications that publish and subscribe to MQTT topics, and connect to an MQTT-based message broker.
- HTTP Library: HTTP client library that you can use to create applications that use REST API to connect to an HTTP server.
- Wi-Fi management library: A common API layer that abstracts port-specific Wi-Fi implementations and simplifies application development.
- Bluetooth Low Energy management library: A standardized API layer that enables you to use Generic Access Profile (GAP) and Generic Attributes (GATT) profiles to create Bluetooth Low Energy applications.
- Device Defender library: Allows your FreeRTOS-based devices to work with AWS IoT Device Defender. Learn more about Device Defender here.
- Device Shadows library: Defines functions to create, update, and delete AWS IoT Device Shadows. Learn more about Device Shadows here.
- OTA agent: Enables you to manage the notification, download, and verification of firmware updates for FreeRTOS devices.
- Greengrass discovery: A library that helps FreeRTOS devices to discover and connect to an AWS IoT Greengrass Core.
- TLS: Transport Layer Security (TLS) interface is an optional wrapper used to abstract cryptographic implementation details of the underlying TLS stack.
- PKCS#11: A cryptographic API layer (OASIS standard) that abstracts key storage, get/set properties for cryptographic objects, and session semantics.

Q. What minimum hardware specifications are required?

If you run all FreeRTOS libraries, including TLS, on the application microcontroller, you may need a microcontroller with >25MHz processing speed and >64KB RAM. If the communication and crypto stack (except for MQTT) is offloaded onto the networking processor, your microcontroller will only need 10MHz processing speed and 16KB RAM. However, these values are just approximations, as factors such as MCU architecture, compiler, and compiler optimization level may impact processing speed and RAM requirements. FreeRTOS needs 128KB of program memory per executable image stored on the microcontroller. For OTA update functionality, two executable images must be stored in program memory at the same time.

Q. What architectures does FreeRTOS support?

FreeRTOS provides IoT Reference Integrations for a wide range of microcontrollers from our partners in the AWS Partner Device Catalog. FreeRTOS includes the FreeRTOS kernel, which supports 40+ architectures, including the latest from RISC-V and ARMv8-M.

Q. How can I get started on FreeRTOS?

You can use the getting started guide for systematic instructions on how to run FreeRTOS on a qualified board.

Q. How can I get technical support?

Use any of the following channels to get support:

FreeRTOS Community Forums

Premium Support

AWS Support

GitHub Issues

Q: What happened to the Amazon FreeRTOS group on AWS Forums?

To create a better forums experience for our customers, we have migrated all content and users from the AWS Forums Amazon FreeRTOS group to the Amazon Web Services category on the FreeRTOS Community Forums. Learn more here.

Q. Is there a user guide?

Yes. You can use the FreeRTOS user guide to get started with connecting FreeRTOS devices to AWS.

Q. Can I use FreeRTOS to connect to other cloud services?

Yes. FreeRTOS is an open-source software, so it can be modified to fit any specific needs of your application.

Q. Can I make changes to the FreeRTOS source code for my project?

Yes. FreeRTOS is an open-source software distributed under the MIT license, so it can be modified to fit any specific needs of your application or project without the permission of AWS.

Q. How much do I pay for using FreeRTOS?

FreeRTOS is free to download and use under an open source MIT license.

Q. How can I explore FreeRTOS without buying hardware?

You can explore FreeRTOS code and functionality on a Windows machine by downloading the libraries and samples ported to run on Windows. This is a set of files referred to as the FreeRTOS simulator for Windows (Windows Simulator). Get started here.

Q. Does FreeRTOS include hardware?

No. FreeRTOS is an open source, real-time operating system for microcontrollers. You can run FreeRTOS on your chosen microcontroller by porting FreeRTOS code and validating the ported code with AWS IoT Device Tester. To make it easier for you, we have provided IoT reference integrations and qualified ports for common microcontrollers in the AWS Partner Device Catalog.

Q. How do I understand FreeRTOS versioning?

The FreeRTOS kernel and each individual library use semantic versioning. In semantic versioning, the version number itself (X.Y.Z) indicates if the release is a major, minor, or point release. An increment of the first version number indicates a major release, an increment of the second version number indicates a minor release, and an increment of the third number indicates a point release.

For FreeRTOS IoT reference integration releases that consist of a group of FreeRTOS libraries, date-based versioning is used. Additionally, downloads that contain the FreeRTOS kernel and additional libraries use date-based versioning. This date-based versioning follows the format YYYYMM.NN for standard releases where Y represents the year, M represents the month, and N represents the release order within the designated month (00 being the first release in a given month). A ‘Major’ denotation indicates the addition of new features and/or significant updates to multiple libraries. For example, ‘201906.00 Major’ implies that it is the first release in June 2019 and contains new features and/or significant updates. By moving semantic versioning down to the individual libraries, you can make your own assessment of the scope and impact of a new release on your application.

FreeRTOS kernel

Q. What is the FreeRTOS kernel?

Developed over a 15-year period and in partnership with the world's leading chip companies, the FreeRTOS kernel is the market-leading, real-time operating system kernel and the de-facto standard solution for microcontrollers and small microprocessors.

Q. Does AWS maintain the FreeRTOS kernel?

Yes. The latest update to v10 of the FreeRTOS kernel, includes support for RISC-V and Armv8-M (Cortex-M33 and Cortex-M23).

Q. What is the difference between the MIT open source license and the (previously used) modified GPL open source license?

Both licenses allow the software to be used for free, even in commercial products, and neither license imposes any obligations when distributing binary (executable) copies. The MIT license provides simplified wording and allows for more permissive use of our source code. With the MIT license, you can still develop and sell commercial products using FreeRTOS (including the kernel) but you are no longer obliged to open source modifications to our source code, meaning you own all the changes you make. The only requirements under MIT is that the copyright notice and permission notice shall be included in all copies or substantial portions of the software (source files).

FreeRTOS Community

Q. Can I contribute code to FreeRTOS?

Yes, you can contribute code to FreeRTOS via GitHub. Please refer to Contributions.md file in GitHub for guidelines.

AWS IoT Device Tester for FreeRTOS

Q. What is AWS IoT Device Tester for FreeRTOS?

AWS IoT Device Tester for FreeRTOS is a Windows/Linux/Mac test automation tool that lets semiconductor vendors self test and qualify FreeRTOS on their microcontroller boards. With AWS IoT Device Tester, semiconductor vendors can verify whether their microcontroller boards can run FreeRTOS and be authenticated by and interoperate with AWS IoT services.

Q. Where do I get AWS IoT Device Tester for FreeRTOS?

You can get AWS IoT Device Tester for FreeRTOS here.

Q. Is AWS IoT Device Tester for FreeRTOS required for qualification and listing in the AWS Partner Device Catalog?

Yes, you can learn more about how to get listed here.

Q. What does AWS IoT Device Tester for FreeRTOS test?

AWS IoT Device Tester for FreeRTOS tests that the combination of a FreeRTOS IoT reference integration with a microcontroller board's porting layer interfaces and underlying device drivers are compatible and can interoperate with AWS IoT services. AWS IoT Device Tester confirms the porting layer interfaces (implemented by semiconductor vendors) for FreeRTOS libraries function correctly on top of the device drivers. Also, AWS IoT Device Tester runs end-to-end tests to confirm the microcontroller board can authenticate and interoperate with AWS IoT services.

Q. How do I get technical support for AWS IoT Device Tester for FreeRTOS?

Use any of the following channels to get support:

Premium Support

Customer Support

GitHub Issues

Q. How can I get my microcontroller-based hardware platform listed in the AWS Partner Device Catalog?

The AWS Device Qualification Program defines the process to get your microcontroller listed on AWS Partner Device Catalog. The high-level overview is as follows: First, you must pass the AWS IoT Device Tester for AWS FreeRTOS tests. Next, log into the AWS Partner Network Portal and upload the AWS IoT Device Tester for FreeRTOS report. Provide reference to your source code for ported FreeRTOS interfaces to make it available to OEMs. Once the ported code and report are verified by AWS and other device related artifacts (such as device image, data sheet, etc.) have been submitted, the device is listed in the AWS Partner Device Catalog.

Q. In which regions is AWS IoT Device Tester for FreeRTOS available?

AWS IoT Device Tester for FreeRTOS is available in all the regions where FreeRTOS is supported.

Q. How much does AWS IoT Device Tester for FreeRTOS cost?

AWS IoT Device Tester for FreeRTOS is free to use. However, you will be responsible for any costs associated with AWS usage as part of qualification tests. On average, a single run of the AWS IoT Device Tester would cost less than a cent. Please refer to AWS IoT Core pricing for associated costs.

FreeRTOS and AWS IoT Greengrass

Q. What is the difference between AWS IoT Greengrass and FreeRTOS?

AWS IoT Greengrass is software that lets you run local compute, messaging, data caching, sync, and ML inference capabilities for connected devices in a secure way. With AWS IoT Greengrass, connected devices can run AWS Lambda functions, keep device data in sync, and communicate with other devices securely – even when not connected to the Internet. Using AWS Lambda, AWS IoT Greengrass ensures your IoT devices can respond quickly to local events, use Lambda functions running on AWS IoT Greengrass Core to interact with local resources, operate with intermittent connections, stay updated with over the air updates, and minimize the cost of transmitting IoT data to the cloud.

FreeRTOS is an open source, real-time operating system for microcontrollers that operates on the edge and does not generally support chipsets that could run AWS IoT Greengrass. These microcontroller devices are found on a variety of IoT endpoints such as fitness trackers, pacemakers, electricity meters, automotive transmissions, and sensor networks. FreeRTOS devices cannot run

AWS IoT Greengrass Core but can trigger the execution of Lambda functions on an AWS IoT Greengrass Core device.

The hardware requirements and operating systems are different on both devices.

	FreeRTOS	AWS IoT Greengrass
Software	Real-time OS with libraries, runs on a microcontroller	Runtime for Linux devices and SDK for AWS IoT Greengrass-aware devices
Hardware Requirements	>64**KB** RAM	>128**MB** of RAM
Category	Embedded systems, IoT endpoints	Edge devices, local gateways
Use Cases	Microcontroller-based devices	Industrial automation systems, wireless routers, smartphones

Q. Does FreeRTOS require the use of AWS IoT Greengrass?

FreeRTOS does not require the use of AWS IoT Greengrass. FreeRTOS runs on IoT endpoints and is often responsible for the ‘sensing’ and ‘acting’ in an IoT topology. FreeRTOS devices can connect directly to the cloud or connect to AWS IoT Greengrass Core devices locally.

Q. How can I connect FreeRTOS devices to AWS IoT Greengrass Core devices?

The AWS IoT Greengrass discovery library is included in the FreeRTOS source code, enabling you to find and connect to an AWS IoT Greengrass Core device. For more information, refer to the FreeRTOS user guide.

FreeRTOS and Bluetooth Low Energy

Q. What is Bluetooth Low Energy support in FreeRTOS?

Bluetooth Low Energy support in FreeRTOS offers a standardized API layer for developers to write Bluetooth Low Energy applications that are portable across FreeRTOS qualified boards. It includes companion Android and iOS SDKs that enable a FreeRTOS device to consume AWS IoT services using an Android or iOS device as proxy. You can use standard Generic Access Profile (GAP) and Generic Attributes (GATT) profiles to write Bluetooth Low Energy applications, and use custom profiles for MQTT over Bluetooth Low Energy, and Wi-Fi provisioning via Bluetooth Low Energy. You can also use other AWS IoT services and features including AWS IoT Device Defender, Device Shadows, and OTA Updates.

Q. Why should I use FreeRTOS Bluetooth Low Energy?

If you are an embedded developer that needs to create a Bluetooth Low Energy application, connect your Bluetooth Low Energy devices to AWS IoT through an Android or iOS proxy, or use AWS IoT features such as AWS IoT Device Shadows, you will benefit from using Bluetooth Low Energy in FreeRTOS. The standardized Bluetooth Low Energy API for FreeRTOS allows you to code portable applications against FreeRTOS-qualified devices. If you decide to use a different microcontroller (e.g. for upgrading the product), you can use your existing Bluetooth Low Energy application code as a base for adding newer features. You can then concentrate on your application code and not worry about connectivity and security libraries underneath, which are not features that differentiate your product.

Q. Which boards are supported by Bluetooth Low Energy in FreeRTOS?

Visit our getting started page for more information on supported hardware.

Q. How do I find the libraries I need?

You can download FreeRTOS source code from the FreeRTOS console and the FreeRTOS GitHub repository, and you can download the companion Android and iOS SDKs from GitHub. FreeRTOS source code has demo examples, and the mobile SDKs have sample applications to help you quickly get started.

Q. Does Bluetooth Low Energy support in FreeRTOS work only with AWS?

No. The FreeRTOS libraries for Bluetooth Low Energy are open source and under the MIT license so developers can modify according to their specific need.

Q. What Bluetooth Low Energy versions are supported?

FreeRTOS supports Bluetooth Low Energy versions 4.2 and above. Bluetooth Low Energy version 4.2 raises the security bar by adding support for Bluetooth Low Energy Secure Connections, an enhanced security feature introduced in Bluetooth Low Energy version 4.2 to authenticate a peer device and create an encrypted channel.

Q. Is Amazon providing the Bluetooth Low Energy stack?

No. FreeRTOS is providing a standardized Bluetooth Low Energy API library that interfaces with a third-party (e.g., MCU vendor) Bluetooth Low Energy stack.

Q. What GATT services does FreeRTOS support for Bluetooth Low Energy enable?

Bluetooth Low Energy support in FreeRTOS enables developers to add any number of standard and custom GATT services, depending on the capabilities of the target hardware. FreeRTOS contains two customer profiles: 1) MQTT over Bluetooth Low Energy, to enable Bluetooth Low Energy devices to use AWS IoT services, and 2) Wi-Fi provisioning over Bluetooth Low Energy, to provision Wi-Fi credentials in an IoT device using Bluetooth Low Energy.

Q. Can the Bluetooth Low Energy proxy take a local action?

Currently, there is no mechanism to intercept messages flowing between a Bluetooth Low Energy device and AWS IoT. The Bluetooth Low Energy proxy only acts as a pass-through device.

However, you can use methods and classes that are provided within the proxy libraries as a starting point and modify these libraries to intercept the messages and take local action.

Q. What are the benefits of using MQTT over Bluetooth Low Energy?

MQTT over Bluetooth Low Energy enables Bluetooth Low Energy devices to connect to AWS IoT via a proxy device, as well as enables you to use other AWS services and features including AWS IoT Device Defender, AWS IoT Device Shadows, and FreeRTOS OTA updates.

Q. Can I use multiple connectivity options from the same device?

Yes. You can use MQTT over Wi-Fi and MQTT over Bluetooth Low Energy simultaneously as long as your device has the memory required to do so.

Q. How can I authenticate my proxy device with AWS IoT?

AWS IoT uses the Amazon Cognito service to authenticate mobile devices with cloud services. However, you can also use X.509 certificates that are supported by FreeRTOS mobile SDKs to authenticate your proxy device with AWS IoT.

FreeRTOS and AWS IoT Device Management

Q. How do I update my devices with new firmware?

You can use the over-the-air (OTA) update feature of FreeRTOS. Within the AWS IoT Device Management console, all you need to do is provide a firmware image, select the devices to update, select a code signing method, and create the FreeRTOS OTA job update. For more information on the OTA update feature and code signing, refer to the FreeRTOS user guide.

Q. What is code signing?

Code signing enables developers to confirm the integrity and origin of firmware images scheduled for OTA deployment to FreeRTOS devices. The process confirms the integrity of firmware images using a cryptographic hash that validates that the code has not been altered or corrupted since it was signed. The process also uses public-key cryptography to sign these images with proof of origin that can be validated on the device. Using the integrated FreeRTOS OTA update device job within the AWS IoT Device Management console, developers can upload a new firmware image, sign that image, and deliver it to a group of devices in the field. Those devices will validate the signature upon download and only install trusted code. Customers can use IAM to provide fine-grained access controls to signing tools, so only designated developers can sign and schedule new firmware updates.

Q. Do I have to use code signing?

No, you can also use your own signing service and upload a signed image directly into Amazon S3. You will need to modify the FreeRTOS OTA agent to accept the signature format that you choose to use.

Q. What hardware supports OTA?

You can find qualified hardware that support FreeRTOS OTA in the AWS Partner Device Catalog.

FreeRTOS Security

Q. How does FreeRTOS secure data in transit?

FreeRTOS uses Transport Layer Security (TLS 1.2) for secure connections to the cloud. The TLS protocol ensures privacy and data integrity between two communicating applications. It ensures that a FreeRTOS device and the cloud server are mutually authenticated using X.509 certificates and that data is encrypted while it is in transit.

Q. How does FreeRTOS secure data within the device (at rest)?

FreeRTOS uses a standard application interface, called PKCS #11, for encryption, digital signatures, and cryptographic object management. Cryptographic objects are kept either in dedicated storage or in the flash memory of the main microcontroller if dedicated storage is not available. If your device requires data encryption at rest, we recommend that you use dedicated cryptographic hardware to protect your encryption keys. Use the PKCS #11 API to access keys and encrypt and decrypt application data.

Q. How can I stay informed of the latest security patches?

Security updates are provided via the FreeRTOS console, the FreeRTOS Security Updates page, and on GitHub.

Q. Where can I report a security concern?

To report a security issue, please visit Vulnerability Reporting for AWS.

Q. How can I update my FreeRTOS devices with the latest security patches?

We recommend that you use the over-the-air (OTA) update feature of FreeRTOS to send security patches to your FreeRTOS devices. Within the AWS IoT Device Management console, you can provide a firmware image, select the devices to update, and create the FreeRTOS OTA job update. The code signing feature will verify the signed image on the device to ensure your device code is not compromised during deployment and updates. For more information on the OTA update feature, refer to the FreeRTOS user guide.

Amazon SageMaker FAQs

General

Q: What is Amazon SageMaker?

Amazon SageMaker is a fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning (ML) models quickly. SageMaker removes the heavy lifting from each step of the machine learning process to make it easier to develop high quality models.

Q: In which regions is Amazon SageMaker available?

For a list of the supported Amazon SageMaker AWS regions, please visit the AWS Region Table for all AWS global infrastructure. Also for more information, see Regions and Endpoints in the AWS General Reference.

Q: What is the service availability of Amazon SageMaker?

Amazon SageMaker is designed for high availability. There are no maintenance windows or scheduled downtimes. SageMaker APIs run in Amazon’s proven, high-availability data centers, with service stack replication configured across three facilities in each AWS region to provide fault tolerance in the event of a server failure or Availability Zone outage.

Q: What security measures does Amazon SageMaker have?

Amazon SageMaker ensures that ML model artifacts and other system artifacts are encrypted in transit and at rest. Requests to the SageMaker API and console are made over a secure (SSL) connection. You pass AWS Identity and Access Management roles to SageMaker to provide permissions to access resources on your behalf for training and deployment. You can use encrypted S3 buckets for model artifacts and data, as well as pass a KMS key to SageMaker notebooks, training jobs, and endpoints, to encrypt the attached ML storage volume.

Q: How does Amazon SageMaker secure my code?

Amazon SageMaker stores code in ML storage volumes, secured by security groups and optionally encrypted at rest.

Q: How am I charged for Amazon SageMaker?

You pay for ML compute, storage, and data processing resources you use for hosting the notebook, training the model, performing predictions, and logging the outputs. Amazon SageMaker allows you to select the number and type of instance used for the hosted notebook, training, and model hosting. You only pay for what you use, as you use it; there are no minimum fees and no upfront commitments. See the Amazon SageMaker pricing page for details.

Q: What if I have my own notebook, training, or hosting environment?

Amazon SageMaker provides a full end-to-end workflow, but you can continue to use your existing tools with SageMaker. You can easily transfer the results of each stage in and out of SageMaker as your business requirements dictate.

Q. What is Amazon SageMaker Studio?

Amazon SageMaker Studio provides a single, web-based visual interface where you can perform all ML development steps. SageMaker Studio gives you complete access, control, and visibility into each step required to build, train, and deploy models. You can quickly upload data, create new notebooks, train and tune models, move back and forth between steps to adjust experiments, compare results, and deploy models to production all in one place, making you much more productive. All ML development activites including notebooks, experiment management, automatic model creation, debugging and profiling, and model drift detection can be performed within the unified SageMaker Studio visual interface.

Q. What is Amazon SageMaker Autopilot?

Amazon SageMaker Autopilot is the industry’s first automated machine learning capability that gives you complete control and visibility into your ML models. SageMaker Autopilot automatically inspects raw data, applies feature processors, picks the best set of algorithms, trains and tunes multiple models, tracks their performance, and then ranks the models based on performance, all with just a few clicks. The result is the best performing model that you can deploy at a fraction of the time normally required to train the model. You get full visibility into how the model was created and what's in it and SageMaker Autopilot integrates with Amazon SageMaker Studio. You can explore up to 50 different models generated by SageMaker Autopilot inside SageMaker Studio so its easy to pick the best model for your use case. SageMaker Autopilot can be used by people without machine learning experience to easily produce a model or it can be used by experienced developers to quickly develop a baseline model on which teams can further iterate.

Q: How is Amazon SageMaker Autopilot different from vertical AI services like Amazon Personalize and Amazon Forecast?

While Amazon Personalize and Amazon Forecast specifically target at personalized recommendation and forecasting use cases, Amazon SageMaker Autopilot is a generic automatic machine learning solution for classification and regression problems, such as fraud detection, churn analysis, and targeted marketing. Personalize and Forecast focus on simplifying end to end experience by offering training and model hosting in a bundle. You can train models using Amazon SageMaker Autopilot and get full access to the models as well as the pipelines that generated the models. They can then deploy the models to the hosting environment of their choice, or further iterate to improve model quality.

Q: What built-in algorithms are supported in Amazon SageMaker Autopilot?

Amazon SageMaker Autopilot supports 2 built-in algorithms at launch: XGBoost and Linear Learner.

Q: Does Amazon SageMaker Autopilot support distributed training?

Yes. All Amazon SageMaker Autopilot built-in algorithms support distributed training out of the box.

Q: Can I stop an Amazon SageMaker Autopilot job manually?

Yes. You can stop a job at any time. When an Amazon SageMaker Autopilot job is stopped, all ongoing trials will be stopped and no new trial will be started.

Build Models

Q: What types of notebooks are supported?

Currently, Jupyter notebooks are supported.

Q. How do Amazon SageMaker Notebooks work?

Amazon SageMaker Notebooks, now in preview, provide one-click Jupyter notebooks that you can start working with in seconds. The underlying compute resources are fully elastic, so you can easily dial up or down the available resources and the changes take place automatically in the background without interrupting your work. SageMaker also enables one-click sharing of notebooks. All code dependencies are automatically captured, so you can easily collaborate with others. They'll get the exact same notebook, saved in the same place.

With SageMaker Notebooks you can sign in with your corporate credentials using SSO and start working with notebooks within seconds. Sharing notebooks within and across teams is easy, since the dependencies needed to run a notebook are automatically tracked in environments that are encapsulated with the notebook as it is shared.

Q. How do Amazon SageMaker Notebooks work with other AWS services?

Amazon SageMaker Notebooks give you access to all SageMaker features, such as distributed training, batch transform, hosting, and experiment management. You can access other services such as datasets in Amazon S3, Amazon Redshift, AWS Glue, Amazon EMR, or AWS Lake Formation from SageMaker Notebooks.

Q. What is Amazon SageMaker Ground Truth?

Amazon SageMaker Ground Truth provides automated data labeling using machine learning. SageMaker Ground Truth will first select a random sample of data and send it to Amazon Mechanical Turk to be labeled. The results are then used to train a labeling model that attempts to label a new sample of raw data automatically. The labels are committed when the model can label the data with a confidence score that meets or exceeds a threshold you set. Where the confidence score falls below your threshold, the data is sent to human labelers. Some of the data labeled by humans is used to generate a new training dataset for the labeling model, and the model is automatically retrained to improve its accuracy. This process repeats with each sample of raw data to be labeled. The labeling model becomes more capable of automatically labeling raw data with each iteration, and less data is routed to humans.

Train Models

Q. What is Amazon SageMaker Experiments?

Amazon SageMaker Experiments helps you organize and track iterations to machine learning models. SageMaker Experiments helps you manage iterations by automatically capturing the input parameters, configurations, and results, and storing them as ‘experiments’. You can work within the visual interface of SageMaker Studio, where you can browse active experiments, search for previous experiments by their characteristics, review previous experiments with their results, and compare experiment results visually.

Q. What is Amazon SageMaker Debugger?

Amazon SageMaker Debugger makes the training process more transparent by automatically capturing real-time metrics during training such as training and validation, confusion matrices, and learning gradients to help improve model accuracy.

The metrics from SageMaker Debugger can be visualized in Amazon SageMaker Studio for easy understanding. SageMaker Debugger can also generate warnings and remediation advice when common training problems are detected. With SageMaker Debugger, you can interpret how a model is working, representing an early step towards model explainability.

Q: What is Managed Spot Training?

Managed Spot Training with Amazon SageMaker lets you train your machine learning models using Amazon EC2 Spot instances, while reducing the cost of training your models by up to 90%.

Q: How do I use Managed Spot Training?

You enable the Managed Spot Training option when submitting your training jobs and you also specify how long you want to wait for Spot capacity. Amazon SageMaker will then use Amazon EC2 Spot instances to run your job and manages the Spot capacity. You have full visibility into the status of your training job, both while they are running and while they are waiting for capacity.

Q: When should I use Managed Spot Training?

Managed Spot Training is ideal when you have flexibility with your training runs and when you want to minimize the cost of your training jobs. With Managed Spot Training, you can reduce the cost of training your machine learning models by up to 90%.

Q: How does Manage Spot Training work?

Managed Spot Training uses Amazon EC2 Spot instances for training, and these instances can be pre-empted when AWS needs capacity. As a result, Managed Spot Training jobs can run in small increments as and when capacity becomes available. The training jobs need not be restarted from scratch when there is an interruption as Amazon SageMaker can resume the training jobs using the latest model checkpoint. The built-in frameworks and the built-in computer vision algorithms with SageMaker enable periodic checkpoints, and you can enable checkpoints with custom models.

Q: Do I need to periodically checkpoint with Managed Spot Training?

We recommend periodic checkpoints as a general best practice for long running training jobs. This prevents your Managed Spot Training jobs from restarting if capacity is pre-empted. When you enable checkpoints, Amazon SageMaker resumes your Managed Spot Training jobs from the last checkpoint.

Q: How do you calculate the cost savings with Managed Spot Training jobs?

Once a Managed Spot Training job is completed, you can see the savings in the AWS management console and also calculate the cost savings as the percentage difference between the duration for which the training job ran and the duration for which you were billed.

Regardless of how many times your Managed Spot Training jobs are interrupted, you are charged only once for the duration for which the data was downloaded.

Q: Which instances can I use with Managed Spot Training?

Managed Spot Training can be used with all instances supported in Amazon SageMaker.

Q: Which AWS regions are supported with Managed Spot Training?

Managed Spot Training is supported on all AWS regions where Amazon SageMaker is currently available.

Q: Are there limits to the size of the dataset I can use for training?

There are no fixed limits to the size of the dataset you can use for training models with Amazon SageMaker.

Q: What data sources can I easily pull into Amazon SageMaker?

You can specify the Amazon S3 location of your training data as part of creating a training job.

Q: What algorithms does Amazon SageMaker use to generate models?

Amazon SageMaker includes built-in algorithms for linear regression, logistic regression, k-means clustering, principal component analysis, factorization machines, neural topic modeling, latent dirichlet allocation, gradient boosted trees, sequence2sequence, time series forecasting, word2vec, and image classification. SageMaker also provides optimized Apache MXNet, Tensorflow, Chainer, PyTorch, Gluon, Keras, Horovod, Scikit-learn, and Deep Graph Library containers. In addition, Amazon SageMaker supports your custom training algorithms provided through a Docker image adhering to the documented specification.

Q: What is Automatic Model Tuning?

Most machine learning algorithms expose a variety of parameters that control how the underlying algorithm operates. Those parameters are generally referred to as hyperparameters and their values affect the quality of the trained models. Automatic model tuning is the process of finding a set of hyperparameters for an algorithm that can yield an optimal model.

Q: What models can be tuned with Automatic Model Tuning?

You can run automatic model tuning in Amazon SageMaker on top of any algorithm as long as it's scientifically feasible, including built-in SageMaker algorithms, deep neural networks, or arbitrary algorithms you bring to SageMaker in the form of Docker images.

Q: Can I use Automatic Model Tuning outside of Amazon SageMaker?

Not at this time. The best model tuning performance and experience is within Amazon SageMaker.

Q: What is the underlying tuning algorithm?

Currently, our algorithm for tuning hyperparameters is a customized implementation of Bayesian Optimization. It aims to optimize a customer specified objective metric throughout the tuning process. Specifically, it checks the object metric of completed training jobs, and leverages the knowledge to infer the hyperparameter combination for the next training job.

Q: Will you recommend specific hyperparameters for tuning?

No. How certain hyperparameters impact the model performance depends on various factors and it is hard to definitively say one hyperparameter is more important than the others and thus needs to be tuned. For built-in algorithms within Amazon SageMaker, we do call out whether or not a hyperparameter is tunable.

Q: How long does a hyperparameter tuning job take?

The length of time for a hyperparameter tuning job depends on multiple factors including the size of the data, the underlying algorithm, and the values of the hyperparameters. Additionally, customers can choose the number of simultaneous training jobs and total number of training jobs. All these choices affect how long a hyperparameter tuning job can last.

Q: Can I optimize multiple objectives simultaneously like a model to be both fast and accurate?

Not at this time. Right now, you need to specify a single objective metric to optimize or change your algorithm code to emit a new metric, which is a weighted average between two or more useful metrics, and have the tuning process optimize towards that objective metric.

Q: How much does Automatic Model Tuning cost?

There is no charge for a hyperparameter tuning job itself. You will be charged by the training jobs that are launched by the hyperparameter tuning job, based on model training pricing.

Q: How do I decide to use Amazon SageMaker Autopilot or Automatic Model Tuning?

Amazon SageMaker Autopilot automates everything in a typical machine learning workflow, including feature preprocessing, algorithm selection, and hyperparameter tuning, while specifically focusing on classification and regression use cases. Automatic Model Tuning, on the other hand, is designed to tune any model, no matter it is based on built-in algorithms, deep learning frameworks, or custom containers. In exchange for the flexibility, you have to manually pick the specific algorithm, determine the hyperparameters to tune, and corresponding search ranges.

Q: What is reinforcement learning?

Reinforcement learning is a machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.

Q: Can I train reinforcement learning models in Amazon SageMaker?

Yes, you can train reinforcement learning models in Amazon SageMaker in addition to supervised and unsupervised learning models.

Q: How is reinforcement learning different from supervised learning?

Though both supervised and reinforcement learning use mapping between input and output, unlike supervised learning where the feedback provided to the agent is correct set of actions for performing a task, reinforcement learning uses a delayed feedback where reward signals are optimized to ensure a long-term goal through a sequence of actions.

Q: When should I use reinforcement learning?

While the goal of supervised learning techniques is to find the right answer based on the patterns in the training data and the goal of unsupervised learning techniques is to find similarities and differences between data points. In contrast, the goal of reinforcement learning techniques is to learn how to achieve a desired outcome even when it is not clear how to accomplish that outcome. As a result, RL is more suited to enabling intelligent applications where an agent can make autonomous decisions such as robotics, autonomous vehicles, HVAC, industrial control, and more.

Q: What type of environments can I use for training reinforcement learning models?

Amazon SageMaker RL supports a number of different environments for training reinforcement learning models. You can use AWS services such as AWS RoboMaker, open source environments or custom environments developed using Open AI Gym interfaces, or commercial simulation environments such as MATLAB and SimuLink.

Q: Do I need to write my own RL agent algorithms to train reinforcement learning models?

No, Amazon SageMaker RL includes RL toolkits such as Coach and Ray RLLib that offer implementations of RL agent algorithms such as DQN, PPO, A3C, and many more.

Q: Can I bring my own RL libraries and algorithm implementation and run in Amazon SageMaker RL?

Yes, you can bring your own RL libraries and algorithm implementations in Docker Containers and run those in Amazon SageMaker RL.

Q: Can I do distributed rollouts using Amazon SageMaker RL?

Yes. You can even select a heterogeneous cluster where the training can run on a GPU instance and the simulations can run on multiple CPU instances.

Deploy Models

Q. What is Amazon SageMaker Model Monitor?

Amazon SageMaker Model Monitor allows developers to detect and remediate concept drift. SageMaker Model Monitor automatically detects concept drift in deployed models and provides detailed alerts that help identify the source of the problem. All models trained in SageMaker automatically emit key metrics that can be collected and viewed in SageMaker Studio. From inside SageMaker Studio you can configure data to be collected, how to view it, and when to receive alerts.

Q: Can I access the infrastructure that Amazon SageMaker runs on?

No. Amazon SageMaker operates the compute infrastructure on your behalf, allowing it to perform health checks, apply security patches, and do other routine maintenance. You can also deploy the model artifacts from training with custom inference code in your own hosting environment.

Q: How do I scale the size and performance of an Amazon SageMaker model once in production?

Amazon SageMaker hosting automatically scales to the performance needed for your application using Application Auto Scaling. In addition, you can manually change the instance number and type without incurring downtime through modifying the endpoint configuration.

Q: How do I monitor my Amazon SageMaker production environment?

Amazon SageMaker emits performance metrics to Amazon CloudWatch Metrics so you can track metrics, set alarms, and automatically react to changes in production traffic. In addition, Amazon SageMaker writes logs to Amazon Cloudwatch Logs to let you monitor and troubleshoot your production environment.

Q: What kinds of models can be hosted with Amazon SageMaker?

Amazon SageMaker can host any model that adheres to the documented specification for inference Docker images. This includes models created from Amazon SageMaker model artifacts and inference code.

Q: How many concurrent real-time API requests does Amazon SageMaker support?

Amazon SageMaker is designed to scale to a large number of transactions per second. The precise number varies based on the deployed model and the number and type of instances to which the model is deployed.

Q: What is Batch Transform?

Batch Transform enables you to run predictions on large or small batch data. There is no need to break down the data set into multiple chunks or managing real-time endpoints. With a simple API, you can request predictions for a large number of data records and transform the data quickly and easily

Q: What is Amazon SageMaker Neo?

Amazon SageMaker Neo enables machine learning models to train once and run anywhere in the cloud and at the edge. SageMaker Neo automatically optimizes models built with popular deep learning frameworks that can be used to deploy on multiple hardware platforms. Optimized models run up to two times faster and consume less than a tenth of the resources of typical machine learning models.

Q: How do I get started with Amazon SageMaker Neo?

To get started with Amazon SageMaker Neo, you log into the Amazon SageMaker console, choose a trained model, follow the example to compile models, and deploy the resulting model onto your target hardware platform.

Q: What are the major components of Amazon SageMaker Neo?

Amazon SageMaker Neo contains two major components – a compiler and a runtime. First, the Neo compiler reads models exported by different frameworks. It then converts the framework-specific functions and operations into a framework-agnostic intermediate representation. Next, it performs a series of optimizations. Then, the compiler generates binary code for the optimized operations and writes them to a shared object library. The compiler also saves the model definition and parameters into separate files. During execution, the Neo runtime loads the artifacts generated by the compiler -- model definition, parameters, and the shared object library to run the model.

Q: Do I need to use Amazon SageMaker to train my model in order to use Amazon SageMaker Neo to convert the model?

No. You can train models elsewhere and use Neo to optimize them for Amazon SageMaker ML instances or AWS IoT Greengrass supported devices.

Q: Which models does Amazon SageMaker Neo support?

Currently, Amazon SageMaker Neo supports the most popular deep learning models that power computer vision applications and the most popular decision tree models used in Amazon SageMaker today. Neo optimizes the performance of AlexNet, ResNet, VGG, Inception, MobileNet, SqueezeNet, and DenseNet models trained in MXNet and TensorFlow, and classification and random cut forest models trained in XGBoost.

Q: Which platforms does Amazon SageMaker Neo support?

Currently, Neo supports SageMaker ML.C5, ML.C4, ML.M5, ML.M4, ML.P3, and ML.P2 instances and AWS DeepLens, Raspberry Pi, and Jetson TX1 and TX2 devices, and Greengrass devices-based Intel® Atom and Intel® Xeon CPUs, ARM Cortex-A CPUs, and Nvidia Maxwell and Pascal GPUs.

Q: Do I need to use a specific version of a framework that is supported on the target hardware?

No. Developers can run models using the Amazon SageMaker Neo container without dependencies on the framework.

Q: How much does it cost to use Amazon SageMaker Neo?

You pay for the use of the Amazon SageMaker ML instance that runs inference using Amazon SageMaker Neo.

Q: In which AWS regions is Amazon SageMaker Neo available?

To see a list of support regions, view the AWS region table.

Amazon Augmented AI FAQs

General

Q: What is Amazon Augmented AI (Amazon A2I)?

A: Amazon Augmented AI (Amazon A2I) is a service that makes it easy to build the workflows required for human review of ML predictions. Amazon A2I brings human review to all developers, removing the undifferentiated heavy lifting associated with building human review systems or managing large numbers of human reviewers.

Using Amazon A2I

Q: Why should I use Amazon A2I?

A: Many machine learning applications require humans to review low confidence predictions to ensure the results are correct. For example, extracting information from scanned mortgage application forms can require human review in some cases due to low-quality scans or poor handwriting. But building human review systems can be time consuming and expensive because it involves implementing complex processes or “workflows”, writing custom software to manage review tasks and results, and in many cases, managing large groups of reviewers.

Amazon A2I makes it easy to build and manage human reviews for machine learning applications. Amazon A2I provides built-in human review workflows for common machine learning use cases, such as content moderation and text extraction from documents, which allows predictions from Amazon Rekognition and Amazon Textract to be reviewed easily. You can also create your own workflows for ML models built on Amazon SageMaker or any other tools. Using Amazon A2I, you can allow human reviewers to step in when a model is unable to make a high confidence prediction or to audit its predictions on an ongoing basis.

Q: How do I get started with Amazon A2I?

A: Amazon A2I provides a managed experience where you can set up an entire human review workflow in a few easy steps. To get started with Amazon A2I, sign in to your AWS Console, and navigate to the Amazon SageMaker console. From there, select Human review workflows under Augmented AI. First, as a part of the human review workflow, you provide a pointer to the S3 bucket where the review results should be stored. Next, you select the appropriate task type and define conditions when a human review should be triggered. Amazon A2I provides pre-built workflows where you only need to enter a few choices and provide instructions on how your objects should be reviewed by humans. Alternatively, you can create your own custom workflow and use your own custom review templates. Once created, the workflow can be used directly in your applications using a generated unique identifier for this workflow.

Q: How can I decide what objects are sent for human review?

A: With A2I, you can define what is an acceptable prediction confidence for your business problem. You can define business rules for the machine learning predictions, based on which a human review is triggered. For Amazon Rekognition image moderation tasks, you can use the confidence score that Amazon Rekognition provides for each label it outputs to trigger human review. For Amazon Textract tasks, you can trigger a human review when specific form keys are missing or when form key detection confidence is low. You can also trigger a human review if, after evaluating all form keys in the text, confidence is lower than your required threshold for any form key. For your own custom workflow, you can write the code for business conditions in AWS Lambda or directly in your client application.

Q: How do I access a human workforce using Amazon A2I?

A: With Amazon A2I, you can choose from three workforce options: (1) Amazon Mechanical Turk; (2) Third party data labeling service providers available through the AWS Marketplace; and (3) Your own employees. See the Amazon A2I developer guide for more information.

Pricing and Availability

Q: How much does A2I cost?

A: Please see the Amazon A2I pricing page for the current pricing information.

Q: In which AWS regions is Amazon A2I available?

A: The AWS Region Table lists all the AWS regions where Amazon A2I is currently available.

Amazon CodeGuru FAQs

General

Q: What is Amazon CodeGuru?

Amazon CodeGuru is a fully managed service that helps you proactively improve code quality and application performance with intelligent recommendations.

Q: What can I do with Amazon CodeGuru?

Leveraging machine-learning models and learned best practices, Amazon CodeGuru provides intelligent recommendations to improve code quality and optimize application performance. Amazon CodeGuru includes CodeGuru Reviewer and CodeGuru Profiler. CodeGuru Reviewer analyzes code pull requests on your code repositories and CodeGuru Profiler analyzes the performance of your applications as they run.

CodeGuru Reviewer automatically detects code issues during code reviews before they reach production, allowing you to proactively detect issues before they are deployed to production and improve overall application performance and quality. CodeGuru Profiler can identify when your application is consuming excessive CPU capacity on a logging routine instead of executing core business logic. CodeGuru Profiler is designed to profile your application continuously in production, with a minimal footprint.

Q: How do I get started with Amazon CodeGuru?

Amazon CodeGuru is publicly available in preview. You can start right now in the Amazon CodeGuru console.

To get started with Amazon CodeGuru Reviewer, log in to the CodeGuru Reviewer console where you can associate an existing code repository on

GitHub or AWS CodeCommit. After a one-time setup, CodeGuru Reviewer begins analyzing code and providing code improvement recommendations directly within the pull request or code repository.

You can also start profiling your applications in minutes. To get started with Amazon CodeGuru Profiler, log in to the CodeGuru Profiler console where you can configure your application. Follow the step-by-step instructions and code provided by CodeGuru Profiler to install a small, low-profile agent in your application. You can let CodeGuru Profiler run continuously so it can proactively catch performance issues in your live applications.

Q: In which AWS Regions is Amazon CodeGuru available?

To see supported AWS Regions, please visit the AWS Region Table for all AWS global infrastructure. For more information, see Regions and Endpoints in the AWS General Reference.

Amazon CodeGuru Reviewer

Q: What is Amazon CodeGuru Reviewer?

Amazon CodeGuru Reviewer is an automated code review service that identifies critical defects and deviation from AWS best practices for Java-based code. It scans the lines of code within a pull request or code repository and provides intelligent recommendations based on standards learned from major open source projects as well as Amazon codebase. CodeGuru Reviewer seamlessly integrates with existing code review workflows on widely-used source control systems, such as AWS CodeCommit and GitHub, and provides actionable recommendations for improving code quality.

Q: What programming languages and source code repositories are supported?

Amazon CodeGuru Reviewer currently supports Java code stored in GitHub and AWS CodeCommit repositories.

Q: What type of issues are detected by Amazon CodeGuru Reviewer?

Amazon CodeGuru Reviewer checks for concurrency issues, potential race conditions, un-sanitized inputs, inappropriate handling of sensitive data such as credentials, resource leaks, and also detects race conditions in concurrent code.

Q: How do I get started with Amazon CodeGuru Reviewer?

Visit the Amazon CodeGuru console to integrate CodeGuru Reviewer recommendations directly within your code pull requests. You can get started by visiting the CodeGuru console and following the steps to associate your AWS CodeCommit and GitHub repositories to start receiving CodeGuru Reviewer's recommendations. CodeGuru Reviewer will need read-only access and the ability to post comments on the Pull Requests. Once enabled, CodeGuru Reviewer will automatically provide intelligent recommendations as comments on your pull requests generated for the connected repositories.

Q: Does Amazon CodeGuru Reviewer access my code?

Amazon CodeGuru Reviewer needs read-only access to your code for the purpose of generating recommendations. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate controls, including encryption in transit, to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see the Data Privacy FAQ for more information.

Q: Does Amazon CodeGuru Reviewer persist a copy of my code?

No, Amazon CodeGuru Reviewer does not store your source code.

Q: How is Amazon CodeGuru Reviewer trained to provide intelligent recommendations?

Amazon CodeGuru Reviewer is trained using rule mining and supervised machine learning models that use a combination of logistic regression and neural networks.

For example, during training for deviation from AWS and industry best practices, CodeGuru Reviewer mines Amazon code bases using search techniques and locality sensitive models for pull requests that include AWS API calls. It looks at code changes intended to improve the quality of the code, and cross-references them against documentation data. The result is the creation of a new set of rules that Reviewer recommends to you as best practices when it reviews your code.

During training for resource and sensitive data leaks, it does a full code analysis for all code paths that use the resource or sensitive data, creates a feature set representing those, and then uses those as inputs for logistic regression models and convolutional neural networks (CNNs).

For both rule-based and machine learning-based models, CodeGuru Reviewer uses the feedback you provide as labels and iteratively improves the quality of code detectors.

Amazon CodeGuru Profiler

Q: What is Amazon CodeGuru Profiler?

Amazon CodeGuru Profiler helps developers easily understand the runtime behavior of their applications, improve performance, and decrease compute costs. CodeGuru Profiler analyzes the application runtime profile and provides intelligent recommendations and visualizations that guide developers on how to improve the performance of the most relevant parts of their code.

Q: I already have extensive logging and debugging integrated into my code. Do I still need to profile?

While “top-down” (logging and debugging) monitoring approach is a good practice, it is impractical for code efficiency analysis. Logging execution time only works for a limited set of scenarios, and it’s time-consuming to implement. This is where “bottom-up” runtime monitoring comes in: Amazon CodeGuru Profiler is designed to collect data on everything that happened in that application’s behavior, regardless of scenarios. CodeGuru Profiler uses a knowledge base of commonly encountered performance inefficiencies to automatically discover code patterns in your live application that impact its performance. Developers can then follow the provided recommendations to fix the issues.

Q: How does CodeGuru Profiler differ from traditional APMs and standalone profilers?

Traditional APMs provide useful data on monitoring, tracing and application performance. Amazon CodeGuru Profiler complements these APM capabilities by providing visualization of the application’s runtime data as well as actionable recommendations for the performance issues it discovers. CodeGuru Profiler enables you to easily see the parts of code that present the biggest opportunity for performance optimization and receive guidance on how to address them without the need to have a deep performance engineering background. Furthermore, some standalone profilers are designed to only run in test environments while CodeGuru Profiler was designed to continuously run in production, under production traffic loads, and without impact to the application. This is useful when troubleshooting operational issues in production, including when running on bare metal hosts.

Q: What types of applications can I profile?

Amazon CodeGuru Profiler works with applications hosted on Amazon EC2, containerized applications running on Amazon ECS and Amazon EKS, as well as serverless applications running on AWS Fargate.

Q: What programming languages are supported?

Amazon CodeGuru Profiler currently supports Java applications.

Q: How does CodeGuru Profiler work?

CodeGuru Profiler consists of three parts: an agent, the profiler service, and intelligent recommendations. The agent runs as an in-process thread as part of your application. It takes data from each of your service instances running the agent and sends them to the profiler service every 5 minutes, which then aggregates them. CodeGuru Profiler then publishes the profile data in interactive flame graphs that enable you to visualize the performance of your application. CodeGuru Profiler also continuously scans the profiled data and compares it against Amazon and performance engineering best practices and proactively alerts you with intelligent recommendations when performance issues are discovered.

Q: What is a profiling group?

A profiling group is a logical grouping created by you. It represents the boundary of one application. For example, in a microservices architecture, a profiling group would aggregate the profiles of the microservices that you have assigned to it, and produce one profile for all of them.

Amazon Comprehend FAQs

General

Q: What is Natural Language Processing?

Q: What is Amazon Comprehend?

Q: What can I do with Amazon Comprehend?

Q: How do I get started with Amazon Comprehend?

Q: What are the most common use cases for Amazon Comprehend?

Q: Do I have to be a natural language processing expert to use Amazon Comprehend?

Q: Is Amazon Comprehend a managed service?

Q: Does Amazon Comprehend learn over time?

Q: In which AWS regions in Amazon Comprehend available?

Q: What security measures does Amazon Comprehend have?

Usage

Q: Where do I store my data?

Q: How do I know if the service can process my data?

Q: How do I know if Amazon Comprehend is giving accurate results?

Q: Can I import or use my own NLP model with Amazon Comprehend?

Pricing

Q: How is Amazon Comprehend priced?

Data privacy

Q: Are text inputs processed by Amazon Comprehend stored, and how are they used by AWS?

Q: Who has access to my content that is processed and stored by Amazon Comprehend?

Q: Do I still own my content that is processed and stored by Amazon Comprehend?

Q: Is the content processed by Amazon Comprehend moved outside the AWS region where I am using Amazon Comprehend?

Q: Can I use Amazon Comprehend in connection with websites, programs or other applications that are directed or targeted to children under age 13 and subject to the Children's Online Privacy Protection Act (COPPA)?

Q: How do I determine whether my website, program, or application is subject to COPPA?

Amazon Comprehend Medical

Q: What is Amazon Comprehend Medical?

Q: What can I do with Amazon Comprehend Medical?

Q: How do I get started with Amazon Comprehend Medical?

Q: Do I have to be a natural language processing (NLP) expert to use Amazon Comprehend Medical?

Q: Does Amazon Comprehend Medical learn over time?

Q: In which AWS regions in Amazon Comprehend Medical available?

Q: What security measures does Amazon Comprehend Medical have?

Q: What else should I know before using the Amazon Comprehend Medical service?

Q: How is Amazon Comprehend Medical priced?

Q: Is the content processed by Amazon Comprehend Medical used for any purpose other than to provide and maintain the service?

Q: Can I use Amazon Comprehend Medical in connection with websites, programs or other applications that are directed or targeted to children under age 13 and subject to the Children's Online Privacy Protection Act (COPPA)?

Q: How do I determine whether my website, program, or application is subject to COPPA?

Amazon Elastic Inference FAQs

General

Q: What is Amazon Elastic Inference?

A: Amazon Elastic Inference (Amazon EI) is an accelerated compute service that allows you to attach just the right amount of GPU-powered inference acceleration to any Amazon EC2 or Amazon SageMaker instance type or Amazon ECS task. This means you can now choose the instance type that is best suited to the overall compute, memory, and storage needs of your application, and then separately configure the amount of inference acceleration that you need.

Q: What are Amazon Elastic inference accelerators?

A: Amazon Elastic inference accelerators are GPU-powered hardware devices that are designed to work with any SageMaker instance type or EC2 instance type or ECS task to accelerate deep learning inference workloads at a low cost. When you launch an EC2 instance or an ECS task with Amazon Elastic Inference, an accelerator is provisioned and attached to the instance over the network. Deep learning tools and frameworks like TensorFlow Serving and Apache MXNet, that are enabled for Amazon Elastic Inference, can automatically detect and offload model computation to the attached accelerator.

Q: What is the difference between the Amazon Elastic inference accelerator family types?

A: The EIA2 accelerators have twice the GPU memory of equivalent EIA1 accelerators. You can determine your GPU memory needs based on your model and tensor input sizes and choose the right accelerator family and type for your needs.

Configuring

Q: How do I provision Amazon Elastic Inference accelerators?

A: You can configure Amazon SageMaker endpoints or Amazon EC2 instances or Amazon ECS tasks with Amazon Elastic Inference accelerators using the AWS management console, AWS command line interface (CLI), or the AWS SDK. There are two requirements for launching EC2 instances with accelerators. First, you will need to provision an AWS PrivateLink VPC Endpoint for the subnets where you plan to launch accelerators. Second, as you launch an instance, you need to provide an instance role with a policy that allows users accessing the instance to connect to accelerators. When you configure an instance to launch with Amazon EI, an accelerator is provisioned in the same Availability Zone behind the VPC endpoint.

Q: What model formats does Amazon Elastic Inference support?

A: Amazon Elastic Inference supports models trained using TensorFlow, Apache MXNet, and ONNX models.

Q: Can I deploy models on Amazon Elastic Inference using TensorFlow or Apache MXNet frameworks?

A: Yes, you can use AWS-enhanced TensorFlow Serving and Apache MXNet libraries to deploy models and make inference calls.

Q: How do I get access to AWS optimized frameworks?

A: The AWS Deep Learning AMIs include the latest releases of TensorFlow Serving and Apache MXNet that are optimized for use with Amazon Elastic Inference accelerators. You can also obtain the libraries via Amazon S3 to build your own AMIs or container images. Please see our documentation for more information.

Q: Can I use CUDA with Amazon Elastic Inference accelerators?

A: No. You can only use either the AWS-enhanced TensorFlow Serving or Apache MXNet libraries as an interface to Amazon Elastic Inference accelerators.

Pricing and billing

Q: How am I charged for Amazon Elastic Inference?

A: You pay only for the Amazon Elastic Inference accelerator hours you use. For more details, see the pricing page.

Q: Will I incur charges for AWS PrivateLink VPC Endpoints for the Amazon Elastic Inference service?

No. You will not incur charges for VPC Endpoints to the Amazon Elastic Inference service, as long as you have at least one instance configured with an accelerator, running in an Availability Zone where a VPC endpoint is provisioned.

Amazon Forecast FAQs

General

Q: What is Amazon Forecast?

Amazon Forecast is a fully managed service that uses machine learning to deliver highly accurate forecasts.

Based on the same technology used at Amazon.com, Amazon Forecast uses machine learning to combine time series data with additional variables to build forecasts. Amazon Forecast requires no machine learning experience to get started. You only need to provide historical data, plus any additional data that you believe may impact your forecasts. For example, the demand for a particular color of a shirt may change with the seasons and store location. This complex relationship is hard to determine on its own, but machine learning is ideally suited to recognize it. Once you provide your data, Amazon Forecast will automatically examine it, identify what is meaningful, and produce a forecasting model capable of making predictions that are up to 50% more accurate than looking at time series data alone.

Amazon Forecast is a fully managed service, so there are no servers to provision, and no machine learning models to build, train, or deploy. You pay only for what you use, and there are no minimum fees and no upfront commitments.

Learn more >>

Q: What is time series data?

A time series is a set of data points that are ordered by some unit of time. Examples of time series are weekly sales of a product, daily inventory levels, and hourly website visits.

Q: What is time series forecasting?

Time series forecasting is a technique that predicts the future time series data based on historical data.

Q: What are key use cases for Amazon Forecast?

Amazon Forecast can be used to forecast any time series data, such as retail demand, manufacturing demand, travel demand, revenue, IT capacity, logistics, and web traffic.

Q: How do I get started with Amazon Forecast?

You can get started with Amazon Forecast using an API or AWS Console. The first step is to upload your data into Amazon Forecast. Once data is uploaded, you can have Amazon Forecast automatically try all different algorithms to train multiple models, then provide the model with the highest forecasting accuracy. You can also manually choose one of the forecasting algorithms to train a model. Once you have the model, Amazon Forecast provides comprehensive accuracy metrics to evaluate the performance of the model. If you are satisfied, you can deploy the model within Amazon Forecast to generate forecasts with a single click or API call. Amazon Forecasts can be retrieved via API, exported in CSV format, or visualized in the console.

Learn more >>

Q: What data does Amazon Forecast require to start forecasting?

Amazon Forecast can provide a forecast given any historical set of time series data. However, customers can also provide meta-data available for each of the time series (e.g., the location of a house when predicting energy consumption) as well as other related time series data (e.g., historical pricing data along with sales data for products).

Learn more >>

Q: What are other datasets available to Amazon Forecast to make more accurate forecasts?

In addition to the data provided by customers, Amazon Forecast provides a set of international holiday calendars which customers can add to their dataset.

Algorithms Customization

Q: Can I customize Amazon Forecast algorithms?

Yes. Customers have the ability to vary recipe specific parameters (e.g. HPO parameters) to optimize their model. For more details, refer to the documentation page.

Pricing and Availability

Q: How much does Amazon Forecast cost?

Refer to Amazon Forecast pricing page to learn more.

Q: What languages does Amazon Forecast support?

Amazon Forecast supports English.

Data Privacy

Q: Who has access to my content that is processed and stored by Amazon Forecast?

Only authorized employees will have access to your content that is processed by Amazon Forecast. Your trust, privacy, and the security of your content are our highest priority, and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information.

Q: Do I still own my content that is processed and stored by Amazon Forecast?

You always retain ownership of your content, and we will only use your content with your consent.

Learn how to get started

Refer to developer guide for instructions on using Amazon Forecast.

Learn more »

Sign up for a free account

Instantly get access to the AWS Free Tier.

Sign up »

Sign up for the Amazon Forecast Preview

Get started building with Amazon Forecast by signing up for preview access.

Sign in »

Amazon Fraud Detector FAQs

Q: What is Amazon Fraud Detector?

Amazon Fraud Detector is a fully managed service that makes it easy to identify potentially fraudulent online activities such as online payment fraud and the creation of fake accounts. Fraud Detector uses machine learning (ML) and 20 years of fraud detection expertise from AWS and Amazon.com to automatically identify potential fraudulent activity in milliseconds. There are no up-front payments or long-term commitments and no infrastructure to manage with Fraud Detector; you pay only for your actual usage.

Q: How does Amazon Fraud Detector work?

First, you select a machine learning model template, which specifies a combination of features and algorithms optimized to detect a specific form of online fraud. Next, you upload your historical fraud data to automatically train, test, and deploy a customized fraud detection model based on your unique information. During this process, a series of models that have learned patterns of fraud from AWS and Amazon's own fraud expertise are used to boost the customer's model performance. The output is a prediction in the form of a score ranging from 0 to 1,000 that predicts the likelihood of fraud risk. At the final stage of the process, you setup decision logic (e.g. rules) to interpret your model's score and assign outcomes such as pass or send transaction to a human investigator for review.

After this framework is created, you can integrate the Fraud Detector API into your website's transactional functions, such as account sign up or during order checkout. Fraud Detector will process these activities in real-time and provide fraud predictions in milliseconds. You can then adjust your end-user experience according based on this prediction.

Q: Is Amazon Fraud Detector generally available?

Amazon Fraud Detector is now available in preview. Please submit a request here to join the preview. We will contact you with instructions once your submission is approved by our team. You will need an AWS Account ID in order to apply.

Q: What specific use cases does Amazon Fraud Detector support?

Fraud Detector is designed for online fraud use cases that require real-time evaluation using machine learning models and rules. For example:

- Online identity fraud
- Payment fraud for online orders
- New account fraud, within an account sign-up process
- Account takeover (when bad actors use stolen credentials to log in to a legitimate customer's account)
- Promotion code abuse
- Seller performance evaluations in online marketplaces

Q: Can I customize Amazon Fraud Detector's configuration for my use case?

Yes. You can customize Amazon Fraud Detector for each use case, using a mix of Fraud Detector ML models, SageMaker models, and rules. You first gather the relevant risk data to be used as inputs for fraud evaluations, such email addresses, phone numbers, and IP addresses. This data will feed into a machine learning model, which outputs a score. You can then use a set of detection rules to interpret the score and other risk data to make decisions, such as approving, or sending orders to fraud analysts for investigation. An example of a simple rule and corresponding outcome could be "IF model_score < 50 & credit_card_country = US THEN approve_order".

Q: How do I use Amazon Fraud Detector to tap into Amazon's fraud detection data and expertise?

With 20 years of fraud experience, Amazon has seen firsthand how bad actors conduct various forms of online fraud. Amazon Fraud Detector helps you tap into this knowledge. During the automated model training process, Fraud Detector uses a series of models that have learned patterns of fraud from AWS and Amazon's own fraud expertise to boost your model's performance.

Q: How does Amazon Fraud Detector use ML to improve my fraud detection?

Amazon Fraud Detector automatically trains, tests, and deploys custom fraud detection machine learning models based on your historical fraud data, with no machine learning experience required. For developers with more machine learning experience, you can add your own models to Amazon Fraud Detector using Amazon SageMaker.

Q: How do I setup fraud detection rules using Amazon Fraud Detector?

Amazon Fraud Detector makes it possible to perform rule-based fraud predictions with or without using machine learning. With Amazon Fraud Detector, you can author detection rules (e.g. “IF model_score < 50 & credit_card_country = US THEN approve_order”) using a simple rule writing language. You can also specify the order that rules trigger during an evaluation using an intuitive interface.

Q: Can my team review fraud evaluations using Amazon Fraud Detector?

Yes, you can review your past fraud evaluations to audit decision logic using the Amazon Fraud Detector console. In the Amazon Fraud Detector console, you can search for past events based on characteristics of the event and/or the detection logic applied, such as the outcome, models or rules used, or event metadata. You can then drill-down into how the detection logic assessed an event.

Q: Does Amazon Fraud Detector share my risk data and risk decisions with other companies?

No. Security and privacy are our top concerns. As a fundamental tenet of earning customer’s trust, AWS will never share customer data.

Amazon Kendra FAQs

Q: What is Amazon Kendra?

A: Amazon Kendra is a highly accurate and easy to use enterprise search service that's powered by machine learning. Kendra enables developers to add search capabilities to their applications so their end users can discover information stored within the vast amount of content spread across their company. This includes data from manuals, research reports, FAQs, HR documentation, customer service guides, and is found across various systems such as file systems, web sites, Box, DropBox, Salesforce, SharePoint, relational databases, Amazon S3, and more. When you type a question, the service uses machine learning algorithms to understand the context and return the most relevant results, whether that be a precise answer or an entire document. For example, you can ask a question like "How much is the cash reward on the corporate credit card?" and Kendra will map to the relevant documents and return a specific answer like "2%". Kendra provides sample code so that you can get started quickly and easily integrate highly accurate search into your new or existing applications.

Q: How does Amazon Kendra work with other AWS services?

A: Amazon Kendra provides ML-powered search capabilities for all unstructured data customers store in AWS. Kendra offers easy integration with popular repository types such as file systems, web sites, Box, DropBox, Salesforce, SharePoint, relational databases, Amazon S3 so that you can index your content with a few clicks, and search your information without writing a single line of code.

Q: What types of questions can I ask Amazon Kendra?

A: Amazon Kendra supports the following common types of questions:

- Factoid questions (who, what, when, where): "Who is Amazon's CEO?" or "When is Prime Day 2019?". These require fact-based answers that may be returned in the form of a single word or phrase. The precise answer, however, must be explicitly stated in the ingested text content.

- Descriptive questions: “How do I connect my Echo Plus to my network?” or “How do I obtain tax benefits for lower income families?” where the answer could be a sentence, passage, or an entire document.
- Keyword searches: “Fitbit waterproof” or “Honda warranty”, where the intent and scope are not very clear. Kendra will leverage its deep learning models to return relevant documents in these cases.

Q: What if my data does not contain the precise answer Amazon Kendra is looking for?

A: When your data does not contain a precise answer to a question, Kendra returns a list of the most relevant documents ranked by its deep learning models.

Q: What types of questions will Amazon Kendra be unable to answer?

A: Kendra does not yet support questions for which the answers require cross-document passage aggregation or calculations.

Q: How do developers get up and running with Amazon Kendra?

A: The Kendra console provides the easiest way to get started. You can point Kendra at unstructured and semi-structured documents like FAQs stored in S3. After ingestion, you can start testing Kendra by typing queries directly in the “search” section of the console. You can then deploy Kendra search in your own application with a few lines of code, or by copying short code samples from the reference implementations to replicate specific functionality.

Q: How can I customize Amazon Kendra to better fit my company’s domain or business specialty?

A: Kendra offers domain specific expertise for IT, pharma, insurance, oil and gas, industrial, financial services, legal, media and entertainment, travel and hospitality, health, human resources, news, telecommunications, mining, food and beverage, and automotive. You can extend beyond this by providing your own synonym lists (coming soon). You just upload a file with your specific terminology and Kendra will use these synonyms to enrich user searches.

Q: What file types does Amazon Kendra support?

A: Kendra supports unstructured and semi-structured data in .html, MS Office (.doc, .ppt), PDF, and text formats.

Q: How does Amazon Kendra handle incremental data updates?

A: Kendra provides two methods of keeping your index up to date. First, connectors provide scheduling to automatically sync your data sources on a regular basis. Second, the Kendra API allows you to build your own connector to send data directly to Kendra from your data source via your existing ETL jobs or applications.

Q: What languages does Amazon Kendra support?

A: Kendra supports US English.

Q: What code changes do I need to make to use Amazon Kendra?

A: Ingesting content does not require coding when using the native connectors. You can also write your own custom connectors to integrate with other data sources, using the Kendra SDK. You can easily deploy search in your application with a few lines of code, or copying short code samples from reference implementations to replicate specific functionality. The SDK provides full control and flexibility of the end user experience.

Q: In what regions is Amazon Kendra available?

A: Kendra is in a public preview. It is currently available in the following AWS regions: Northern Virginia, Oregon, and Ireland. See the AWS Region Table for more details.

Q: How does Amazon Kendra improve its machine learning models?

A: Amazon Kendra's machine learning models are regularly retrained and tuned for each customer by analyzing and incorporating end-user search patterns and feedback. Improvements to a customer’s models are not shared with other customers.

Q: Does Amazon Kendra use my content to improve machine learning models for other customers?

A: Content processed by Amazon Kendra is not used to develop or improve the quality of Amazon Kendra for other customers or other Amazon machine-learning/artificial-intelligence technologies.

Amazon Lex FAQs

General

Q: What is Amazon Lex?

Q. How can I get started with Amazon Lex?

Q. What are the most common use cases for Amazon Lex?

Q. How does Amazon Lex work with other AWS services?

Q. Do I have to be a machine learning expert to use Amazon Lex?

Q. In which AWS regions is Amazon Lex available?

Q. What is the maximum bandwidth supported on Amazon Lex?

Q: Is Amazon Lex a managed service?

Q. When do I use Amazon Polly vs. Amazon Lex?

Q. Does Amazon Lex get more intelligent over time?

Q: I was in the Amazon Lex preview program. Now that Amazon Lex is GA, what happens to my account?

Bot Structure

Q: How do I create a bot in Amazon Lex?

Q. Can I implement business logic on the client?

Q. How can I validate user input?

Q. What is an Intent?

Q. What is an utterance?

Q. What are slots?

Q. What are prompts?

Q. How is an action fulfilled?

Q. How do I monitor and track my bot?

Build and Test

Q: What happens when I ‘build’ a bot?

Q. How can I test an Amazon Lex bot?

Deployment

Q. How can I create Amazon Lex bots for mobile?

Q. How can I make Amazon Lex bots available on messaging services?

Q. Do I have to submit my bot for certification prior to deployment?

Q. Can I have an Amazon Lex bot version deployed for use by end users while I continue to develop on a different version?

Q. Can I choose different versions while deploying to different messaging services?

Q. What is the maximum duration of speech input?

Q. Can I configure for speech input and text output?

Q. How many languages are supported on Amazon Lex?

Q. What audio formats does Amazon Lex support?

Q. Can I use Amazon Lex in VPC?

Q. Can I access Amazon Lex bots locally i.e. without an Internet connection?

Amazon Alexa

Q. How is this different from Alexa Skills Kit?

Q. Do I need a wake word to invoke an Amazon Lex intent?

Q. Can an Amazon Lex bot respond using Alexa’s voice?

Q. Can I create an Alexa Skill from an Amazon Lex bot?

Q: When exporting my Amazon Lex bot schema to use in an Alexa skill, are my AWS Lambda functions exported and included in the bot schema?

Q: I have created an Alexa Skill from an Amazon Lex bot using the schema export feature. Which Alexa platforms support the Amazon Lex bot schema?

Data and Security

Q: Are voice and text inputs processed by Amazon Lex stored, and how are they used by AWS?

Q: Can I delete voice and text inputs stored by Amazon Lex?

Q: Who has access to my content that is processed and stored by Amazon Lex?

Q: Do I still own my content that is processed and stored by Amazon Lex?

Q: Is the content processed by Amazon Lex moved outside the AWS region where I am using Amazon Lex?

Q: Can I use Amazon Lex in connection with websites, programs or other applications that are directed or targeted to children under age 13 and subject to the Children's Online Privacy Protection Act (COPPA)?

Q: How do I determine whether my website, program, or application is subject to COPPA?

SDK

Q. What SDKs are supported for Amazon Lex?

Q. Can I use SDKs to build bots?

AWS Mobile Hub Integration

Q. Which enterprise connectors are supported on Amazon Lex?

Support and Billing

Q. What support is provided for Amazon Lex?

Q. How does Amazon Lex count the number of requests?

Q. How much does Amazon Lex cost?

Q. Does Amazon Lex participate in the AWS Free Tier?

Q: I was in the Amazon Lex preview program. Now that Amazon Lex is GA, what happens to my account?

Amazon Personalize FAQs

General

Q: Why should I use Amazon Personalize?

A: Amazon Personalization has been empirically proven to increase key user engagement and revenue metrics in diverse industries. A market study of 1.5 billion shopping sessions across e-commerce businesses revealed 11.5% incremental revenue attributed to recommendations. For example, 30% of page views on Amazon.com are driven by recommendations.

Q: What are the key use cases supported by Amazon Personalize

A: Amazon Personalize supports the following key use cases:

- Personalized recommendations
- Similar items
- Personalized reranking i.e. rerank a list of items for a user
- Personalized promotions/notifications

Q: What are some of the common business applications for Amazon Personalize

A: Amazon Personalize can be used to personalize the end-user experience over any digital channel. Examples include product recommendations for e-commerce, news articles and content recommendation for publishing, media and social networks, hotel recommendations for travel websites, credit card recommendations for banks, and match recommendations for dating sites. These recommendations and personalized experiences can be delivered over websites, mobile apps, or email/messaging. Amazon Personalize can also be used to customize the user experience when user interaction is over a physical channel, e.g., a meal delivery company could personalize weekly meal to users in a subscription plan.

Using Amazon Personalize

Q: How do I get started with Amazon Personalize?

A: Developers get started by creating an account and accessing the Amazon Personalize developer console which walks them through an intuitive set-up wizard. Developers have the option of using a JavaScript API and Server-Side SDKs to send real-time activity stream data to Amazon Personalize or bootstrapping the service using a historical log of user events. Developers can also import their catalog (item dataset) and user data via Amazon S3. Then, with only a few API calls, developers can train a personalization model, either by letting the service choose the right algorithm for their dataset with AutoML or manually choosing one of the several algorithm options available. Once trained, the models can be deployed with a single API call and can then be used by production applications. When deployed, developers call the service from their production services to get real-time recommendations, and Amazon Personalize will automatically scale to meet demand.

Q: What data do I have to provide to Amazon Personalize?

A: Developers should provide the following data to Amazon Personalize:

- User activity stream or event data - User interaction data on the website/application is captured in the form of events and is sent to Amazon Personalize often via an integration that involves a single line of code. This includes key events such as click, buy, watch, add-to-shopping cart, like etc. When onboarding to the service, developers can also provide a historical log of all event/activity stream data, if available.
- Catalog (item) data - This can be any type of catalog including books, videos, news articles or products. This involves item ids and meta-data associated with each item. This data is optional.
- User data - User profile data including user demographic data such as gender and age. This data is optional.

Amazon Personalize will train and deploy a model based on this data. Developers can then use a simple inference API to get individualized recommendations at run-time and generate a personalized experience for the end users according to the type of personalization model (e.g. user personalization, related items or personalized reranking).

Q: How do I apply/export Amazon Personalize recommendations to my business workflows or applications?

A: Amazon Personalize provide customers two inference APIs: getRecommendations and getPersonalizedRanking. These APIs return a list of recommended itemIDs for a user, a list of similar items for an item or a reranked list of items for a user. The itemID can be a product identifier, videoID etc. The customers are then expected to use these itemIDs to generate the end user experience through steps, such as fetching image and description, and then rendering a display. In some cases, customers might integrate with AWS or third party email delivery services, or notification services etc. to generate the end user experience.

Q: How is the effectiveness of a personalization measured in Amazon Personalize?

A: The effectiveness of a personalization solution can be measured by first specifying a business goal the customer wants to optimize and then measuring the impact on this goal via an A/B test. These goals typically are click-thru rate, revenue per user, revenue per session, average time spent per session, retention rates etc. Amazon Personalize also provides offline metrics for the models trained on the customer data.

Pricing

Q: What does Amazon Personalize cost?

A: Refer to the Amazon Personalize pricing page to learn more.

Amazon Polly FAQs

General

Q. What is Amazon Polly?

Q. Why should I use Amazon Polly?

Q. What features are available?

Q: What are Speech Marks?

Q. What are the most common use cases for this service?

Q. How does this product work with other AWS products?

Q. What are the advantages of a cloud-based Text-to-Speech solution over an on-device one?

Q. How do I get started with Amazon Polly?

Q. In which regions is the service available?

Q. Which programming languages are supported?

Q. Which audio formats are supported?

Q. Which languages are supported?

Q: Does Amazon Polly have AWS service limits?

Q: Is Amazon Polly a HIPAA Eligible Service?

Q: How do I get started with Amazon Polly Brand Voice?

Q: What is the cost and timeline to build a Brand Voice?

Billing

Q. How much does Amazon Polly cost?

Q. Can I use the service for generating static voice prompts that will be replayed multiple times?

Q. Can I use the service to generate content that will be used in mass notification systems (for example on train station)?

Q. If I request 1,000 characters to be synthesized and request Speech Marks with the same 1,000 characters, will I be charged for 2,000 characters?

Q. Does Amazon Polly participate in the AWS Free Tier?

Q. Do your prices include taxes?

Data Privacy

Q. Are text inputs processed by Amazon Polly stored, and how are they used by AWS?

Q: Who has access to my content that is processed and stored by Amazon Polly?

Q: Do I still own my content that is processed and stored by Amazon Polly?

Q: Is the content processed by Amazon Polly moved outside the AWS region where I am using Amazon Polly?

Q: Can I use Amazon Polly in connection with websites, programs or other applications that are directed or targeted to children under age 13 and subject to the Children’s Online Privacy Protection Act (COPPA)?

Amazon Rekognition FAQs

General

Q: What is Amazon Rekognition?

Amazon Rekognition is a service that makes it easy to add powerful visual analysis to your applications. Rekognition Image lets you easily build powerful applications to search, verify, and organize millions of images. Rekognition Video lets you extract motion-based context from stored or live stream videos and helps you analyze them.

Rekognition Image is an image recognition service that detects objects, scenes, and faces; extracts text; recognizes celebrities; and identifies inappropriate content in images. It also allows you to search and compare faces. Rekognition Image is based on the same proven, highly scalable, deep learning technology developed by Amazon’s computer vision scientists to analyze billions of images daily for Prime Photos.

Rekognition Image uses deep neural network models to detect and label thousands of objects and scenes in your images, and we are continually adding new labels and facial recognition features to the service. With Rekognition Image, you only pay for the images you analyze and the face metadata you store.

Rekognition Video is a video recognition service that detects activities; understands the movement of people in frame; and recognizes objects, celebrities, and inappropriate content in videos stored in Amazon S3 and live video streams from Acuity. Rekognition Video detects persons and tracks them through the video even when their faces are not visible, or as the whole person might go in and out of the scene. For example, this could be used in an application that sends a real-time notification when someone delivers a package to your door. Rekognition Video allows you also to index metadata like objects, activities, scene, celebrities, and faces that make video search easy.

Q: What is deep learning?

Deep learning is a sub-field of Machine Learning and a significant branch of Artificial Intelligence. It aims to infer high-level abstractions from raw data by using a deep graph with multiple processing layers composed of multiple linear and non-linear transformations. Deep learning is loosely based on models of information processing and communication in the brain. Deep learning replaces handcrafted features with ones learned from very large amounts of annotated data. Learning occurs by iteratively estimating hundreds of thousands of parameters in the deep graph with efficient algorithms.

Several deep learning architectures such as convolutional deep neural networks (CNNs), and recurrent neural networks have been applied to computer vision, speech recognition, natural language processing, and audio recognition to produce state-of-the-art results on various tasks.

Amazon Rekognition is a part of the Amazon AI family of services. Amazon AI services use deep learning to understand images, turn text into lifelike speech, and build intuitive conversational text and speech interfaces.

Q: Do I need any deep learning expertise to use Amazon Rekognition?

No. With Amazon Rekognition, you don’t have to build, maintain or upgrade deep learning pipelines.

To achieve accurate results on complex computer vision tasks such as object and scene detection, face analysis, and face recognition, deep learning systems need to be tuned properly and trained with massive amounts of labeled ground truth data. Sourcing, cleaning, and labeling data accurately is a time-consuming and expensive task. Moreover, training a deep neural network is computationally expensive and often requires custom hardware built using Graphics Processing Units (GPU).

Amazon Rekognition is fully managed and comes pre-trained for image and video recognition tasks, so that you don’t have invest your time and resources on creating a deep learning pipeline. Amazon Rekognition continues to improve the accuracy of its models by building upon the latest research and sourcing new training data. This allows you to focus on high-value application design and development.

Q: What are the most common use cases for Amazon Rekognition?

The most common use-cases for Rekognition Image include:

- Searchable Image Library
- Face-Based User Verification
- Sentiment Analysis
- Facial Recognition
- Image Moderation

The most common use-cases for Rekognition Video include:

- Search Index for video archives
- Easy filtering of video for explicit and suggestive content

Q: How do I get started with Amazon Rekognition?

If you are not already signed up for Amazon Rekognition, you can click the "Try Amazon Rekognition" button on the Amazon Rekognition page and complete the sign-up process. You must have an Amazon Web Services account; if you do not already have one, you will be prompted to create one during the sign-up process. Once you are signed up, try out Amazon Rekognition with your own images and videos using the Amazon Rekognition Management Console or download the Amazon Rekognition SDKs to start creating your own applications. Please refer to our step-by-step Getting Started Guide for more information.

Q: What APIs does Amazon Rekognition offer?

Amazon Rekognition Image offers APIs to detect objects and scenes, detect and analyze faces, recognize celebrities, detect inappropriate content, and search for similar faces in a collection of faces, along with APIs to manage resources. Rekognition Image also offers APIs to compare faces and extract text, while Rekognition Video also offers APIs to track persons and manage live stream video from Acuity. For details, please refer to the Amazon Rekognition API Reference.

Q: What image and video formats does Amazon Rekognition support?

Amazon Rekognition Image currently supports the JPEG and PNG image formats. You can submit images either as an S3 object or as a byte array. Amazon Rekognition Video operations can analyze videos stored in Amazon S3 buckets. The video must be encoded using the H.264 codec. The supported file formats are MPEG-4 and MOV. A codec is software or hardware that compresses data for faster delivery and decompresses received data into its original form. The H.264 codec is commonly used for the recording, compression and distribution of video content. A video file format may contain one or more codecs. If your MOV or MPEG-4 format video file does not work with Rekognition Video, check that the codec used to encode the video is H.264.

Q: What file sizes can I use with Amazon Rekognition?

Amazon Rekognition Image supports image file sizes up to 15MB when passed as an S3 object, and up to 5MB when submitted as an image byte array. Amazon Rekognition Video supports up to 8 GB files and up to 2 hour videos when passed through as an S3 file.

Q: How does image resolution affect the quality of Rekognition Image API results ?

Amazon Rekognition works across a wide range of image resolutions. For best results we recommend using VGA (640x480) resolution or higher. Going below QVGA (320x240) may increase the chances of missing faces, objects, or inappropriate content; although Amazon Rekognition accepts images that are at least 80 pixels in both dimensions.

Q: How small can an object be for Amazon Rekognition Image to detect and analyze it?

As a rule of thumb, please ensure that the smallest object or face present in the image is at least 5% of the size (in pixels) of the shorter image dimension. For example, if you are working with a 1600x900 image, the smallest face or object should be at least 45 pixels in either dimension.

Q: How can I get Amazon Rekognition predictions reviewed by humans?

Amazon Rekognition is directly integrated with Amazon Augmented AI (Amazon A2I) so you can easily route low confidence predictions from Amazon Rekognition Image to human reviewers. Using the Amazon Rekognition API for content moderation or the Amazon A2I console, you can specify the conditions under which Amazon A2I routes predictions to reviewers, which can be either a confidence threshold or a random sampling percentage. If you specify a confidence threshold, Amazon A2I routes only those predictions that fall below the threshold for human review. You can adjust these thresholds at any time to achieve the right balance between accuracy and cost-effectiveness. Alternatively, if you specify a sampling percentage, Amazon A2I routes a random sample of the predictions for human review. This can help you implement audits to monitor the prediction accuracy regularly. Amazon A2I also provides reviewers with a web interface consisting of all the instructions and tools they need to complete their review tasks. For more information about implementing human review with Amazon Rekognition, see the Amazon A2I webpage.

Q: How does video resolution affect the quality of Rekognition Video API results?

The system is trained to recognize faces larger than 32 pixels (on the shortest dimension), which translate into a minimum size for a face to be recognized that varies from approximately 1/7 of the screen smaller dimension at QVGA resolution to 1/30 at HD 1080p resolution. For example, at VGA resolution, users should expect lower performances for faces smaller than 1/10 of the screen smaller dimension.

Q: What else can affect the quality of the Rekognition Video APIs ?

Besides video resolution, heavy blur, fast moving persons, lighting conditions, pose may affect the quality of the APIs.

Q: What is the preferred user video content that is suitable for Rekognition Video APIs?

This API works best with consumer and professional videos taken with frontal field of view in normal color and lighting conditions. This API is not tested for black and white, IR or extreme lighting condition. Applications that are sensitive to false alarms are advised to discard outputs with confidence score below a selected (application-specific) confidence score.

Q: In which AWS regions is Amazon Rekognition available?

For a list of all regions where Amazon Rekognition is available, see the AWS Region table.

Object and Scene Detection

Q: What is a label?

A label is an object, scene, or concept found in an image based on its contents. For example, a photo of people on a tropical beach may contain labels such as ‘Person’, ‘Water’, ‘Sand’, ‘Palm Tree’, and ‘Swimwear’ (objects), ‘Beach’ (scene), and ‘Outdoors’ (concept).

Q: What is a confidence score and how do I use it?

A confidence score is a number between 0 and 100 that indicates the probability that a given prediction is correct. In the tropical beach example, if the object and scene detection process returns a confidence score of 99 for the label ‘Water’ and 35 for the label ‘Palm Tree’, then it is more likely that the image contains water but not a palm tree.

Applications that are very sensitive to detection errors (false positives) should discard results associated with confidence scores below a certain threshold. The optimum threshold depends on the application. In many cases, you will get the best user experience by setting minimum confidence values higher than the default value.

Q: What is Object and Scene Detection?

Object and Scene Detection refers to the process of analyzing an image or video to assign labels based on its visual content. Amazon Rekognition Image does this through the DetectLabels API. This API lets you automatically identify thousands of objects, scenes, and concepts and returns a confidence score for each label. DetectLabels uses a default confidence threshold of 50. Object and Scene detection is ideal for customers who want to search and organize large image libraries, including consumer and lifestyle applications that depend on user-generated content and ad tech companies looking to improve their targeting algorithms.

Q: Can Amazon Rekognition detect object locations and return bounding boxes?

Yes, Amazon Rekognition can detect the location of many common objects such as 'Person', 'Car', 'Gun', or 'Dog' in both images and videos. You get the coordinates of the bounding rectangle for each instance of the object found, as well as a confidence score. For more details on the API response structure for object bounding boxes, please refer to the documentation.

Q: Does Amazon Rekognition provide information on the relationship between detected labels?

Yes, for every label found, Amazon Rekognition returns the parent labels if they exist. Parents are returned in hierarchical order - the leftmost label is the immediate parent or synonym, while following labels are parents of parents. For example, when a 'Car' is identified, Amazon Rekognition returns two parent labels 'Vehicle' (parent), and 'Transportation' (parent's parent). For more details on how to use hierarchical taxonomy relations, please refer to the documentation.

Q: What types of labels does Amazon Rekognition support?

Rekognition supports thousands of labels belonging to common categories including, but not limited to:

- People and Events: 'Wedding', 'Bride', 'Baby', 'Birthday Cake', 'Guitarist', etc.

- Food and Drink: ‘Apple’, ‘Sandwich’, ‘Wine’, ‘Cake’, ‘Pizza’, etc.
- Nature and Outdoors: ‘Beach’, ‘Mountains’, ‘Lake’, ‘Sunset’, ‘Rainbow’, etc.
- Animals and Pets: ‘Dog’, ‘Cat’, ‘Horse’, ‘Tiger’, ‘Turtle’, etc.
- Home and Garden: ‘Bed’, ‘Table’, ‘Backyard’, ‘Chandelier’, ‘Bedroom’, etc.
- Sports and Leisure: ‘Golf’, ‘Basketball’, ‘Hockey’, ‘Tennis’, ‘Hiking’, etc.
- Plants and Flowers: ‘Rose’, ‘Tulip’, ‘Palm Tree’, ‘Forest’, ‘Bamboo’, etc.
- Art and Entertainment: ‘Sculpture’, ‘Painting’, ‘Guitar’, ‘Ballet’, ‘Mosaic’, etc.
- Transportation and Vehicles: ‘Airplane’, ‘Car’, ‘Bicycle’, ‘Motorcycle’, ‘Truck’, etc.
- Electronics: ‘Computer’, ‘Mobile Phone’, ‘Video Camera’, ‘TV’, ‘Headphones’, etc.

Q: How is Object and Scene Detection different for video analysis?

Rekognition Video enables you to automatically identify thousands of objects - such as vehicles or pets - and activities - such as celebrating or dancing - and provides you with timestamps and a confidence score for each label. It also relies on motion and time context in the video to accurately identify complex activities, such as “blowing a candle” or “extinguishing fire”.

Q: I can’t find the label I need. How do I request a new label?

Please send us your requests through AWS Customer Support. Amazon Rekognition continuously expands its catalog of labels based on customer feedback.

Q: How can you check if Amazon Rekognition has updated its models?

Amazon Rekognition returns a LabelModelVersion parameter that lets you know whether the model has been updated. Object and Scene detection models are updated frequently based on customer feedback.

Amazon Rekognition Custom Labels

Q: Can I use Custom Labels for analyzing faces, customized text detection, or finding unsafe image content?

No. Custom Labels is meant for finding objects and scenes in images. Custom Labels is not designed for analyzing faces, customized text detection, or finding unsafe image content. You should use other Rekognition APIs for these tasks. Please refer to the documentation for face analysis, Text detection, and Moderation.

Q: How many images are needed to train a custom model?

The number of images required to train a custom model depends on the variability of the custom labels you want the model to predict and the quality of the training data. For example, a distinct logo overlaid on an image can be detected with 1-2 training images, while a more subtle logo required to be detected under many variations (scale, viewpoint, deformations) may need in the order of tens to hundreds of training examples with high quality annotations. If you already have a high number of labeled images, we recommend training a model with as many images as you have available. Please refer to the documentation for limits on maximum training dataset size.

Although hundreds of images may sometimes be required to train a custom model with high accuracy, with Custom Labels you can first train a model with tens of images per label, review your test results to understand where it does not work, and accordingly add new training images and train again to iteratively improve your model.

Q: How many inference compute resources should I provision for my custom model?

The number of parallel inference compute resources needed depends on how many images you need to process at a given point in time. The throughput of a single resource will depend factors including the size of the images, the complexity of those images (how many detected objects are visible), and the complexity of your custom model. We recommend that you monitor the frequency at which you need provision your custom model, and the number of images that need to be processed at a single time, in order to schedule provisioning of your custom model most efficiently.

If you expect to process images periodically (e.g. once a day or week, or scheduled times during the day), you should Start provisioning your custom model at a scheduled time, process all your images, and then Stop provisioning. If you don’t stop provisioning, you will be charged even if no images are processed.

Q: My training has failed. Will I be charged?

No. You will not be charged for the compute resources if your training fails.

Unsafe Content Detection

Q: What is Unsafe Content Detection?

Amazon Rekognition’s Unsafe Content Detection is a deep-learning based easy to use API for detection of explicit or suggestive adult content, violent content, weapons, and visually disturbing content in image and videos. Beyond flagging an image or video based on presence of unsafe content, Amazon Rekognition also returns a hierarchical list of labels with confidence scores. These labels indicate specific sub-categories of the type of unsafe content detected (such as violence), thus providing more granular control to developers to filter and manage large volumes of user generated content (UGC). This API can be used in moderation workflows for applications such as social and dating sites, photo sharing platforms, blogs and forums, apps for children, e-commerce site, entertainment and online advertising services.

Q: What types of unsafe content does Amazon Rekognition detect?

Amazon Rekognition detects the following types of explicit and suggestive adult content in images and videos:

Explicit Nudity:

- Nudity
- Graphic Male Nudity
- Graphic Female Nudity

- Sexual Activity
- Illustrated Nudity Or Sexual Activity
- Adult Toys

Suggestive:

- Female Swimwear Or Underwear
- Male Swimwear Or Underwear
- Partial Nudity
- Revealing Clothes

Amazon Rekognition returns a hierarchy of labels, as well as a confidence score for each detected label. For instance, given an inappropriate image, Rekognition may return “Explicit Nudity” with a confidence score as a top level label. Developers could just use this to flag content. In the same response, Rekognition also returns second level of granularity by providing additional context like “Graphic Male Nudity” with its own confidence score. Developers could use this information to build more complex filtering logic to serve different geographies and demographics.

In addition, Amazon Rekognition can detect the following types of violent and visually disturbing content:

Violence:

- Graphic Violence Or Gore
- Physical Violence
- Weapons Violence
- Weapons
- Self Injury

Visually Disturbing:

- Emaciated Bodies
- Corpses

- Hanging

Please note that the Unsafe Image Detection API is not an authority on, or in any way purports to be an exhaustive filter of explicit or suggestive adult content, violent content, weapons, and visually disturbing content. Furthermore, this API does not detect whether an image includes illegal content (such as child pornography) or unnatural adult content.

Q: Can Amazon Rekognition's Unsafe Content Detection API detect other inappropriate content besides adult, violent and visually disturbing content?

Currently, Rekognition only supports the labels we have outlined above. We are working to continuously add and improve labels based on feedback from our customers.

If you require other types of inappropriate content to be detected in images, please reach out to us using the feedback process outlined later in this section.

Q: How can I know which model version I am currently using?

Amazon Rekognition makes regular improvement to its models. To keep track of model version, you can use the 'ModerationModelVersion' field in the API response.

Q: How can I ensure that Amazon Rekognition meets accuracy goals for my unsafe image or video detection use case?

Amazon Rekognition's Unsafe Content Detection models have been and tuned and tested extensively, but we recommend that you measure the accuracy on your own data sets to gauge performance.

You can use the 'MinConfidence' parameter in your API requests to balance detection of content (recall) vs the accuracy of detection (precision). If you reduce 'MinConfidence', you are likely to detect most of the inappropriate content, but are also likely to pick up content that is not actually explicit or suggestive. If you increase 'MinConfidence' you are likely to ensure that all your detected content is actually explicit or suggestive but some inappropriate content may not be tagged. For examples on how to use 'MinConfidence' for images, please refer to the documentation here.

Q: How can I give feedback to Rekognition to improve its Unsafe Content Detection?

Please send us your requests through AWS Customer Support. Amazon Rekognition continuously expands the types of inappropriate content detected based on customer feedback. Please note that illegal content (such as child pornography) will not be accepted through this process.

Facial Analysis

Q: What is Facial Analysis?

Facial analysis is the process of detecting a face within an image and extracting relevant face attributes from it. Amazon Rekognition Image takes returns the bounding box for each face detected in an image along with attributes such as gender, presence of sunglasses, and face landmark points. Rekognition Video will return the faces detected in a video with timestamps and, for each detected face, the position and a bounding box along with face landmark points.

Q: What face attributes can I get from Amazon Rekognition?

Amazon Rekognition returns the following facial attributes for each face detected, along with a bounding box and confidence score for each attribute:

- Gender
- Smile
- Emotions
- Eyeglasses
- Sunglasses
- Eyes open
- Mouth open
- Mustache
- Beard

- Pose
- Quality
- Face landmarks

Q: What is face pose?

Face pose refers to the rotation of a detected face on the pitch, roll, and yaw axes. Each of these parameters is returned as an angle between -180 and +180 degrees. Face pose can be used to find the orientation of the face bounding polygon (as opposed to a rectangular bounding box), to measure deformation, to track faces accurately, and more.

Q: What is face quality?

Face quality describes the quality of the detected face image using two parameters: sharpness and brightness. Both parameters are returned as values between 0 and 1. You can apply a threshold to these parameters to filter well-lit and sharp faces. This is useful for applications that benefit from high-quality face images, such as face comparison and face recognition.

Q: What are face landmarks?

Face landmarks are a set of salient points, usually located on the corners, tips or mid points of key facial components such as the eyes, nose, and mouth. Amazon Rekognition DetectFaces API returns a set of face landmarks that can be used to crop faces, morph one face into another, overlay custom masks to create custom filters, and more.

Q: How many faces can I detect in an image?

You can detect up to 100 faces in an image using Amazon Rekognition.

Q: How is Facial Analysis different for video analysis?

With Rekognition Video, you can locate faces across a video and analyze face attributes, such as whether the face is smiling, eyes are open, or showing emotions. Rekognition Video will return the detected faces with timestamps and, for each detected face, the position and a bounding box along with landmark points such as left eye, right eye, nose, left corner of the mouth, and right corner of the mouth. This position and time information can be used to easily track user sentiment over time and deliver additional functionality such as automatic face frames, highlights, or crops.

Q: In addition to Video resolution, what else can affect the quality of the Rekognition Video APIs?

Besides video resolution, the quality and representative faces, part of the face collections to search, has major impact. Using multiple face instances per person with variations like beard, glasses, poses (profile and frontal) will significantly improve the performance. Typically very fast moving people and blurred videos may experience lower quality.

Face Comparison

Q: What is Face Comparison?

Face Comparison is the process of comparing one face to one or more faces to measure similarity. Using the CompareFaces API, Amazon Rekognition Image lets you measure the likelihood that faces in two images are of the same person. The API compares a face in the source input image with each face detected in the target input image and returns a similarity score for each comparison. You also get a bounding box and confidence score for each face detected. You can use face comparison to verify a person's identity against their personnel photo on file in near real-time.

Q: Can I use a source image with more than one face?

Yes. If the source image contains multiple faces, CompareFaces detects the largest face and uses it to compare with each face detected in the target image.

Q: How many faces can I compare against?

You can compare one face in the source image with up to 15 detected faces in the target image.

Facial Recognition

Q: What is Facial Recognition?

Facial recognition is the process of identifying or verifying a person's identity by searching for their face in a collection of faces. Using facial recognition, you can easily build applications such as multi-factor authentication for bank payments, automated building entry for employees, and more.

Q: What is a face collection and how do I create one?

A face collection is a searchable index of face feature vectors, owned and managed by you. Using the CreateCollection API, you can easily create a collection in a supported AWS region and get back an Amazon Resource Name (ARN). Each face collection has a unique CollectionId associated with it.

Q: How do I add faces to or delete faces from a face collection?

To add a face to an existing face collection, use the IndexFaces API. This API accepts an image in the form of an S3 object or image byte array and adds a vector representation of the faces detected to the face collection. IndexFaces also returns a unique FaceId and face bounding box for each of the faces added.

To delete a face from an existing face collection, use the DeleteFaces API. This API operates on the face collection supplied (using a CollectionId) and removes the entries corresponding to the list of FaceIds. For more information on adding and deleting faces, please refer to our Managing Collections example.

Q: How do I search for a face within a face collection?

Once you have created an indexed collection of faces, you can search for a face within it using either an image (SearchFaceByImage) or a FaceId (SearchFaces). These APIs take in an input face and return a set of faces that match, ordered by similarity score with the highest similarity first. For more details, please refer to our Searching Faces example.

Q: How is Facial Recognition different for video analysis?

Rekognition Video allows you to perform real time face searches against collections with tens of millions of faces. First, you create a face collection, where you can store faces, which are vector representations of facial features. Rekognition then searches the face collection for visually similar faces throughout your video. Rekognition will return a confidence score for each of the faces in your video, so you can display likely matches in your application.

Q: In addition to Video resolution what else can affect the quality of the Video APIs ?

Besides video resolution, the quality and representative faces part of the face collections to search has major impact. Using multiple face instances per person with variations like beard, glasses, poses (profile and frontal) will significantly improve the performance. Typically very fast moving people may experience low recall. In addition, blurred videos may also experience lower quality.

Celebrity Recognition

Q: What is Celebrity Recognition?

Amazon Rekognition's Celebrity Recognition is a deep learning based easy-to-use API for detection and recognition of individuals who are famous, noteworthy, or prominent in their field. The RecognizeCelebrities API has been built to operate at scale and recognize celebrities across a number of categories, such as politics, sports, business, entertainment, and media. Our Celebrity Recognition feature is ideal for customers who need to index and search their digital image libraries for celebrities based on their particular interest.

Q: Who can be identified by the Celebrity Recognition API?

Amazon Rekognition can only identify celebrities that the deep learning models have been trained to recognize. Please note that the RecognizeCelebrities API is not an authority on, and in no way purports to be, an exhaustive list of celebrities. The feature has been designed to include as many celebrities as possible, based on the needs and feedback of our customers. We are constantly adding new names, but the fact that Celebrity Recognition does not recognize individuals that may be deemed prominent by any other groups or by our customers is not a reflection of our opinion of their celebrity status. If you would like to see additional celebrities identified by Celebrity Recognition, please submit feedback.

Q: Can a celebrity identified through the Amazon Rekognition API request to be removed from the feature?

Yes. If a celebrity wishes to be removed from the feature, he or she can send an email to AWS Customer Support and we will process the removal request.

Q: What sources are supported to provide additional information about a Celebrity ?

The API supports an optional list of sources to provide additional information about the celebrity as a part of the API response. We currently provide the IMDB URL, when it is available. We may add other sources at a later date.

Q: How is Celebrity Recognition different for video analysis?

With Rekognition Video, you can detect and recognize when and where well known persons appear in a video. The time-coded output includes the name and unique id of the celebrity, bounding box coordinates, confidence score, and URLs pointing to related content for the celebrity, for example, the celebrity's IMDB link. The celebrity is also detected even if sometimes the face becomes occluded in the video. This feature allows you to index and search digital video libraries for use cases related to your specific marketing and media needs.

Q: In addition to Video resolution, what else can affect the quality of the Rekognition Video APIs?

Very fast moving celebrities and blurred videos can affect the quality of the Rekognition Video APIs. In addition, heavy makeup and camouflage common for actors/actresses, can also affect the quality.

Text Detection

Q: What is Text Detection?

Text detection is a capability of Amazon Rekognition that allows you to detect and recognize text within an image or a video, such as street names, captions, product names, overlaid graphics, video subtitles, and vehicular license plates. Text detection is specifically built to work with real-world images and videos, rather than document images. Amazon Rekognition’s DetectText API takes in an image and returns the text label and a bounding box for each detected string of characters, along with a confidence score. For example, in image sharing and social media applications, you can enable visual search based on an index of images that contain the same text labels. In security applications, you can identify vehicles based on license plate numbers from images taken by traffic cams. Similarly, for videos, using the StartTextDetection and GetTextDetection APIs, you can detect text and get confidence scores and timestamps for each detection. In media and entertainment applications, you can create text metadata to support search for relevant content, such as news, sport scores, commercials, and captions. You can also review the detected text for policy or compliance violations e.g. an email address or phone number that has been overlaid by spammers.

Q: What type of text does Amazon Rekognition Text Detection support?

Text Detection is specifically built to work with real-world images and videos rather than document images. It supports text in most Latin scripts and numbers embedded in a large variety of layouts, fonts and styles, and overlaid on background objects at various orientation as banners and posters. Text detection recognizes up to 50 sequences of characters per the image or video frame and lists them as words and lines. Text detection supports text rotated by up to -90 to +90 degrees from the horizontal axis.

Q: Can I limit text detection to specific regions in an image or video frame?

Yes, you can use text detection filtering options to specify up to 10 Regions of Interest (ROIs) in the API request. Amazon Rekognition will only return text that falls within these regions.

Q: Can I filter text detections by word confidence or bounding box size?

Yes, in the API request you can use the text detection filtering options to specify thresholds for minimum confidence scores or minimum bounding box dimensions.

Q: How can I give feedback to Rekognition to improve its text recognition?

Please send us your requests through AWS Customer Support. Amazon Rekognition continuously expands the types of text content recognized based on customer feedback.

Video Analytics

Q: How do Amazon Rekognition Video asynchronous APIs work?

Rekognition Video processes a video stored in an Amazon S3 bucket. The design pattern is an asynchronous set of operations. You start video analysis by calling a Start operation such as StartLabelDetection. The completion status of the request is published to an Amazon Simple Notification Service topic. To get the completion status from the Amazon SNS topic, you can use an Amazon Simple Queue Service queue or an AWS Lambda function. Once you have the completion status, you call a Get operation such as GetLabelDetection to get the results of the request.

Q: What is Person Tracking?

With Rekognition Video, you can track each person within a shot and through the video across shots. Rekognition Video detects persons even when the camera is in motion and, for each person, returns a bounding box and the face, along with face attributes and timestamps. For retail applications, this allows to generate customer insights, such as how customers move across aisles in a shopping mall or how long they are waiting in checkout lines.

Q: How can I analyze videos in real time?

In streaming mode, you can search faces against a collection with tens of millions of faces in real time. Rekognition Video face detection and face recognition APIs natively integrate with video stream from Amazon Kinesis

Video Streams, a service that enables developers to transmit thousands of live feeds and associated metadata. For security applications, this makes real-time identification of Persons of Interest easy and accurate.

Q: Does Amazon Rekognition Video work with Amazon Kinesis Video Streams?

Rekognition Video uses a Kinesis Video Stream as input, to process a video stream. The analysis results are output from Rekognition Video to a Kinesis data stream and finally read by your client application. Rekognition Video provides a stream processor you can use to start and manage the analysis of streaming video. To learn more, please refer to Working with Streaming Videos.

Billing

Q: How does Amazon Rekognition count the number of images processed?

For APIs that accept images as inputs, Amazon Rekognition counts the actual number of images analyzed as the number of images processed. DetectLabels, DetectModerationLabels, DetectFaces, IndexFaces, RecognizeCelebrities, and SearchFaceByImage belong to this category. For the CompareFaces API, where two images are passed as input, only the source image is counted as a unit of images processed.

For API calls that don’t require an image as an input parameter, Amazon Rekognition counts each API call as one image processed. SearchFaces belongs to this category.

The remaining Amazon Rekognition APIs - ListFaces, DeleteFaces, CreateCollection, DeleteCollection, and ListCollections - do not count towards images processed.

Q: How does Amazon Rekognition count the number of minutes of videos processed?

For archived videos, Amazon Rekognition counts the minutes of video that is successfully processed by the API and meters them for billing. For Live stream videos you get charged in chunks of every five seconds of video that we successfully process.

Q: Which APIs does Amazon Rekognition charge for?

Amazon Rekognition Image charges for the following APIs: DetectLabels, DetectModerationLabels, DetectText, DetectFaces, IndexFaces, RecognizeCelebrities, SearchFaceByImage, CompareFaces, and SearchFaces. Amazon Rekognition Video charges are based on duration of video in minutes, successfully processed by StartLabelDetection, StartFaceDetection, StartFaceDetection, StartTextDetection, StartContentModeration, StartPersonTracking, StartCelebrityRecognition, StartFaceSearch and StartStreamProcessor APIs.

Q: How much does Amazon Rekognition cost?

Please see the Amazon Rekognition Pricing Page for current pricing information.

Q: Will I be charged for the feature vectors I store in my face collections?

Yes. Amazon Rekognition charges $0.01 per 1,000 face vectors per month. For details, please see the pricing page.

Q: Does Amazon Rekognition participate in the AWS Free Tier?

Yes. As part of the AWS Free Usage Tier, you can get started with Amazon Rekognition for free. Upon sign-up, new Amazon Rekognition customers can analyze up to 5,000 images for free each month for the first 12 months. You can use all Amazon Rekognition APIs with this free tier, and also store up to 1,000 faces without any charge. In addition, Amazon Rekognition Video customers can analyze 1,000 minutes of Video free, per month, for the first year.

Q: Do your prices include taxes?

For details on taxes, please see Amazon Web Services Tax Help.

AWS Integration

Q: Does Amazon Rekognition Video work with images stored on Amazon S3?

Yes. You can start analyzing images stored in Amazon S3 by simply pointing the Amazon Rekognition API to your S3 bucket. You don't need to move your data. For more details of how to use S3 objects with Amazon Rekognition API calls, please see our Detect Labels exercise.

Q: Can I use Amazon Rekognition with images stored in an Amazon S3 bucket in another region?

No. Please ensure that the Amazon S3 bucket you want to use is in the same region as your Amazon Rekognition API endpoint.

Q: How do I process multiple image files in a batch using Amazon Rekognition?

You can process your Amazon S3 images in bulk using the steps described in our Amazon Rekognition Batch Processing example on GitHub.

Q: How can I use AWS Lambda with Amazon Rekognition?

Amazon Rekognition provides seamless access to AWS Lambda and allows you bring trigger-based image analysis to your AWS data stores such as Amazon S3 and Amazon DynamoDB. To use Amazon Rekognition with AWS Lambda, please follow the steps outlined here and select the Amazon Rekognition blueprint.

Q: Does Amazon Rekognition work with AWS CloudTrail?

Yes. Amazon Rekognition supports logging the following actions as events in CloudTrail log files: CreateCollection, DeleteCollection, CreateStreamProcessor, DeleteStreamProcessor, DescribeStreamProcessor, ListStreamProcessors, and ListCollections. For more details on the Amazon Rekognition API calls that are integrated with AWS CloudTrail, see Logging Amazon Rekonition API Calls with AWS CloudTrail.

Data Privacy

Q: Are image and video inputs processed by Amazon Rekognition stored, and how are they used by AWS?

Amazon Rekognition may store and use image and video inputs processed by the service solely to provide and maintain the service and, unless you opt out as provided below, to improve and develop the quality of Amazon Rekognition and other Amazon machine-learning/artificial-intelligence technologies. Use of your content is important for continuous improvement of your Amazon Rekognition customer experience, including the development and training of related technologies. We do not use any personally identifiable information that may be contained in your content to target products, services or marketing to you or your end users. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information. You may opt out of having your image and video inputs used to improve or develop the quality of Amazon Rekognition and other Amazon machine-learning/artificial-intelligence technologies by contacting AWS Support.

Q: Can I delete image and video inputs stored by Amazon Rekognition?

Yes. You can request deletion of image and video inputs associated with your account by contacting AWS Support. Deleting image and video inputs may degrade your Amazon Rekognition experience.

Q: Who has access to my content that is processed and stored by Amazon Rekognition?

Only authorized employees will have access to your content that is processed by Amazon Rekognition. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information.

Q: Do I still own my content that is processed and stored by Amazon Rekognition?

You always retain ownership of your content and we will only use your content with your consent.

Q: Is the content processed by Amazon Rekognition moved outside the AWS region where I am using Amazon Rekognition?

Any content processed by Amazon Rekognition is encrypted and stored at rest in the AWS region where you are using Amazon Rekognition. Unless you opt out as provided below, some portion of content processed by Amazon Rekognition may be stored in another AWS region solely in connection with the continuous improvement and development of your Amazon Rekognition customer experience and other Amazon machine-learning/artificial-intelligence technologies. You can request deletion of image and video inputs associated with your account by contacting AWS Support. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information. Your content will not be stored in another AWS region if you contact AWS Support to opt out of having your content used to improve and develop the quality of Amazon Rekognition and other Amazon machine-learning/artificial-intelligence technologies.

Q: Can I use Amazon Rekognition in connection with websites, programs or other applications that are directed or targeted to children under age 13 and subject to the Children's Online Privacy Protection Act (COPPA)?

Yes, subject to your compliance with the Amazon Rekognition Service Terms, including your obligation to provide any required notices and obtain any required verifiable parental consent under COPPA, you may use Amazon

Rekognition in connection with websites, programs, or other applications that are directed or targeted, in whole or in part, to children under age 13.

Q: How do I determine whether my website, program, or application is subject to COPPA?

For information about the requirements of COPPA and guidance for determining whether your website, program, or other application is subject to COPPA, please refer directly to the resources provided and maintained by the United States Federal Trade Commission. This site also contains information regarding how to determine whether a service is directed or targeted, in whole or in part, to children under age 13.

Q: Is Amazon Rekognition a HIPAA Eligible Service?

Amazon Rekognition is a HIPAA Eligible Service covered under the AWS Business Associate Addendum (AWS BAA). If you have an AWS BAA in place, Amazon Rekognition will use, disclose, and maintain your Protected Health Information (PHI) only as permitted by the terms of your AWS BAA.

Access Control

Q: How do I control user access for Amazon Rekognition?

Amazon Rekognition is integrated with AWS Identity and Access Management (IAM). AWS IAM policies can be used to ensure that only authorized users have access to Amazon Rekognition APIs. For more details, please see the Amazon Rekognition Authentication and Access Control page.

Report Abuse

Q: How can I report potential Amazon Rekognition abuse?

If you suspect that Amazon Rekognition is being used in manner that is abusive or illegal, or infringes on your rights or the rights of other people, please report this use and AWS will investigate the issue.

Amazon SageMaker Ground Truth FAQs

General

Q: What is Amazon SageMaker Ground Truth?

A: Amazon SageMaker Ground Truth makes it easy for you to efficiently and accurately label the datasets required for training machine learning systems. SageMaker Ground Truth can automatically label a portion of the dataset based on the labels done manually by human labelers. You can choose to use a crowdsourced Amazon Mechanical Turk workforce of over 500,000 labelers, your own employees , or one of the third party data labeling service providers listed on AWS Marketplace, pre-screened by Amazon. SageMaker Ground Truth uses innovative algorithms and user experience (UX) techniques to improve the accuracy of human labeling. Over time, the model gets progressively better by continuosly learning from the labels created by humans, for increased automatic labeling.

Q: What is Automated Data Labeling?

A: Automated data labeling is labeling of data using machine learning. Amazon SageMaker Ground Truth will first select a random sample of data and send it to humans to be labeled. The results are then used to train a labeling model that attempts to label a new sample of raw data automatically. The labels are committed when the model can label the data with a confidence score that meets or exceeds a high threshold. Where the confidence score falls below this threshold, the data is sent to human labelers. Some of the data labeled by humans is used to generate a new training dataset for the labeling model, and the model is automatically retrained to improve its accuracy. This process repeats with each sample of raw data to be labeled. The labeling model becomes more capable of automatically labeling raw data with each iteration, and less data is routed to humans.

Using Amazon SageMaker Ground Truth

Q: Why should I use Amazon SageMaker Ground Truth?

A: Prior to building, training, and deploying machine learning models, you need data. Successful models are built on high-quality training data, and collecting and labeling the training datasets involves a lot of time and effort. To build the training datasets, human labelers need to evaluate a large number of images or other data types, and then identify and label particular objects in each data type. These labeling tasks are distributed across many human labelers, adding significant overhead and cost. If there are incorrect labels, the system will learn from the bad information and make inaccurate predictions.

Amazon SageMaker Ground Truth solves this problem by making it easy to efficiently perform highly accurate data labeling using data stored in Amazon S3, using a combination of automated data labeling and human-performed labeling.

Q: How do I get started with Amazon SageMaker Ground Truth?

A: Amazon SageMaker Ground Truth provides a managed experience where you can set up an entire data labeling job with just a few steps. To get started with Amazon SageMaker Ground Truth, you sign into the AWS Management Console and navigate to the SageMaker console. From there, select Labeling jobs under Ground Truth. Here you can create a labeling job. First as part of the labeling job creation flow, you provide a pointer to the S3 bucket that contains your dataset to be labeled. Ground Truth offers templates for common labeling tasks where you only need to click a few choices and provide minimal instructions on how to get your data labeled. Alternatively, you can create your own custom template. As the last step of creating a labeling job, you select one of the three human workforce options: (1) a public crowdsourced workforce, (2) a curated set of third party data labeling service providers , or (3) bring your own workers. You also have the option to enable automated data labeling.

Q: How are my training datasets managed using Amazon SageMaker Ground Truth?

A: Amazon SageMaker Ground Truth manages the metadata, associated labels, and a taxonomy of your labels and datasets. You can easily use the AWS SDK through a SageMaker Notebook or the Ground Truth console within the SageMaker console to query and manage your datasets and labels. Visit the Amazon SageMaker Ground Truth documentation for more information.

Q: How does Amazon SageMaker Ground Truth help with increasing the accuracy of my training datasets?

A: Amazon SageMaker Ground Truth offers the following features to help you increase the accuracy of data labeling performed by humans:

(a) *Annotation consolidation:* This counteracts the error/bias of individual workers by sending each data object to multiple workers and then consolidating their responses (called "annotations") into a single label. It then takes their annotations and compares them using an annotation consolidation algorithm. This algorithm first detects outlier annotations that are disregarded. It then performs a weighted consolidation of the annotations, assigning higher weights to more reliable annotations. The output is a single label for each object.

(b) *Annotation interface best practices:* These are features of the annotation interfaces that enable workers to perform their tasks more accurately. Human workers are prone to error and bias, and well-designed interfaces improve worker accuracy. One best practice is to display brief instructions along with good and bad label examples in a fixed side panel. Another best practice is to darken the area outside of the box bounding boundary when workers are drawing the bounding box on an image.

Q: How does Amazon SageMaker Ground Truth ensure that my data is protected and secure?

A: By default, Amazon SageMaker Ground Truth encrypts your data at rest and in transit. In addition, access to your data can be controlled using AWS Identity and Access Management (IAM). Ground Truth does not store or make copies of your data outside of your AWS environment, and your data remains in your control. Further, Ground Truth supports compliance standards such as General Data Protection Regulation (GDPR), and provides comprehensive logging and auditing capabilities using Amazon CloudWatch and Amazon CloudTrail. Visit the Amazon SageMaker Ground Truth documentation for more information.

Q: How do I access a human workforce using Amazon SageMaker Ground Truth?

A: From SageMaker Ground Truth, you can choose any of the three workforce options namely (1) Public crowdsourced workforce through Amazon Mechanical Turk; (2) Third party data labeling service providers available through AWS Marketplace; and (3) Your own employees. Visit the Amazon SageMaker Ground Truth documentation for more information.

Using Third Party Data Labeling Service Providers

Q: Can Amazon SageMaker Ground Truth data labeling service provider process confidential data?

A: Yes, Amazon SageMaker Ground Truth data labeling service providers can process confidential data. The Standard Service Agreement between AWS customers and the third party data labeling service provider contains some basic protections for your confidential information. Please review those terms before sharing any confidential information with the service provider. The terms are located on the listing page for the service provider on AWS Marketplace.

Q: What security standards are Amazon SageMaker Ground Truth data labeling service providers required to meet?

A: Data labeling service providers are required to go through SOC 2 compliance and certification on an annual basis. The SOC 2 report is a description of the service provider’s control environment based on the American Institute of Certified Public Accountants (AICPA) Trust Services Criteria - Security, Availability, Processing Integrity, Confidentiality, and Privacy.

In addition to SOC 2, service providers are required to maintain these additional security controls to help keep customer data secure.

Technology Controls:
Service providers are required to utilize the appropriate software to block any attempts made to download or copy files/data from their system and prevent unauthorized access to their systems. Service providers are also required to prohibit their workforce from storing or copying customer task-related data.

Network Security Controls:
We require the service provider’s network to be designed to prevent remote access to customer's task-related data. Further, peer-to-peer file sharing software is blocked on the provider's network, and the firewall should be designed in a way to provide high availability.

Employee Controls:
Service providers are required to ensure they have Non-Disclosure Agreements (NDAs) with their employees. Service providers are required to adopt stringent policies to prevent any information leakage and prevent employees from transmitting information by any means: paper, USBs, mobile phones, or any other media.

Physical Access Controls:
Service providers are required to maintain physical access control measures to prevent unauthorized access to their production site. These may include turnstiles with biometric authentication, employee badge identification, etc.

Q: How does AWS help ensure service providers meet these security standards?

A: AWS requests that service providers furnish their SOC 2 certification reports prior to being listed in the marketplace and confirms:

Authenticity (if service provider auditor is certified by the AICPA);

Report period (SOC 2 certification validity date); and

Production site (the physical site where the service provider workforce will work on Amazon SageMaker Ground Truth labeling tasks).

Q: What is the frequency of review of service provider security standards?

A: The security standards from every service provider are reviewed annually to ensure they meet the mandatory requirements.

Q: Are there any exceptions to the AWS review?

A: No. If the service provider fails to meet security standards, then their listing will be removed from AWS Marketplace. De-listing will be completed within 24 hours and all active customers will be notified by email.

Q: If the service provider offers data labeling services through multiple production sites, do all sites need to go through the review process?

A: Yes, all sites need to meet the required security standards.

Q: What happens if there is a data breach at the service provider production site?

A: The service provider will inform AWS and affected customers within 24 hours of detecting any actual or suspected unauthorized access, collection, acquisition, use, transmission, disclosure, corruption, or loss of customer information. The service provider will remedy each security incident promptly and provide AWS and affected customers written details about the internal investigation.

Pricing and Availability

Q: How much does Amazon SageMaker Ground Truth cost?

A: Please see the SageMaker Ground Truth pricing page for the current pricing information.

Q: In which AWS regions is Amazon SageMaker Ground Truth available?

A: The AWS Region Table lists all the AWS regions where Amazon SageMaker Ground Truth is currently available.

Amazon Textract FAQs

General

Q: What is Amazon Textract?

A: Amazon Textract is a document analysis service that detects and extracts text, structured data, such as fields of interest and their values, and tables from images and scans of documents. Amazon Textract's machine learning models have been trained on millions of documents so that virtually any document type you upload is automatically recognized and processed for text extraction. When information is extracted from documents, the service returns a confidence score for each element it identifies so that you can make informed decisions about how you want to use the results. For instance, if you are extracting information from tax documents you can set custom rules to flag any extracted information with a confidence score lower than 95%. Also, all extracted data are returned with bounding box coordinates, which is a rectangular frame that fully encompasses each piece of data identified, so that you can quickly identify where a word or number appears on a document. You can access these features with the Amazon Textract API, in the AWS Management Console, or using the AWS command-line interface (CLI).

Q: What are the most common use cases for Amazon Textract?

A: The most common use cases for Amazon Textract include:

- Import Documents and Forms into Business Applications
- Create Smart Search Indexes
- Build Automated Document Processing Workflows
- Maintain Compliance in Document Archives
- Extract Text for Natural Language Processing (NLP)
- Text Extraction for Document Classification

Q: What type of text can Amazon Textract detect and extract?

A: Amazon Textract can detect Latin-script characters from the standard English alphabet and ASCII symbols.

Q: What document formats does Amazon Textract support?

A: Amazon Textract currently supports PNG, JPEG, and PDF formats. For synchronous APIs, you can submit images either as an S3 object or as a byte array. For asynchronous APIs, you can submit S3 objects.

Q: How do I get started with Amazon Textract?

A: To get started with Amazon Textract, you can click the "Get Started with Amazon Textract", button on the Amazon Textract page. You must have an Amazon Web Services account; if you do not already have one, you will be prompted to create one during the process. Once you are signed in to your AWS account, try out Amazon Textract with your own images or PDF documents using the Amazon Textract Management Console. You can also download the Amazon Textract SDKs to start creating your own applications. Please refer to our step-by-step Getting Started Guide for more information.

Q: What APIs does Amazon Textract offer?

A: Amazon Textract offers APIs that detect and extract text from scanned images of documents, extracts structured data such as tables, and performs key-value pairing on extracted text. Amazon Textract performs OCR using the Detect Document Text API, but goes a step further in the document analyzing process and also performs key-value pair detection so that text extractions remain organized in their intended structure. The Analyze Document API can detect text, fields, values, their relationships, tables, and other entities within a document along with their associated confidence scores. With the Analyze Document API, developers can automatically capture structured data from a wide variety of documents including tax forms, financial reports, medical records, and loan applications. For details, please refer to the Amazon Textract API reference.

Q: How do I use the confidence score Amazon Textract provides?

A: A confidence score is a number between 0 and 100 that indicates the probability that a given prediction is correct. With Amazon Textract, all extracted text and structured data are returned with bounding box coordinates, which is a rectangular frame that fully encompasses each piece of data identified. This allows you to identify the score for each extracted entity so that you can make informed decisions on how you want to use the results.

Q: How can I get Amazon Textract predictions reviewed by humans?

A: Amazon Textract is directly integrated with Amazon Augmented AI (A2I) so you can easily get low confidence predictions from Amazon Textract reviewed by humans. Using Amazon Textract's API for form data extraction and the Amazon A2I console, you can specify the conditions under which Amazon A2I routes predictions to reviewers, which can be either a confidence threshold or a random sampling percentage. If you specify a confidence threshold, Amazon A2I routes only those predictions that fall below the threshold for human review. You can adjust these thresholds at any time to achieve the right balance between accuracy and cost-effectiveness. Alternatively, if you specify a sampling percentage, Amazon A2I routes a random sample of the predictions for human review. This can help you implement audits to monitor the prediction accuracy regularly. Amazon A2I also provide reviewers a web interface consisting of all the instructions and tools they need to complete their review tasks. For more information about implementing human review with Amazon Textract, see the Amazon A2I website.

Q: How can I get the best results from Amazon Textract?

A: Amazon Textract uses machine learning to read virtually any type of document, in order to extract text and structured information. Keep the following tips in mind in order to get the best results:

- Make sure your document uses a language supported by Amazon Textract (Currently English).
- Provide as high quality an image as you can, ideally at least 150 DPI.
- If your document is already in one of the file formats that Amazon Textract supports (PDF, JPG, PNG), don't convert or downsample it before uploading it to Amazon Textract.
- Amazon Textract's table feature works best when the tables in your document

are visually separated from surrounding elements on the page (e.g. not overlaid on an image or complex pattern), and the text within the table is upright (e.g. not rotated relative to other text on the page)

You can get started with analyzing you own documents with Amazon Textract with just a few clicks in the Amazon Textract Management Console. If you have trouble achieving high accuracy with receipts, identification, or industrial diagrams, please contact us on amazon-textract@amazon.com for assistance.

Q: In which AWS regions is Amazon Textract available?

A: Amazon Textract is currently available in the US East (Northern Virginia), US East (Ohio), US West (Oregon), US West (N. California), EU (Ireland), EU (London), and Asia Pacific (Sydney).

Q: Does Amazon Textract work with AWS CloudTrail?

A: Yes. Amazon Textract supports logging of the following actions as CloudTrail events - DetectDocumentText, AnalyzeDocument, StartDocumentTextDetection, StartDocumentAnalysis, GetDocumentTextDetection, and GetDocumentAnalysis. For more details, please see Logging Amazon Textract API Calls with AWS CloudTrail.

Billing

Q: How does Amazon Textract count the number of pages processed?

A: An image (PNG or JPEG) counts as a single page. For PDFs, each page in the document is counted as a page processed.

Q: Which APIs am I charged for with Amazon Textract?

A: Refer to the Amazon Textract pricing page to learn more about pricing.

Q: How much does Amazon Textract cost?

A: Amazon Textract charges you based on the number of pages and images processed. For more information, visit the pricing page.

Q: Does Amazon Textract participate in the AWS Free Tier?

A: Yes. As part of the AWS Free Usage Tier, you can get started with Amazon Textract for free. New customers can analyze up to 1,000 pages per month using the Detecting Document Text API and up to 100 pages per month using the Analyze Document API, for the first three months.

Q: Do your prices include taxes?

A: For details on taxes, please see Amazon Web Services Tax Help.

Data Privacy

Q. Are document and image inputs processed by Amazon Textract stored, and how are they used by AWS?

A: Amazon Textract may store and use document and image inputs processed by the service solely to provide and maintain the service and to improve and develop the quality of Amazon Textract and other Amazon machine-learning/artificial-intelligence technologies. Use of your content is necessary for continuous improvement of your Amazon Textract customer experience, including the development and training of related technologies. We do not use any personally identifiable information that may be contained in your content to target products, services or marketing to you or your end users. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information.

Q. Can I delete images and documents stored by Amazon Textract?

A: Yes. You can request deletion of document and image inputs associated with your account by contacting AWS Support. Deleting image and document inputs may degrade your Amazon Textract experience.

Q: Who has access to my content that is processed and stored by Amazon Textract?

A: Only authorized employees will have access to your content that is processed by Amazon Textract. Your trust, privacy, and the security of your content are our highest priority, and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information.

Q: Do I still own my content that is processed and stored by Amazon Textract?

A: Yes. You always retain ownership of your content, and we will only use your content with your consent.

Q: Is Amazon Textract HIPAA eligible?

Yes, AWS has expanded its HIPAA compliance program to include Amazon Textract as a HIPAA eligible service. If you have an executed Business Associate Agreement (BAA) with AWS, you can use Amazon Textract to extract text including protected health information (PHI) from images.

Learn more about HIPAA Compliance »

Amazon Translate FAQs

General

Q: What is Amazon Translate?

Amazon Translate is a Neural Machine Translation (MT) service for translating text between supported languages. Powered by deep learning methods, the service provides high quality and affordable machine translation, enabling developers to translate company and user-authored content, or build applications requiring support across multiple languages. The service can be used via an API, enabling either real-time or batch translation of text from the source language to the target language.

Q: What languages are covered?

Amazon Translate supports translation between the following 54 languages: Afrikaans, Albanian, Amharic, Arabic, Azerbaijani, Bengali, Bosnian, Bulgarian, Chinese Simplified, Chinese Traditional, Croatian, Czech, Danish, Dari, Dutch, English, Estonian, Farsi (Persian), Finnish, French, Canadian French, Georgian, German, Greek, Hausa, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Latvian, Malay, Norwegian, Pashto, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Somali, Spanish, Swahili, Swedish, Tagalog, Tamil, Thai, Turkish, Ukrainian, Urdu, and Vietnamese. See this documentation page for more details.

Q: Why should I use Amazon Translate?

You should use Amazon Translate because it enables you to reach more customers, communicate with them more effectively, and decrease your TCO. Many businesses have large volumes of content, user or company authored; the only way to make all of it accessible in multiple languages in a timely manner is to use Machine Translation. Because Amazon Translate costs a fraction of the cost of human translation (0.05% at $15/1M characters for Amazon Translate vs $30K for human translation on average), businesses can now afford to translate content they could not before.

For Language Service Providers (LSP) and value-added resellers, Amazon Translate supports business growth and expansion. With Amazon Translate, LSPs can increase productivity by as much as 50% and produce larger volumes of translation, freeing professional translators to focus on high-end creative content. Resellers can broaden their service portfolio without building new infrastructure or hiring staff.

Q: What are the most common use cases for Amazon Translate?

Amazon Translate is a great solution in cases where the volume of content is high, speed is critical, and a certain level of translation imperfection (usually minor) is acceptable. For example, if you need to extract insights from large volumes of text in many languages, enable customers to search your application in their language of choice, make user-authored content such as forums and support content accessible in languages other than the source, get the gist out of responses to questionnaires and surveys, or publish a first draft – you can use Amazon Translate's raw output.

With light human post-editing, Amazon Translate can be applied to enabling customer service agents to support anyone, and translating company authored information such as specifications, comparisons of alternatives, FAQs, and support content. With more extensive post-editing, you can also use Amazon Translate to translate high-value, branded content, such as advertising and marketing materials, contracts, etc.

Using Amazon Translate

Q: How can I use the service?

The easiest way to get started with Amazon Translate is to use the console to translate some text. You can also call the service directly from the AWS Command Line Interface, or use one of the SDKs in the programming language of your choice to integrate with your applications. Either way, you can start using Amazon Translate for multilingual text capabilities to translate text with just a few lines of code.

You can pass source text to the API and indicate the source and target languages. Amazon Translate return the text translated into the target language. There are three main ways to use the API – first, you can integrate the API into your application to localize highly dynamic application components such as multi-participant chat, for example. Second, you can string it with other services to enable language-independent processing. For example, Database services such as Amazon Relational Database Service (RDS) can be called through AWS Lambda blueprints to enable website localization of moderately-dynamic content such as user generated reviews and forum posts. Finally, you can translate batches of documents. For example, financial services companies can translate and monitor news articles in any language; legal teams can discover materials in multiple languages related to a lawsuit (known as eDiscovery); patent attorneys can search patent repositories anywhere in the world in IP cases.

Q: Does the service provide automatic source language detection?

A: Amazon Translate takes plain text input and language flags to indicate the language of the source text and desired target. If the source language is unknown, Amazon Translate will identify the source language using Amazon Comprehend behind the scenes, and report that language back along with the translation to the target language.

Q: What kind of inputs does the service support?

Amazon Translate supports plain text input in UTF-8 format.

Q: What are the limits on the API?

Amazon Translate real-time service calls are limited to 5,000 bytes per API call. We provide instructions on how to break up large documents into sections and paragraphs so that customers can translate text of any length. See instructions here.

Amazon Translate asynchronous Batch Translation service accepts a batch of up to 5 GB in size per API call with each document not exceeding 1 MB in size and the number of documents in the S3 bucket folder not exceeding 1 million per batch.

The Amazon Translate service is highly scalable. The default limits can be found here.

Q: Am I required to attribute the translation to Amazon? To Machine Translation?

You are not required to attribute translations, but we do suggest that you attribute the translation to Machine Translation to inform your own customers.

Q: Where can I get technical support? How do I submit feedback?

For technical support, please contact AWS Customer Service. You can submit feedback through Customer Service, or by going to the Amazon Translate console and selecting the feedback option.

Pricing & Availability

Q: What does it cost?

Refer to the Amazon Translate pricing page to learn more.

Q: What AWS regions are available for Amazon Translate?

Please refer to the AWS Global Infrastructure Region Table.

Amazon Translate Batch Translation is available in US East 1 (Northern Virginia), US East 2 (Ohio), US West 2 (Oregon), EU West 1 (Ireland), EU West 2 (London), EU Central 1 (Frankfurt), and Asia Pacific North East 2 (Seoul).

Q: Are requests in which no translation occurs charged for?

Requests where the source language equals the target language (whether user designated or automatically identified), and when an error occurs and no translation is returned, are not charged for. Requests where the content is non-translatable (e.g., “&*^%((**&(^”) are charged for.

Data Privacy

Q. Are text inputs processed by Amazon Translate stored, and how are they used by AWS?

Amazon Translate may store and use text inputs processed by the service solely to provide and maintain the service and to improve and develop the quality of Amazon Translate and other Amazon machine-learning/artificial-intelligence technologies. Use of your content is important for continuous improvement of your Amazon Translate customer experience, including the development and training of related technologies. We do not use any personally identifiable information that may be contained in your content to target products, services or marketing to you or your end users. Your trust, privacy, and the security of your content are our highest priority, and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information.

Q: Who has access to my content that is processed and stored by Amazon Translate?

Only authorized employees will have access to your content that is processed by Amazon Translate. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information.

Q: Do I still own my content that is processed and stored by Amazon Translate

You always retain ownership of your content, and we will only use your content with your consent.

Q: Is the content processed by Amazon Translate moved outside the AWS region where I am using Amazon Translate?

Any content processed by Amazon Translate is encrypted and stored at rest in the AWS region where you are using Amazon Translate. Some portion of content processed by Amazon Translate may be stored in another AWS region solely in connection with the continuous improvement and development of your Amazon Translate customer experience and other Amazon machine-learning/artificial-intelligence technologies. Your trust, privacy, and the security of your content are our highest priority, and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information.

Q: Can I use Amazon Translate in connection with websites, programs or other applications that are directed or targeted to children under age 13 and subject to the Children's Online Privacy Protection Act (COPPA)?

Yes, subject to your compliance with the Amazon Translate Service Terms, including your obligation to provide any required notices and obtain any required verifiable parental consent under COPPA, you may use Amazon Translate in connection with websites, programs, or other applications that are directed or targeted, in whole or in part, to children under age 13.

Q: How do I determine whether my website, program, or application is subject to COPPA?

For information about the requirements of COPPA and guidance for determining whether your website, program, or other application is subject to COPPA, please refer directly to the resources provided and maintained by the

United States Federal Trade Commission. This site also contains information regarding how to determine whether a service is directed or targeted, in whole or in part, to children under age 13.

Amazon Transcribe FAQs

General

Q: What is Amazon Transcribe?

Amazon Transcribe is an AWS service that makes it easy for customers to convert speech-to-text. Using Automatic Speech Recognition (ASR) technology, customers can choose to use Amazon Transcribe for a variety of business applications, including transcription of voice-based customer service calls, generation of subtitles on audio/video content, and conduct (text based) content analysis on audio/video content.

Q: How does Amazon Transcribe interact with other AWS products?

Amazon Transcribe converts audio input into text, which opens the door for various text analytics applications on voice input. For instance, by using Amazon Comprehend on the converted text data from Amazon Transcribe, customers can perform sentiment analysis or extract entities and key phrases. Similarly, by integrating with Amazon Translate and Amazon Polly, customers can accept voice input in one language, translate it into another and generate voice output, effectively enabling multi-lingual conversations. It is also possible to integrate Amazon Transcribe with Amazon Elasticsearch to index and perform text based search across audio/video library.

Q: What else should I know before using Amazon Transcribe service?

Amazon Transcribe service is designed to handle a wide range of speech and acoustic characteristics, including variations in volume, pitch, and speaking rate. The quality and content of the audio signal (including but not limited to factors such as background noise, overlapping speakers, accented speech, or switches between languages within a single audio file) may affect the accuracy of service output. We are constantly updating the service to improve its ability to accommodate additional acoustic variation and content types.

Using Amazon Transcribe

Q: How will developers access Transcribe?

The easiest way to get started with Amazon Transcribe is to submit a job using the console to transcribe an audio file. You can also call the service directly from the AWS Command Line Interface, or use one of the supported SDKs of your choice to integrate with your applications. Either way, you can start using Amazon Transcribe to generate automated transcripts for your audio files with just a few lines of code.

Q: What kind of inputs does Amazon Transcribe support?

Amazon Transcribe supports both 16 kHz and 8kHz audio streams, and multiple audio encodings, including WAV, MP3, MP4 and FLAC.

Q: Does Amazon Transcribe support real-time transcriptions?

Yes. Amazon Transcribe enables users to open a bidirectional stream over HTTP2. Users can send an audio stream to the service while receiving a text stream in return in real time.

Q: What encoding does real-time transcription support?

Streaming transcription currently supports 16-bit Linear PCM encoding.

Q: What languages does Amazon Transcribe support?

For information on language support, please refer to this documentation page.

Q: What devices does Amazon Transcribe work with?

Amazon Transcribe for the most part is device agnostic. In general, Amazon Transcribe works with any device that includes an on-device microphone such as phones, PCs, tablets, and IoT devices (e.g. car audio systems). Amazon Transcribe API will be able to detect the quality of the audio stream being input at the device (8kHz VS 16kHz) and will appropriately select the acoustic models for converting speech-to-text. Furthermore, developers can call Transcribe API through their applications to access speech-to-text conversion capability.

Q: Are there size restrictions on the audio content that Amazon Transcribe can process?

Amazon Transcribe service calls are limited to 4 hours (or 2GB) per API call for our batch service. The streaming service can accommodate open connections up to 4 hours long.

Q: What programming languages does Amazon Transcribe support?

Amazon Transcribe batch service supports .NET, Go, Java, Javascript, PHP, Python and Ruby.
Amazon Transcribe real-time service supports Java SDK, Ruby SDK, and C++ SDK. Additional SDK support are coming. For more details, visit the Resources page.

Q: My custom vocabulary words are not being recognized! What can I do?

The speech recognition output depends on a number of factors in addition to custom vocabulary entries, so there can be no assurance that if a term is included in the custom vocabulary, it will be correctly recognized.
However, the most frequent reason is that a custom word lacks the correct pronunciation. If you haven't provided a pronunciation for your custom word, please try to create one. If you already have provided one, double-check its correctness, or include other pronunciation variants if necessary. This can be done by creating multiple entries in the custom vocabulary file that differ in the pronunciation field.

Q: Why do I see too many custom words in my output?

Custom vocabularies are optimized for a small list of targeted words; larger vocabularies may lead to over-generation of custom words, especially when they contain words that are pronounced in a similar way. If you have a large list, please try reducing it to rare words and words that are actually expected to occur in your audio files. If you have a large vocabulary covering multiple use cases, split it into separate lists for different use cases. The words that are short and sound similar to many other words, may lead to over-generation (too many custom words appearing in the output). It is preferable to combine these words with surrounding words and list them as hyphen-separated phrases. For example, the custom word “A.D.” could be included as part of a phrase such as ‘A.D.-converter’.

Q: There are two ways of giving pronunciations, IPA or SoundsLike fields in the custom vocabulary table. Which one is better?

IPA allows for more precise pronunciations. You should provide IPA pronunciations if you are able to generate IPA (e.g., from a lexicon that has IPA pronunciations or an online converter tool).

Q: I'd like to use IPA but I'm not a linguistic expert. Is there an online tool I can use?

Several standard dictionaries, such as the Oxford English Dictionary or the Cambridge Dictionary (including their online versions) provide pronunciations in IPA. There are also online converters (e.g. easypronunciation.com or tophonetics.com for English) — however, note that in most cases these tools are based on underlying dictionaries and may not generate correct IPA for some words, such as proper names. Amazon Transcribe does not endorse any third-party tools.

Q: Do I need to use different IPA standards that are specific to a different accents of the same language? (e.g. US English versus British English)?

You should use the IPA standard that is appropriate for the audio files you will be processing — e.g., if you are expecting to process audio from British English speakers, use the British English pronunciation standard. The set of allowed IPA symbols may differ for the different languages and dialects supported by Amazon Transcribe; please make sure that your pronunciations contain only the allowed characters. Details on the IPA character sets can be found in the documentation: https://docs.aws.amazon.com/transcribe/latest/dg/how-vocabulary.html#charsets

Q: How can I provide the pronunciation using SoundsLike field in the custom vocabulary table?

You can break a word or phrase down into smaller pieces and provide a pronunciation for each piece using the standard orthography of the language to mimic the way that the word sounds. For example, in English you can provide pronunciation hints for the phrase *Los-Angeles* like this: *loss-ann-gel-es*. The hint for the word Etienne would look like this: *eh-tee-en*. You separate each part of the hint with a hyphen (-). You can use any of the allowed characters for the input language.

Q: How do two different ways of providing acronyms (with periods and without periods but with pronunciations) work?

If you use an acronym containing periods, the spelling pronunciation will be generated internally. If you do not use periods, please provide the pronunciation in the pronunciation field. For some acronyms, it is not obvious whether they have a spelling pronunciation or a word-like pronunciation (e.g., NATO is often pronounced 'n eɪ t oʊ' (*nay-toh*) rather than 'ɛn eɪ ti oʊ' (N. A. T. O.)).

Q: Where can I find examples of how to use custom pronunciations?

You can find sample input formats and examples in the documentation: https://docs.aws.amazon.com/transcribe/latest/dg/how-vocabulary.html.

Q: What happens if I use the wrong IPA? If I am uncertain, am I better off not inputting any IPA?

The system will use the pronunciation you provide; this should increase the likelihood of the word being recognized correctly if the pronunciation is correct and matches what was spoken. If you are not certain you are generating correct IPA, please run a comparison by processing your audio files with a vocabulary that contains your IPA pronunciations, and with a vocabulary that only contains the words (and, optionally, display-as forms). If you do not provide any pronunciations the service will use an approximation, which may or may not work better than your input.

Q: When using DisplayAs forms, can I display character sets unrelated to the original language being transcribed? (e.g. output "Street" as "街道").

Yes. While phrases may only use a restricted set of characters for the specific language, UTF-8 characters apart from \t (TAB) are permitted in the DisplayAs column.

Q: Is Automatic content redaction available with both batch and streaming APIs for Transcribe?

No, it is only available for batch APIs at this time.

Q: What languages are supported for Automatic content redaction?

US-English (en-US) is supported at this time.

Q: Does Automatic content redaction also redact sensitive personal information from the source audio?

No, this feature does not remove sensitive personal information from the source audio. You can however redact personal information from the source audio yourself using the start and end timestamps that are provided in the redacted transcripts for each instance of an identified PII utterance.

Q: Can I use Automatic content redaction for redacting personal information from the existing text transcripts?

No, Automatic content redaction only works on audio file as an input.

Q: What else should I know before using Automatic content redaction?

Automatic content redaction is designed to identify and remove personally identifiable information (PII), but due to the predictive nature of machine learning, it may not identify and remove all instances of PII in a transcript generated by the service. You should review any output provided by Automatic content redaction to ensure it meets your needs.

Pricing & Availability

Q: What does it cost?

Refer to the Amazon Transcribe Pricing page to learn more.

Q: What AWS regions are available for Amazon Transcribe?

Please refer to the AWS Global Infrastructure Region Table.

Data Privacy

Q. Are voice inputs processed by Amazon Transcribe stored, and how are they used by AWS?

Amazon Transcribe may store and use voice inputs processed by the service solely to provide and maintain the service and to improve and develop the quality of Amazon Transcribe and other Amazon machine-learning/artificial-intelligence technologies. Use of your content is important for continuous improvement of your Amazon Transcribe customer experience, including the development and training of related technologies. We do not use any personally identifiable information that may be contained in your content to target products, services, or marketing to you or your end users. Your trust, privacy, and the security of your content are our highest priority, and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information. You may opt out of having your content used to improve and develop the quality of Amazon Transcribe and other Amazon machine-learning/artificial-intelligence technologies by contacting AWS Support.

Q. Can I delete voice inputs stored by Amazon Transcribe?

Yes. You can request deletion of voice inputs associated with your account by contacting AWS Support. Deleting voice inputs may degrade your Amazon Transcribe experience.

Q: Who has access to my content that is processed and stored by Amazon Transcribe?

Only authorized employees will have access to your content that is processed by Amazon Transcribe. Your trust, privacy, and the security of your content are our highest priority, and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information.

Q: Do I still own my content that is processed and stored by Amazon Transcribe?

You always retain ownership of your content, and we will only use your content with your consent.

Q: Is the content processed by Amazon Transcribe moved outside the AWS region where I am using Amazon Transcribe?

Any content processed by Amazon Transcribe is encrypted and stored at rest in the AWS region where you are using Amazon Transcribe. Some portion of content processed by Amazon Transcribe may be stored in another AWS region solely in connection with the continuous improvement and development of your Amazon Transcribe customer experience and other Amazon machine-learning/artificial-intelligence technologies. If you opt out of having your content used to develop the quality of Amazon Transcribe and other Amazon machine-learning/artificial-intelligence technologies by contacting AWS Support, your content will not be stored in another AWS region. You can request deletion of voice inputs associated with your account by contacting AWS Support. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information.

Q: Can I use Amazon Transcribe in connection with websites, programs or other applications that are directed or targeted to children under age 13 and subject to the Children's Online Privacy Protection Act (COPPA)?

Yes, subject to your compliance with the Amazon Transcribe Service Terms, including your obligation to provide any required notices and obtain any required verifiable parental consent under COPPA, you may use Amazon Transcribe in connection with websites, programs, or other applications that are directed or targeted, in whole or in part, to children under age 13.

Q: How do I determine whether my website, program, or application is subject to COPPA?

For information about the requirements of COPPA and guidance for determining whether your website, program, or other application is subject to COPPA, please refer directly to the resources provided and maintained by the United States Federal Trade Commission. This site also contains information regarding how to determine whether a service is directed or targeted, in whole or in part, to children under age 13.

Amazon Transcribe Medical

Q. What is Amazon Transcribe Medical?

Amazon Transcribe Medical is an automatic speech recognition (ASR) service that makes it easy for developers to add medical speech-to-text capabilities to their applications. Using Amazon Transcribe Medical, you can quickly and accurately transcribe medical dictation and conversational speech into text for a variety of purposes, such as recording physician notes or processing in downstream text analytics to extract meaningful insights.

Q. What can I do with Amazon Transcribe Medical?

Amazon Transcribe Medical uses advanced machine learning models to accurately transcribe medical speech into text. Transcribe Medical can general text transcripts that can be used to support a variety of use cases, spanning clinical documentation workflow and drug safety monitoring (pharmacovigilance) to subtitling for telemedicine and even contact center analytics in the healthcare and life sciences domains.

Q. Do I need to be an expert in automatic speech recognition (ASR) to use Amazon Transcribe Medical?

No, you don’t need any ASR or machine learning expertise to use Amazon Transcribe Medical. You only need to call Transcribe Medical’s API, and the service will handle the required machine learning in the backend to transcribe medical speech to text.

Q. How do I get started with Amazon Transcribe Medical?

You can get started with Amazon Transcribe Medical from the AWS Management console or by using the SDK. Please refer to this technical documentation page for details.

Amazon Transcribe Medical provides a free tier so that you can test the service. Refer to this pricing page for more information.

Q. Which languages does Amazon Transcribe Medical support?

Amazon Transcribe Medical currently supports medical transcription in US English.

Q. Which medical specialties does Amazon Transcribe Medical support?

Amazon Transcribe Medical supports transcription for primary care, covering specialties such as family medicine, internal medicine, pediatrics, and OB-GYN.

Q. In which AWS regions is Amazon Transcribe Medical available?

Amazon Transcribe Medical is currently available in US-East (N. Virginia), US East (Ohio), US West (Oregon), Canada (Central), EU (Ireland), and Asia Pacific (Sydney).

Q. How is Amazon Transcribe Medical priced?

Refer to the Amazon Transcribe Medical pricing page to learn more about pricing details.

Q. Is Amazon Transcribe Medical HIPAA eligible?

Yes.

Q. Is the content processed by Amazon Transcribe Medical used for any purpose other than to provide the service?

Amazon Transcribe Medical does not use content processed by the service for any reason other than to provide and maintain the service. Content processed by the service is not used to develop or improve the quality for Amazon Transcribe Medical or any other Amazon machine-learning/artificial-intelligence technologies.

Q. Does Amazon Transcribe Medical learn over time?

Yes, Amazon Transcribe Medical uses machine learning and is continuously being trained to make it better for customer use cases. Amazon Transcribe medical does not store or use customer data used with the service to train the models

Q. What else should I know before using the Amazon Transcribe Medical service?

Amazon Transcribe Medical is not a substitute for professional medical advice, diagnosis, or treatment. You and your end users are responsible for exercising your and their own discretion, experience, and judgment in determining the correctness, completeness, timeliness, and suitability of any information provided by Amazon Transcribe Medical. You and your end users are solely responsible for any decisions, advice, actions, and/or inactions based on the use of Amazon Transcribe Medical. You are responsible for reviewing any output provided by Amazon Transcribe Medical to ensure it meets your needs.

AWS Deep Learning Containers FAQs

Q: What are AWS Deep Learning Containers?

AWS Deep Learning Containers (AWS DL Containers) give machine learning and deep learning practitioners optimized Docker environments to train and deploy models in their pipelines and workflows across Amazon Sagemaker, Amazon EC2, Amazon ECS, and Amazon EKS. AWS DL Containers are available as Docker images for training and inference with TensorFlow, PyTorch, and MXNet on Amazon ECR.

Q: Why should I use AWS DL Containers?

Building, testing, maintaining, and optimizing Docker images for deep learning requires a sustained investment in time and resources by data scientists, machine learning developers, and practitioners. Instead of focusing on building and improving models, practitioners have to spend valuable resources in undifferentiated tasks. These tasks can include installing packages, debugging compatibility issues, optimizing for performance, and integrating and testing with Amazon Sagemaker, Amazon EC2, Amazon ECS, and Amazon EKS. AWS DL Containers offer fully tested and optimized deep learning Docker environments that require no installation, configuration, or maintenance. Deep learning practitioners looking to train and serve models in TensorFlow, PyTorch, or Apache MXNet, get what they need packaged and optimized in these Docker images.

Q. How does this service relate to/work with other AWS services?

AWS DL Containers are built, tested, and optimized to be used in Amazon Sagemaker, Amazon EC2, Amazon ECS, and Amazon EKS. Docker images for AWS DL Containers are available on Amazon ECR. For training and inference of deep learning models using GPUs, AWS DL Containers require the underlying Amazon Machine Image (AMI) to have the appropriate GPU drivers installed. DL Containers are built to work with the default GPU AMIs available in Amazon SageMaker, Amazon ECS, and Amazon EKS.

Q. How do AWS DL Containers work with AWS Deep Learning AMIs?

AWS Deep Learning AMIs are EC2 Amazon Machine Images (AMIs) built and optimized for building, training, and inference of machine learning and deep learning models. For more information, see AWS Deep Learning AMIs. For more information about using AWS DL Containers in EC2, see the documentation.

Q. Do I need to pay to use AWS DL Containers?

AWS DL Containers are available at no additional charge. You pay only for the Amazon Sagemaker, Amazon EC2, Amazon ECS, Amazon EKS, and other AWS resources that you use.

Q. How do I access Docker images for AWS DL Containers?

You can access Docker images for AWS DL Containers from repositories in Amazon ECR. For more information, see the documentation for a list of available Docker images.

AWS DeepComposer FAQ

General

Q: What is AWS DeepComposer?
AWS DeepComposer is the world’s first musical keyboard powered by machine learning to enable developers of all skill levels to learn Generative AI while creating original music outputs. DeepComposer consists of a USB keyboard that connects to the developer’s computer, and the DeepComposer service, accessed through the AWS Management Console. DeepComposer includes tutorials, sample code, and training data that can be used to start building generative models.

Q: How is AWS DeepComposer different from other musical keyboards in the market?
AWS DeepComposer is the world’s first musical keyboard designed specifically to work with the DeepComposer service to teach developers Generative AI. AWS DeepComposer gives developers a simple way to learn and experiment with Generative AI algorithms, train models, and compose musical outputs.

Q: What level of musical knowledge do I need?
No musical knowledge is required to use DeepComposer. DeepComposer provides you a quick and easy way to get started by providing sample melodies such as Twinkle, Twinkle, Little Star, or Ode to Joy. You can use these sample melodies as an input to generate an entirely new musical output, with a 4 part accompaniment.

Q: What are Generative AI algorithms?
Viewed by some as the most interesting machine learning idea in a decade, Generative AI allows computers to learn the underlying pattern of a given problem and use this knowledge to generate new content from input (such as image, music, and text). In contrast to more commonly used machine learning models that learn to differentiate, for example between images of cats and dogs (by identifying traits that set them apart), a Generative AI model based on

cat images would learn the features that are common across cats, and use that knowledge to generate all-new images of what it believes are cats. This difference is significant because with the advancement in Generative AI algorithms, machines can automatically discover and learn the patterns in data and generate new data based on the data they were trained on.

Q: How can I get an AWS DeepComposer keyboard?
AWS DeepComposer keyboard will be available for pre-order in the US via amazon.com. More information to follow.

Q: Do I have to purchase the DeepComposer keyboard to use the AWS DeepComposer Service?
You will have the best experience with the DeepComposer keyboard, through integration with keyboard buttons that can control recording of musical phrases and generation of new musical outputs when working with the DeepComposer cloud service. For those without the DeepComposer keyboard, the management console includes an on-screen virtual keyboard that allows developers to input musical notes in a similar fashion.

Q: Which geographic regions will AWS DeepComposer be available in?
The musical keyboard is available only in the US. However, you can use the virtual keyboard DeepComposer management console provides from anywhere in the world by signing into US East (N. Virginia) Region.

Q: Which AWS regions will AWS DeepComposer be available in?
Customers can access the AWS DeepComposer console from the US East (N. Virginia) Region.

Product Details

Q: What are the product specifications of the AWS DeepComposer keyboard?

- **Item weight:** 1.68 pounds
- **Product dimensions:** 18.1x4.9x1.2 inches
- **Shipping weight:** 2.3 pound
- **Features:** 32 velocity-sensitive keys, 1 endless encoder, 3 rotary knobs and 11 function buttons with LED back lit, USB powered

Q: What pre-trained genre models are available at launch?
DeepComposer will come with four pre-trained genre models: rock, pop, jazz, and classical.

Getting Started

Q: How do I get started with AWS DeepComposer?
There are two ways you can get started. First, you can create music using one of the sample models that have been pre-trained on musical inputs from multiple genres, such as pop and classic. You input a melody using the keyboard and DeepComposer performs inferences to generate an output with a 4 part accompaniment in that genre. You can also take the learning further by creating your own custom models using one of DeepComposer's publicly available SageMaker notebooks. You can also bring your own dataset and build a customized machine learning notebook in SageMaker to learn and create your own Generative AI models.

Q: Do I need to be connected to internet to run the models?
Yes. DeepComposer is a cloud service, connection to the internet is required to run inference against models for musical creations.

Q: Will I have to bring my own dataset in order to train models?
No. DeepComposer comes with pre-trained genre models to help you get started with Generative AI technologies.

Q: Can I bring my own dataset?
Yes. You can bring your own music dataset in MIDI format and create your own custom models in SageMaker.

Q: How can I run my own custom models?
You can run your custom models within DeepComposer console where you'll be able to optimize for hyperparameters and select your dataset.

Q: Can I save and export my musical outputs (such as generated compositions) or share them with the community?

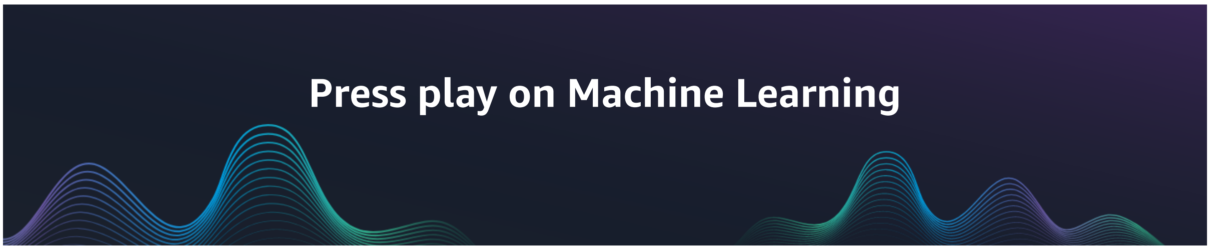
Yes, you can save and export your musical creations in MIDI for additional processing using external tools, or in wav or mp3 format for sharing. Choose either the ‘Download MIDI’ or ‘Submit to SoundCloud’ button on DeepComposer console to export and save your creations.

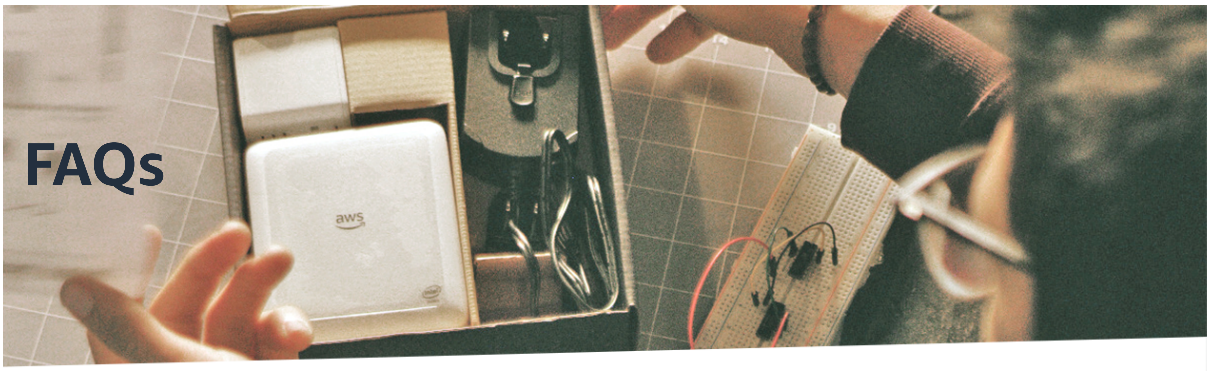
Q: Can I save and export my input melodies?

No, not at this time. You will be able to save and export your input melodies after GA of the service.

Q: How do I submit my creations to SoundCloud?

You can submit your creations to SoundCloud by choosing the Submit to SoundCloud button on the DeepComposer console. You will be required to log into your SoundCloud account for permissions.





FAQs

General

Q: What is AWS DeepLens? >>

Q: What is the AWS DeepLens (2019 Edition)? >>

Q: How is AWS DeepLens different from other video cameras in the market? >>

Q: What sample projects are available at launch? >>

Q: What geographic regions is AWS DeepLens available? >>

Q: Does AWS DeepLens include Alexa? >>

Q: How can I get a AWS DeepLens? >>

Q: In what regions will the AWS DeepLens console be available? >>

Product Details

Q: What are the product specifications of the device? >>

Q: What deep learning frameworks can I run on the device? >>

Q: What kind of performance can I expect with AWS DeepLens? >>

Q: What MXNet network architecture layers does AWS DeepLens support? >>

Getting Started

Q: What comes in the box and how do I get started? >>

Q: Can I train my models on the device? >>

Q: What AWS services are integrated with AWS DeepLens? >>

Q: Can I SSH into AWS DeepLens? >>

Q: What programming languages are supported by AWS DeepLens? >>

Q: Do I need to be connected to internet to run the models? >>

Q: Can I run my own custom models on AWS DeepLens? >>

Q: Why do I have "v1.1" marked on the bottom of my device? >>

AWS DeepRacer FAQs

General

Q: What is AWS DeepRacer
AWS DeepRacer is the fastest way to get rolling with reinforcement learning (RL), literally, with a fully autonomous 1/18th scale race car driven by reinforcement learning, 3D racing simulator, and a global racing league. Developers can train, evaluate, and tune RL models in the online simulator, deploy their models onto AWS DeepRacer for a real-world autonomous experience and compete in the AWS DeepRacer League for a chance to win the AWS DeepRacer Championship Cup.

Q: What are the new features of AWS DeepRacer?
AWS DeepRacer customers can take on their next machine learning challenge using AWSDeepRacer with the launch of multi-car racing and object avoidance capabilities in the AWSDeepRacer console. Customers can now build models for object avoidance and dual-car head-to-head races by experimenting with multiple sensor inputs and the latest reinforcement learningalgorithms and neural network configurations. Developers can build reinforcement learning modelsready to deploy to AWS DeepRacer Evo in the 2020 season of the AWS DeepRacer League.

Q: How is AWS DeepRacer different from other robotic scale cars in the market?
AWS DeepRacer is the first autonomous scale car specifically developed to help developers get hands-on with reinforcement learning. AWS DeepRacer gives developers a simple way to learn RL, experiment with new RL algorithms and simulation-to-real domain transfer methods, and experience RL in the real world.

Q: What is the relationship between AWS DeepRacer and AWS Robocar Rally?
After the AWS DeepRacer launch, all Robocar Rally events will be replaced with

AWS DeepRacer events. Robocar Rally inspired AWS DeepRacer, but unlike Robocar Rally, AWS DeepRacer focuses on RL and not on behavioral cloning.

Q: How can I get an AWS AWS DeepRacer?
You can access the AWS DeepRacer 3D racing simulator in the AWS DeepRacer console to train your models, evaluate them, and take part in the AWS DeepRacer League. The AWS DeepRacer car is available for purchase on on Amazon.com.

Q: How can I get an AWS AWS DeepRacer Evo?
AWS DeepRacer Evo will be available in 2020. If you would like to be notified when it is ready to ship, you can sign up to the AWS DeepRacer Evo interest list.

Q: Which geographic regions will AWS DeepRacer be available in?
AWS Customers can access the AWS AWS DeepRacer simulator from the US East (N. Virginia) Region. The AWS DeepRacer car is available to order on Amazon.com for customers in the USA.

Q: Which geographic regions will AWS DeepRacer be available in?
AWS Customers can access the AWS DeepRacer simulator from the US East (N. Virginia) Region. The AWS DeepRacer car is available to pre-order on Amazon.com for customers in the USA.

Q: What is the AWS DeepRacer League?
The AWS DeepRacer League is the world's first global autonomous racing league for developers.AWS Customers can use their AWS DeepRacer RL models to compete in a global championship,racing for prizes, glory, and a chance to lift the Championship Cup at re:Invent 2020. There are two ways to participate in the AWS DeepRacer League in 2020 – racing in person racing in the SummitCircuit, or online in the Virtual Circuit.

Learn more about the league in the dedicated FAQ section below

Getting Started

Q: How do I get started with AWS DeepRacer simulator?
The AWS DeepRacer simulator provides a tutorial to get you started with reinforcement learning and training your first model. You will then also be able to evaluate and tune your models, before racing them in the AWS DeepRacer League. The AWS DeepRacer Developer Documentation provides additional details on building your first model and also how to improve your models.

Q: How can customers get started with multi-car racing and object avoidance?
Customers can login into the AWS DeepRacer console from anywhere in the world. Here they can experiment with new sensor configurations in the Garage section of the console and build reinforcement learning models for head to head and object avoidance racing.

Q: How can I start my own Community Race?
Customers can create their own virtual races the Community Races section of the AWS DeepRacer console.

Q: Do I need an AWS DeepRacer car in order to use the AWS DeepRacer simulator?
No. You can train models, evaluate them without owning an AWS DeepRacer car. Furthermore, you can race your models in both the AWS DeepRacer League Virtual and Summit Circuits without owning an AWS DeepRacer car.

Q: Does my DeepRacer need to be connected to the internet to race autonomously?
No. AWS DeepRacer uses the deployed RL model and input from the camera to run inference locally. AWS DeepRacer must be connected to the same Wi-Fi network as the device used to start and stop autonomous driving. Details on how to set up your vehicle can be found in the Developer Documentation.

AWS DeepRacer League

Q: How do I race in the DeepRacer League?

There are two ways to participate in the AWS DeepRacer League in 2020, in-person at AWSSummits globally or online in the 3D racing simulator:

1. Virtual Circuit: Join the AWS DeepRacer League from anywhere in the world via the AWSDeepRacer console and put your skills to the test on virtual tracks monthly in time trial, objectavoidance, and head-to-head formats. Race for prizes and glory and a chance to be eligible to winan expenses paid trip to the AWS DeepRacer Championship Cup at re:Invent 2020. The racer withthe fastest time at the end of the monthly race in each racing format, will be eligible to win a trip tocompete in the Championship Cup at re:Invent 2020.

2. Summit Circuit: Join us at any of the thirty-seven in-person race events globally, where we willhelp you build and train a model at a workshop, or you can bring one you have trained at home. Youcan then put your model to the test and compete on the track at the event in time trial and head-to-head formats. There are no limits on how many of the 37 events you can participate in. Simplyregister for and attend the AWS Summits you want to race at, and we'll see you there!

The League will conclude when all qualifying participants come together at AWS re:Invent 2020: Joinus in Las Vegas for a last chance to qualify for the knockout rounds. At the knockout rounds,participants will race in a series of elimination rounds, with the top racers advancing to the final todetermine the winner of the 2020 AWS DeepRacer Championship Cup.

Q: How many times can I enter the League?
There is no limit on the number of races or race formats you can enter, in-person or virtually. In fact,participating in multiple racing events (at a summit or select AWS events, or a virtual monthlychallenge), increases your chances of winning one of the prizes to advance to re:Invent 2020, andrace for the AWS DeepRacer Championship Cup. You can race in any combination of Summit Circuit races and Virtual Circuit races, plus you can participate in multiple races and formats in each Circuit.

Q: What are the prizes?
Details of the AWS DeepRacer League prizes can be found here. For full details please read the AWS DeepRacer League Terms and Conditions.

Q: Do I have to build my own model to race?
No, you do not need to build your own model to compete in the AWS DeepRacer League. The model selector at Summit Circuit races gives you the opportunity to participate using one of the pre-trained models we have available. For the Virtual Circuit, there are tutorials inside the AWS DeepRacer simulator on how to build your own model, as well as pre-trained sample models you can use to get started.

Q: How much time do you get to race on a track in the Summit Circuit?
You have three minutes on the track to attempt to get a valid lap time and enter the leaderboard. Intime trial racing, a valid lap is one where your car can complete a lap around the track without goingoff the track more than three times. In head-to-head racing, a valid lap is one where your car can complete a lap before our AWS DeepRacer X bot car crosses the finish line. Full rules can be foundin the AWS DeepRacer League Terms and Conditions.

Q: How long does a Virtual Circuit race last?
Each virtual race will last one month (from the 1st business day of the month to the last day of the month).

Q: How many Virtual Circuit races are there?
Starting on March 2nd, the Virtual Circuit will have 8 monthly races, with three racing formats in each(time trial, object avoidance, and head-to-head). Each race will be revealed in the console on thefirst business day of each month leading up to re:Invent 2020 (March – October).

Q: How many Summit Circuit events are there?
There are 37 in person races at AWS Summits and select amazon events. For a full race schedule please visit the schedule and standings page.

Q: How do I score points?
In 2020, the AWS DeepRacer league will no longer track cumulative points. Winners will be awarded based on single race performance only.

Q: What do I get if I win a race in either the Summit Circuit or Virtual Circuit?
The winner of each race will win an expenses paid trip to compete in the AWS DeepRacer knockout rounds at re:Invent 2020, where they will race to compete in the AWS DeepRacer Championship Cup Final. The top 5 in each racing format will win AWS DeepRacer cars, where applicable. For more information on prizes, please read the AWS DeepRacer League rules.

Product Details

Q: What are the product specifications of the AWS DeepRacer device?

Car: 1/18th scale 4WD with monster truck chassis
CPU: Intel Atom™ Processor
Memory: 4GB RAM
Storage: 32GB (expandable)
Wi-Fi: 802.11ac
Camera: 4 MP camera with MJPEG
Software: Ubuntu OS 16.04.3 LTS, Intel® OpenVINO™ toolkit, ROS Kinetic
Drive battery: 7.4V/1100mAh lithium polymer
Compute battery: 13600mAh USB-C PD
Ports: 4x USB-A, 1x USB-C, 1x Micro-USB, 1x HDMI
Sensors: Integrated accelerometer and gyroscope

Additional sensors on AWS DeepRacer Evo are LiDAR and 1 stereo 4 MP camera with MJPEG.

Q: If I already own an AWS DeepRacer car will it be able to master multi-car racing and object avoidance?
Developers who already own a DeepRacer car can buy an easy-to-install sensor kit from Amazon.com, also available in 2020, to give their car the same capabilities as AWS DeepRacer Evo. The AWS DeepRacer Sensor Kit includes stereo camera and LIDAR, and you can sign up for the interest list here.

Q: What does the LIDAR and stereo camera do?
LIDAR stands for light detection and ranging. It provides a continuous light beam providing the reinforcement model with inputs about whether a car is fast approaching from behind. The stereo camera adds depth perception to allow the car to detect objects in the road and be more responsive to its environment. Combining these sensory inputs with advanced algorithms, and updated reward functions, developers can build models that will not only detect obstacles (including other cars), but will also decide when to overtake to beat the other car to the finish line.

Q: Which AWS services are integrated with AWS DeepRacer?
AWS DeepRacer integrates with Amazon SageMaker for reinforcement learning model training, AWS RoboMaker to provide the racing simulator, Amazon Kinesis Video Streams for video streaming of virtual simulation footage, Amazon S3 for model storage, and Amazon CloudWatch for log capture.

Q: How do I add new tracks to the AWS DeepRacer simulator?
Currently, developers cannot add additional tracks to the AWS DeepRacer simulator. AWS DeepRacer will release a number of new racing tracks throughout 2020.

Q: Can I train my models on the AWS DeepRacer device?
No. Training an RL model requires feedback regarding the outcome of actions taken by the model. This feedback loop exists in the AWS DeepRacer simulator, but not in the real-world.

Q: Can I train my models locally on my own machine as opposed to the AWS Cloud?

Currently AWS DeepRacer does not support local training.

Q: Can I train AWS DeepRacer RL models directly on Amazon SageMaker?
Yes. You can use the AWS DeepRacer Distributed Training SageMaker Notebook to create and train RL models. You will be able to deploy these models to your AWS DeepRacer manually, but won’t yet be able to import them into the AWS DeepRacer console.

Amazon CloudWatch FAQs

General

Q: What is Amazon CloudWatch?

Amazon CloudWatch is a monitoring service for AWS cloud resources and the applications you run on AWS. You can use Amazon CloudWatch to collect and track metrics, collect and monitor log files, and set alarms. Amazon CloudWatch can monitor AWS resources such as Amazon EC2 instances, Amazon DynamoDB tables, and Amazon RDS DB instances, as well as custom metrics generated by your applications and services, and any log files your applications generate. You can use Amazon CloudWatch to gain system-wide visibility into resource utilization, application performance, and operational health. You can use these insights to react and keep your application running smoothly.

To get started with monitoring, you can use Automatic Dashboards with built-in AWS best practices, explore account and resource-based view of metrics and alarms, and easily drill-down to understand the root cause of performance issues.

Q: What can I use to access CloudWatch?

Amazon CloudWatch can be accessed via API, command-line interface, AWS SDKs, and the AWS Management Console.

Q: Which operating systems does Amazon CloudWatch support?

Amazon CloudWatch receives and provides metrics for all Amazon EC2 instances and should work with any operating system currently supported by the Amazon EC2 service.

Q: What access management policies can I implement for CloudWatch?

Amazon CloudWatch integrates with AWS Identity and Access Management (IAM) so that you can specify which CloudWatch actions a user in your AWS Account can perform. For example, you could create an IAM policy that gives only certain users in your organization permission to use GetMetricStatistics. They could then use the action to retrieve data about your cloud resources.

You can't use IAM to control access to CloudWatch data for specific resources. For example, you can't give a user access to CloudWatch data for only a specific set of instances or a specific LoadBalancer. Permissions granted using IAM cover all the cloud resources you use with CloudWatch. In addition, you can't use IAM roles with the Amazon CloudWatch command line tools.

Q: What is Amazon CloudWatch Logs?

Amazon CloudWatch Logs lets you monitor and troubleshoot your systems and applications using your existing system, application and custom log files.

With CloudWatch Logs, you can monitor your logs, in near real time, for specific phrases, values or patterns. For example, you could set an alarm on the number of errors that occur in your system logs or view graphs of latency of web requests from your application logs. You can then view the original log data to see the source of the problem. Log data can be stored and accessed indefinitely in highly durable, low-cost storage so you don’t have to worry about filling up hard drives.

Q: What kinds of things can I do with CloudWatch Logs?

CloudWatch Logs is capable of monitoring and storing your logs to help you better understand and operate your systems and applications. You can use CloudWatch Logs in a number of ways.

Real time application and system monitoring: You can use CloudWatch Logs to monitor applications and systems using log data. For example, CloudWatch Logs can track the number of errors that occur in your application logs and send you a notification whenever the rate of errors exceeds a threshold you specify. CloudWatch Logs uses your log data for monitoring; so, no code changes are required.

Long term log retention: You can use CloudWatch Logs to store your log data indefinitely in highly durable and cost effective storage without worrying about hard drives running out of space. The CloudWatch Logs Agent makes it easy to quickly move both rotated and non rotated log files off of a host and into the log service. You can then access the raw log event data when you need it.

Q: What platforms does the CloudWatch Logs Agent support?

The CloudWatch Logs Agent is supported on Amazon Linux, Ubuntu, CentOS, Red Hat Enterprise Linux, and Windows. This agent will support the ability to monitor individual log files on the host.

Q: Does the CloudWatch Logs Agent support IAM roles?

Yes. The CloudWatch Logs Agent is integrated with Identity and Access Management (IAM) and includes support for both access keys and IAM roles.

Q: What is Amazon CloudWatch Logs Insights?

Amazon CloudWatch Logs Insights is an interactive, pay-as-you-go, and integrated log analytics capability for CloudWatch Logs. It helps developers, operators, and systems engineers understand, improve, and debug their applications, by allowing them to search and visualize their logs. Logs Insights is fully integrated with CloudWatch, enabling you to manage, explore, and analyze your logs. You can also leverage CloudWatch Metrics, Alarms and Dashboards with Logs to get full operational visibility into your applications. This empowers you to understand your applications, make improvements, and find and fix problems quickly, so that you can continue to innovate rapidly. You can write queries with aggregations, filters, and regular expressions to derive actionable insights from your logs. You can also visualize timeseries data, drill down into individual log events, and export your query results to CloudWatch Dashboards.

Q: How can I get started with CloudWatch Logs Insights?

You can immediately start using Logs Insights to run queries on all your logs being sent to CloudWatch Logs. There is no setup required and no infrastructure to manage. You can access Logs Insights from the AWS Management Console or programmatically through your applications by using the AWS SDK.

Q: What is CloudWatch Container Insights?

CloudWatch Container Insights is a feature for monitoring, troubleshooting, and alarming on your containerized applications and microservices. Container Insights simplifies the isolation and analysis of performance issues impacting your container environment. DevOps and systems engineers have access to automatic dashboards in the CloudWatch console, giving them end-to-end operational visibility of metrics, logs, and distributed traces summarizing the performance and health of their Amazon Elastic Container Service for Kubernetes (EKS), Amazon Elastic Container Service (ECS), AWS Fargate, and Kubernetes clusters by pods/tasks, containers, and services.

Q: How can I get started with CloudWatch Container Insights?

You can get started collecting detailed performance metrics, logs, and metadata from your containers and clusters in just a few clicks by following these steps in the CloudWatch Container Insights documentation.

Q: What is Amazon CloudWatch Anomaly Detection?

Amazon CloudWatch Anomaly Detection applies machine-learning algorithms to continuously analyze single time series of systems and applications, determine a normal baseline, and surface anomalies with minimal user intervention. It allows you to create alarms that auto-adjust thresholds based on natural metric patterns, such as time of day, day of week seasonality or changing trends. You can also visualize metrics with anomaly detection bands on dashboards, monitoring, isolating, and troubleshooting unexpected changes in your metrics.

Q: How can I get started with Amazon CloudWatch Anomaly Detection?

It is easy to get started with Anomaly Detection. In the CloudWatch console, go to *Alarms *in the navigation pane to create an alarm, or start with *Metrics *to overlay the metric's expected values onto the graph as a band. You can also enable Anomaly Detection using the AWS CLI, AWS SDKs, or AWS

CloudFormation templates. To learn more, please visit the CloudWatch Anomaly Detection documentation and pricing pages.

Q: What is Amazon CloudWatch Contributor Insights?

Amazon CloudWatch now includes Contributor Insights, which analyzes time-series data to provide a view of the top contributors influencing system performance. Once set up, Contributor Insights runs continuously without needing additional user intervention. This helps developers and operators more quickly isolate, diagnose, and remediate issues during an operational event.

Q: How can I get started with CloudWatch Contributor Insights?

It is easy to get started with Contributor Insights. In the CloudWatch console, go to Contributor Insights in the navigation pane to create a Contributor Insights rule. You can also enable Contributor Insights using the AWS CLI, AWS SDKs, or AWS CloudFormation templates. Contributor Insights is available in all commercial AWS Regions. To learn more, please visit the documentation on CloudWatch Contributor Insights.

Q: What is Amazon CloudWatch ServiceLens?

Amazon CloudWatch ServiceLens is a new feature that enables you to visualize and analyze the health, performance, and availability of your applications in a single place. CloudWatch ServiceLens ties together CloudWatch metrics and logs as well as traces from AWS X-Ray to give you a complete view of your applications and their dependencies. This enables you to quickly pinpoint performance bottlenecks, isolate root causes of application issues, and determine users impacted. CloudWatch ServiceLens enables you to gain visibility into your applications in three main areas: Infrastructure monitoring (using metrics and logs to understand the resources supporting your applications), transaction monitoring (using traces to understand dependencies between your resources), and end user monitoring (using canaries to monitor your endpoints and notify you when your end user experience has degraded).

Q: How can I get started with CloudWatch ServiceLens?

It’s easy to get started. If you already use AWS X-Ray, you can access CloudWatch ServiceLens on the CloudWatch console by default. If you do not yet use AWS X-Ray, you can get started by enabling AWS X-Ray on your applications using the X-Ray SDK. Amazon CloudWatch ServiceLens is available in all public AWS Regions where AWS-X-Ray is available. To learn more, visit the documentation on Amazon CloudWatch ServiceLens.

Q: What is Amazon CloudWatch Synthetics?

Amazon CloudWatch Synthetics allows you to monitor application endpoints more easily. It runs tests on your endpoints every minute, 24x7, and alerts you as soon as your application endpoints don’t behave as expected. These tests can be customized to check for availability, latency, transactions, broken or dead links, step by step task completions, page load errors, load latencies for UI assets, complex wizard flows, or checkout flows in your applications. You can also use CloudWatch Synthetics to isolate alarming application endpoints and map them back to underlying infrastructure issues to reduce mean time to resolution.

Q: How can I get started with CloudWatch Synthetics?

It's easy to get started with CloudWatch Synthetics. You can write your first passing canary in minutes. Amazon CloudWatch Synthetics is available in preview in the following public AWS Regions: US East (N. Virginia), US East (Ohio), and EU (Ireland). To learn more, visit the documentation on Amazon CloudWatch Synthetics.

Pricing

Q: How much does Amazon CloudWatch cost?

Please see our pricing page for the latest information.

Q: Does the Amazon CloudWatch monitoring charge change depending on which type of Amazon EC2 instance I monitor?

All Amazon EC2 instance types automatically send key health and performance metrics to CloudWatch at no cost. If you enable EC2 Detailed Monitoring, you will be charged for custom metrics based on the number of metrics sent to CloudWatch for the instance. The number of metrics sent for an instance is dependent on the instance type - see available CloudWatch Metrics for Your Instances for details.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. Learn more.

Q: Why does my AWS monthly bill for CloudWatch appear different between July 2017 and previous months?

Prior to July 2017, charges for CloudWatch were split under two different sections in your AWS bill and Cost and Usage Reports. For historical reasons, charges for CloudWatch Alarms, CloudWatch Metrics, and CloudWatch API usage were reported under the “Elastic Compute Cloud” (EC2) detail section of your bill, while charges for CloudWatch Logs and CloudWatch Dashboards are reported under the “CloudWatch” detail section. To help consolidate and simplify your monthly AWS CloudWatch usage and billing, we moved the charges for your CloudWatch Metrics, Alarms, and API usage from the “EC2” section of your bill to the “CloudWatch” section, effectively bringing together all of your CloudWatch monitoring charges under the “CloudWatch” section. Note that this has no impact to your total AWS bill amount. Your bill and Cost and Usage Reports will now simply display charges for CloudWatch under a single section.

Additionally, there is a Billing Metric in CloudWatch named “Estimated Charges” that can be viewed as Total Estimated Charge or broken down By Service. The “Total Estimated Charge” metric will not change. However, the “EstimatedCharges” metric broken down by Service will change for dimension ServiceName equal to “AmazonEC2” and dimension ServiceName equal to “AmazonCloudWatch”. Due to the billing consolidation, you may see that your AmazonEC2 billing metric decrease and AmazonCloudWatch billing metric increase as usage and billing charges get moved out of EC2 and into CloudWatch.

Q: How is CloudWatch Logs Insights priced?

Logs Insights is priced per query and charges based on the amount of ingested log data scanned by the query. For additional details about pricing, you can see CloudWatch pricing.

Q: Does CloudWatch Logs Insights charge me for cancelled queries?

Yes, if you cancel a query manually, you are charged for the amount of ingested log data scanned up to the point at which you cancelled the query.

Q: Does CloudWatch Logs Insights charge me for failed queries?

No, you are not charged for failed queries.

Q: How is CloudWatch Container Insights priced?

CloudWatch Container Insights automatically collects custom metrics from performance events ingested as CloudWatch Logs from your container environment. More details on pricing is available on the CloudWatch pricing page.

AWS resource & custom metrics monitoring

Q: What can I measure with Amazon CloudWatch Metrics?

Amazon CloudWatch allows you to monitor AWS cloud resources and the applications you run on AWS. Metrics are provided automatically for a number of AWS products and services, including Amazon EC2 instances, EBS volumes, Elastic Load Balancers, Auto Scaling groups, EMR job flows, RDS DB instances, DynamoDB tables, ElastiCache clusters, RedShift clusters, OpsWorks stacks, Route 53 health checks, SNS topics, SQS queues, SWF workflows, and Storage Gateways. You can also monitor custom metrics generated by your own applications and services.

Q: What is the retention period of all metrics?

CloudWatch launched High Resolution Custom Metrics on July 26, 2017. This enables you to publish and store custom metrics down to 1-second resolution. Extended retention of metrics was launched on November 1, 2016, and enabled storage of all metrics for customers from the previous 14 days to 15 months. CloudWatch retains metric data as follows:

- Data points with a period of less than 60 seconds are available for 3 hours. These data points are high-resolution custom metrics.
- Data points with a period of 60 seconds (1 minute) are available for 15 days
- Data points with a period of 300 seconds (5 minute) are available for 63 days
- Data points with a period of 3600 seconds (1 hour) are available for 455 days (15 months)

Data points that are initially published with a shorter period are aggregated together for long-term storage. For example, if you collect data using a period of 1 minute, the data remains available for 15 days with 1-minute resolution. After 15 days this data is still available, but is aggregated and is retrievable only with a resolution of 5 minutes. After 63 days, the data is further aggregated and is available with a resolution of 1 hour. If you need availability of metrics longer than these periods, you can use the GetMetricStatistics API to retrieve the datapoints for offline or different storage.

The feature is currently available in US East (N. Virginia), US West (Oregon), US West (N. California), EU (Ireland), EU (Frankfurt), S. America (São Paulo), Asia Pacific (Singapore), Asia Pacific (Tokyo), Asia Pacific (Seoul), Asia Pacific (Mumbai), Asia Pacific (Sydney), EU (London), Canada (Central), US East (Ohio), and China (Beijing).

Q: What is the minimum resolution for the data that Amazon CloudWatch receives and aggregates?

The minimum resolution supported by CloudWatch is 1-second data points, which is a high-resolution metric, or you can store metrics at 1-minute granularity. Sometimes metrics are received by Cloudwatch at varying intervals, such as 3-minute or 5-minute intervals. If you do not specify that a metric is high resolution, by setting the StorageResolution field in the PutMetricData API request, then by default CloudWatch will aggregate and store the metrics at 1-minute resolution.

Depending on the age of data requested, metrics will be available at the resolutions defined in the retention schedules above. For example, if you request for 1-minute data for a day from 10 days ago, you will receive the 1440 data points. However, if you request for 1-minute data from 5 months back, the UI will automatically change the granularity to 1-hour and the GetMetricStatistics API will not return any output.

Q: Can I delete any metrics?

CloudWatch does not support metric deletion. Metrics expire based on the retention schedules described above.

Q: Will I lose the metrics data if I disable monitoring for an Amazon EC2 instance?

No. You can always retrieve metrics data for any Amazon EC2 instance based on the retention schedules described above. However, the CloudWatch console limits the search of metrics to 2 weeks after a metric is last ingested to ensure that the most up to date instances are shown in your namespace.

Q: Can I access the metrics data for a terminated Amazon EC2 instance or a deleted Elastic Load Balancer?

Yes. Amazon CloudWatch stores metrics for terminated Amazon EC2 instances or deleted Elastic Load Balancers for 15 months.

Q: Why does the graphing of the same time window look different when I view the metrics in 5 minute and 1 minute periods?

If you view the same time window in a 5 minute period versus a 1 minute period, you may see that data points are displayed in different places on the graph. For the period you specify in your graph, Amazon CloudWatch will find all the available data points and calculates a single, aggregate point to represent the entire period. In the case of a 5 minute period, the single data point is placed at the beginning of the 5 minute time window. In the case of a 1 minute period, the single data point is placed at the 1 minute mark. We recommend using a 1 minute period for troubleshooting and other activities that require the most precise graphing of time periods.

Q: What is a Custom Metric?

You can use Amazon CloudWatch to monitor data produced by your own applications, scripts, and services. A custom metric is any metric you provide to Amazon CloudWatch. For example, you can use custom metrics as a way to monitor the time to load a web page, request error rates, number of processes or threads on your instance, or amount of work performed by your application. You can get started with custom metrics by using the PutMetricData API, our sample monitoring scripts for Windows and Linux, CloudWatch collectd plugin, as well as a number of applications and tools offered by AWS partners.

Q: What resolution can I get from a Custom Metric?

A custom metric can be one of the following:

- Standard resolution, with data having one-minute granularity
- High resolution, with data at a granularity of one second

By default, metrics are stored at 1-minute resolution in CloudWatch. You can define a metric as high-resolution by setting the StorageResolution parameter to 1 in the PutMetricData API request. If you do not set the optional StorageResolution parameter, then CloudWatch will default to storing the metrics at 1-minute resolution.

When you publish a high-resolution metric, CloudWatch stores it with a resolution of 1 second, and you can read and retrieve it with a period of 1 second, 5 seconds, 10 seconds, 30 seconds, or any multiple of 60 seconds.

Custom metrics follow the same retention schedule listed above.

Q: What metrics are available at high resolution?

Currently, only custom metrics that you publish to CloudWatch are available at high resolution. High-resolution custom metrics are stored in CloudWatch at 1-second resolution. High resolution is defined by the StorageResolution parameter in the PutMetricData API request, with a value of 1, and is not a required field. If you do not specify a value for the optional StorageResolution field, then CloudWatch will store the custom metric at 1-minute resolution by default.

Q: Are high-resolution custom metrics priced differently than regular custom metrics?

No, high-resolution custom metrics are priced in the same manner as standard 1-minute custom metrics.

Q: When would I use a Custom Metric over having my program emit a log to CloudWatch Logs?

You can monitor your own data using custom metrics, CloudWatch Logs, or both. You may want to use custom metrics if your data is not already produced in log format, for example operating system processes or performance measurements. Or, you may want to write your own application or script, or one provided by an AWS partner. If you want to store and save individual measurements along with additional detail, you may want to use CloudWatch Logs.

Q: What statistics can I view and graph in CloudWatch?

You can retrieve, graph, and set alarms on the following statistical values for Amazon CloudWatch metrics: Average, Sum, Minimum, Maximum, and Sample Count. Statistics can be computed for any time periods between 60-seconds and 1-day. For high-resolution custom metrics, statistics can be computed for time periods between 1-second and 3-hours.

Q: What is CloudWatch Application Insights for .NET and SQL Server?

Amazon CloudWatch Application Insights for .NET and SQL Server is a capability that you can use to easily monitor your .NET and SQL Server applications. It helps identify and set up key metrics and logs across your application resources and technology stack, i.e. database, web (IIS) and application servers, OS, load balancers, queues, etc. It constantly monitors these telemetry data to detect and correlate anomalies and errors, to notify you of any problems in your application. To aid in troubleshooting, it creates automatic dashboards to visualize problems it detects which includes correlated metric anomalies and log errors, along with additional insights to point you to their potential root-cause.

Q: What are the benefits of using CloudWatch Application Insights for .NET and SQL Server?

- **Automatically recognize application metrics and logs**: It scans your application resources, provides a list of recommended metrics and logs to monitor, and sets them up automatically, making it easier to set up monitoring for your applications.
- **Intelligent problem detection**: It uses built-in rules and machine learning algorithms to dynamically monitor and analyze symptoms of a problem across your application stack and detect application problems. It helps you reduce the overhead of dealing with individual metric spikes, or events, or log exceptions, and instead get notified on real problems, along with contextual information these problems.
- **Faster troubleshooting**: It assesses the detected problems to give you insights on them, such as the possible root-cause of the detected problem and list of metrics and logs impacted because of the problem. You can provide feedback on generated insights to make the problem detection engine specific to your use-case.

Q: How do I get started with monitoring using CloudWatch Application Insights for .NET and SQL Server?

On-board application: Specify the application you want to monitor by choosing the AWS Resource Group associated with it.

Identify application components: It analyzes your application resources to identify application components (standalone resources, or groups of related resources such as auto scaling groups and load balancer groups). You can also customize components by grouping resources for better insights and easy onboarding.

Enable monitoring: For your application components, you can specify the technology tier i.e. IIS front-end, .NET worker tier, etc. Based on your selection it provides a recommended set of metrics and logs that can be customized based on your needs. Once you save these “monitors”, Application Insights for .NET and SQL Server sets up CloudWatch to collect these on your behalf.

Once on-boarded, Application Insights for .NET and SQL Server uses a combination of pre-built rules and machine learning models to start identifying application problems. It creates automated dashboards on CloudWatch with the list of problems detected, and a detailed view for these problems along with related anomalies and errors.

Log monitoring

Q: What log monitoring does Amazon CloudWatch provide?

CloudWatch Logs lets you monitor and troubleshoot your systems and applications using your existing system, application and custom log files.

With CloudWatch Logs, you can monitor your logs, in near real time, for specific phrases, values or patterns. For example, you could set an alarm on the number of errors that occur in your system logs or view graphs of latency of web requests from your application logs. You can then view the original log data to see the source of the problem. Log data can be stored and accessed for up to as long as you need in highly durable, low-cost storage so you don’t have to worry about filling up hard drives.

Q: What are Amazon CloudWatch Vended Logs?

Amazon CloudWatch Vended logs are logs that are natively published by AWS services on behalf of the customer. VPC Flow logs is the first Vended log type that will benefit from this tiered model. However, more AWS Service log types will be added to Vended Log type in the future.

Q: Is CloudWatch Logs available in all regions?

Please refer to Regional Products and Services for details of CloudWatch Logs service availability by region.

Q: How much does CloudWatch Logs cost?

Please see our pricing page for the latest information.

Q: What kinds of things can I do with my logs and Amazon CloudWatch?

CloudWatch Logs is capable of monitoring and storing your logs to help you better understand and operate your systems and applications. When you use CloudWatch Logs with your logs, your existing log data is used for monitoring, so no code change are required. Here are a two examples of what you can do with Amazon CloudWatch and your logs:

Real time Application and System Monitoring: You can use CloudWatch Logs to monitor applications and systems using log data in near real time. For example, CloudWatch Logs can track the number of errors that occur in your application logs and send you a notification whenever the rate of errors exceeds a threshold you specify. Amazon CloudWatch uses your log data for monitoring and consequently it doesn't involve any code changes from you.

Long Term Log Retention: You can use CloudWatch Logs to store your log data for as long as you need in highly durable and cost effective storage without worrying about hard drives running out of space. The CloudWatch Logs Agent makes it easy to quickly move both rotated and non rotated log files off of a host and into the log service. You can then access the raw log event data when you need it.

Q: What types of data can I send to Amazon CloudWatch Logs from my EC2 instances running Microsoft SQL Server and Microsoft Windows Server?

You can configure the EC2Config service to send a variety of data and log files to CloudWatch including: custom text logs, Event (Application, Custom, Security, System) logs, Event Tracing (ETW) logs, and Performance Counter (PCW) data. Learn more about the EC2Config service here.

Q: How frequently does the CloudWatch Logs Agent send data?

The CloudWatch Logs Agent will send log data every five seconds by default and is configurable by the user.

Q: What log formats does CloudWatch Logs support?

CloudWatch Logs can ingest, aggregate and monitor any text based common log data or JSON-formatted logs.

Q: What if I configure the CloudWatch Logs Agent to send non-text log data?

The CloudWatch Logs Agent will record an error in the event it has been configured to report non text log data. This error is recorded in the /var/logs/awslogs.log.

Q: How do I start monitoring my logs with CloudWatch Logs?

You can monitor log events as they are sent to CloudWatch Logs by creating Metric Filters. Metric Filters turn log data into Amazon CloudWatch Metrics for graphing or alarming. Metric Filters can be created in the Console or the CLI. Metric Filters search for and match terms, phrases or values in your log events. When a Metric Filter finds one of the terms, phrases or values in your log events, it counts it in an Amazon CloudWatch Metric that you choose. For example, you can create a Metric Filter to search for and count the occurrence of the word “Error” in your log events. Metric Filters can also extract values from space delimited log events, such as the latency of web requests. You can also use conditional operators and wildcards to create exact matches. The Amazon CloudWatch Console can help you test your patterns before creating Metric Filters.

Q: What is the syntax of Metric Filter patterns?

A Metric Filter pattern can contain search terms or a specification of your common log or JSON event format.

For example, if you want to search for the term Error, the pattern for the metric filter would just be the term Error. Multiple search terms can be included to search for multiple terms. For example, if you wanted to count events which contained the terms Error and Exception you would use the pattern Error Exception. If you wanted to match the term Error Exception exactly, you would put double quotes around the search term, "Error Exception". You can specify as many search terms as you like.

CloudWatch Logs can also be used to extract values from a log event in common log or JSON format. For example, you could track the bytes transferred from your Apache access logs. You can also use conditional operators and wildcards to match and extract the data you are interested in. To use the extraction feature of Metric Filters, log events must be space delimited and use a starting and ending double quote """, or, a starting square brace "[" and a closing square brace "]"square, to enclose fields. Alternatively, they can be JSON-formatted log events. For the full details of the syntax and examples, please see the Developer Guide for Metric Filters.

Q: How do I know that a Metric Filter pattern I specified will match my log events?

CloudWatch Logs lets you test the Metric Filter patterns you want before you create a Metric Filter. You can test your patterns against your own log data that is already in CloudWatch Logs or you can supply your own log events to test. Testing your pattern will show you which log events matched the Metric Filter pattern and, if extracting values, what the extracted value is in the test data. Metric Filter testing is available for use in the console and the CLI.

Q: Can I use regular expressions with my log data?

Amazon CloudWatch Metric Filters does not support regular expressions. To process your log data with regular expressions, consider using Amazon Kinesis and connect the stream with a regular expression processing engine.

Log management

Q: How do I retrieve my log data?

You can retrieve any of your log data using the CloudWatch Logs console or through the CloudWatch Logs CLI. Log events are retrieved based on the Log Group, Log Stream and time with which they are associated. The CloudWatch Logs API for retrieving log events is GetLogEvents.

Q: How do I search my logs?

You can use the CLI to retrieve your log events and search through them using command line grep or similar search functions.

Q: How long does CloudWatch Logs store my log data?

You can store your log data in CloudWatch Logs for as long as you want. By default, CloudWatch Logs will store your log data indefinitely. You can change the retention for each Log Group at any time.

Log analytics

Q: What permissions do I need to access Logs Insights?

To access Logs Insights, your IAM policy must include permissions for logs:DescribeLogGroups and logs:FilterLogEvents.

Q: What logs can I query with CloudWatch Logs Insights?

You can use Logs Insights to query all logs being sent to CloudWatch. Logs Insights automatically discovers the logs fields from logs from AWS services such as Lambda, CloudTrail, Route53, and VPC Flow Logs; and any application log that generates log events in JSON format. Additionally, for all log types, it generates 3 system fields @message, @logStream, and @timestamp for all logs sent to CloudWatch. @message contains the raw unparsed log event, @logStream contains the name of the source that generated the log event, and @timestamp contains the time at which the log event was added to CloudWatch.

Q: Which query language does CloudWatch Logs Insights support?

Logs Insights introduces a new purpose-built query language for log processing. The query language supports a few simple, but powerful query commands. You can write commands to retrieve one or more log fields, find log events that match one or more search criteria, aggregate your log data, and extract ephemeral fields from your text-based logs. The query language is easy to learn, and Logs Insights offers in-product help in the form of sample queries, command descriptions, and query auto-completion to help you get started. You can find additional details about the query language here.

Q: What are the service limits for CloudWatch Logs Insights?

The service limits are documented here.

Q: What regions is CloudWatch Logs Insights available in?

Logs Insights is available in US West (Oregon), US West (N. California), US East (Ohio), US East (N. Virginia), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), EU (Frankfurt), EU (Ireland), EU (London), EU (Paris), South America (São Paulo).

Q: What type of queries does CloudWatch Logs Insights support?

You can write queries containing aggregations, filters, regular expressions, and text searches. You can also extract data from log events to create ephemeral fields, which can be further processed by the query language to help you access the information you are looking for. The query language supports string, numeric, and mathematical functions, such as concat, strlen, trim, log, and sqrt, among others. You can also use boolean and logical expressions, and aggregate functions such as min, max, sum, average, and percentile, among others. You can find additional details about the query language and supported functions here.

Q: What query commands and functions can I use with CloudWatch Logs Insights?

You can find a list of query commands here. You can find a list of supported functions here.

Q: What data visualizations can I use with CloudWatch Logs Insights?

You can use visualizations to identify trends and patterns that occur over time within your logs. Logs Insights supports visualizing data using line charts and stacked area charts. It generates visualizations for all queries containing one or more aggregate functions, where data is grouped over a time interval specified using the bin() function. You can find additional details about visualizing timeseries data here.

Q: Can I use regular expressions with CloudWatch Logs Insights?

You can use Java-style regular expressions with Logs Insights. Regular expressions can be used in the filter command. You can find examples of queries with regular expressions using the in-product help or here.

Q: How do I escape special characters with CloudWatch Logs Insights queries?

You can use backticks to escape special characters. Log field names that contain characters other than alphanumeric characters, @, and . require escaping with backticks.

Q: Why do certain log fields have a "@" sign and others don't?

System fields generated by Logs Insights begin with @. Logs Insights currently generates 3 system fields @message which contains the raw, unparsed log event as sent to CloudWatch, @logStream which contains the name of the source that generated the log event, and @timestamp which contains the time when the log event was added to CloudWatch.

Q: Can I query historical logs with CloudWatch Logs Insights?

Logs Insights enables you to query log data that was added to CloudWatch Logs on or after November 5, 2018.

Q: Can I search for log events from a specific log stream?

You can search for log events from a specific log stream by adding the query command
filter @logStream = "log_stream_name" to your log query.

Q: Today I use an AWS Partner ISV solution to analyze my logs from CloudWatch. What does CloudWatch Logs Insights change for me?

CloudWatch Logs already supports integration options with other AWS Services such as Amazon Kinesis, Amazon Kinesis Data Firehose, Amazon Elasticsearch and AWS Partner ISV solutions such as Splunk, Sumo Logic, and DataDog, among others, to provide you with choice and flexibility across all environments, for your custom log processing, enrichment, analytics, and visualization needs. In addition, the query capabilities of CloudWatch Logs Insights are available for programmatic access through the AWS SDK, to facilitate AWS ISV Partners to build deeper integrations, advanced analytics, and additional value on top of CloudWatch Logs Insights.

Q: How will I benefit from having access to query capabilities of CloudWatch Logs Insights through an AWS ISV Partner solution?

ISV Partner integrations with CloudWatch Logs Insights enable you to bring in your log data into one place and have the ability to analyze using the tools and frameworks of your choice in a high performance, cost-effective way, without having to move large amounts of data. It also provides you with faster access to your logs by removing the associated data transfer latencies and eliminates the operational complexities of configuring and maintaining certain data transfers.

Alarms

Q: What types of CloudWatch Alarms can be created?

You can create an alarm to monitor any Amazon CloudWatch metric in your account. For example, you can create alarms on an Amazon EC2 instance CPU utilization, Amazon ELB request latency, Amazon DynamoDB table throughput, Amazon SQS queue length, or even the charges on your AWS bill.

You can also create an alarm on custom metrics that are specific to your custom applications or infrastructure. If the custom metric is a high-resolution metric, you have the option of creating high-resolution alarms that alert as soon as 10-second or 30-second periods.

Please reference the CloudWatch pricing page to learn more.

Q: What actions can I take from a CloudWatch Alarm?

When you create an alarm, you can configure it to perform one or more automated actions when the metric you chose to monitor exceeds a threshold you define. For example, you can set an alarm that sends you an email, publishes to an SQS queue, stops or terminates an Amazon EC2 instance, or executes an Auto Scaling policy. Since Amazon CloudWatch alarms are integrated with Amazon Simple Notification Service, you can also use any notification type supported by SNS.

Q: What thresholds can I set to trigger a CloudWatch Alarm?

When you create an alarm, you first choose the Amazon CloudWatch metric you want it to monitor. Next, you choose the evaluation period (e.g., five minutes or one hour) and a statistical value to measure (e.g., Average or Maximum). To set a threshold, set a target value and choose whether the alarm will trigger when the value is greater than (>), greater than or equal to (>=), less than (<), or less than or equal to (<=) that value.

Q: My CloudWatch Alarm is constantly in the Alarm state, what did I do wrong?

Alarms continue to evaluate metrics against your chosen threshold, even after they have already triggered. This allows you to view its current up-to-date state at any time. You may notice that one of your alarms stays in the ALARM state for a long time. If your metric value is still in breach of your threshold, the alarm will remain in the ALARM state until it no longer breaches the threshold. This is normal behavior. If you want your alarm to treat this new level as OK, you can adjust the alarm threshold accordingly.

Q: How long can I view my Alarm history?

Alarm history is available for 14 days. To view your alarm history, log in to CloudWatch in the AWS Management Console, choose Alarms from the menu at left, select your alarm, and click the History tab in the lower panel. There you will find a history of any state changes to the alarm as well as any modifications to the alarm configuration.

Dashboards

Q: What is CloudWatch Dashboards?

Amazon CloudWatch Dashboards allow you to create, customize, interact with, and save graphs of AWS resources and custom metrics.

Q: How do I get started with CloudWatch Dashboards?

To get started, visit the Amazon CloudWatch Console and select “Dashboards”. Click the “Create Dashboard” button. You can also copy the desired view from Automatic Dashboards by clicking on Options -> “Add to Dashboard”.

Q: What are the advantages of Automatic Dashboards?

Automatic Dashboards are pre-built with AWS service recommended best practices, remain resource aware, and dynamically update to reflect the latest state of important performance metrics. You can now filter and troubleshoot to a specific view without adding additional code to reflect the latest state of your AWS resources. Once you have identified the root cause of a performance issue, you can quickly act by going directly to the AWS resource.

Q: Do the dashboards support auto refresh?

Yes. Dashboards will auto refresh while you have them open.

Q: Can I share my dashboard?

Yes, a dashboard is available to anyone with the correct permissions for the account with the dashboard.

Events

Q: What is CloudWatch Events?

Amazon CloudWatch Events (CWE) is a stream of system events describing changes in your AWS resources. The events stream augments the existing CloudWatch Metrics and Logs streams to provide a more complete picture of the health and state of your applications. You write declarative rules to associate events of interest with automated actions to be taken.

Q: What services emit CloudWatch Events?

Currently, Amazon EC2, Auto Scaling, and AWS CloudTrail are supported. Via AWS CloudTrail, mutating API calls (i.e., all calls except Describe*, List*, and Get*) across all services are visible in CloudWatch Events.

Q: What can I do once an event is received?

When an event matches a rule you've created in the system, you can automatically invoke an AWS Lambda function, relay the event to an Amazon Kinesis stream, notify an Amazon SNS topic, or invoke a built-in workflow.

Q: Can I generate my own events?

Yes. Your applications can emit custom events by using the PutEvents API, with a payload uniquely suited to your needs.

Q: Can I do things on a fixed schedule?

CloudWatch Events is able to generate events on a schedule you set by using the popular Unix cron syntax. By monitoring for these events, you can implement a scheduled application.

Q: What is the difference between CloudWatch Events and AWS CloudTrail?

CloudWatch Events is a near real time stream of system events that describe changes to your AWS resources. With CloudWatch Events, you can define rules to monitor for specific events and perform actions in an automated manner. AWS CloudTrail is a service that records API calls for your AWS account and delivers log files containing API calls to your Amazon S3 bucket or a CloudWatch Logs log group. With AWS CloudTrail, you can look up API activity history related to creation, deletion and modification of AWS resources and troubleshoot operational or security issues.

Q: What is the difference between CloudWatch Events and AWS Config?

AWS Config is a fully managed service that provides you with an AWS resource inventory, configuration history, and configuration change notifications to enable security and governance. Config rules help you determine whether configuration changes are compliant. CloudWatch Events is for reacting in near real time to resource state changes. It doesn’t render a verdict on whether the changes comply with policy or give detailed history like Config/Config Rules do. It is a general purpose event stream.

AWS Auto Scaling FAQs

General

Q: What is AWS Auto Scaling? »

Q. What are the benefits of AWS Auto Scaling? »

Q. When should I use AWS Auto Scaling? »

Q. How can I get started with AWS Auto Scaling? »

Scaling Options

Q. What are the different ways that I can scale AWS resources? »

Q. When should I use AWS Auto Scaling vs. Amazon EC2 Auto Scaling? »

Q. When should I use AWS Auto Scaling vs. Auto Scaling for individual services? »

Q. What is Predictive Scaling? »

Q. Which services can I use Predictive Scaling with? »

Q. How can I use Predictive Scaling with target tracking? »

Q. What is a scaling plan? »

Q. Can I configure a scaling plan without Predictive Scaling? »

Q. How much historical data does Predictive Scaling need to generate the scaling plan? »

Q. How much into the future does Predictive Scaling forecast the traffic? »

Q. Can I configure Predictive Scaling to provision instances before an actual spike in traffic? »

Q. How much does Predictive Scaling cost? »

Q. How is AWS Auto Scaling different than the scaling capabilities for individual services? »

Features

Q. What can I scale with AWS Auto Scaling? »

Q. How does AWS Auto Scaling make scaling recommendations? »

Q. How do I select an application stack within AWS Auto Scaling? »

Q. How does AWS Auto Scaling discover what resources can scale? »

Availability and Pricing

Q. Which regions is AWS Auto Scaling available in? »

Q. How much does AWS Auto Scaling cost? »

AWS Chatbot FAQs

General

Q: What is AWS Chatbot?
AWS Chatbot makes it easy to securely integrate multiple AWS services with your Slack channels and Amazon Chime chat rooms for ChatOps. With AWS Chatbot, you can receive notifications about operational events, security findings, or budget alerts right in your chat room where your entire team can see and discuss them. You can run commands from Slack to retrieve diagnostic information, invoke AWS Lambda functions, or create AWS Support cases.

Q: What kind of notifications can I get with AWS Chatbot?
You can use AWS Chatbot to get notifications with CloudWatch alarms, Health events, Security Hub findings, Budgets alerts, and CloudFormation stack events. For the full list of supported services, refer to the AWS Chatbot documentation.

Q: What kind of commands can I run with AWS Chatbot?
AWS Chatbot supports read-only commands for most AWS services. You can also initiate workflows by invoking Lambda functions and create AWS Support cases. AWS Chatbot commands use the already-familiar AWS Command Line Interface syntax.

Q: How do I get started with AWS Chatbot?
To get started with AWS Chatbot, go to the AWS Chatbot console, perform a configuration with Slack or Chime, and add AWS Chatbot to your channels or chat rooms.

Q: When should I use AWS Chatbot?
AWS Chatbot helps your entire team to stay updated on and respond to operational events, security findings, or budget alerts for applications running in your AWS accounts. If your team uses a team chat application supported by AWS Chatbot, you can configure AWS Chatbot to publish notifications and run

commands in a team channel or chat room where your entire team can see and quickly act on them. For example, you can set up CloudWatch alarms to go into a “Cloud DevOps” chat room where DevOps engineers can see alarms, discuss them, and retrieve diagnostic information immediately after events occur.

Q: What’s the difference between AWS Chatbot and Amazon Lex?
Amazon Lex provides the advanced deep learning capabilities of automatic speech recognition (ASR) for converting speech to text and natural language understanding (NLU) to recognize intent and build lifelike interactions. This lets you quickly and easily build your own sophisticated, natural language, conversational bots or “chatbots.” AWS Chatbot is a pre-built interactive agent designed to monitor and interact with your AWS resources (ChatOps). With AWS Chatbot you can securely receive alerts and request diagnostic information from services such as Amazon CloudWatch and AWS Security Hub in your Slack channel or Amazon Chime chatroom.

Q: How much does AWS Chatbot cost?
AWS Chatbot is available at no additional charge. You only pay for the AWS resources that are used with AWS Chatbot (e.g., SNS topics, CloudWatch alarms, etc.)

Q: In which regions is AWS Chatbot available?
AWS Chatbot is a global service and can be used in all commercial AWS regions. You can combine SNS topics from multiple regions in a single AWS Chatbot configuration. Please refer to the Regional Product and Services table for details about AWS resource availability.

Chat client integrations

Q: What chat clients does AWS Chatbot support?
AWS Chatbot supports Slack and Amazon Chime. Running commands is currently only supported in Slack.

Q: How does AWS Chatbot integrate with Amazon Chime?
AWS Chatbot integrates with Amazon Chime via webhooks.

Q: How does AWS Chatbot integrate with Slack?
AWS Chatbot integrates with Slack via an AWS Chatbot Slack app that you can install to your Slack workspace from the AWS Chatbot console. The installation is performed via a click-through OAuth 2.0 flow in a browser and takes just a few seconds.

Q: What is an AWS Chatbot configuration?
An AWS Chatbot configuration is a mapping of a Slack channel or an Amazon Chime chat room with SNS topics and an IAM role.

Notifications from AWS services

Q: How does AWS Chatbot integrate with AWS services?
AWS Chatbot integrates with supported AWS services via SNS topics. You need to configure the service to publish notifications to an SNS topic and then create an AWS Chatbot configuration that maps the SNS topic to a Slack channel or an Amazon Chime chat room.

Q: How does AWS Chatbot work with Amazon CloudWatch Events?
To use CloudWatch Events for notifications from supported services with AWS Chatbot, use an SNS topic as a target for a CloudWatch Event Rule and then use that SNS topic in an AWS Chatbot configuration. For the full list of supported services, refer to the AWS Chatbot documentation.

Q: How can I see more details about the notification I received in a channel or chat room?
You can click the title of the notification to navigate to the AWS Management Console page for the notification source. For example, if you click on the title on an AWS Budgets notification, you will be taken to the budget details page for that specific budget where you can review and analyze your budget performance.

Q: Can I use AWS Chatbot to receive arbitrary notifications?
No, AWS Chatbot only supports notifications from the services listed in the documentation. Events from unsupported sources will not be delivered to chat rooms.

Q: Can I use SNS topics from multiple AWS accounts within a single AWS Chatbot configuration?
No, only SNS topics from the AWS account that hosts the AWS Chatbot configuration can be used, however, you can create Chatbot configurations in other AWS accounts and map those configurations to a single chat room. Because each AWS Chatbot configuration is linked to a separate AWS account, the configurations will be independent of each other.

Q: Can I use SNS topics from multiple regions within an AWS Chatbot configuration?
Yes, you can use SNS topics from multiple public AWS regions in the same AWS Chatbot configuration.

Q: How can I filter notifications coming via AWS Chatbot?
You can filter notifications using an SNS filter policy or CloudWatch Event Rules for events that support filtering. For other events, filtering is not available.

Q: Can I add custom formatting to AWS Chatbot notifications?
No, the AWS Chatbot notifications formatting is not customizable.

Q: Are there rate limits for AWS Chatbot?
Yes, AWS Chatbot is subject to rate limits from Slack and Amazon Chime. Refer to Slack Web API documentation and Amazon Chime webhook documentation for specific details.

Q: What should I do if the AWS service I want notifications from is not supported by AWS Chatbot?
Until AWS Chatbot supports that service, you will not be able to use it with AWS Chatbot. Please submit a request using the Feedback button in the footer of the AWS Chatbot console for consideration.

Q: How can I unsubscribe from AWS Chatbot notifications in a channel or chat room?
To unsubscribe a channel or chat room from AWS Chatbot notifications, you can remove the respective configuration. If you want to unsubscribe only some notifications from the channel or chat room, you can remove specific SNS topics from the AWS Chatbot configuration.

Running commands and actions

Q: How do I run a command using AWS Chatbot?
To run a command in a Slack channel, first create a Slack channel configuration using the AWS Chatbot console. To start interacting with AWS Chatbot in Slack, type @aws followed by a command using the standard AWS Command Line Interface syntax. For example, to get a list and a chart of CloudWatch Alarms, type @aws cloudwatch describe-alarms. Please refer to AWS Chatbot documentation for the limitations compared to the AWS CLI.

Q: What services are supported by AWS Chatbot?
AWS Chatbot supports commands for most AWS services and its permissions scope is defined by the IAM role used in your AWS Chatbot configurations. Regardless of the IAM role permissions, access to certain services and commands, such as AWS IAM and AWS KMS, is disabled to prevent exposing credentials in Slack channels. Please refer to AWS Chatbot documentation for details on permissions.

Q: Can I interact with AWS Chatbot using direct messages in Slack?
Direct messages are not currently supported. You can create a private channel with just yourself and AWS Chatbot and use it as a channel for direct message communication.

Q: What is a notification action?
Notification actions are shortcuts that enable you to take a quick action by clicking a button on notifications sent by AWS Chatbot. For example, CloudWatch Alarm notifications for Lambda functions and API Gateway stages have “Show logs” and “Show error logs” buttons that will display the logs for the affected resource in the Slack channel.

Q: In which chat applications can I use commands and actions?
Currently, you can use commands and actions in Slack.

Security

Q: What is the purpose of the AWS Chatbot IAM role?
AWS Chatbot configurations use IAM roles that the AWS Chatbot service assumes when making API calls and running commands on behalf of AWS Chatbot users.

Q: What policies are included in the AWS Chatbot policy templates?
Refer to the AWS Chatbot documentation for the details.

AWS CloudFormation FAQs

General

Q: What is AWS CloudFormation?

AWS CloudFormation is a service that gives developers and businesses an easy way to create a collection of related AWS and third party resources and provision them in an orderly and predictable fashion.

Q: What can developers now do with AWS CloudFormation that they could not before?

AWS CloudFormation automates and simplifies the task of repeatedly and predictably creating groups of related resources that power your applications. Creating and interconnecting all resources your application needs to run is now as simple as creating a single EC2 or RDS instance.

Q: How is AWS CloudFormation different from AWS Elastic Beanstalk?

These services are designed to complement each other. AWS Elastic Beanstalk provides an environment to easily deploy and run applications in the cloud. It is integrated with developer tools and provides a one-stop experience for you to manage the lifecycle of your applications. AWS CloudFormation is a convenient provisioning mechanism for a broad range of AWS and third party resources. It supports the infrastructure needs of many different types of applications such as existing enterprise applications, legacy applications, applications built using a variety of AWS resources and container-based solutions (including those built using AWS Elastic Beanstalk).

AWS CloudFormation supports Elastic Beanstalk application environments as one of the AWS resource types. This allows you, for example, to create and manage an AWS Elastic Beanstalk–hosted application along with an RDS database to store the application data. In addition to RDS instances, any other supported AWS resource can be added to the group as well.

Q: What new concepts does AWS CloudFormation introduce?

AWS CloudFormation introduces two concepts: The template, a JSON or YAML-format, text-based file that describes all the AWS resources you need to deploy to run your application and the stack, the set of AWS resources that are created and managed as a single unit when AWS CloudFormation instantiates a template.

Q: What resources does AWS CloudFormation support?

To see a complete list of supported AWS resources and their features, visit the Supported AWS Services page in the Release History of the documentation.

The AWS CloudFormation Registry and AWS CloudFormation custom resources enable management of additional AWS and third party resources.

Q: Can I manage individual AWS resources that are part of an AWS CloudFormation stack?

Yes. AWS CloudFormation does not get in the way; you retain full control of all elements of your infrastructure. You can continue using all your existing AWS and third-party tools to manage your AWS resources.

Q: What are the elements of an AWS CloudFormation template?

AWS CloudFormation templates are JSON or YAML-formatted text files that are comprised of five types of elements:

1. An optional list of template parameters (input values supplied at stack creation time)

2. An optional list of output values (e.g. the complete URL to a web application)

3. An optional list of data tables used to lookup static configuration values (e.g., AMI names)

4. The list of AWS resources and their configuration values

5. A template file format version number

With parameters, you can customize aspects of your template at run time, when the stack is built. For example, the Amazon RDS database size, Amazon EC2 instance types, database and web server port numbers can be passed to AWS CloudFormation when a stack is created. Each parameter can have a default value and description and may be marked as “NoEcho” in order to hide the actual value you enter on the screen and in the AWS CloudFormation event logs. When you create an AWS CloudFormation stack, the AWS Management Console will automatically synthesize and present a pop-up dialog form for you to edit parameter values.

Output values are a very convenient way to present a stack’s key resources (such as the address of an Elastic Load Balancing load balancer or Amazon RDS database) to the user via the AWS Management Console, or the command line tools. You can use simple functions to concatenate string literals and value of attributes associated with the actual AWS resources.

Q: How does AWS CloudFormation choose actual resource names?

You can assign logical names to AWS resources in a template. When a stack is created, AWS CloudFormation binds the logical name to the name of the corresponding actual AWS resource. Actual resource names are a combination of the stack and logical resource name. This allows multiple stacks to be created from a template without fear of name collisions between AWS resources.

Q: Why can’t I name all my resources?

Although AWS CloudFormation allows you to name some resources (such as Amazon S3 buckets), CloudFormation doesn’t allow this for all resources. Naming resources restricts the reusability of templates and results in naming conflicts when an update causes a resource to be replaced. To minimize these issues, CloudFormation will support resource naming on a case by case basis.

Q: Can I install software at stack creation time using AWS CloudFormation?

Yes. AWS CloudFormation provides a set of application bootstrapping scripts that enable you to install packages, files, and services on your EC2 instances by simply describing them in your CloudFormation template. For more details and a how-to see Bootstrapping Applications via AWS CloudFormation.

Q: Can I use AWS CloudFormation with Chef?

Yes. AWS CloudFormation can be used to bootstrap both the Chef Server and Chef Client software on your EC2 instances. For more details and a how-to see Integrating AWS CloudFormation with Chef.

Q: Can I use AWS CloudFormation with Puppet?

Yes. AWS CloudFormation can be used to bootstrap both the Puppet Master and Puppet Client software on your EC2 instances. For more details and a how-to see Integrating AWS CloudFormation with Puppet.

Q: Does AWS CloudFormation support Amazon EC2 tagging?

Yes. Amazon EC2 resources that support the tagging feature can also be tagged in an AWS template. The tag values can refer to template parameters, other resource names, resource attribute values (e.g. addresses), or values computed by simple functions (e.g., a concatenated a list of strings).AWS CloudFormation automatically tags Amazon EBS volumes and Amazon EC2 instances with the name of the AWS CloudFormation stack they are part of.

Q: Do I have access to the Amazon EC2 instance, or Auto Scaling Launch Configuration user-data fields?

Yes. You can use simple functions to concatenate string literals and attribute values of the AWS resources and pass them to user-data fields in your template. Please refer to our sample templates to learn more about these easy to use functions.

Q: What happens when one of the resources in a stack cannot be created successfully?

By default, the “automatic rollback on error” feature is enabled. This will cause all AWS resources that AWS CloudFormation created successfully for a stack up to the point where an error occurred to be deleted. This is useful when, for example, you accidentally exceed your default limit of Elastic IP addresses, or you don’t have access to an EC2 AMI you’re trying to run. This feature enables you to rely on the fact that stacks are either fully created, or not at all, which simplifies system administration and layered solutions built on top of AWS CloudFormation.

Q: Can stack creation wait for my application to start up?

Yes. AWS CloudFormation provides a *WaitCondition* resource that acts as a barrier, blocking the creation of other resources until a completion signal is received from an external source such as your application, or management system.

Q: Can I save my data when a stack is deleted?

Yes. AWS CloudFormation allows you to define deletion policies for resources in the template. You can specify that snapshots be created for Amazon EBS volumes or Amazon RDS database instances before they are deleted. You can also specify that a resource should be preserved and not deleted when the stack is deleted. This is useful for preserving Amazon S3 buckets when the stack is deleted.

Q: Can I update my stack after it has been created?

Yes. You can use AWS CloudFormation to modify and update the resources in your existing stacks in a controlled and predictable way. By using templates to manage your stack changes, you have the ability to apply version control to your AWS infrastructure just as you do with the software running on it.

Q: Can I create stacks in a Virtual Private Cloud (VPC)?

Yes. CloudFormation supports creating VPCs, Subnets, Gateways, Route Tables and Network ACLs as well as creating resources such as Elastic IPs, Amazon EC2 Instances, EC2 Security Groups, Auto Scaling Groups, Elastic Load Balancers, Amazon RDS Database Instances and Amazon RDS Security Groups in a VPC.

Getting Started

Q: How do I sign up for AWS CloudFormation?

To sign up for AWS CloudFormation, click Create Free Account on the AWS CloudFormation product page. After signing up, please refer to the AWS CloudFormation documentation, which includes our Getting Started Guide.

Q: Why am I asked to verify my phone number when signing up for AWS CloudFormation?

AWS CloudFormation registration requires you to have a valid phone number and email address on file with AWS in case we ever need to contact you. Verifying your phone number takes only a few minutes and involves receiving an automated phone call during the registration process and entering a PIN number using the phone key pad.

Q: How do I get started after I have signed up?

The best way to get started with AWS CloudFormation is to work through the Getting Started Guide, which is included in our technical documentation. Within a few minutes, you will be able to deploy and use one of our sample templates that illustrate how to create the infrastructure needed to run applications such as Tracks, WordPress, and others.

Q: Are there sample templates that I can use to check out AWS CloudFormation?

Yes, AWS CloudFormation includes sample templates that you can use to test drive the offering and explore its functionality. Our sample templates illustrate how to interconnect and use multiple AWS resources in concert, following best practices for multiple Availability Zone redundancy, scale out, and alarming. To get started, all you need to do is go to the AWS Management Console, click **Create Stack**, and follow the steps to select and launch one of our samples. Once created, select your stack in the console and review the **Template** and **Parameter** tabs to look at the details of the template file used to create the respective stack.

AWS CloudFormation Registry

Q: What is the AWS CloudFormation Registry?

The AWS CloudFormation Registry is a managed service that lets you register, use, and discover AWS and third party resource providers. Third party resource providers must be registered before they can be used to provision resources with AWS CloudFormation templates. Please refer to our documentation for details.

Q: What are resource providers in AWS CloudFormation?

A resource provider is a set of resource types with specifications and handlers that control the lifecycle of underlying resources via create, read, update, delete and list operations. You can use resource providers to model and provision resources using CloudFormation. For example, AWS::EC2::Instance is a resource type from the Amazon EC2 provider. You can use this type to model and provision an Amazon EC2 instance using CloudFormation. Using the AWS CloudFormation Registry, you can build and use resource providers to model and provision third party resources such as SaaS monitoring, team productivity, or source code management resources.

Q: What is the difference between AWS and third party resource providers?

The difference between AWS and third party resource providers is their origin. AWS resource providers are built and maintained by Amazon and AWS to manage AWS resources and services. For example, three AWS resource providers help you manage Amazon DynamoDB, AWS Lambda, and Amazon EC2 resources. These providers contain resource types such as AWS::DynamoDB::Table, AWS::Lambda::Function, and AWS::EC2::Instance. For a complete reference, go to our documentation.

Third party resource providers are built by another company, organization, or the developer community. They can help you manage both AWS and non-AWS resources such as AWS application resources and non-AWS SaaS software services such as monitoring, team productivity, incident management, or version control management tools.

Q: What is a resource schema?

In a resource provider, a resource type is expressed using a CloudFormation Resource Schema to define its properties and attributes. This schema is also used to validate the definition of a resource type.

Q: How do I develop resource providers or resource types?

Use the AWS CloudFormation CLI to build resource providers. You start by defining a simple declarative schema for your resources, which includes permissions required and relationships to other resources. You then use the CloudFormation CLI to generate the scaffolding for resource lifecycle handlers (Create, Read, Update, Delete and List) along with test stubs for unit and integration testing.

Q: How do I register a resource provider?

You can either use the use the open source AWS CloudFormation CLI or directly call the RegisterType and related Registry APIs that are available via the AWS SDKs and AWS CLI. For more details, visit our documentation. AWS resource providers are available out of the box and do not require any additional registration steps before use.

Billing

Q: How much does AWS CloudFormation cost?

There is no additional charge for using AWS CloudFormation with resource providers in the following namespaces: AWS::*, Alexa::*, and Custom::*. In this case you pay for AWS resources (such as Amazon EC2 instances, Elastic Load Balancing load balancers, etc.) created using AWS CloudFormation as if you created them manually. You only pay for what you use, as you use it; there are no minimum fees and no required upfront commitments.

When you use resource providers with AWS CloudFormation outside the namespaces mentioned above, you incur charges per handler operation.

Handler operations are create, update, delete, read, or list actions on a resource. For more information, please refer to our pricing page.

Q: Will I be charged for resources that were rolled back during a failed stack creation attempt?

Yes. Charges for AWS resources created during template instantiation apply irrespective of whether the stack as a whole could be created successfully or not.

Limits and Restrictions

Q: Are there limits to the number of templates or stacks?

There are no limits to the number of templates. Each AWS CloudFormation account is limited to a maximum of 200 stacks. Complete our request for a higher limit here, and we will respond to your request within two business days.

Q: Are there limits to the size of description fields?

Template, Parameter, Output, and Resource description fields are limited to 4096 characters.

Q: Are there limits to the number of parameters or outputs in a template?

You can include up to 60 parameters and 60 outputs in a template.

Regions and Endpoints

Q: What are the AWS CloudFormation service access points in each region?

Endpoints for each region are available in the technical documentation.

Q: What are the AWS regions where AWS CloudFormation is currently available?

Please refer to Regional Products and Services for details of CloudFormation availability by region.

AWS CloudTrail FAQs

General

Q: What is AWS CloudTrail? >>

Q: What are the benefits of CloudTrail? >>

Q: Who should use CloudTrail? >>

Getting started

Q: If I am a new AWS customer or existing AWS customer and don’t have CloudTrail setup, do I need to enable or setup anything to view my account activity? >>

Q: Does the CloudTrail Event History show all account activity within my account? >>

Q: What search filters can I use to view my account activity? >>

Q: Can I use the lookup-events CLI command even if I don’t have a trail configured? >>

Q: What additional CloudTrail features are available by setting up CloudTrail and creating a trail? >>

Q: Can I restrict access for users in my account from seeing the CloudTrail Event History? >>

Q: Is there any cost associated with CloudTrail Event History being enabled on my account upon creation? >>

Q: Can I turn CloudTrail Event History off for my account? >>

Services and region support

Q: What services are supported by CloudTrail? >>

Q: Are API calls made from the AWS Management Console recorded? >>

Q: Where are my log files stored and processed before they are delivered to my Amazon S3 bucket? >>

Applying a trail to all regions

Q: What is applying a trail to all regions? >>

Q: What are the benefits of applying a trail to all regions? >>

Q: How do I apply a trail to all regions? >>

Q: What happens when I apply a trail to all regions? >>

Q: Can I apply an existing trail to all regions? >>

Q: How long will it take for CloudTrail to replicate the trail configuration to all regions? >>

Multiple trails

Q: How many trails can I create in an AWS region? >>

Q: What is the benefit of creating multiple trails in an AWS region? >>

Q: Does CloudTrail support resource level permissions? >>

Security and expiration

Q: How can I secure my CloudTrail log files? >>

Q: Where can I download a sample S3 bucket policy and an SNS topic policy? >>

Q: How long can I store my activity log files? >>

Event payload, timeliness, and delivery frequency

Q: What information is available in an event? >>

Q: How long does it take CloudTrail to deliver an event for an API call? >>

Q: How often will CloudTrail deliver log files to my Amazon S3 bucket? >>

Q: Can I be notified when new log files are delivered to my Amazon S3 bucket? >>

Q: What happens if CloudTrail is turned on for my account but my Amazon S3 bucket is not configured with the correct policy? >>

Data events

Q: What are Data events? >>

Q: How can I consume Data events? >>

Q: What are Amazon S3 Data events? How do I record them? >>

Q: What are AWS Lambda Data Events? How do I record them? >>

CloudTrail Insights

Q: What are CloudTrail Insights events? >>

Q: What type of activity does AWS CloudTrail Insights help identify? >>

Q: How does CloudTrail Insights work with other AWS services that use anomaly detection? >>

Q: Do I need to have AWS CloudTrail set up in order for CloudTrail Insights to work? >>

Q: What kinds of events does CloudTrail Insights monitor? >>

Q: How do I get started? >>

Log file aggregation

Q: I have multiple AWS accounts. I would like log files for all the accounts to be delivered to a single S3 bucket. Can I do that? >>

Integration with CloudWatch Logs

Q: What is CloudTrail integration with CloudWatch Logs? >>

Q: What are the benefits of CloudTrail integration with CloudWatch Logs? >>

Q: How do I turn on CloudTrail integration with CloudWatch Logs? >>

Q: What happens when I turn on CloudTrail integration with CloudWatch Logs? >>

Q: In which AWS regions is CloudTrail integration with CloudWatch Logs supported? >>

Q: How does CloudTrail deliver events containing account activity to my CloudWatch Logs? >>

Q: What charges do I incur once I turn on CloudTrail integration with CloudWatch Logs? >>

CloudTrail log file encryption using AWS Key Management Service (KMS)

Q: What is the benefit of CloudTrail log file encryption using Server-side Encryption with KMS? >>

Q: I have an application that ingests and processes CloudTrail log files. Do I need to make any changes to my application? >>

Q: How do I configure CloudTrail log file encryption? >>

Q: What charges do I incur once I configure encryption using SSE-KMS? >>

CloudTrail log file integrity validation

Q: What is CloudTrail log file integrity validation? >>

Q: What is the benefit of CloudTrail log file integrity validation? >>

Q: How do I enable CloudTrail log file integrity validation? >>

Q: What happens once I turn on the log file integrity validation feature? >>

Q: Where are the digest files delivered to? >>

Q: How can I validate the integrity of a log file or digest file delivered by CloudTrail? >>

Q: I aggregate all my log files across all regions and multiple accounts into one single Amazon S3 bucket. Will the digest files be delivered to the same Amazon S3 bucket? >>

AWS CloudTrail processing library

Q: What is AWS CloudTrail Processing Library? >>

Q: What functionality does CloudTrail Processing Library provide? >>

Q: What software do I need to start using the CloudTrail Processing Library? >>

Pricing

Q: How do I get charged for AWS CloudTrail? >>

Q: If I have only one trail with management Events, and apply it to all regions, will I incur charges? >>

Q: If I enable data events on an existing trail with free management events, will I get charged? >>

Partners

Q: How do the AWS partner solutions help me analyze the events recorded by CloudTrail? >>

Other

Q: Will turning on CloudTrail impact the performance of my AWS resources, or increase API call latency? >>

AWS Compute Optimizer FAQs

General

Q: What is AWS Compute Optimizer?

AWS Compute Optimizer helps you identify the optimal AWS compute resources, such as Amazon EC2 instance type using machine learning on historical utilization metrics. AWS Compute Optimizer provides a set of APIs and a console experience to help you reduce costs and increase workload performance by recommending the optimal AWS compute resources for your AWS workloads.

Q: What can I do with AWS Compute Optimizer?

AWS Compute Optimizer delivers intuitive and easily actionable AWS compute resource recommendations to help you quickly identify optimal AWS compute resources for your workloads without requiring specialized expertise or investing substantial time and money. The AWS Compute Optimizer console provides you with a global, cross-account view of all compute resources analyzed by AWS Compute Optimizer and recommendations so that you can quickly identify the most impactful optimization opportunities.

Q: How can I get started with AWS Compute Optimizer?

To sign up for AWS Compute Optimizer, go to the AWS Compute Optimizer console and click “opt in”. You must have an Amazon Web Services account to access this service. Once you opt in, AWS Compute Optimizer immediately starts analyzing your AWS resources and starts delivering recommendations. When you first opt in AWS Compute Optimizer, it may take up to 12 hours to fully analyze the AWS compute resources in your account.

Q: What data does AWS Compute Optimizer use for my recommendations?

When you opt in AWS Compute Optimizer, you authorize the service to use AWS resource configuration data and CloudWatch metrics. This data is required because AWS Compute Optimizer needs to identify the resources to assess, and it needs sufficient metrics history before it makes recommendations.

Q: When should I use AWS Compute Optimizer and when should I use AWS Cost Explorer?

You should use AWS Cost Explorer if you want to identify under-utilized EC2 instances that may be downsized on an instance by instance basis within the same instance family, and you want to understand the potential impact on your AWS bill by taking into account your RIs and Savings Plans. Cost Explorer offers recommendations for all commercial regions (outside of China) and supports the A, T, M, C, R, X, Z, I, D, H instance families.

You should use AWS Compute Optimizer if you want to look at instance type recommendations beyond downsizing within an instance family. You can use AWS Compute Optimizer to get downsizing recommendations within or across instance families, upsizing recommendations to remove performance bottlenecks, and recommendations for EC2 instances that are parts of an Auto Scaling group. AWS Compute Optimizer provides you additional capabilities to enhance recommendation quality and the user experience, such as using machine learning to identify workload types and automatically choose workload-specific recommendation methodology for them. You should also use AWS Compute Optimizer if you want to understand the performance risks and how your workload would perform on various EC2 instance options to evaluate the price-performance trade-off for your workloads. AWS Compute Optimizer is available in US East (N. Virginia and Ohio), US West (Oregon and N. California), South America (Sao Paulo), Asia Pacific (Mumbai, Seoul, Singapore, Sydney, and Tokyo), Canada (Central), and Europe (Ireland, Frankfurt, London, and Stockholm), and supports the M, C, R, T and X instance families.

Recommendations

Q: How many recommended options does AWS Compute Optimizer deliver for each AWS resource?

AWS Compute Optimizer delivers up to 3 recommended options for each AWS compute resource analyzed.

Q: Does AWS Compute Optimizer deliver recommendations for all AWS resources?

AWS Compute Optimizer delivers recommendations for selected types of EC2 instances and EC2 auto scaling groups. For detail, see EC2 and EC2 Auto Scaling FAQs.

Q: How much data does AWS Compute Optimizer analyze to generate recommendations?

AWS Compute Optimizer analyzes metrics from the past 14 days to generate recommendations.

EC2 instance recommendations

Q: What types of EC2 instance recommendations does AWS Compute Optimizer support?

AWS Compute Optimizer supports EC2 instance type and size recommendations for standalone EC2 instances of M, C, R, T, and X instance families.

Q: What data does AWS Compute Optimizer use for my EC2 instance recommendations?

AWS Compute Optimizer needs at least 30 hours of metrics before it makes recommendations for standalone EC2 instances. AWS Compute Optimizer analyzes default CloudWatch metrics for EC2 instances, such as CPU utilization and network I/O metrics.

Q: Does AWS Compute Optimizer analyze my EC2 instance memory metrics?

If you use CloudWatch agent to publish memory utilization, AWS Compute Optimizer automatically analyzes memory metrics published by the CloudWatch Agent in the “CWAgent” namespace.

Q: What happens if I don’t have memory metrics available for my EC2 instances?

If metrics for a hardware resource, such as memory, are not available, AWS Compute Optimizer will attempt to avoid making a recommendation that downsizes that dimension.

Q: How does AWS Compute Optimizer determine performance risk for recommended EC2 instance options?

Performance risk indicates the likelihood of the recommended option does not meet the performance requirements of your workload. The higher the performance risk is, the more effort you may need to spend to validate whether the recommended EC2 instance type meets the performance requirements of your workload.

Q: How does AWS Compute Optimizer help me to understand recommended EC2 instance options?

AWS Compute Optimizer projects the would-be CPU and memory utilization of your EC2 instance had you used the recommended option, so that you can understand how your workload would have performed on the recommended options.

Q: Does AWS Compute Optimizer consider EC2 instance pricing information when delivering recommendations?

After AWS Compute Optimizer identifies a list of optimal AWS resources for your workload, it incorporates a variety of pricing dimensions, such as on-demand pricing, along with expected performance risk to rank the recommendations. AWS Compute Optimizer does not consider transient pricing factors, such as spot pricing.

Auto scaling group recommendations

Q: What types of auto scaling group recommendations does AWS Compute Optimizer support?

AWS Compute Optimizer provides EC2 instance type and size recommendations for EC2 Auto Scaling groups with a fixed group size, meaning desired, minimum, and maximum are all set to the same value and have no scaling policy attached. Additionally, all Auto Scaling group member instances must be of type M, C, R, T, or X instance families. At this time Compute Optimizer does not support Auto Scaling groups configured with mixed instances policy.

Q: What data does AWS Compute Optimizer use for my auto scaling group recommendations?

AWS Compute Optimizer needs at least 30 hours of metrics before it makes recommendations for auto scaling groups. AWS Compute Optimizer analyzes default CloudWatch metrics of each member EC2 instances, such as CPU utilization and network I/O metrics, as well as auto scaling group configuration, such as scaling policy and associated launch template.

AWS service integration

Q: Does AWS Compute Optimizer integrate with AWS Organizations?

Yes, AWS Compute Optimizer integrates with AWS Organizations to allow you see all your recommendations within your organization. In order to use this feature, your organization must have “all features” enabled, and you must login as the master account of your organization.

AWS Config FAQs

General

Q: What is AWS Config? >>

Q: What is a Config Rule? >>

Q: What is a Conformance Pack? >>

Q: What are the benefits of AWS Config? >>

Q: How can AWS Config help with audits? >>

Q: Who should use AWS Config and Config Rules? >>

Q: Who should use AWS Config Conformance Packs? >>

Q: Does the service guarantee that my configurations are never out of compliance? >>

Q: Does the service prevent users from taking non-compliant actions? >>

Q: Can rules be evaluated prior to provisioning a resource? >>

Q: How does AWS Config work with AWS CloudTrail? >>

Q: Can I monitor compliance information of multiple accounts and regions via a central account? >>

Getting started

Q: How do I get started with this service? >>

Q: How do I access my resources' configuration? >>

Q: Do I turn on AWS Config regionally or globally? >>

Q: Can AWS Config aggregate data across different AWS accounts? >>

Q: Is API activity on AWS Config itself logged by AWS CloudTrail? >>

Q: What time and timezones are displayed in the timeline view of a resource? What about daylight savings? >>

Config Rules

Q: What is a resource's configuration? >>

Q: What is a rule? >>

Q: How are rules created? >>

Q: How many rules can I create? >>

Q: How are rules evaluated? >>

Q: What is an evaluation? >>

Q: What does compliance mean? >>

Q: What information does the Config Rules dashboard provide? >>

Conformance Packs

Q: When should I use AWS Config Rules versus Conformance Packs? >>

Q: When should I use AWS Config Conformance Packs vs. AWS Security Hub compliance standards? >>

Q: How do I get started with Conformance Packs? >>

Q: Is there any cost associated with using this feature in AWS Config? >>

Multi-account, multi-region data aggregation

Q: What is multi-account, multi-region data aggregation? >>

Q: Can I use the data aggregation capability to centrally provision Config rules across multiple accounts? >>

Q: How do I enable data aggregation in my account? >>

Q: What is an aggregator? >>

Q: What information does the Aggregated view provide? >>

Q: I am not an AWS Organizations customer. Can I still use the data aggregation capability? >>

Q: I only have a single account, can I still take advantage of the data aggregation capability? >>

Q: In what regions is the multi-account, multi-region data aggregation capability available? >>

Q: What if I have an account that includes a region not supported by this feature? >>

Services and region support

Q: What AWS resources types are covered by AWS Config? >>

Q: What regions is AWS Config available in? >>

Resource configuration

Q: What is a configuration item? >>

Q: What is a custom configuration item? >>

Q: What are AWS Config relationships and how are they used? >>

Q: Does AWS Config record every state a resource has been in? >>

Q: Does AWS Config record configuration changes that did not result from API activity on that resource? >>

Q: Does AWS Config record configuration changes to software within EC2 instances? >>

Q: Does AWS Config continue to send notifications if a resource that was previously non-compliant is still non-compliant after a periodic rule evaluation? >>

Q: Can I flag or exempt resources from being evaluated by Config rules? >>

Pricing

Q: How will I be charged for AWS Config? >>

Q: Does the pricing for AWS Config rules include the costs for AWS Lambda functions? >>

Q: I want to change the Lambda function for my custom AWS Config rule. What is the recommended approach? >>

Partner solutions

Q: What AWS Partner solutions are available for AWS Config? >>

AWS Control Tower FAQs

General

What is AWS Control Tower?
AWS Control Tower offers the easiest way to set up and govern a new, secure, multi-account AWS environment. It establishes a landing zone that is based on best-practices blueprints, and enables governance using guardrails you can choose from a pre-packaged list. The landing zone is a well-architected, multi-account baseline that follows AWS best practices. Guardrails implement governance rules for security, compliance, and operations.

Who should use AWS Control Tower?
AWS Control Tower is for customers who want to create a new, multi-account AWS environment with best practices. It offers prescriptive guidance to govern your AWS environment at scale. It gives you control over your environment without sacrificing the speed and agility AWS provides for builders. You will benefit from Control Tower if you are building a new AWS environment, starting out on your journey on AWS, starting a new cloud initiative, or are completely new to AWS.

What are the benefits of AWS Control Tower?
With AWS Control Tower, distributed teams are able to provision new AWS accounts quickly, while cloud IT has the peace of mind knowing that all accounts are aligned with centrally established, company-wide policies. AWS Control Tower provides a single location to easily set up your new well-architected multi-account environment and govern your AWS workloads with rules for security, operations, and internal compliance. You can automate the setup of your AWS environment with best-practices blueprints for multi-account structure, identity, access management, and account provisioning workflow. For ongoing governance, you can select and apply pre-packaged policies enterprise-wide or to specific groups of accounts.

What features does AWS Control Tower provide?
AWS Control Tower automates the creation of a landing zone with best-practices blueprints that configure AWS Organizations for a multi-account structure, provide identity management using AWS SSO Directory, provide federated access using AWS Single Sign-On (AWS SSO), create a central log archive using AWS CloudTrail and AWS Config, enable security audits across accounts using AWS SSO, implement network configurations using Amazon VPC, and define workflows for provisioning accounts using AWS Service Catalog.

Control Tower offers “guardrails” for ongoing governance of your AWS environment. Guardrails provide governance controls by preventing deployment of resources that don’t conform to selected policies or detecting non-conformance of provisioned resources. AWS Control Tower automatically implements guardrails using multiple building blocks such as AWS CloudFormation to establish a baseline, AWS Organizations service control policies (SCPs) to prevent configuration changes, and AWS Config rules to continuously detect non-conformance.

AWS Control Tower offers a dashboard for continuous oversight of your multi-account environment. You get visibility into provisioned accounts across your enterprise. Control Tower dashboards provide reports on detective and preventive guardrails you have enabled on your accounts. And they give you status on any resources that don’t comply with policies you have enabled through guardrails.

Can I use Control Tower to meet industry compliance standards (such as HIPAA, PCI, SOC-1, SOC-2)?
Out-of-the-box guardrails offered by AWS Control Tower are not intended to meet regulatory compliance standards (such as HIPAA, PCI, SOC-1, SOC-2). Control Tower guardrails represent a set of AWS best-practices policies for governing your AWS environment through rules such as disallowing configuration changes to log archive, and requiring account activity to be logged using AWS CloudTrail. Over time, Control Tower will continue to offer additional functionality such as custom guardrails to enable AWS customers to implement policies that support their regulatory compliance, based on the AWS shared security model.

Availability

In which AWS Regions is AWS Control Tower available?

To see a current list of regions where AWS Control Tower is available, please visit the AWS Regional Table.

How much does AWS Control Tower cost?

There is no additional charge to use AWS Control Tower. You only pay for AWS services enabled by AWS Control Tower, e.g., AWS Service Catalog and AWS CloudTrail. You also pay for AWS Config rules that are set up by AWS Control Tower to implement guardrails.

Does AWS Control Tower create a new AWS Organizations account structure?

Yes, AWS Control Tower creates a new organization that starts with your existing AWS account as the master account. You cannot deploy AWS Control Tower on accounts with an existing AWS Organizations master account. As a result, AWS Control Tower is intended to automate a brand new landing zone with a separate master payer account.

Can I deploy AWS Control Tower on my existing AWS Organizations accounts?

You cannot yet deploy AWS Control Tower on an existing account that is a member of AWS Organizations. AWS Control Tower requires a standalone account that is not a member of AWS Organizations for setup. In the near future, you will be able to deploy Control Tower to an existing AWS Organizations account structure.

Can I use my existing directory with AWS Control Tower?

AWS Control Tower sets up AWS SSO with a native default directory. After the landing zone setup, you can configure AWS SSO with a supported directory such as AWS Managed Microsoft AD.

Is there an API available for AWS Control Tower?

No. You can use AWS Control Tower through the management console to perform all necessary operations.

AWS Solution and Service Comparisons

How is AWS Control Tower different than the AWS Landing Zone solution?
Control Tower is an AWS native service providing a pre-defined set of blueprints and guardrails to help customers implement a landing zone for new AWS accounts. AWS Landing Zone is an AWS solution offered through AWS Solution Architect, Professional Services, or AWS Partner Network (APN) Partners providing a fully configurable, customer-managed landing zone implementation. Customers can use either the Landing Zone solution or AWS Control Tower to create a foundational AWS environment based on best practice blueprints implemented through AWS Service Catalog. Control Tower is designed to provide an easy, self-service setup experience and an interactive user interface for ongoing governance with guardrails. While Control Tower automates creation of a new landing zone with pre-configured blueprints (e.g., AWS SSO for directory and access), the AWS Landing Zone solution provides a configurable setup of a landing zone with rich customization options through custom add-ons (e.g., Active Directory, Okta Directory) and ongoing modifications through a code deployment and configuration pipeline.

When should I use AWS Landing Zone and when should I use AWS Control Tower?
You should use Control Tower if you are looking for a self-service experience to set up a new AWS environment based on a landing zone with pre-configured blueprints and then interactively govern your accounts with pre-configured guardrails. You will benefit from Control Tower if you are building a new offering, have teams starting out on their journey to AWS, are starting a new cloud initiative, or are completely new to AWS. You should use the AWS Landing Zone solution if you are looking to set up a configurable landing zone with rich customization options through custom add-ons (e.g., Active Directory, Okta Directory) and change management through a code deployment and configuration pipeline.

Can AWS Control Tower help me operate my infrastructure?
Control Tower helps you deploy a multi-account AWS environment based on best practices, however, the customer is still responsible for day-to-day operations. Enterprises that need help operating regulated infrastructure in the cloud should consider a certified MSP partner or AWS Managed Services (AMS).

AMS is best-suited for enterprises that want to move regulated workloads to the cloud quickly and do not yet have the required AWS skillsets needed for compliant operations, or want to keep AWS talent focused on application migration and modernization instead of the undifferentiated heavy lifting of infrastructure operations.

Is there a migration path from AWS Landing Zone to AWS Control Tower?
Yes, in the near future, you will be able to migrate your existing accounts created with the AWS Landing Zone solution to AWS Control Tower. The migration path will occur in several phases to ensure compatibility between Control Tower and your AWS Landing Zone solution starting with ability to deploy Control Tower to an existing Organizations, followed by enabling custom guardrails and custom blueprints for Control Tower.

How does AWS Control Tower interoperate with AWS Organizations?
AWS Control Tower offers an abstracted, automated, and prescriptive experience on top of AWS Organizations. It automatically sets up AWS Organizations as the underlying AWS service to organize accounts and implement preventive guardrails using Service Control Policies (SCPs). Using AWS Organizations, you can further create and attach granular SCPs that centrally control the use of AWS services and resources across multiple AWS accounts.

How is AWS Control Tower different from AWS Security Hub?
AWS Security Hub is the primary destination for security and compliance professionals. It provides a comprehensive and timely view of the overall security and compliance posture of their AWS environment and take necessary actions. AWS Control Tower is the primary destination for cloud administrators. While AWS Security Hub is primarily detective in nature - i.e., assesses and reports on the security and compliance posture of an existing environment, AWS Control Tower is primarily preventive - i.e., helps set up a new AWS landing zone and enforces controls to prevent provisioning of resources that do not conform to applied policies.

How does AWS Control Tower interoperate with AWS Service Catalog?
AWS Control Tower automatically sets up AWS Service Catalog as the underlying AWS service to enable provisioning of new accounts through an account factory. While AWS Control Tower provides central governance at an

account level, AWS Service Catalog can further provide granular governance at a resource level. AWS Service Catalog also lets you provision infrastructure and application stacks that have been pre-approved by IT for use inside your accounts.

How does AWS Control Tower interoperate with AWS Systems Manager?
You can use AWS Control Tower to set up and govern your AWS environment, and then use AWS Systems Manager to handle the ongoing day to day operations of that environment. Systems Manager provides a unified user interface so you can view operational data from multiple AWS services and automate operational tasks across your AWS resources. With Systems Manager, you can group resources, like Amazon EC2 instances, Amazon S3 buckets, or Amazon RDS instances, by application, view operational data for monitoring and troubleshooting, and take action on your groups of resources.

Get an overview of AWS Control Tower

See overview »

Check out AWS Control Tower Pricing

Learn more »

Read the blog post

Read more »

FAQs

- How do I sign in?
- Can I sign-in with biometric authentication?
- Where can I download the app?
- What services are supported?
- Is MFA supported?
- Can I create resources?
- Can I download S3 objects?
- Can I view my current AWS usage charges?
- What versions of iOS and Android are supported?
- Does the Console Mobile application support tablets?
- I lost my Mobile Device. What should I do?
- Can I provide feedback?

Q: How do I sign in?

The The Console Mobile Application supports several authentication methods, including owner/root credentials, IAM user credentials, and AWS access keys. An owner account is the AWS login that created the account. An IAM user is an identity that has been created by an administrator through the IAM service. Note that IAM users need to also provide their account alias, which can be found at the top of the web console sign-in screen. AWS access keys are used to sign programmatic requests that the app makes to AWS.

For security reasons, we recommend that you secure your device with a passcode and that you follow an AWS best practice by creating and using an IAM user's credentials to log in to the app. If you lose your device, an IAM user can be deactivated to prevent unauthorized access. Root accounts cannot be deactivated.

Click here to learn more about the different types of AWS security credentials.

Back to top

Q: Can I sign-in with biometric authentication?

You can setup Touch ID or Face ID on supported iOS devices running Console Mobile Application v2.0+. Biometric authentication is currently not supported on Android.

Back to top

Q: Where can I download the app?

Download the Console Mobile Application from Amazon Appstore, Google Play, and iTunes.

Back to top

Q: What services are supported?

The Console Mobile Application supports AWS Billing and Cost Management, AWS CloudFormation, Amazon CloudWatch, Amazon DynamoDB, AWS Elastic Beanstalk, Amazon Elastic Compute Cloud (Amazon EC2), Elastic Load Balancing, AWS OpsWorks, AWS Personal Health Dashboard, Amazon Relational Database Service (Amazon RDS), Amazon Route 53, Amazon Simple Storage Service (Amazon S3), Amazon Virtual Private Cloud (Amazon VPC). The Console Mobile Application does not support the AWS GovCloud (US) region. For a full description, see the AWS Console Mobile App page. We plan to add new features to The Console Mobile Application. Tell us what you need using the feedback link in the app.

Back to top

Q: Is MFA supported?

Yes. We recommend using either a hardware MFA device or a virtual MFA on a separate mobile device for the greatest level of account protection.

Back to top

Q: Can I create resources

You cannot create resources in the current version. Please use the feedback link in the Console Mobile Application's menu to tell us what you need.

Back to top

Q: Can I download S3 objects?

You can use The Console Mobile Application to generate a pre-signed URL for an S3 object. A pre-signed URL grants time-limited permission to download the object. Read more about pre-signed URLs here.

In order to open a pre-signed URL for an S3 object in your device's browser, use the app to navigate to the S3 object's detail page and tap "View in browser". Your device configuration will determine what actions are possible with the object.

Back to top

Q: Can I view my current AWS usage charges?

Yes, you can view your current usage charges in the app. Simply visit your Billing Preferences page and select the checkbox to Receive Billing Alerts. In order to view usage charges, your identity must have permission to view CloudWatch.

Back to top

Q: What versions of iOS and Android are supported?

iOS 7.0+ and Android 4.0+ are supported.

Back to top

Q: Does the Console Mobile application support tablets?

The Console Mobile application is optimized for iOS and Android mobile devices with a screen size < 7”, but it works on larger screen sizes as well.

Back to top

Q: I lost my Mobile Device. What should I do?

We strongly recommend that in addition to using a password or biometric lock on your mobile device, you use an IAM user to manage AWS resources. If you lose your mobile device, you can remove the IAM user's access.

Back to top

Q: Can I provide feedback?

Yes! Click the Feedback button in the Console Mobile Application's menu. We’re eager to hear about your experience.

Back to top

AWS License Manager FAQs

How does AWS License Manager work?

AWS License Manager provides you with the flexibility and control to manage license usage to match your organizational structure and processes. AWS License Manager can be set in different configurations to address specific business needs. In general, there are three distinct phases:

Define licensing rules: Administrators work with the relevant stakeholders (for example, business or compliance teams) in your organization to carefully review licensing agreements, and create licensing rules in AWS License Manager. Licensing rules contain settings that are configured to emulate the terms of your enterprise agreement.

Enforce licensing rules: After the rules are created, they can be applied in several different ways to track license usage and compliance. Administrators can attach the rules to the organization's specific Amazon Machine Images (AMIs), create AWS CloudFormation templates, use Amazon EC2 launch templates, or simply attach them to applications in AWS Service Catalog. After the rules are created and attached to the relevant instances to be deployed, end users in your organization can seamlessly launch AWS resources such as EC2 instances with the certainty they are licensed correctly. Administrators can track the usage through AWS License Manager's built-in dashboard. AWS License Manager flags any resources that are not compliant with the predefined rules.

Discover usage of software installed on AWS and on-premises: AWS License Manager integrates seamlessly with AWS Systems Manager, allowing discovery of any software installed on your AWS resources. With AWS Systems Manager, you can manage instances running on AWS and in your on-premises data center through a single interface. AWS Systems Manager securely communicates with a lightweight agent installed on your servers to execute management tasks. This helps you manage resources for Windows and Linux operating systems running on Amazon EC2 or on-premises. After the instances are attached to AWS License Manager, administrators can search for any operating system or application software through a single pane that covers AWS resources and on-premises servers. You can associate your licensing rules to the discovered software and track all the applications through the built-in dashboard.

How does AWS License Manager help me stay compliant?

AWS License Manager reduces the risk of non-compliance by increasing transparency and enforcing and tracking licensing rules that administrators define. AWS License Manager provides built-in dashboards that can be used when considering new license purchases, reporting to procurement, and in vendor audits. However, customers are responsible for compliance and assume the responsibility of carefully understanding and adding rules into AWS License Manager based on their licensing agreements. While AWS cannot participate in audits, AWS License Manager’s rich reports provide valuable insights that allow for more accuracy and transparency.

What types of software licenses can I track using AWS License Manager?

With AWS License Manager, you can track software that is licensed based on virtual cores(vCPUs), physical cores or sockets. This includes a variety of software products from vendors including Microsoft, Oracle, IBM, and SAP. Common use cases include tracking Oracle databases, Microsoft Windows Server, and SQL Server licenses that can be licensed by physical and virtual cores.

Which AWS services are supported through AWS License Manager?

AWS License Manager integrates with various AWS launch mechanisms such as AWS CloudFormation, EC2 Launch Templates, and Service Catalog. You can use License Manager to track licenses across your EC2 instances using default tenancy, Dedicated Instances, Dedicated Hosts, Spot Instances and Spot Fleet, and Auto Scaling groups.

Does AWS License Manager support tracking license usage outside of AWS?

Yes, you can use AWS License Manager and the AWS Systems Manager agent to track licenses outside of AWS, including on-premises servers.

How does AWS License Manager work with AWS Organizations?

AWS License Manager integrates with AWS Organizations seamlessly. Administrators can sign in to their organizational master account and link all their organizational accounts, thereby being able to manage and control license usage centrally across their organization. They will also be able to discover software across all their organizational accounts, centrally using the master account.

How can I leverage the benefits of License Manager with BYOL products purchased from AWS Marketplace?

Using AWS License Manager allows you to associate licensing rules to AWS Marketplace BYOL AMI products and benefit from centralized license management tracking and compliance. AWS License Manager doesn't change the way you obtain or activate your BYOL AMI's in the Marketplace. For example, if you launch an EC2 instance, you provide the license key obtained directly from the seller to activate the software.

Are there additional charges attached with using AWS License Manager?

There is no charge for using AWS License Manager. You only pay for the resources created in your account. These include EC2 instances but can also include an S3 bucket for storing software based on AWS Systems Manager, Amazon Athena queries, and AWS Glue jobs for enabling the centralized discovery of the Systems Manager data, and Amazon SNS notifications.

FAQs

Web Console | Resource Groups | Tag Editor

Web Console

- What is the AWS Management Console?
- How do I sign into the Management Console?
- How can I find a service in the AWS Management Console?
- Can I create shortcuts for the services I use?
- What can I expect to find under "Build a Solution"?
- What can I expect to find under "Learn to Build"?
- Can I provide feedback?
- When does my session expire?
- What browsers does the Management Console support?
- Does the Management Console share the same functions as the AWS CLI and APIs?

Q: What is the AWS Management Console?

The AWS Management Console provides a simple web interface for Amazon Web Services. You can log in using your AWS account name and password. If you've enabled AWS Multi-Factor Authentication, you will be prompted for your device's authentication code.

Q: How do I sign into the Management Console?

You can sign into the management console using your AWS or IAM account credentials at https://console.aws.amazon.com/console/home. For the AWS GovCloud (US) region, you can sign into the management console using your IAM account credentials at https://console.amazonaws-us-gov.com.

Q: How can I find a service in the AWS Management Console?

There are several ways for you to locate and navigate to the services you need. On Console Home, you can utilize the search functionality, select services from the *Recently visited services* section, or expand the *All services* section to browse through the list of all the services offered by AWS.

At any time, you can also select the Services menu in the top level navigation bar, which also includes the search functionality and the list of all services, either grouped, or arranged alphabetically.

Q: Can I create shortcuts for the services I use?

Yes. You can add service shortcuts to the top level toolbar in the Console. Select the pin icon beside the Resource Groups menu and drag and drop the service links you want to save as shortcuts. You have the option to display the service icon alone, the service name alone, or both together.

Q: What can I expect to find under "Build a solution"?

The *Build a Solution* section, on the AWS Console Home page, features various simplified and automated workflows and wizards that utilize and introduce you to different AWS services in order to create the resources required to build your intended solution.

Q: What can I expect to find under "Learn to Build"?

The *Learn to Build* section, on the AWS Console Home page, presents training and learning resources for various solutions and use cases that might interest you. Selecting one of the categories presented in this section, you can expect to find materials such as introduction videos, webinar recordings, simple tutorials, project guides, self-paced labs, documentation, and other resources.

Q: Can I provide feedback?

Yes! Click the **Feedback** button at the bottom of the console. We're eager to hear about your experience with the new console.

Q: When does my session expire?

For security purpose, a login session will expire in 12 hours when you sign into the AWS Management Console with your AWS or IAM account credentials. To resume your work after the session expires, we ask you to click the "**Click login to continue**" button and login again. The duration of federated sessions varies depending on the federation API (GetFederationToken or AssumeRole) and the administrator's preference. Please go to our Security Blog to learn more about building a secure delegation solution to grant temporary access to your AWS account.

Q: What browsers does the Management Console support?

Important: As of May 1, 2016, the AWS Management Console no longer supports versions of Internet Explorer older than version 11. We recommend migrating to a more recent browser version to ensure the best possible experience and security.

Please contact us if you have any questions.

Browser	Version	Service
Google Chrome	Latest 3 Versions	All services
Mozilla Firefox	Latest 3 Versions	All services
Microsoft Internet Explorer	11	All services
Microsoft Edge	Latest 3 Versions	All services
Apple Safari	Latest 3 Versions	All services

Q: Does the Management Console share the same functions as the AWS CLI and APIs?

All IaaS AWS administration, management, and access functions in the AWS Console are available in the AWS API and CLI. New AWS IaaS features and services provide full AWS Console functionality through the API and CLI at launch or within 180 days of launch.

AWS Managed Services FAQs

What use cases does AWS Managed Services solve?

How can AWS Managed Services help enterprises accelerate cloud adoption?

Does AWS Managed Services manage applications?

How do APN Partners and AWS Managed Services work together?

Will AWS Managed Services work with existing ITSMs?

What industry standards does AWS Managed Services comply with?

What kind of workloads does AWS Managed Services support?

How do I deploy applications on AWS Managed Services?

Does AWS Managed Services manage on-premises resources or other public clouds?

Will I still have access to my resources?

Can I access my managed environments from my corporate network?

In which AWS regions is AWS Managed Services available?

What languages are supported by AWS Managed Services?

How much does AWS Managed Services cost?

Launch a Fully Managed Message Broker in Minutes

Extend managed service benefits to your message broker. Migrate to Amazon MQ to reduce your operational load.

Get started for free »

AWS OpsWorks for Chef Automate FAQs

General

Q: What is AWS OpsWorks for Chef Automate? >>

Q: How is OpsWorks for Chef Automate different from OpsWorks Stacks? >>

Q: Who should use OpsWorks for Chef Automate? >>

Q: How can I access OpsWorks for Chef Automate? >>

Q: In which regions is OpsWorks for Chef Automate available? >>

Q: Are there any limits to OpsWorks for Chef Automate? >>

Q: What network requirements must my servers meet to work with OpsWorks for Chef Automate? >>

Q: What is Chef and how does OpsWorks for Chef Automate use it? >>

Q: What is Chef Automate? >>

Q: How do I use the Chef Automate console? >>

Q: I am an AWS OpsWorks Stacks customer. Should I migrate to OpsWorks for Chef Automate? >>

Q: How can I migrate from OpsWorks Stacks to OpsWorks for Chef Automate? >>

Q: Which versions of Chef are supported? >>

Q: Which cloud resources power my AWS OpsWorks for Chef Automate server? >>

Q: How can I back up my Chef server? >>

Q: How many backups can I keep for every Chef server? >>

Q: How can I restore my Chef server to an earlier point in time? >>

Q: Which resources can I connect to my Chef server? >>

Q: How do I register nodes with the Chef server? >>

Q: How can I obtain Chef related training? >>

Maintenance Window

Q: How can I keep the underlying Chef server running and up-to-date? >>

Q: What is an OpsWorks for Chef Automate maintenance window? >>

Q: How do I set up a maintenance window? >>

Q: What kinds of version updates will be performed by OpsWorks for Chef Automate? >>

Q: When and how can I perform major version updates? >>

Q: How does AWS OpsWorks for Chef Automate apply updates? >>

Q: Will my Chef server be available during the maintenance window? >>

Q: How will I be notified of the availability of new OpsWorks for Chef Automate versions? >>

Q: Where can I find details about changes between platform versions? >>

Q: How often are platform version updates released? >>

Getting Started

Q: How do I get started with OpsWorks for Chef Automate? >>

Configuration and Management

Q: How do I create Chef cookbooks and recipes? >>

Q: Can I use community cookbooks from the Chef Supermarket? >>

Q: How do I upgrade my Chef nodes to a newer release version? >>

Q: Does my OpsWorks for Chef Automate server support community tools like Knife and Test Kitchen? >>

Q: Is there a sample cookbook that I can use to check out OpsWorks for Chef Automate? >>

Security

Q: Is it possible to use AWS Identity and Access Management (IAM) with OpsWorks for Chef Automate? >>

Q: How do I create IAM users? >>

Q: Do I have root access to my OpsWorks for Chef Automate server EC2 instance? >>

Q: Where can I find more information about security and running applications on AWS? >>

Q: Can I get a history of OpsWorks for Chef Automate API calls made on my account for security analysis and troubleshooting purposes? >>

Billing

Q: How much do the AWS resources powering my application on OpsWorks for Chef Automate server cost? >>

Q: Am I billed for EC2 instances and on-premises servers that are connected to my OpsWorks for Chef Automate server? >>

Q: How do I view the cost of AWS resources that have been used by my OpsWorks for Chef Automate server? >>

Support

Q: Does AWS Support cover OpsWorks for Chef Automate? >>

Q: What other support options are available? >>

AWS Organizations FAQs

General

What is AWS Organizations?

Which central governance and management capabilities does AWS Organizations enable?

Which regions are AWS Organizations available in?

How do I get started?

What’s the difference between AWS Organizations and AWS Control Tower?

Core concepts

What is an organization?

What is an AWS account?

What is a master account?

What is a member account?

What is an administrative root?

What is an organizational unit (OU)?

What is a policy?

Organizing AWS accounts

Can I define and manage my organization regionally?

Can I change which AWS account is the master account?

How do I add an AWS account to my organization?

Can an AWS account be a member of more than one organization?

How can I access an AWS account that was created in my organization?

Can I set up multi-factor authentication (MFA) on the AWS account that I create in my organization programmatically?

Can I move an AWS account that I have created using AWS Organizations to another organization?

Can I remove an AWS account that I created using Organizations and make it a standalone account?

How many AWS accounts can I manage in my organization?

How can I remove an AWS member account from an organization?

How can I create an organizational unit (OU)?

How can I add a member AWS account to an OU?

Can an AWS account be a member of multiple OUs?

Can an OU be a member of multiple OUs?

How many levels can I have in my OU hierarchy?

Control management

At what levels of my organization can I apply a policy?

How can I attach a policy?

Are policies inherited through hierarchical connections in my organization?

What types of policies does AWS Organizations support?

What is a Service Control Policy (SCP)?

What does an SCP look like?

If I attach an empty SCP to an AWS account, does that mean that I allow all AWS service actions in that AWS account?

What are the effective permissions if I apply an SCP to my organization and my principals also have IAM policies?

Can I simulate the effect of an SCP on an AWS account?

Can I create and manage an organization without enforcing an SCP?

Billing

What does AWS Organizations cost?

Who pays for usage incurred by users under an AWS member account in my organization?

Will my bill reflect the organizational unit structure that I created in my organization?

Integrated AWS services

Why should I enable an AWS service integrated with AWS Organizations?

Which AWS services are currently integrated with AWS Organizations?

How do I enable an AWS service integration?

AWS Service Catalog FAQs

General

Q: What is AWS Service Catalog?

AWS Service Catalog allows IT administrators to create, manage, and distribute catalogs of approved products to end users, who can then access the products they need in a personalized portal. Administrators can control which users have access to each product to enforce compliance with organizational business policies. Administrators can also setup adopted roles so that End users only require IAM access to AWS Service Catalog in order to deploy approved resources. AWS Service Catalog allows your organization to benefit from increased agility and reduced costs because end users can find and launch only the products they need from a catalog that you control.

Q: Who should use AWS Service Catalog?

AWS Service Catalog was developed for organizations, IT teams, and managed service providers (MSPs) that need to centralize policies. It allows IT administrators to vend and manage AWS resource and services. For large organizations, it provides a standard method of provisioning cloud resources for thousands of users. It is also suitable for small teams, where front-line development managers can provide and maintain a standard dev/test environment.

Q: How do I get started with AWS Service Catalog?

In the AWS Management Console, choose AWS Service Catalog in Management Tools. In the AWS Service Catalog console, administrators can create portfolios, add products, and grant users permissions to use them with just a few clicks. End users logged into the AWS Service Catalog console can see and launch the products that administers have created for them.

Q: What can end users to do with AWS Service Catalog that they could not do before?

End users have a simple portal in which to discover and launch products that comply with organizational policies and budget constraints.

Q: What is a portfolio?

A portfolio is a collection of products, with configuration information that determines who can use those products and how they can use them. Administrators can create a customized portfolio for each type of user in an organization and selectively grant access to the appropriate portfolio. When an administrator adds a new version of a product to a portfolio, that version is automatically available to all current portfolio users. The same product can be included in multiple portfolios. Administrators also can share portfolios with other AWS accounts and allow the administrators of those accounts to extend the portfolios by applying additional constraints. By using portfolios, permissions, sharing, and constraints, administrators can ensure that users are launching products that are configured properly for the organization's needs.

Q: What is a product?

A product is a service or application for end users. A catalog is a collection of products that the administrator creates, adds to portfolios, and provides updates for using AWS Service Catalog. A product can comprise one or more AWS resources, such as Amazon Elastic Compute Cloud (Amazon EC2) instances, storage volumes, databases, monitoring configurations, and networking components. It can be a single compute instance running AWS Linux, a fully configured multitier web application running in its own environment, or anything in between.

Administrators distribute products to end users in portfolios. Administrators create catalogs of products by importing AWS CloudFormation templates. These templates define the AWS resources that the product needs to work, the relationships between components, and the parameters that the end user chooses when launching the product to configure security groups, create key pairs, and perform other customizations.

An end user with access to a portfolio can use the AWS Management Console to find a standard dev/test environment product, for example, in the form of an AWS CloudFormation template, then manage the resulting resources using the AWS CloudFormation console. For information about creating a product, see “How do I create a product?” in the Administrator FAQ.

Q: Is AWS Service Catalog a regionalized service?

Yes. AWS Service Catalog is fully regionalized, so you can control the regions in which data is stored. Portfolios and products are a regional construct which will need to be created per region and are only visible/usable on the regions in which they were created.

Q: In which Regions is AWS Service Catalog available?

For a full list of supported AWS Regions, see the AWS Region Table.

Q: Are APIs available? Can I use the CLI to access AWS Service Catalog?

Yes, APIs are available and enabled through the CLI. Actions from the management of Service Catalog artifacts through to provisioning and terminating are available. You can find more information in the AWS Service Catalog documentation or download the latest AWS SDK or CLI.

Q: Can I privately access AWS Service Catalog APIs from my Amazon Virtual Private Cloud (VPC) without using public IPs?

Yes, you can privately access AWS Service Catalog APIs from your Amazon Virtual Private Cloud (VPC) by creating VPC Endpoints. With VPC Endpoints, the routing between the VPC and AWS Service Catalog is handled by the AWS network without the need for an Internet gateway, NAT gateway, or VPN connection. The latest generation of VPC Endpoints used by AWS Service Catalog are powered by AWS PrivateLink, an AWS technology enabling the private connectivity between AWS services using Elastic Network Interfaces (ENI) with private IPs in your VPCs. To learn more about AWS PrivateLink, visit the AWS PrivateLink documentation.

Q: Does AWS Service Catalog offer a Service Level Agreement (SLA)?

Yes. The AWS Service Catalog SLA provides for a service credit if a customer's monthly uptime percentage is below our service commitment in any billing cycle.

IT administrator

Q: How do I create a portfolio?

You create portfolios in the AWS Service Catalog console. For each portfolio, you specify the name, a description, and owner.

Q: How do I create a product?

To create a product, you first create an AWS CloudFormation template by using an existing AWS CloudFormation template or creating a custom template. Next, you use the AWS Service Catalog console to upload the template and create the product. When creating products, you can provide additional information for the product listing, including a detailed product description, version information, support information, and tags.

Q: Why would I use tags with a portfolio?

Tags are useful for identifying and categorizing AWS resources that are provisioned by end users. You can also use tags in AWS Identity and Access Management (IAM) policies to allow or deny access to IAM users, groups, and roles or to restrict operations that can be performed by IAM users, groups, and roles. When you add tags to your portfolio, the tags are applied to all instances of resources provisioned from products in the portfolio.

Q: How do I make a portfolio available to my users?

You publish portfolios that you've created or that have been shared with you to make them available to IAM users in the AWS account. To publish a portfolio, you add IAM users, groups, or roles to the portfolio from the AWS Service Catalog console by navigating to the portfolio details page. When you add users to a portfolio, they can browse and launch any of the products in the portfolio. Typically, you create multiple portfolios with different products and access permissions customized for specific types of end users. For example, a portfolio for a development team will likely contain different products from a portfolio targeted at the sales and marketing team. A single product can be published to multiple portfolios with different access permissions and provisioning policies.

Q: Can I share my portfolio with other AWS accounts?

Yes. You can share your portfolios with users in one or more other AWS accounts. When you share your portfolio with other AWS accounts, you retain ownership and control of the portfolio. Only you can make changes, such as adding new products or updating products. You, and only you, can also “unshare” your portfolio at any time. Any products, or stacks, currently in use will continue to run until the stack owner decides to terminate them.

To share your portfolio, you specify the account ID you want to share with, and then send the Amazon Resource Number (ARN) of the portfolio to that account. The owner of that account can create a link to this shared portfolio, and then assign IAM users from that account to the portfolio. To help end users with discovery, you can curate a directory of portfolios.

Q: Can I customize the experience for end users when they use a product?

Yes. You can tailor a product’s user experience for specific end users. The AWS CloudFormation template contains input parameters that drive the user experience. You can define business-level input parameters (such as “How many users do you need to support?” or “Are you going to store PII data?”) or infrastructure-level input parameters (such as “Which Amazon EC2 instance type?”) depending on the user. When the AWS CloudFormation template is deployed, the user is asked these questions and can select from a constrained list of answers for each question. Depending on the answers, the template may be deployed using different Amazon Elastic Compute Cloud (EC2) instances and different AWS resources.

Q: Can I create a product from an existing Amazon EC2 AMI?

Yes. You can use an existing Amazon EC2 AMI to create a product by wrapping it in an AWS CloudFormation template.

Q: Can I use products from the AWS Marketplace?

Yes. You can subscribe to a product in the AWS Marketplace and use the copy to Service Catalog action to copy your Marketplace product directly to Service Catalog. Also you can use the Amazon EC2 AMI for the product to create an AWS Service Catalog product. To do that, you wrap the subscribed product in an AWS CloudFormation template. For more details on how to copy or package your AWS Marketplace products, please click here.

Q: How do I control access to portfolios and products?

To control access to portfolios and products, you assign IAM users, groups, or roles on the Portfolio details page. Providing access allows users to see the products that are available to them in the AWS Service Catalog console.

Q: Can I provide a new version of a product?

Yes. You can create new product versions in the same way you create new products. When a new version of a product is published to a portfolio, end users can choose to launch the new version. They can also choose to update their running stacks to this new version. AWS Service Catalog does not automatically update products that are in use when an update becomes available.

Q: Can I provide a product and retain full control over the associated AWS resources?

Yes. You have full control over the AWS accounts and roles used to provision products. To provision AWS resources, you can use either the user's IAM access permissions or your pre-defined IAM role. To retain full control over the AWS resources, you specify a specific IAM role at the product level. AWS Service Catalog uses the role to provision the resources in the stack.

Q: Can I restrict the AWS resources that users can provision?

Yes. You can define rules that limit the parameter values that a user enters when launching a product. These rules are called template constraints because they constrain how the AWS CloudFormation template for the product is deployed. You use a simple editor to create template constraints, and you apply them to individual products.

AWS Service Catalog applies constraints when provisioning a new product or updating a product that is already in use. It always applies the most restrictive constraint among all constraints applied to the portfolio and the product. For example, consider a scenario where the product allows all EC2 instances to be launched and the portfolio has two constraints: one that allows all non-GPU type EC2 instances to be launched and one that allows only t1.micro and m1.small EC2 instances to be launched. For this example, AWS Service Catalog applies the second, more restrictive constraint (t1.micro and m1.small).

Q: Can I use a YAML language CloudFormation template in Service Catalog?

Yes, we currently support both JSON and YAML language templates.

Q: Can I connect my ServiceNow instance to AWS Service Catalog?

Yes. The AWS Service Catalog Connector for ServiceNow allows ServiceNow administrators to view AWS Service Catalog portfolios and products, align them to organizational structures such as teams, grant access to users, and connect ServiceNow workflows to provisioning requests. ServiceNow administrators can configure the connector to work with existing or new AWS accounts and roles. ServiceNow users can browse and request provisioning of AWS Service Catalog products, which can include AWS Marketplace software products that have been copied to AWS Service Catalog. This simplifies AWS product provisioning for ServiceNow users and provides ServiceNow administrators governance and oversight over AWS products.

The AWS-supplied connector is available in the ServiceNow Store for the Helsinki (H), Istanbul (I), Jakarta (J), and Kingston (K) versions of ServiceNow.

Q: Can I use Terraform with AWS Service Catalog?

You can leverage the AWS Service Catalog Terraform Reference Architecture. This reference architecture provides an example for using AWS Service Catalog products, an AWS CloudFormation custom resource, and Terraform to provision resources on AWS.

Q: Can I connect Jira Service Desk to AWS Service Catalog?

Yes. The AWS Service Catalog Connector for Jira Service Desk (JSD) allows JSD administrators to view AWS Service Catalog portfolios and products, align them to organizational structures such as teams, grant access to users, and connect JSD workflows to provisioning requests. JSD administrators can configure the connector to work with existing or new AWS accounts and roles. JSD users can browse and request provisioning of AWS Service Catalog products, which can include AWS Marketplace software products that have been copied to AWS Service Catalog. This simplifies AWS product provisioning for JSD users and provides JSD administrators governance and oversight over AWS products. The AWS-supplied connector for JSD version 1.0.4 is available at no charge in the Atlassian Marketplace.

End user

Q: How do I find out which products are available?

You can see which products are available by logging in to the AWS Service Catalog console and searching the portal for products that meet your needs, or you can navigate to the full product list page. You can sort to find the product that you want.

For each product, you can view a Product details page that displays information about the product, including the version, whether a newer version of the product is available, a description, support information, and tags associated with the product. The Product details page might also indicate whether the product will be provisioned using your access permissions (Self) or an administrator-specified role (role-arn).

Q: How do I deploy a product?

When you find a product that meets your requirements in the portal, choose Launch. You will be guided through a series of questions about how you plan to use the product. The questions might be about your business needs or your infrastructure requirements (such as “Which EC2 instance type?”). When you have provided the required information, you'll see the product in the AWS Service Catalog console. While the product is being provisioned, you will see that it is "in progress." After provisioning is complete, you will see "complete" and information, such as endpoints or Amazon Resource Names (ARNs), that you can use to access the product.

Q: Can I see which products I am using?

Yes. You can see which products you are using in the AWS Service Catalog console. You can see all of the stacks that are in use, along with the version of the product used to create them.

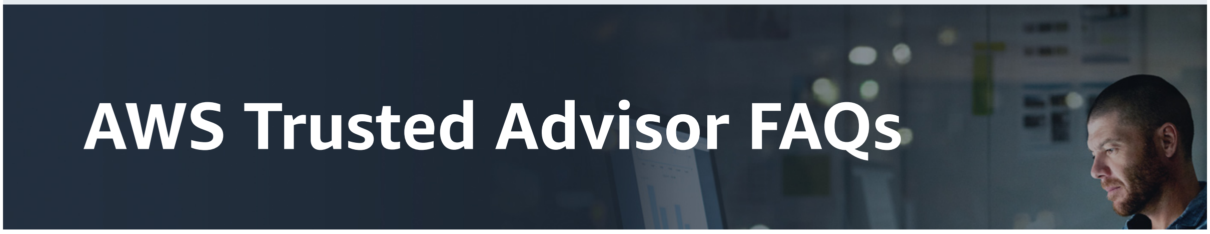
Q: How do I update my products when a new version becomes available?

When a new version of a product is published, you can use the Update Stack command to use that version. If you are currently using a product for which there is an update, it continues to run until you close it, at which point you can choose to use the new version.

Q: How do I monitor the health of my products?

You can see the products that you are using and their health state in the AWS Service Catalog console.

AWS Support



AWS Trusted Advisor FAQs

AWS Trusted Advisor General FAQs

Q: What is AWS Trusted Advisor?

Q: How do I access Trusted Advisor?

Q: What made you choose the current checks/recommendations over others?

Q: Does Trusted Advisor monitor my usage? Can Amazon see what I’m doing with AWS?

Q: What does Trusted Advisor check?

Q: How does the Trusted Advisor notification feature work?

Q: How does the "Recent Changes" feature work?

Q: How does the "Exclude Items" function work?

Q: What is an "Action Link"?

Q: How do I manage the access to the Trusted Advisor console? What is the new IAM policy?

Q: How do I access AWS Trusted Advisor via API?

Q: How often can I refresh my Trusted Advisor result?

Q: How do Trusted Advisor activities affect my Amazon CloudTrail logs?

Q: Which Trusted Advisor checks and features are available to all AWS customers?

Q: Why aren’t my CloudWatch event rules and metric alarms for the EC2 On-Demand Instances check working?

Service Limit Check Questions

Q: What service limits do you check?

Q: What are the default service limits?

Q: How can I get the Service Limit data with command-line tools?

Reserved Instance Optimization Check Questions

Q: What data set are you using to make a Reserved Instance recommendation?

Q: Does the recommendation consider volume discounts?

Q: I just purchased a new Reserved Instance. Why isn’t it showing up in the recommendation?

Q: How do you calculate the optimized number of Reserved Instances?

Q: Do you include other Reserved Instance types in the recommendation?

Q: Why are there separate sections for 1 year and 3 year Reserved Instances?

Q: Are all instance types included in the recommendation?

Q: I use spot instance. Do you include spot rates in the calculation?

Q: I have third party Reserved Instances from the Reserved Instance Marketplace. Do you include those in the results?

Q: Does the recommendations include any money I made if I sell my existing Reserved Instance to purchase the recommended Partial Upfront Reserved Instances?

Q: What defines the alert criteria for this check?

Q: What is the recommended action when the check goes yellow?

Q: Where can I learn more about Reserved Instances?

Q: What does each field in the check result mean?

Q: Why do I see a blue question mark for this recommendation on the Trusted Advisor Console?

Q: Why can I not refresh this recommendation every 5 minutes?

AWS Well-Architected Tool FAQs

General

Q: What is the AWS Well-Architected Tool?
The AWS Well-Architected Tool lets you review your workloads against current AWS best practices and obtain advice on how to architect your workloads for the cloud. This tool uses the AWS Well-Architected Framework.

Q: What is the AWS Well-Architected Framework?
The AWS Well-Architected Framework enables customers and partners to review their architectures using a consistent approach and provides guidance to improve designs over time. Learn more.

Q: What are the pillars of the AWS Well-Architected Framework?
The general design principles and specific AWS best practices and guidance are organized into five conceptual areas. These conceptual areas are the pillars of the AWS Well-Architected Framework. These five pillars are operational excellence, security, reliability, performance efficiency, and cost optimization.

Q: What is a workload?
A workload is the collection of resources and code that make up a cloud application.

Q: What can I do with the AWS Well-Architected Tool?
You can review workloads against best practices across the five architectural pillars of operational excellence, security, reliability, performance efficiency, and cost optimization. The AWS Well-Architected Tool delivers a list of issues found in your workloads and step-by-step guidance to make improvements. You can also view the issues found across the portfolio of workloads.

Q: How do I get started with the AWS Well-Architected Tool?
You can access the AWS Well-Architected Tool by signing in to the AWS Management Console with your AWS account. After accessing the AWS Well-Architected Tool, you can define the workload for which you want to perform a Well-Architected review and then answer questions across each of the five architectural pillars. The tool will evaluate your responses and provide an improvement plan with a prioritized list of issues for the workload.

Q: When should I perform a Well-Architected workload review?
We recommend performing a workload review at major milestones in your development cycle.

Availability & Pricing

Q. Which AWS Regions is the AWS Well-Architected Tool available in?
Visit the AWS Region Table for the current list of regions for the AWS Well-Architected Tool.

Q. How much does the AWS Well-Architected Tool cost?
There is no additional charge for the AWS Well-Architected Tool. You pay only for your underlying AWS resources.

Partner Access

Q. How do I get help with workload reviews and improvement initiatives?
Locate a partner who has been trained on the AWS Well-Architected Framework.

Q: I'm an APN Partner, how do I get involved?
Learn about the AWS Well-Architected Partner Program.

Amazon Elastic Transcoder FAQs

General

Q: What is Amazon Elastic Transcoder?

Amazon Elastic Transcoder is a highly scalable, easy to use and cost effective way for developers and businesses to convert (or “transcode”) video and audio files from their source format into versions that will playback on devices like smartphones, tablets and PCs.

Q: What can I do with Amazon Elastic Transcoder?

You can use Amazon Elastic Transcoder to convert video and audio files into supported output formats optimized for playback on desktops, mobile devices, tablets, and televisions. In addition to supporting a wide range of input and output formats, resolutions, bitrates, and frame rates, Amazon Elastic Transcoder also offers features for automatic video bit rate optimization, generation of thumbnails, overlay of visual watermarks, caption support, DRM packaging, progressive downloads, encryption and more. For more details, please visit the Product Details page.

Q: Why should I use Amazon Elastic Transcoder?

Amazon Elastic Transcoder manages all the complexity of running media transcoding in the AWS cloud. Amazon Elastic Transcoder enables you to focus on your content, such as the devices you want to support and the quality levels you want to provide, rather than managing the infrastructure and software needed for conversion. Amazon Elastic Transcoder scales to handle the largest encoding jobs. As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources that you use. We offer a free tier that enables you to explore the service and transcode up to up to 20 minutes of SD video or 10 minutes of HD video a month free of charge. To see terms and additional information on the free tier program, please visit the AWS Free Usage Tier page.

Q: How do I get started with Amazon Elastic Transcoder?

You can sign up for Amazon Elastic Transcoder through the AWS Management Console. You can then use the console to create a pipeline, set up an IAM role, and create your first transcoding job. To help you test Amazon Elastic Transcoder, the first 20 minutes of SD content (or 10 minutes of HD content) transcoded each month is provided free of charge. Once you exceed the number of minutes in this free usage tier, you will be charged at the prevailing rates. We do not watermark the output content or otherwise limit the functionality of the service, so you can use it and truly get a feel for its capabilities. To see terms and additional information on the free tier program, please visit the AWS Free Usage Tier page. If you do not have an AWS account, you can create one by clicking the Sign Up button at the top of this page.

Q: How do I use Amazon Elastic Transcoder?

To use Amazon Elastic Transcoder you need to have at least one media file in an Amazon S3 bucket. The easiest way to use Amazon Elastic Transcoder is to try it through the console. Create a transcoding pipeline that connects the input Amazon S3 bucket to the output Amazon S3 bucket. Create a transcoding job that will transcode your media file, choose a transcoding preset (a template), and submit the job. Your transcoded file will appear in your output bucket once it has been processed.

Q: What tools and libraries work with Amazon Elastic Transcoder?

Amazon Elastic Transcoder uses a JSON API, and we provide SDKs for Python, Node.js, Java, .NET, PHP, and Ruby. The new AWS Command Line Interface also supports Amazon Elastic Transcoder. You can see a full list of our SDKs here.

Q: Can I use the AWS Management Console with Amazon Elastic Transcoder?

Yes. Amazon Elastic Transcoder has a console that is accessed through the AWS Management Console. You can use our console to create pipelines, jobs, and presets as well as manage and view existing pipelines and jobs.

Q: How do I get my media files into Amazon S3?

There are many ways to get content into Amazon S3, from the simple web-based uploader in the AWS Management Console to programmatic approaches through APIs. For very large files, you may wish to use AWS Import/Export, AWS Direct Connect, or file-acceleration solutions available in the AWS Marketplace. For more information please refer to the Amazon S3 documentation and the AWS Digital Media website.

Q: How do I retrieve my media files from Amazon S3?

You can retrieve files from Amazon S3 programmatically, using the AWS Management Console or a third party tool. You can also mark Amazon S3 objects as public and download them directly from Amazon S3.

Q: Can I use a Content Distribution Network (CDN) to distribute my media files?

Yes. You can easily use CDNs to distribute your content; for example, you can use Amazon CloudFront to distribute your content to end-users with low latency, high data transfer speeds, and no commitments. You can use an output bucket that contains your transcoded content in Amazon S3 as the origin server for Amazon CloudFront. For more information, please visit the detail page for Amazon CloudFront.

Q: How long does it take to transcode a job?

Jobs start processing in the order in which they are received in a pipeline. Once a job is ready to be transcoded, many variables affect the speed of transcoding, for example, the input file size, resolution, and bitrate. For example, if you were to submit a 10 minute video using the iPhone 4 preset, it would take approximately 5 minutes. If a large number of jobs are received they are backlogged (queued). Please note that the transcoding speed may be different between regions.

Q: When will my job be ready?

You can use Amazon SNS notifications to be informed of job status changes. For example, you can be notified when your job starts to transcode and when it has finished transcoding. For more information on Amazon SNS notifications, please see the detail page on Amazon SNS.

Q: How many jobs are processed at once?

Pipelines operate independently from one another. Each pipeline processes jobs in parallel up to a default limit set for that pipeline. Within a job, each individual output also progresses in parallel. For more information on limits and capacity, visit the limits section in the Elastic Transcoder Developer Guide. You can request higher limits by opening a support case.

Q: How many jobs can I submit?

Currently, we allow a maximum of 100,000 jobs per pipeline. Once you exceed this limit, you will receive a 429 Rate Limit Exception. If you require this limit to be raised, please contact us here.

Q: Can I create multiple outputs per job?

Each transcoding job relates to a single input file and can create one or more output files. For example, you may wish to create audio only, low- and high-resolution renditions of the same input file and could do so as part of a single transcoding job. The number of outputs per job is limited. For more information on Amazon Elastic Transcoder limits, please refer to the documentation.

Multiple outputs are charged individually: each output is charged as a separate transcode.

Q: How do I generate clips?

You can create a clip from your source media in your transcoding job. You specify a start time and a duration (both specified as HH:mm:ss.SSS or sssss.SSS.) To cut off the start of a file, you would just specify a start time. You can generate different length clips (or transcode the entire file) for each different output in your transcoding job. You will be charged based on the output duration of your transcode, so if you have a five-minute input file and you create a one-minute output from it, you will only be charged for one minute of transcoding. Please remember that fractional minutes are rounded up, so if you create a clip that is one minute and thirty seconds in duration, you will be charged for two minutes of transcoding.

Q: How do I stitch clips?

You can specify two or more input files that need to be stitched to create a single output file in your transcoding job. Input files are stitched in the order they are specified. So if you want to add a bumper to your video, specify the bumper file as the first input and your video file as the second input. For each input, you can specify a Start Time and a Duration, which allows you to stitch together only the parts of each input that you want included in the output. You will be charged for the output duration of your transcode, so if you are stitching two five-minute input files to create a ten-minute output, you will be charged for ten minutes of transcoding.

Q: What is a transcoding pipeline, what can I use it for, and how many can I have?

A pipeline is a queue-like structure that manages your transcoding jobs. A pipeline can process multiple jobs simultaneously, and generally starts to process jobs in the order in which you added them to the pipeline. Jobs often finish in a different order based on job specifications. It is up to you how you wish to use pipelines. Some examples include submitting jobs to different pipelines based on the priority or the duration of a transcode, or using different pipelines for your development, test and production environments. The number of pipelines per AWS account is limited. For more information on Amazon Elastic Transcoder limits, please refer to the documentation.

Q: What are transcoding presets?

A preset is a template that contains the settings that you want Amazon Elastic Transcoder to apply during the transcoding process, for example, the codec and the resolution that you want in the transcoded file. When you create a job, you specify which preset you want to use. We provide presets that create media files that play on any device and presets that target specific devices. For maximum compatibility, choose a “breadth preset” that creates output that plays on a wide range of devices. For optimum quality and file size, choose an “optimized preset” that creates output for a specific device or class of devices.

Q: What do I do if none of your transcoding presets work for me?

You can create your own custom presets based on an existing preset. Once you create your own custom preset, it is available across your AWS account for the Amazon Elastic Transcoder service within a specific region. For more information on presets, please refer to the Amazon Elastic Transcoder Developer Guide. The number of pipelines per AWS account is limited. For more information on Amazon Elastic Transcoder limits, please refer to the documentation.

Q: Why do I need to assign a role to a transcoding pipeline?

Amazon Elastic Transcoder uses AWS Identity and Access Management (IAM) roles to enable you to securely control access to your media assets. The IAM role sets a policy that defines what permissions you have for accessing Amazon S3 resources. You can assign different roles to different pipelines, and an IAM administrator can create specific roles for use with Amazon Elastic Transcoder. More information about IAM can be found here.

Q: How can I configure roles to be more restrictive?

You can use the AWS Management Console to edit and create new IAM roles. IAM roles that are created by Amazon Elastic Transcoder are visible in the AWS Management Console and can also be edited.

Q: How do I use notifications?

Amazon Elastic Transcoder uses Amazon SNS to notify you of specific events. You can choose to be notified about jobs that start to process, jobs that complete, warnings, and errors. Each event type is assigned to an SNS topic, and you can use the same topic or different topics for each event. The Amazon Elastic Transcoder console will create an SNS topic for you or you can specify an existing one.

Q: Why should I use notifications?

Notifications are a much more efficient way to check transcoding status than polling the API. Notifications provide a way to be notified on specific events that occur in the system. For example, you can be notified on a completed event. This is useful if you want to know when a job has finished transcoding and this is far more efficient than calling the 'List Jobs By Status' or 'Read Job' API at regular intervals.

Q: Why does my job keep failing?

The most common reason for jobs to fail is that the input file is corrupted in some way. If you receive an error about the format not being supported, we are unable to decode your source file and we’d love for you to tell us more about on our Discussion Forum. We need the following information to assist with diagnosis: AWS Account ID, Region and Job ID. For a list of error codes, please refer to the documentation.

Q: How can I generate more than one thumbnail per job?

You can specify a thumbnail creation interval in seconds to create one thumbnail every n seconds. To create thumbnails in more than one size, you need to create different jobs.

Q: Can I reserve a transcoder for my exclusive use?

Amazon Elastic Transcoder provides a shared transcoding service and does not enable a transcoder to be reserved or allocated to an individual customer.

Q: Do I need to pay license fees?

We have licensed relevant intellectual property from the applicable patent pools for transcoding content. Like any other transcoder, customers are responsible for evaluating and, if necessary, securing licenses for distribution of content in various formats.

Q: Do you support live encoding?

Amazon Elastic Transcoder is a file-based transcoding service and does not support live transcoding.

Q: Are there limits to the service?

The number of transcoding pipelines, transcoding presets and outputs per job have limits. Most of these limits can be adjusted on a customer-by-customer basis. For the current limits, please refer to the documentation.

Q: How do I increase service limits?

If you require an increase in the service limits, please contact us here and provide all the information requested on the form. We will then contact you to discuss your requirements.

Q: Where is Amazon Elastic Transcoder available?

Amazon Elastic Transcoder is available in the following AWS regions: US East (N Virginia), US West (Oregon), US West (N California), EU (Ireland), Asia Pacific (Tokyo), Asia Pacific (Singapore), Asia Pacific (Sydney), and Asia Pacific (Mumbai).

The service operates standalone in each region, so jobs created in one region may not be transferred to another region.

You can create a transcoding pipeline in one region that would specify Amazon S3 buckets in another region. However, if you choose to do this, you should be aware that you will incur Amazon S3 transfer costs when content is read from or written out to an Amazon S3 bucket in a region other than the one where the transcoding work is taking place.

Q: Can I pass metadata when creating a job?

You have the option to attach up to 10 custom metadata key-value pairs to your Elastic Transcoder jobs. This metadata will be included in the job notifications and when reading the job via the API or console. You provide this information in the “UserMetadata” field on the Job object.

Format Support

Q: What input formats do you support?

We support popular web, consumer and professional media formats. Examples include 3GP, AAC, AVI, FLV, MP4 and MPEG-2. If there is a format that you’ve found does not work, please let us know through our forum.

Q: Where can I find a comprehensive list of support formats?

We add new input formats on an ongoing basis, so such a list would age quickly. Please take advantage of our free tier and console to try a format not mentioned above and if you run into problems, please let us know!

Q: When creating MP4 files, do you support "fast start"?

We locate the MOOV atom for an MP4 at the start of the file so that your player can start playback immediately without waiting for the entire file to finish downloading.

Q: Do you support Apple ProRes or digital cinematography formats?

We do not support reading Apple ProRes files or raw camera formats like ARRI and RED at this time.

Q: What video formats can I transcode into?

We support the following video codecs: H.264, VP9, VP8, MPEG-2, and animated GIF. File formats supported include MPEG-2 TS container (for HLS), fmp4 (for Smooth Streaming and MPEG-DASH), MP4, WebM, FLV, MPG, and MXF (XDCAM-compatible). For information on file formats that are supported by specific codecs, please visit the Product Details page.

Q: What audio formats can I transcode into?

We support the following audio codecs: AAC, MP3, MP2, PCM, FLAC, and Vorbis. Audio-only file formats supported include MP3, MP4, FLAC, OGA, OGG, and WAV. For information on file formats that are supported by specific codecs, please visit the Product Details page.

Q: How is album art supported for audio files?

Album art is supported in MP4 files containing AAC audio, in MP3 files, and in FLAC files. Album art is not supported for OGA, OGG, WAV, WebM or MPEG-2 TS outputs. You can specify whether album art from the source file is passed through to the output, removed, or whether new album art should replace it or be appended to it.

Q: How do I create an audio file from a video file?

To strip out video and create an output that only contains the audio track, run a transcoding job with your input file and use one of the system transcoding presets that contains Audio in its name. Alternatively, you can create your own audio only custom transcoding preset. The output file will only contain the audio portion of the input file.

Q: Do you support surround sound formats?

The audio portion of the transcoded output from Amazon Elastic Transcoder is two-channel AAC, MP3 or Vorbis.

Q: Do you support audio channel remapping?

If the source file contains multi-channel audio, the output will contain the first two channels, which are frequently left and right audio tracks. For the MXF container, we support multiple modes of packaging the audio into the file, including optional insertion of motor only shots (MOS).

Q: Can I generate XDCAM-compatible video?

Yes, the easiest way to generate XDCAM-compatible outputs is to specify one of the XDCAM system presets when creating a transcoding job. You can also create a custom preset by choosing the MXF container with MPEG-2 video and PCM audio.

Q: Do you support closed captions?

Yes, you can add, remove, or preserve captions as you transcode your video from one format to another.

Supported input formats:
Embedded: CEA-608, CEA-708 (MPEG-2 only) and mov-text
Sidecar captions: DFXP, EBU-TT, SCC, SMPT, SRT, TTML, WebVTT

Supported output formats:
Embedded captions: mov-text (MP4), and CEA-708 (MP4 and MPEG-TS)
Sidecar captions: DFXP, EBU-TT, SCC, SMPT, SRT, TTML, and WebVTT

CEA-708 captions are embedded in the H.264 SEI user data of the stream.

Q: Can you support multiple caption tracks?

Yes, you can add one track per language.

Q: How do I create content for HLS output?

There are two steps:

1. Create a transcoding job containing outputs for each variation using one of our supplied system presets or your own, based on the MPEG-2 TS container and H.264 and AAC codecs. The lowest rate stream should be an audio only stream.

2. Specify that the transcoding job create a playlist that references the outputs. You should order your bit rates from lowest to highest, with the audio only stream last, since this order will be maintained in the generated playlist file. Once your transcoding job has completed, the output bucket will contain a proper arrangement of your master and individual M3U8 playlists, and MPEG-2 TS media stream fragments.

Note: When selecting the HLSv4 option, your outputs should be matched to audio-only and video-only presets. For system presets, these can be identified by words "Audio" or "Video" as part of their name. For example, “System preset: HLS Video – 600k,” would match with the HLSv4 option whereas "System preset: HLS – 600k,” would be used with the HLSv3 option.

Q: How do I create content for Smooth Streaming?

There are two steps:

1. Create a transcoding job containing outputs for each variation using one of our supplied system presets or your own, based on the fragmented MP4 container and H.264 and AAC codecs.

2. Specify that the transcoding job create a playlist that references the outputs. Once your transcoding job has completed, the output bucket specified by the transcoding pipeline will contain your manifest ISM file, client ISMC file, and fragmented MP4 media files.

Q: How do I create content for MPEG-DASH streaming?

There are two steps:

1. Create a transcoding job containing the video-only outputs (with the desired resolutions and bitrates) and the audio-only output using either the system presets or your own customized presets, based on the fragmented MP4 container with H.264 video and AAC audio.

2. Create an MPEG-DASH playlist for the transcoding job by selecting MPEG-DASH as the Playlist Format. Specify the outputs that this playlist will reference. Once your transcoding job has completed, the output bucket specified by the transcoding pipeline will contain your manifest MPD file, and the fragmented MP4 media files.

Q: Should I use the HLSv3 or the HLSv4 option?

HLS version 3 has been supported natively on iOS 2+ devices since July 2008 and on Android 4.0+ since Oct. 2011. HLS version 4 has been supported natively on iOS 5+ devices since Oct. 2011 and on Android 4.4+ since Sept. 2013.

If you able to reach your target devices with HLS version 4, you will be able to generate playlists that use byte range requests, late-binding audio, and I-frame only playback. Playlists with byte range requests are able to use just one file per bit rate, eliminating the need to manage thousands of small segment files. Late-binding audio allows the audio to be streamed separately from the video, eliminating redundant audio storage. I-frame only playback enables trick-play modes used to enhance fast-forward, rewind, and seeking through the video.

Q: Can I stream HLS directly from S3?

Yes, you can play your HLS renditions directly from S3 by pointing the player to the M3U8 playlist. We recommend you use a CDN such as Amazon CloudFront, which provides a better end user experience with improved scalability and performance. See Configuring On-Demand Apple HTTP Live Streaming (HLS).

Q: Do I need a streaming server to deliver my Smooth Streaming content?

Usually playing back Smooth Streaming requires an IIS origin server, and you cannot stream directly from S3. However, if you distribute your content with CloudFront you can simply configure a CloudFront Smooth Streaming distribution, eliminating the need for a streaming server. See Configuring On-Demand Smooth Streaming.

Q: Why is the codec parameter that I want to change not exposed by the API?

In designing Amazon Elastic Transcoder, we wanted to create a service that was simple to use. Therefore, we expose the most frequently used codec parameters. If there is a parameter that you require, please let us know by letting us know through our forum.

Q: What settings do I use to preserve the dimensions of my video?

Use the following settings in your custom preset:
MaxWidth: auto; MaxHeight: auto; SizingPolicy: ShrinkToFit; PaddingPolicy: NoPad; DisplayAspectRatio: auto

Q: How do I scale my output to a specified width and set the height to preserve the aspect ratio of the source content?

Use the following settings in your custom preset:
MaxWidth: [Desired Width]; MaxHeight: auto; SizingPolicy: Fit; PaddingPolicy: NoPad; DisplayAspectRatio: auto

Q: How do I limit the height or width of a video without stretching the output to fit my set limit while preserving the input aspect ratio?

Use the following settings in your custom preset:
MaxWidth: [Desired Width Limit]; MaxHeight: [Desired Height Limit]; SizingPolicy: ShrinkToFit; PaddingPolicy: NoPad; DisplayAspectRatio: auto

Q: What settings should I use to create a preset that causes the output video to fill the screen without distortion, if necessary cropping some of the edges ("center cut")?

Use the following settings in your custom preset:
MaxWidth: [Desired Width]; MaxHeight: [Desired Height]; SizingPolicy: Fill; PaddingPolicy: NoPad; DisplayAspectRatio: auto

Q: What settings should I use to create a preset that causes the output video to fill the screen without cropping any image area, if necessary distorting the image ("squeeze" or "stretch")?

Use the following settings in your custom preset:
MaxWidth: [Desired Width]; MaxHeight: [Desired Height]; SizingPolicy: Stretch; PaddingPolicy: NoPad; DisplayAspectRatio: auto

Q: How do I make my watermark scale with my video?

In the watermark settings of your transcoding preset, set the HorizontalAlign, VerticalAlign, and Target parameters as desired. Then set the HorizontalOffset and VerticalOffset with relative parameters. For example, to place the watermark 10% away from the edges, set both values to 10%.

Q: How do I avoid distorting my watermark?

If you do not want your watermark to be distorted when the video output is resized, set the SizingPolicy to ShrinkToFit while setting MaxWidth and MaxHeight to 100%. With these settings, Elastic Transcoder will never up-sample, expand, or distort your watermark.

Q: What are the settings for placing my watermark over the active video region rather than over the matte?

To place your watermark so that it is always over the active video content, use relative size for the MaxWidth and MaxHeight settings, and set the Target to be Content. For example, to fix the watermark size to 10% of the active output video size, set both MaxWidth and MaxHeight to 10%.

Q: How do I use multiple watermarks?

Presets specify placement settings for up to four watermarks. Each setting has an associated watermark ID. You can create a job with up to four watermarks by specifying an array of watermarks in the job creation call. Each element of the array specifies the Id of the watermark setting to use, and the watermark image file.

Q: Can I generate NTSC or PAL outputs?

Yes, you can generate both NTSC and PAL compliant outputs. The easiest way to generate NTSC and PAL compliant outputs is to specify the NTSC or PAL system preset when creating a transcoding job. Via the console, this is done by the preset drop down for each output of your transcoding job.

Pricing

Q: How much does Amazon Elastic Transcoder cost to use?

Pricing for Amazon Elastic Transcoder is described here. Our pricing does not require any commitment or minimum volume of jobs. We also offer a free tier that enables you to explore the service and transcode up to up to 20 minutes of audio-only output, 20 minutes of SD video output and 10 minutes of HD video output a month free of charge. To see terms and additional information on the free tier program, please visit the AWS Free Usage Tier page.

Q: How are jobs charged?

Transcoding jobs are charged according to the duration of the content. For example, media that lasts 60 minutes costs twice as much as media that lasts 30 minutes. High definition (HD) content costs twice as much as standard definition (SD). Audio-only output is priced lower than standard definition (SD) output. The minimum charge for a job is one minute. We do not charge for thumbnail generation, for API calls, or for Amazon S3 transfer within the same region. For more information, please refer to the Amazon Elastic Transcoder pricing page.

Q: How are fractional minutes charged?

Fractional minutes are rounded up. For example, if your output duration is less than a minute, you are charged for one minute. If your output duration is 1 minutes and 10 seconds, you are charged for 2 minutes.

Q: Do you charge for failed jobs?

Our policy is to forgive customers for failed jobs unless the number of failed jobs becomes excessive.

Q: Is it cheaper to use multiple outputs per job than to use separate jobs?

When you use multiple outputs per job, transcoding costs remain the same as if you had submitted multiple jobs for each output. However, the processing time will be quicker for larger jobs since the source file is only being transferred from your S3 bucket to Amazon Elastic Transcoder once.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Security

Q: Are my media assets secure?

You are in complete control of your media assets because they are stored in your own Amazon S3 buckets. You use IAM roles to grant us access to your specific Amazon S3 bucket.

Q: Can I set S3 permissions and storage options?

Amazon Elastic Transcoder enables you to specify which users, groups, and canonical IDs you want to grant access to your transcoded files, thumbnails and playlists, as well as the type of access that you want them to have. You can also specify whether to store transcoded content using Standard or Reduced Redundancy Storage. Please refer to Amazon Elastic Transcoder documentation for further information.

Q: Can I use encrypted input media files or encrypt my output files?

Yes. You can use encrypted mezzanine files as input to Amazon Elastic Transcoder, or protect your transcoded files by letting the service encrypt the output. Supported options range from fully managed integration with Amazon S3's Server-Side Encryption, to keys that you manage on your own and protect using AWS Key Management Service (KMS). Furthermore, encryption support is not limited to your video files. You can protect thumbnails, captions, and even watermarks.

Q: Do you support DRM?

Yes, we support packaging for Microsoft PlayReady DRM. Our Smooth Streaming packaging is compatible with the Microsoft PIFF 1.1, and our HLSv3 packaging is compatible with the Discretix 3.0.1 specification for Microsoft PlayReady.

Q: Can I get a history of all Amazon Elastic Transcoder API calls made on my account for security, operational or compliance auditing?

Yes. To start receiving a history of all Elastic Transcoder API calls made on your account, you simply turn on AWS CloudTrail in CloudTrail's AWS Management Console. For more information, visit the AWS CloudTrail home page.

Q: Do I need to setup AWS KMS before using the Elastic Transcoder encryption and DRM packaging features?

Yes. You must first create a master AWS KMS key and add the role used by Elastic Transcoder as an authorized user of that key. Elastic Transcoder uses your KMS master key to protect the data encryption keys that it exchanges with you.

Q: Can I save the keys used to encrypt my HLS streams to S3?

Yes. If you elect to store your keys in S3, Elastic Transcoder will write your keys to the same folders as your playlist files, and your keys will be protected using Server-Side Encryption with Amazon S3-Managed Encryption Keys (SSE-S3).

Q: Can I rotate the keys used for HLS with AES-128 encryption?

Key rotation is not supported. All renditions and file segments share the same key.

Amazon Elastic Transcoder

Product Details

Pricing

Getting Started

Developer Resources

FAQs

RELATED LINKS

Documentation

Management Console

Release Notes

Discussion Forum

Amazon Kinesis Video Streams FAQs

General

Q: What is Amazon Kinesis Video Streams?

Amazon Kinesis Video Streams makes it easy to securely stream media from connected devices to AWS for storage, analytics, machine learning (ML), playback, and other processing. Kinesis Video Streams automatically provisions and elastically scales all the infrastructure needed to ingest streaming media from millions of devices. It durably stores, encrypts, and indexes media in your streams, and allows you to access your media through easy-to-use APIs. Kinesis Video Streams enables you to quickly build computer vision and ML applications through integration with Amazon Rekognition Video, Amazon SageMaker, and libraries for ML frameworks such as Apache MxNet, TensorFlow, and OpenCV. For live and on-demand playback, Kinesis Video Streams provides fully-managed capabilities for HTTP Live Streaming (HLS) and Dynamic Adaptive Streaming over HTTP (DASH). Kinesis Video Streams also supports ultra-low latency two-way media streaming with WebRTC, as a fully managed capability.

Q: What is time-encoded data?

Time-encoded data is any data in which the records are in a time series, and each record is related to its previous and next records. Video is an example of time-encoded data, where each frame is related to the previous and next frames through spatial transformations. Other examples of time-encoded data include audio, RADAR, and LIDAR signals. Amazon Kinesis Video Streams is designed specifically for cost-effective, efficient ingestion, and storage of all kinds of time-encoded data for analytics and ML use cases.

Q: What are common use cases for Kinesis Video Streams?

Kinesis Video Streams is ideal for building media streaming applications for camera-enabled IoT devices and for building real-time computer vision-enabled

ML applications that are becoming prevalent in a wide range of use cases such as the following:

Smart Home

With Kinesis Video Streams, you can easily stream video and audio from camera-equipped home devices such as baby monitors, webcams, and home surveillance systems to AWS. You can then use the streams to build a variety of smart home applications ranging from simple media playback to intelligent lighting, climate control systems, and security solutions.

Smart City

Many cities have installed large numbers of cameras at traffic lights, parking lots, shopping malls, and just about every public venue, capturing video 24/7. You can use Kinesis Video Streams to securely and cost-effectively ingest, store, playback, and analyze this massive volume of media data to help solve traffic problems, help prevent crime, dispatch emergency responders, and much more.

Industrial Automation

You can use Kinesis Video Streams to collect a variety of time-encoded data such as RADAR and LIDAR signals, temperature profiles, and depth data from industrial equipment. You can then analyze the data using your favorite machine learning framework including Apache MxNet, TensorFlow, and OpenCV for industrial automation use cases like predictive maintenance. For example, you can predict the lifetime of a gasket or valve and schedule part replacement in advance, reducing downtime and defects in a manufacturing line.

Q: What does Amazon Kinesis Video Streams manage on my behalf?

Amazon Kinesis Video Streams is a fully managed service for media ingestion, storage, and processing. It enables you to securely ingest, process, and store video at any scale for applications that power robots, smart cities, industrial automation, security monitoring, machine learning (ML), and more. Kinesis Video Streams also ingests other kinds of time-encoded data like audio, RADAR, and LIDAR signals. Kinesis Video Streams provides you SDKs to install on your devices to make it easy to securely stream media to AWS. Kinesis Video Streams automatically provisions and elastically scales all the infrastructure needed to ingest media streams from millions of devices. It also durably stores, encrypts, and indexes the media streams and provides easy-to-use APIs so that applications can retrieve and process indexed media fragments based on tags and timestamps. Kinesis Video Streams provides a library to integrate ML frameworks such as

Apache MxNet, TensorFlow, and OpenCV with video streams to build machine learning applications. Kinesis Video Streams is integrated with Amazon Rekognition Video, enabling you to build computer vision applications that detect objects, events, and people.

Key concepts

Q: What is a video stream?

A video stream is a resource that enables you to capture live video and other time-encoded data, optionally store it, and make the data available for consumption both in real time and on a batch or ad-hoc basis. When you choose to store data in the video stream, Kinesis Video Streams will encrypt the data, and generate a time-based index on the stored data. In a typical configuration, a Kinesis video stream has only one producer publishing data into it. The Kinesis video stream can have multiple consuming applications processing the contents of the video stream.

Q: What is a fragment?

A fragment is a self-contained sequence of media frames. The frames belonging to a fragment should have no dependency on any frames from other fragments. As fragments arrive, Kinesis Video Streams assigns a unique fragment number, in increasing order. It also stores producer-side and server-side time stamps for each fragment, as Kinesis Video Streams-specific metadata.

Q: What is a producer?

A producer is a general term used to refer to a device or source that puts data into a Kinesis video stream. A producer can be any video-generating device, such as a security camera, a body-worn camera, a smartphone camera, or a dashboard camera. A producer can also send non-video time-encoded data, such as audio feeds, images, or RADAR data. One producer can generate one or more video streams. For example, a video camera can push video data to one Kinesis video stream and audio data to another.

Q: What is a consumer?

Consumers are your custom applications that consume and process data in Kinesis video streams in real time, or after the data is durably stored and time-indexed when low latency processing is not required. You can create these consumer applications to run on Amazon EC2 instances. You can also use other Amazon AI services such as Amazon Rekognition, or third party video analytics providers to process your video streams.

Q: What is a chunk?

Upon receiving the data from a producer, Kinesis Video Streams stores incoming media data as chunks. Each chunk consists of the actual media fragment, a copy of media metadata sent by the producer, and the Kinesis Video Streams-specific metadata such as the fragment number, and server-side and producer-side timestamps. When a consumer requests media data through the GetMedia API operation, Kinesis Video Streams returns a stream of chunks, starting with the fragment number that you specify in the request.

Q: How do I think about latency in Amazon Kinesis Video Streams?

There are four key contributors to latency in an end-to-end media data flow.

- Time spent in the device's hardware media pipeline: This pipeline can comprise of the image sensor and any hardware encoders as appropriate. In theory, this can be as little as a single frame duration. In practice it rarely is. All encoders in order to work effectively for media encoding (compression) will accumulate several frames to construct a fragment. This process and any corresponding motion compensation algorithms will add anywhere from one second to several seconds of latency on the device before the data is packaged for transmission.
- Latency incurred on actual data transmission on the internet: The quality of the network throughput and latency can vary significantly based on where the producing device is located.
- Latency added by the Kinesis Video Streams as it receives data from the producer device: The incoming data is made available immediately on the GetMedia API operation for any consuming application. If you choose to retain data, then Kinesis Video Streams will ensure that the data is encrypted using AWS Key Management Service (AWS KMS) and generate a time-based index on the individual fragments in the video stream. When you access this retained data using the GetMediaforFragmentList API, Kinesis Video Streams fetches the fragments from durable storage, decrypt the data, and make it available for the consuming application.

- Time latency on data transmission back to the consumer: There can be consuming devices on the internet or other AWS regions that request the media data. The quality of the network throughput and latency can vary significantly based on where the consuming device is located.

Publishing data to streams

Q: How do I publish data to my Kinesis video stream?

You can publish media data to a Kinesis video stream via the PutMedia operation, or use the Kinesis Video Streams Producer SDKs in Java, C++, or Android. If you choose to use the PutMedia operation directly, you will be responsible for packaging the media stream according to the Kinesis Video Streams data specification, handle the stream creation, token rotation, and other actions necessary for reliable streaming of media data to the AWS cloud. We recommend using the Producer SDKs to make these tasks simpler and get started faster.

Q: What is the Kinesis Video Streams PutMedia operation?

Kinesis Video Streams provides a PutMedia API to write media data to a Kinesis video stream. In a PutMedia request, the producer sends a stream of media fragments. As fragments arrive, Kinesis Video Streams assigns a unique fragment number, in increasing order. It also stores producer-side and server-side time stamps for each fragment, as Kinesis Video Streams-specific metadata.

Q: What is the Kinesis Video Streams Producer SDK?

The Amazon Kinesis Video Streams Producer SDK are a set of easy-to-use and highly configurable libraries that you can install and customize for your specific producers. The SDK makes it easy to build an on-device application that securely connects to a video stream, and reliably publishes video and other media data to Kinesis Video Streams. It takes care of all the underlying tasks required to package the frames and fragments generated by the device's media pipeline. The SDK also handles stream creation, token rotation for secure and uninterrupted streaming, processing acknowledgements returned by Kinesis Video Streams, and other tasks.

Q: In which programming platforms is the Kinesis Video Streams Producer SDK available?

Kinesis Video Streams Producer SDK's core is built in C, so it is efficient and portable to a variety of hardware platforms. Most developers will prefer to use the C, C++ or Java versions of the Kinesis Video Streams producer SDK. There is also an Android version of the producer SDK for mobile app developers who want to stream video data from Android devices.

Q: What should I be aware of before getting started with the Kinesis Video Streams producer SDK?

The Kinesis Video Streams producer SDK does all the heavy lifting of packaging frames and fragments, establishes a secure connection, and reliably streams video to AWS. However there are many different varieties of hardware devices and media pipelines running on them. To make the process of integration with the media pipeline easier, we recommend having some knowledge of: 1) the frame boundaries, 2) the type of a frame used for the boundaries, I-frame or non I-frame, and 3) the frame encoding time stamp.

Reading data from streams

Q: What is the GetMedia API?

You can use the GetMedia API to retrieve media content from a Kinesis video stream. In the request, you identify stream name or stream Amazon Resource Name (ARN), and the starting chunk. Kinesis Video Streams then returns a stream of chunks in order by fragment number. When you put media data (fragments) on a stream, Kinesis Video Streams stores each incoming fragment and related metadata in what is called a "chunk." The GetMedia API returns a stream of these chunks starting from the chunk that you specify in the request.

Q: What is the GetMediaForFragmentList API?

You can use the GetMediaForFragmentList API to retrieve media data for a list of fragments (specified by fragment number) from the archived data in a Kinesis video stream. Typically a call to this API operation is preceded by a call to the ListFragments API.

Q: What is the ListFragments API?

You can use the ListFragments API to return a list of Fragments from the specified video stream and start location - using the fragment number or timestamps - within the retained data.

Q: How long can I store data in Kinesis Video Streams?

You can store data in their streams for as long as you like. Kinesis Video Streams allows you to configure the data retention period to suit your archival and storage requirements.

Q: What is the Kinesis Video Streams parser library?

The Kinesis Video Streams parser library makes it easy for developers to consume and process the output of Kinesis Video Streams GetMedia operation. Application developers will include the library in their video analytics and processing applications that operate on video streams. The applications themselves will run on your EC2 instances, although they can be run elsewhere. The library has features that make it easy to get a frame-level object and its associated metadata, extract and collect Kinesis Video Streams-specific metadata attached to fragments, and consecutive fragments. You can then build custom applications that can more easily use the raw video data for your use cases.

Q: If I have a custom processing application that needs to use the frames (and fragments) carried by the Kinesis video stream, how do I do that?

In general, if you want to consume video streams and then manipulate them to fit your custom application's needs, then there are two key steps to consider. First, get the bytes in a frame from the formatted stream vended by the GetMedia API. You can use the stream parser library to get the frame objects. Next, get the metadata necessary to decode a frame such as the pixel height, width, codec id, and codec private data. Such metadata is embedded in the track elements. The parser library makes extracting this information easier by providing helper classes to collect the track information for a fragment.

The steps after this are highly application dependent. You may wish to decode frames, format them for a playback engine, transcode them for content distribution, or feed them into a custom deep learning application format. The Kinesis Video Streams stream parser library is open-sourced so that you can extend it for your specific use cases.

Playing back video from streams

Q: How do I playback the video captured in my own application?

You can use Amazon Kinesis Video Streams’ HTTP Live Streams (HLS) and Dynamic Adaptive Streaming over HTTP (DASH) capabilities to playback the ingested video in fragmented MP4 or MPEG_TS packaged format. HLS and DASH are industry-standard, HTTP-based media streaming protocols. As you capture video from devices using Amazon Kinesis Video Streams, you can use the HLS or DASH APIs to playback live or recorded video. This capability is fully managed, so you do not have to build any cloud-based infrastructure to support video playback. For low-latency playback and two-way media streaming, see the FAQs on WebRTC–based streaming.

Q: How do I get started with Kinesis Video Streams HLS or DASH APIs?

To view a Kinesis video stream using HLS or DASH, you first create a streaming session using GetHLSStreamingSessionURL or GetDASHStreamingSessionURL APIs. This action returns a URL (containing a session token) for accessing the HLS or DASH session, which you can then use in a media player or a standalone application to playback the stream. You can use a third-party player (such as Video.js or Google Shaka Player) to display the video stream, by providing the HLS or DASH streaming session URL, either programmatically or manually. You can also play back video by entering the HLS or DASH streaming session URL in the Location bar of the Apple Safari or Microsoft Edge browsers. Additionally, you can use the video players for Android (Exoplayer) and iOS (AVMediaPlayer) for mobile apps.

Q: What are the basic requirements to use the Kinesis Video Streams HLS APIs?

An Amazon Kinesis video stream has the following requirements for providing data through HLS:

- The media must contain h.264 or h.265 encoded video and, optionally, AAC encoded audio. Specifically, the codec ID of track 1 should be V_MPEG/ISO/AVC for h.264 or V_MPEG/ISO/HEVC for h.265. Optionally, the codec ID of track 2 should be A_AAC.
- The video track of each fragment must contain codec private data in the Advanced Video Coding (AVC) for h.264 format or HEVC for h.265 format (MPEG-4 specification ISO/IEC 14496-15). For information about adapting stream data to a given format, see NAL Adaptation Flags.
- Data retention must be greater than 0.

- The audio track (if present) of each fragment must contain codec private data in the AAC format (AAC specification ISO/IEC 13818-7)

Q: What are the basic requirements to use the Kinesis Video Streams DASH APIs?

An Amazon Kinesis video stream has the following requirements for providing data through DASH:

- The media must contain h.264 or h.265 encoded video and, optionally, AAC or G.711 encoded audio. Specifically, the codec ID of track 1 should be V_MPEG/ISO/AVC (for h.264) or V_MPEGH/ISO/HEVC (for H.265). Optionally, the codec ID of track 2 should be A_AAC (for AAC) or A_MS/ACM (for G.711).
- The video track of each fragment must contain codec private data in the Advanced Video Coding (AVC) for H.264 format and HEVC for H.265 format. For more information, see MPEG-4 specification ISO/IEC 14496-15. For information about adapting stream data to a given format, see NAL Adaptation Flags.
- Data retention must be greater than 0.
- The audio track (if present) of each fragment must contain codec private data in the AAC format (AAC specification ISO/IEC 13818-7) or the MS Wave format.

Q: What are the available playback modes for HLS or DASH streaming in Kinesis Video Streams?

There are two different playback modes supported by both HLS and DASH: Live and On Demand.

LIVE: For live sessions, the HLS media playlist is continually updated with the latest fragments as they become available. When this type of session is played in a media player, the user interface typically displays a "live" notification, with no scrubber control for choosing the position in the playback window to display.

ON DEMAND: For on-demand, the HLS media playlist contains all the fragments for the session, up to the number that is specified in MaxMediaPlaylistFragmentResults. The playlist can only be retrieved once for each session.
Additionally, HLS also supports playback in LIVE_REPLAY mode. In this mode, the HLS media playlist is updated similarly to how it is updated for LIVE mode except that it starts by including fragments from a given start time. This mode is useful

for cases when you want to start playback from a point in the past from stored media and continue into live streaming.

Q: What is the delay in the playback of video using the API?

The latency for live playback is typically between 3 and 5 seconds, but this could vary. We strongly recommend running your own tests and proof-of-concepts to determine the target latencies. There are a variety of factors that impact latencies, including the use case, how the producer generates the video fragments, the size of the video fragment, the player tuning, and network conditions both streaming into AWS and out of AWS for playback. For low-latency playback, see the FAQs on WebRTC–based streaming.

Q: What are the relevant limits to using HLS or DASH?

A Kinesis video stream supports a maximum of ten active HLS or DASH streaming sessions. If a new session is created when the maximum number of sessions is already active, the oldest (earliest created) session is closed. The number of active GetMedia connections on a Kinesis video stream does not count against this limit, and the number of active HLS sessions does not count against the active GetMedia connection limit. See Kinesis Video Streams Limits for more details.

Q: What's the difference between Kinesis Video Streams and AWS Elemental MediaLive?

AWS Elemental MediaLive is a broadcast-grade live video encoding service. It lets you create high-quality video streams for delivery to broadcast televisions and internet-connected multiscreen devices, like connected TVs, tablets, smart phones, and set-top boxes. The service functions independently or as part of AWS Media Services.

Amazon Kinesis Video Streams makes it easy to securely stream video from connected devices to AWS for real-time and batch-driven machine learning (ML), video playback, analytics, and other processing. It enables customers to build machine-vision based applications that power smart homes, smart cities, industrial automation, security monitoring, and more.

Q: Am I charged to use this capability?

Kinesis Video Streams uses a simple pay as you go pricing. There are no upfront costs and you only pay for the resources you use. Kinesis Video Streams pricing is based on the data volume (GB) ingested, volume of data consumed (GB) including through the HLS or DASH APIs, and the data stored (GB-Month) across all the video streams in your account. Please see the pricing page for more details.

Low-latency two-way media streaming with WebRTC

Q: What is WebRTC and how does Kinesis Video Streams support this capability?

WebRTC is an open technology specification for enabling real-time communication (RTC) across browsers and mobile applications via simple APIs. It leverages peering techniques for real-time data exchange between connected peers and provides low media streaming latency required for human-to-human interaction. WebRTC specification includes a set of IETF protocols including Interactive Connectivity Establishment (ICE RFC5245), Traversal Using Relay around NAT (TURN RFC5766), and Session Traversal Utilities for NAT (STUN RFC5389) for establishing peer-to-peer connectivity, in addition to protocol specifications for real-time media and data streaming. Kinesis Video Streams provides a standards compliant WebRTC implementation, as a fully-managed capability. You can use this capability to securely live stream media or perform two-way audio or video interaction between any camera IoT device and WebRTC compliant mobile or web players. As a fully-managed capability, you do not have to build, operate, or scale any WebRTC related cloud infrastructure such as signaling or media relay servers to securely stream media across applications and devices.

Q: What does Amazon Kinesis Video Streams manage on my behalf to enable live media streaming with WebRTC?

Kinesis Video Streams provides managed end-points for WebRTC signaling that allows applications to securely connect with each other for peer-to-peer live media streaming. Next, it includes managed end-points for TURN that enables media relay via the cloud when applications cannot stream peer-to-peer media. It also includes managed end-points for STUN that enables applications to discover their public IP address when they are located behind a NAT or a firewall. Additionally, it provides easy to use SDKs to enable camera IoT devices with WebRTC capabilities. Finally, it provides client SDKs for Android, iOS, and for Web applications to integrate Kinesis Video Streams WebRTC signaling, TURN, and STUN capabilities with any WebRTC compliant mobile or web player.

Q: What can I build using Kinesis Video Streams WebRTC capability?

With Kinesis Video Streams WebRTC, you can easily build applications for live media streaming or real-time audio or video interactivity between camera IoT devices, web browsers, and mobile devices for usecases such as helping parents keep an eye on their baby's room, enable home-owners use a video doorbell to check who's at the door, allow owners of camera-enabled robot vacuums to remotely control the robot by viewing the live camera stream on a mobile phone, and much more.

Q: How do I get started with Kinesis Video Streams WebRTC capability?

You can get started by building and running the sample applications in the Kinesis Video Streams SDKs for WebRTC available for Web browsers, Android or iOS based mobile devices, and for Linux, Raspbian, and MacOS based IoT devices. You can also run a quick demo of this capability in the Kinesis Video Streams management console by creating a signaling channel, and running the demo application to live stream audio and video from your laptop's built-in camera and microphone.

Q: What is a Signaling Channel?

A signaling channel is a resource that enables applications to discover, set up, control, and terminate a peer-to-peer connection by exchanging signaling messages. Signaling messages are metadata that two applications exchange with each other to establish peer-to-peer connectivity. This metadata includes local media information such as media codecs and codec parameters, and possible network candidate paths for the two applications to connect with each other for live streaming.

Q: How do applications use a signaling channel to enable peer-to-peer connectivity?

Streaming applications can maintain persistent connectivity with a signaling channel and wait for other applications to connect to them or they can connect to a signaling channel only when they need to live stream media. The signaling channel enables applications to connect with each other in a one to few model using the concept of one master connecting to multiple viewers. The application that initiates the connection assumes the responsibility of a master via the ConnectAsMaster API and wait for viewers. Upto 10 applications can then connect to that signaling channel by assuming the viewer responsibility via the ConnectAsViewer API. Once connected to the signaling channel, the master and viewer applications can send each other signaling messages to establish peer-t0-peer connectivity for live media streaming.

Q: How do applications live stream peer-to-peer media when they are located behind a NAT or a firewall?

Applications use Kinesis Video Streams STUN end point to discover their public IP address when they are located behind a NAT or a firewall. An application provides its public IP address as a possible location where it can receive connection requests from other applications for live streaming. The default option for all WebRTC communication is direct peer-to-peer connectivity but if the NAT or firewall does now allow direct connectivity (e.g. in case of symmetric NATs), applications can connect to the Kinesis Video Streams TURN end points for relaying media via the cloud. The GetIceServerConfig API provides the necessary TURN end point information that applications can use in their WebRTC configuration. This configuration allows applications to use TURN relay as a fallback when they are unable to establish a direct peer-to-peer connection for live streaming.

Q: How does Kinesis Video Streams secure the live media streaming with WebRTC?

End to end encryption is a mandatory feature of WebRTC, and Kinesis Video Streams enforces it on all the components, including signaling and media or data streaming. Regardless of whether the communication is peer-to-peer or relayed via Kinesis Video Streams TURN end points, all WebRTC communications are securely encrypted through standardized encryption protocols. The signaling messages are exchanged using secure Websockets (WSS), data streams are encrypted using Datagram Transport Layer Security (DTLS), and media streams are encrypted using Secure Real-time Transport Protocol (SRTP).

Console

Q: What is the Kinesis Video Streams management console?

The Kinesis Video Streams management console enables you to create, update, manage, and monitor your video streams. It console can also playback your media streams live or on an on-demand basis, as long as the content in the streams is in the supported media type. Using the player controls, you can view the live stream, skip forwards or backwards 10 seconds, use the date and time picker to rewind to a point in the past when you have set the corresponding retention period for the video stream. The Kinesis Video Streams management console's video playback capabilities are offered as a quick diagnostic tool for development and test scenarios for developers as they build solutions using Kinesis Video Streams.

Q: What media type does the console support?

The only supported video media type for playback in the Kinesis Video Streams management console is the popular H.264 format. This media format has wide support on devices, hardware and software encoders and playback engines. While, you can ingest any variety of video, audio, or other custom time-encoded data types for your own consumer applications and use cases, the management console will not perform playback of those other data types.

Q: What is the delay in the playback of video on the Kinesis Video Streams management console?

For a producer that is transmitting video data into the video stream, you will experience a 2 - 10 second lag in the live playback experience in the Kinesis Video Streams management console. The majority of the latency is added by the producer device as it accumulates frames into fragments before it transmits data over the internet. Once the data enters into the Kinesis Video Streams endpoint and you request playback, the console will get H.264 media type fragments from the durable storage, trans-package the fragments into a media format suitable for playback across different internet browsers. The trans-packaged media content will then be transferred to your location where you requested the playback from over the internet.

Encryption

Q: What Is Server-Side Encryption for Kinesis Video Streams?

Server-side encryption is a feature in Kinesis Video Streams that automatically encrypts data before it's at rest by using an AWS KMS customer master key (CMK) that you specify. Data is encrypted before it is written to the Kinesis Video Streams storage layer, and it is decrypted after it is retrieved from storage. As a result, your data is always encrypted at rest within the Kinesis Video Streams service.

Q: How do I get started with server-side encryption?

Server-side encryption is always enabled on Kinesis video streams. If a user-provided key is not specified when the stream is created, the default key (provided by Kinesis Video Streams) is used.

A user-provided AWS KMS master key must be assigned to a Kinesis Video Streams stream when it is created. You can't later assign a different key to a stream using the UpdateStream API.

You can assign a user-provided AWS KMS master key to a Kinesis video stream in two ways: When creating a Kinesis video stream in the console, specify the AWS KMS master key in the Encryption section on the Create new Kinesis Video stream page. Or when creating a Kinesis Video Streams stream using the CreateStream API, specify the key ID in the KmsKeyId parameter.

Q: How much does it cost to use server-side encryption?

When you apply server-side encryption, you are subject to AWS KMS API usage and key costs. Unlike custom AWS KMS master keys, the (Default) aws/kinesis-video customer master key (CMK) is offered free of charge. However, you still pay for the API usage costs that Kinesis Video Streams incurs on your behalf. API usage costs apply for every CMK, including custom ones. Kinesis Video Streams calls AWS KMS approximately every 45 minutes when it is rotating the data key. In a 30-day month, the total cost of AWS KMS API calls that are initiated by a Kinesis Video Streams stream should be less than a few dollars. This cost scales with the number of user credentials that you use on your data producers and consumers because each user credential requires a unique API call to AWS KMS.

Pricing and billing

Q: Is Amazon Kinesis Video Streams available in AWS Free Tier?

No. Amazon Kinesis Video Streams is not available in AWS Free Tier.

Q: How much does Kinesis Video Streams cost?
Kinesis Video Streams uses a simple pay as you go pricing. There is neither upfront cost nor minimum fees and you only pay for the resources you use. Kinesis Video Streams pricing is based on the data volume (GB) ingested, volume of data consumed (GB), and data stored (GB-Month) across all the video streams in your account.

Furthermore, Kinesis Video Streams will only charge for media data it successfully received, with a minimum chunk size of 4 KB. For comparison, a 64 kbps audio sample is 8 KB in size, so the minimum chunk size is set low enough to accommodate the smallest of audio or video streams.

Q: How does Kinesis Video Streams bill for data stored in streams?

Kinesis Video Streams will charge you for total amount of data durably stored under any given stream. The total amount of stored data per video stream can be controlled using retention hours.

Q: How am I charged for using Kinesis Video Streams WebRTC capability?

For using the Amazon Kinesis Video Streams WebRTC capability, you are charged based on the number of signaling channels that are active in a given month, number of signaling messages sent and received, and TURN streaming minutes used for relaying media. A signaling channel is considered active in a month if at any time during the month a device or an application connects to it. TURN streaming minutes are metered in 1 minute increments. Please see the pricing page for more details.

Service Level Agreement

Q: What does the Amazon Kinesis Video Streams SLA guarantee?

Our Amazon Kinesis Video Streams SLA guarantees a Monthly Uptime Percentage of at least 99.9% for Amazon Kinesis Video Streams.

Q: How do I know if I qualify for a SLA Service Credit?

You are eligible for a SLA credit for Amazon Kinesis Video Streams under the Amazon Kinesis Video Streams SLA if more than one Availability Zone in which you are running a task, within the same region has a Monthly Uptime Percentage of less than 99.9% during any monthly billing cycle.

For full details on all of the terms and conditions of the SLA, as well as details on how to submit a claim, please see the Amazon Kinesis Video Streams SLA details page.

AWS Elemental MediaConnect FAQs

What is AWS Elemental MediaConnect?

AWS Elemental MediaConnect is a reliable, secure, and flexible transport service for live video. Using MediaConnect, you can build live video processing workflows and securely share content with your partners and customers, all using the AWS Cloud. MediaConnect takes your high-quality video from one source and sends it to one or more destinations. MediaConnect adds a quality-of-service layer over standard IP transport, enabling uninterrupted live video connections by using packet recovery, and it supports broadcast-grade standards for video encryption. By using the video replication feature, you can distribute your streams to multiple workflows within your own account or share your content with other AWS accounts. MediaConnect gives you visibility into your live video's operational performance with real-time statistics and alerts. MediaConnect can be deployed in minutes, and is integrated with other AWS services such as AWS Elemental MediaLive.

How do I get started using MediaConnect?

To use MediaConnect, you start by creating a flow using either the MediaConnect console or the MediaConnect API. A flow logically connects one video source with many destinations. You specify three details about your flow: 1) the transport protocol to use, 2) the type of encryption, and 3) the destinations. MediaConnect will return an ingest endpoint where you can send your live video as a single unicast transport stream. MediaConnect will then replicate and distribute the stream to every destination specified, whether inside or outside AWS.

Who can use MediaConnect?

MediaConnect is designed for video broadcasters, content owners, and content aggregators who expect a professional-grade solution that encompasses high quality of service, strong encryption, flexible routing, monitoring, and sharing capabilities. Before MediaConnect, customers found it difficult to send high-quality video to the cloud due to reliability, security, and cost concerns. With MediaConnect, content producers can send high-quality video feeds to AWS without sacrificing reliability or security while reducing costs and improving agility compared to satellite and fiber. Video distributors get a reliable way to receive live video in the AWS Cloud. MediaConnect can be used with either Direct Connect (DX) or using transmission over the public internet.

Why should I choose MediaConnect over infrastructure providers such as satellite or fiber?

Satellite and fiber solutions have long lead times, have limited agility, require up-front commitments, and lack the scale to meet demand as workloads change over time. MediaConnect is a pay-as-you-go service with predictable costs where you can get started in minutes, not months. Because MediaConnect is an AWS service, it’s simple to build end-to-end workflows using other AWS services, and you benefit from integrated billing, monitoring, and support. In addition, MediaConnect is more flexible: start, stop, and change your configurations on demand.

Why should I choose MediaConnect over existing solutions to improve IP transport quality of service?

MediaConnect has advantages over other existing solutions to improve IP transport quality of service because MediaConnect a) performs better over long distances, b) supports multiple protocols, c) offers both contribution and business-to-business distribution (on-premises-to-cloud, cloud-to-cloud, and cloud-to-on-premises) workflows, and d) doesn’t require customers to build their own solutions to enable security or sharing.

Is MediaConnect a standalone service or is it dependent on other services?

MediaConnect can function as a standalone service or within a larger video workflow that includes other AWS services such as AWS Media Services. AWS Media Services is a family of services that form the foundation of cloud-based workflows. They offer the capabilities that you need to create, package, and deliver live or on-demand video to consumers around the world, all accessible through the AWS Management Console and APIs.

How does MediaConnect provide reliability for my live video flows?

MediaConnect lets you select from several video-specific, network-level protocols to maintain video quality. These protocols send additional information over the network so that packet loss doesn’t impact live video quality. Specifically, MediaConnect supports both an open standard (the Society of Motion Picture and Television Engineers SMPTE-2022 Part 1 standard for Forward Error Correction for Real-Time Video/Audio Transport Over IP Networks), and a commercial protocol (Zixi), that is commonly used by broadcasters. MediaConnect also manages all the infrastructure for you, and recovers from any issues automatically.

How does MediaConnect provide security for my live video?

MediaConnect gives you the option to apply encryption to your video workflows. It supports trusted security methodologies like AES and Zixi encryption and decryption. Each output can be encrypted with a unique key as it leaves MediaConnect, giving you control over your content’s security. MediaConnect also lets you securely share your content with your customers and partners using entitlements which grant access to your video content.

How does MediaConnect provide operational visibility for my live video?

MediaConnect gives you metrics on both the health of the network carrying your video and the health of your video streams. On the network side, MediaConnect provides packet loss metrics that quantify the end-to-end network performance. MediaConnect will suggest when network conditions indicate that you might want to reduce the bitrate of the input stream to get a higher quality of service, or adjust the dimensions of the error correction parameter at the source of your video flow. Additional alarms (using the TR-101 290 broadcast standard) identify issues with the transport stream timing going across the network end-to-end. This tells you when you might want to investigate settings on your source encoder or network and adjust them to maximize quality.

What is the difference between AWS Elemental MediaConnect and AWS Ground Station?

AWS Elemental MediaConnect is a reliable, secure, and flexible transport service for live video that enables broadcasters and content owners to cost-effectively build video workflows and securely share live content with partners and customers. Whether you manage live 24x7 channels, stream live events, or need disaster recovery capabilities, MediaConnect gives you the tools to build your applications in the AWS Cloud.

AWS Ground Station enables you to control and ingest data from orbiting satellites without having to buy or build satellite ground station infrastructure. AWS Ground Station does this by integrating the ground station equipment like antennas, digitizers, and modems into our AWS Regions around the world. It's as easy as onboarding your satellites and scheduling time to communicate with them. You have the option of conducting all of your satellite operations on the AWS Cloud, including the storing and processing of your satellite data and delivering your products using AWS services, or use AWS Ground Station just to downlink your satellite data and transport it to your own processing center.

Do I need to supply my own on-premises encoder to use MediaConnect?

MediaConnect ingests content from on-premises locations or from the cloud. For live video sources that originate on premises, such as from a broadcast center or an event location, a customer-supplied encoder is required to create the contribution video.

What is contribution video? What bitrates and formats can I use with MediaConnect?

Contribution video is lightly compressed, high-quality video, often referred to as mezzanine video, originating from a contribution encoder or playout system. MediaConnect supports the AVC (Advanced Video Codec) and HEVC (High Efficiency Video Codec) compression formats, and bitrates up to 80 Mbps (megabits per second).

Calculate your costs

Learn more about pricing.

Learn more »

Sign up for an account

Instantly get access to AWS Elemental MediaConnect.

Sign up »

Start building in the console

Get started building with AWS Elemental MediaConnect in the AWS Console.

Get started »

AWS Elemental MediaConvert FAQs

What is AWS Elemental MediaConvert?

AWS Elemental MediaConvert is a file-based video processing service that formats and compresses offline content for delivery to televisions or connected devices. With AWS Elemental MediaConvert's high-quality video transcoding, you can create on-demand video assets for playback on virtually any device. The service combines advanced video and audio capabilities with a simple web services interface and pay-as-you-go pricing. With AWS Elemental MediaConvert, you can focus on delivering compelling media experiences without having to worry about the complexity of building and operating your own video processing infrastructure.

What is file-based video processing?

In a video processing workflow, a file-based video transcoding solution processes video files, creating compressed versions of the original content to reduce its size, change its format, or increase playback device compatibility. A file-based video transcoding solution can convert any video input source, ranging from high-quality studio masters to videos captured on a mobile device, and produce content ready for distribution to viewers. File-based transcoding processes media as fast as the underlying infrastructure allows, which may be faster or slower than real-time depending on the input type, output format(s), and transcoder settings. File-based transcoding solutions are also expected to handle packaging and protection of content by preparing assets for on-demand delivery to multiple device types with integrated digital rights management (DRM).

Who can use AWS Elemental MediaConvert?

AWS Elemental MediaConvert is designed to meet the needs of all types of video providers. Video is increasingly important to companies from small businesses to global enterprises, as well as government agencies, nonprofit organizations, and schools, all of which can use AWS Elemental MediaConvert to improve the efficiency and effectiveness of their video operations. Companies in the media and entertainment industry, including film and TV studios, broadcast networks, pay TV channels and system operators, programming distributors, internet service providers, online video platforms, and professional sports leagues and teams, can all benefit from AWS Elemental MediaConvert as part of their broadcast, streaming, and over-the-top video offerings.

Why should I choose AWS Elemental MediaConvert over an on-premises file-based encoding solution?

On-premises solutions have long lead times to become production-ready, have limited agility, and lack the scale to meet demand as workloads change over time. These solutions require customers to own, host, and manage all of the required infrastructure, which equates to significant capital investment up front, and may require interfacing with multiple vendors for billing and support, which is prohibitive for individuals and small companies. Long term, the total cost of ownership only increases when factoring in ongoing expense to update, patch, and maintain on-premises systems. AWS Elemental MediaConvert, on the other hand, is a pay-as-you-go service with predictable costs that offers a high degree of availability, reliability, and scalability without the burden of infrastructure management. Because AWS Elemental

MediaConvert is an AWS service, it’s simple for video providers to build end-to-end workflows using other AWS services, and benefit from integrated billing, monitoring, and support.

Why should I choose AWS Elemental MediaConvert over other file-based cloud encoding services?
AWS Elemental MediaConvert gives you access to a full broadcast-grade feature set. Other cloud-based video processing services often do not provide the premium capabilities that content owners and distributors require. With AWS Elemental MediaConvert you can take advantage of features to create broadcast-quality video including static graphic overlays, audio loudness normalization, ad insertion via SCTE-35 support, manifest decoration, DRM integration, and broadcast and OTT closed captioning. AWS Elemental MediaConvert leverages video processing technology built from the ground up by AWS Elemental and used by video providers of all kinds today. It includes ingest and output support for the highest quality codecs (MPEG-2, AVC, HEVC including support for 10-bit 4:2:2 color sampling), extensive adaptive bitrate packaging formats (CMAF, HLS, DASH, and MSS), and processing and conversion of HDR content (HDR 10 and HLG BT.2020). With AWS Elemental MediaConvert, you can get started with simple jobs while gaining access to a deep set of configurable parameters to precisely control video output quality as required. AWS Elemental MediaConvert charges jobs based on the duration and the characteristics of the outputs generated, which means that you always only pay for what you use. Finally, AWS Elemental MediaConvert scales elastically with the rate of incoming jobs so you benefit from short turnaround times even when your incoming load varies.

Is AWS Elemental MediaConvert a standalone service or is it dependent on other services?
AWS Elemental MediaConvert can function as a standalone service or within a larger video workflow that includes other AWS Media Services.

AWS Media Services are a family of services that form the foundation of cloud-based video workflows, which offer customers the capabilities they need to create, package, and deliver video, all accessible through the AWS Management Console and APIs. Working together with other AWS services, AWS Media Services offer a complete solution for processing and delivery of live or on-demand video content to consumers around the world, cost efficiently and with high quality.

Why should I choose AWS Elemental MediaConvert over Amazon Elastic Transcoder?

AWS Elemental MediaConvert should be the first option you consider for file-based video processing. MediaConvert provides a comprehensive suite of transcoding features that addresses the needs of the majority of use cases. It is optimized to reduce turnaround time and improve scalability which allows you to process more files in parallel. And you benefit from a flexible pricing structure where you only pay for functionality that you use. However, if you need to create WebM video, MP3 audio, or animated GIF files you will need to use Amazon Elastic Transcoder. New capabilities will continue to be added to MediaConvert and these features may be supported in future updates.

Can I use AWS Elemental MediaConvert for live video?

AWS Elemental MediaConvert is a file-based transcoding service and does not support live video. For encoding live video for broadcast and streaming to any device, you should use AWS Elemental MediaLive.

AWS Elemental MediaLive FAQs

What is AWS Elemental MediaLive?
AWS Elemental MediaLive is a cloud-based live video encoding service that offers fast, reliable and easy-to-use delivery of high-quality live video streams without the need to manage infrastructure. AWS Elemental MediaLive streamlines live video operations by automating the configuration and management of ingest and encoding components for highly reliable delivery of live streams. The service provides broadcast quality features, configurable capability and support for industry standard formats and technologies. Combining broadcast-grade encoding capabilities with the scale and elasticity of AWS, video providers can efficiently deliver live stream to their audiences and focus on their content and differentiated viewing experiences.

What is real-time live video encoding?
In a video processing workflow, an encoder compresses a video stream—taking high-quality video as an input, and outputting smaller-sized versions—with as little loss as possible to the resulting picture quality. While this is a complicated task when working with pre-recorded video files, for live video it is made even more difficult as the video processing needs to work in real-time: the encoder must be powerful enough to

produce exactly one second of video every second it runs without fail so that viewers see an uninterrupted video stream.

Who can use AWS Elemental MediaLive?

AWS Elemental MediaLive is designed to meet the needs of all types of video providers. Video is increasingly important to companies from small businesses to global enterprises, as well as government agencies, nonprofit organizations, and schools, all of which can use AWS Elemental MediaLive to improve the efficiency and effectiveness of their video operations. Companies in the media and entertainment industry, including film and TV studios, broadcast networks, pay TV channels and system operators, programming distributors, internet service providers, online video platforms, and professional sports leagues and teams, can all benefit from AWS Elemental MediaLive as part of their broadcast, streaming, and over-the-top video offerings.

Why should I choose AWS Elemental MediaLive over an on-premises live encoding solution?

AWS Elemental MediaLive provides a secure, flexible and highly available live encoding solution. It enables push button deployment of live channels and handles resource provisioning, service orchestration, scaling, healing, resiliency failover, monitoring, and reporting. AWS Elemental MediaLive enables users to set up a channel in minutes and does all the heavy lifting behind the scenes to provision and start the various resources required.

Why should I choose AWS Elemental MediaLive over other cloud-based live encoding services?

AWS Elemental MediaLive gives you access to a broadcast-grade feature sets, complete control over encoding settings, and is built with industry-leading technology for encoding that supports standard codecs (MPEG-2, AVC, HEVC, etc.), resolutions (SD, HD), and broadcast features (ad insertion using the SCTE35 standard, captioning, audio descriptors, loudness correction, etc.). AWS Elemental MediaLive is easy to use and scales with AWS resources globally.

Does AWS Elemental MediaLive support statistical multiplexing (statmux)?
Yes. Statmux for MediaLive enables broadcasters and content owners to implement flexible and scalable workflows in AWS, generating content for distribution to headends via traditional broadcast methods. Content owners and broadcasters can use MediaLive to support distribution systems that rely on statistical multiplexing. Combined with the advanced video encoding features and built-in resiliency of MediaLive, Statmux extracts more bandwidth capacity from the network, ensures reliable 24/7 operations, and reduces total cost of ownership for linear video delivery when deploying hundreds of channels.

Is AWS Elemental MediaLive a standalone service or is it dependent on other services?
AWS Elemental MediaLive can work as a standalone service or within a larger video workflow that includes other AWS Media Services. AWS Media Services are a family of services that form the foundation of cloud-based video workflows, which offer customers the capabilities they need to create, package, and deliver video, all accessible through the AWS Management Console and APIs. Working together with other AWS services, AWS Media Services offer a complete solution for processing and delivery of live or on-demand video content to consumers around the world, cost efficiently and with high quality.

How does AWS Elemental MediaLive ensure the security of channels in progress?
AWS Elemental MediaLive automatically protects video content as it moves between components by natively employing AWS security capabilities. The service uses customer identity and access management (IAM) roles and security groups within their own AWS environments.

You can also add input security groups to whitelist IP addresses for input types that push content to the service.

How does AWS Elemental MediaLive ensure reliability of service?
AWS Elemental MediaLive automates the provisioning, configuration and management of AWS-based resources for ingesting and encoding live video. The service gives customers highly reliable live video encoding without managing infrastructure. When you create a channel in AWS Elemental MediaLive, the service deploys redundant infrastructure in two AWS availability zones (AZs). Each component is monitored for health and the service detects any degraded resources and replaces them with new ones.

How is AWS Elemental MediaLive billed?
Pricing is based on a straightforward per-minute model that simplifies budgeting and allows users to forecast exactly what they will spend on each channel. The pricing scales as more inputs/outputs are selected and is pay-as-you-go based on the codec (MPEG-2, AVC, HEVC); resolution (SD, HD, UHD), bitrate (less than 10Mbps, between 10 and 20Mbps, and over 20Mbps); and frame-rate (less than 30fps, between 30 and 60 fps, and over 60fps) you use. There are no minimum commitments or long-term contracts. In addition to on demand pricing, there is also a monthly option with an annual commitment for 24x7 channels. Visit the AWS Elemental MediaLive Pricing page for more information.

What are the resolutions that correspond to SD, HD, and UHD?
SD is less than 1280x720; HD is equal to, or greater than 1280x720, up to and including 1920x1080, and UHD is greater than 1920x1080; up to 4096x2160.

What options does AWS Elemental MediaLive support for ingesting live video?

AWS Elemental MediaLive will accept video using any of the following standards: RTP with forward error correction (FEC); RTMP (in a push or pull mode); and HLS. The supported codecs are MPEG-2, h.264/AVC, and h.265/HEVC. The service can also be paired with AWS Elemental Live encoding appliances that can be used on-premises at production facilities or remote event venues to ingest content for delivery as an input to your channels. These appliances support a range of compressed and uncompressed live input sources, including SDI, HDMI, ASI, MPEG-TS over IP, and SDI over IP.

What options does AWS Elemental MediaLive support for outputting live video?

AWS Elemental MediaLive supports HLS, RTP, RTMP/S, and Microsoft Smooth Streaming (MSS) streaming outputs. It will also archive to file. AWS Elemental MediaLive can deliver to an origin and just-in-time packaging service like AWS Elemental MediaPackage or a 3rd party packager, and it can deliver to AWS Elemental MediaStore for simple origination.

How do I use AWS Elemental MediaLive with my other workflow vendors?

AWS Elemental MediaLive provides a robust and flexible API that ecosystem partners such as content management systems (CMS) and packagers can integrate with. AWS Elemental MediaLive runs within customers' AWS accounts, so other services, components or software that runs or interfaces with customer-maintained VPCs are also easily incorporated in video workflows.

How do I use AWS Elemental MediaLive with AWS Elemental MediaPackage?

AWS Elemental MediaLive can work with a range of just-in-time packing products including AWS Elemental MediaPackage. To configure a AWS Elemental MediaLive channel with AWS Elemental Mediapackage, simply create a channel with AWS Elemental MediaPackage to get a destination address, then select HLS WebDAV as the output for your AWS Elemental Media Live channel profile and add the destination address. With AWS Elemental MediaPackage you can create output groups for multiple delivery protocols like HLS and DASH, add DRM and content protection, and a live archive window for DVR-like features.

Which DRM providers does AWS Elemental MediaLive support?
AWS Elemental MediaLive does not support DRM providers on its own. It can be used with AWS Elemental MediaPackage and a published DRM API to work with several DRM providers. This way, a combination of AWS Elemental MediaLive and AWS Elemental MediaPackage can efficiently encrypt and protect multiple channels using multiple DRM standards and multiple DRM providers.

AWS Elemental MediaPackage FAQs

What is AWS Elemental MediaPackage?
AWS Elemental MediaPackage is a highly scalable, video origination and just-in-time packaging service that helps video providers securely, reliably and cost-efficiently package and deliver live video streams. Video providers can improve the viewing experience, easily integrate advanced, broadcast-grade capabilities, increase workflow resiliency, and better protect and monetize their multiscreen content. The service uses just-in-time packaging to cost effectively output multiple standards-based streaming protocols and DRM types in different combinations to support an array of multiscreen devices. It supports consistent quality of service by elastically scaling to meet demand and manage failover within a highly available managed service. AWS Elemental MediaPackage does not limit customers' choices of video players, CDNs or ad providers, and works seamlessly with other AWS services to build a solution for high-quality, resilient live streaming for 24/7 channels or live events.

What is just-in-time packaging and origination?
In a video processing workflow, a just-in-time packaging and origination product customizes live video streams or VOD assets for delivery in a format compatible with the device making the request. An

advanced origin is used to convert incoming content on-the-fly from a single format to multiple delivery formats while applying DRM standards, allowing it to serve streaming video content in response to requests from users to devices such as tablets, smartphones, connected TVs, or set-top boxes.

Who can use AWS Elemental MediaPackage?

AWS Elemental MediaPackage is designed to meet the needs of all types of video providers. Video is increasingly important to companies from small businesses to global enterprises, as well as government agencies, nonprofit organizations, and schools, all of which can use AWS Elemental MediaPackage to improve the efficiency and effectiveness of their video operations. Companies in the media and entertainment industry, including film and TV studios, broadcast networks, pay TV channels and system operators, programming distributors, internet service providers, online video platforms, and professional sports leagues and teams, can all benefit from AWS Elemental MediaPackage as part of their broadcast, streaming, and over-the-top video offerings.

How does AWS Elemental MediaPackage compare to Cloud Streaming Services?

AWS Elemental MediaPackage is available in multiple regions, while with other cloud services, some options are only available in a subset of all the AWS regions on which their service is deployed. AWS Elemental MediaPackage can record all the renditions in an adaptive bitrate (ABR) stream and supports up to 4K resolution with high frame rate using HEVC, while other cloud services restrict recordings to only the highest bitrate. Competing services may also provide a broad set of limitations on the delivered content, including limiting resolution to 1080p and frame rate to 30 frames per second. AWS Elemental MediaPackage also supports a wide range of OTT ABR standards and provides standards-based subtitles for all OTT formats, which is broader support than other cloud services.

How does AWS Elemental MediaPackage compare to third-party packaging software running on AWS?
AWS Elemental MediaPackage packages and archives in the same workflow using the same formats that are streamed; with other providers, you need a separate workflow. AWS Elemental MediaPackage provides more control over what is exposed to subscribers, helping restrict access. In addition, AWS Elemental MediaPackage offers an embedded redundancy model, removing the complexity of implementing and managing instances from the customer, and separates ingest from egress for even greater scalability and redundancy. It is natively designed as a service, not as software running on a virtual server, and customers don't have to monitor the load on their EC2 instances or manually scale to accommodate more channels and more end-user connections.

How does AWS Elemental MediaPackage compare to packaging services from Content Delivery Networks (CDNs)?
AWS Elemental MediaPackage offers highly customizable packaging options and parameters, providing flexibility for streaming protocols, segment sizes, manifest manipulation, subtitles, and other metadata handling along with broad DRM support. With some CDNs, the options for packaging are limited to a given set of parameters. Unlike CDN-based packaging services that only work with their specific CDN, AWS Elemental MediaPackage is CDN agnostic, offering customers an easy way to implement a multi-CDN strategy and improve audience quality of service using third-party tools.

Can I use a CDN other than Amazon CloudFront with AWS Elemental MediaPackage?
Yes. A customer can connect any CDN that delivers content in pull mode as a CDN output from AWS Elemental MediaPackage. Using Amazon CloudFront provides the benefit of staying in the AWS Cloud, and saves on data transfer rates compared to external 3rd-party CDNs. However, AWS Elemental MediaPackage is designed to work with Amazon CloudFront and non-Amazon CDNs, giving customers the ability to run multi-CDN or hybrid-CDN strategies.

Is AWS Elemental MediaPackage a standalone service or is it dependent on other services?
AWS Elemental MediaPackage can function as a standalone service or within a larger video workflow that includes other AWS Media Services. AWS Media Services are a family of services that form the foundation of cloud-based video workflows, which offer customers the capabilities they need to create, package, and deliver video, all accessible through the AWS Management Console and APIs. Working together with other AWS services, AWS Media Services offer a complete solution for processing and delivery of live or on-demand video content to consumers around the world, cost efficiently and with high quality.

Does AWS Elemental MediaPackage work with AWS Elemental MediaLive?
AWS Elemental MediaLive is deeply integrated with AWS Elemental MediaPackage so customers can easily combine live encoding with content origination, dynamic packaging, and live-to-VOD capabilities. To configure an AWS Elemental MediaLive channel with AWS Elemental MediaPackage, simply create a channel with AWS Elemental MediaPackage to get a destination address, then select HLS WebDAV as the output for your AWS Elemental MediaLive channel profile and add the destination address. With AWS Elemental MediaPackage you can

create output groups for multiple delivery protocols like HLS and DASH, add DRM and content protection, and a live archive window for DVR-like features.

What is the difference between AWS Elemental MediaPackage and AWS Elemental MediaStore?
AWS Elemental MediaPackage provide just-in-time package and DVR-like features as well as origination for live video streams. If a customer does not require packaging to different or multiple formats, DRM, or DVR-like features, customers can use AWS Elemental MediaStore as a pass-through video origination and storage service that offers the high performance and immediate consistency required for delivering media combined with the security and durability that AWS offers across its services.

Does AWS Elemental MediaPackage work with on-premises AWS Elemental Live encoders, or the AWS Elemental Cloud platform?
AWS Elemental MediaPackage supports HLS as an input over HTTPS. On-premises AWS Elemental customers can benefit from the improved scalability and resiliency of AWS Elemental MediaPackage as compared to an on-premises origin, even if they run encoding on site. If you use AWS Elemental Live appliances on-premises, you can use the authenticated WebDAV HLS output to feed AWS Elemental MediaPackage. This approach also works with AWS Elemental Live in the Elemental Cloud platform.

How can I connect AWS Elemental MediaPackage to a DRM Key Management System?
AWS Elemental MediaPackage has a published DRM API based on the Content Protection Information Exchange (CPIX) standard that makes

integrating with DRM Key providers easier. Many providers, like Verimatrix, Irdeto, BuyDRM, castLabs EZDRM, and Conax have already implemented the API, with others coming on board soon.

AWS Elemental MediaStore FAQs

What is AWS Elemental MediaStore?
AWS Elemental MediaStore is a media origin and storage service that offers the performance, predictable low latency, and consistency required for delivery and processing workloads like live streaming video. The service provides a write-behind cache, designed for performance, in front of object storage. It is an inexpensive method for pass-through and low-latency segmented content delivery, with predictable pay-as-you-go pricing.

Who can use AWS Elemental MediaStore?
AWS Elemental MediaStore is designed to meet the needs of all types of content providers. Video is increasingly important to companies from small businesses to global enterprises, as well as government agencies, nonprofit organizations, and schools, all of which can use AWS Elemental MediaStore to improve the efficiency and effectiveness of their video operations. Companies in the media and entertainment industry, including film and TV studios, broadcast networks, pay TV channels and system operators, programming distributors, internet service providers, online video platforms, and professional sports leagues and teams, can all benefit from AWS Elemental MediaStore as part of their broadcast, streaming, and over-the-top video offerings.

Why should I use AWS Elemental MediaStore instead of managing storage infrastructure myself?

AWS Elemental MediaStore provides the performance required for demanding content processing and delivery workloads without needing to worry about building custom infrastructure or implementing best practices in object storage, so customers can focus on their workflow rather than worrying about their storage infrastructure.

How does AWS Elemental MediaStore improve performance?

When you write content to AWS Elemental MediaStore, it is automatically held in a replicated cache for the first few minutes after creation, and again after each update. This replicated cache gives performance, predictable low latency, and consistency, even with the high request loads and with the frequent updates common with files like streaming video manifests during live video streams.

What is the difference between AWS Elemental MediaStore and AWS Elemental MediaPackage?

AWS Elemental MediaPackage provides just-in-time packaging and live-to-VOD features as well as origination for live streams. If multiple formats and DRMs are required, or DVR-like features, you can use AWS Elemental MediaPackage. If the live streams are already in the correct formats and have any required DRM applied, you can use AWS Elemental MediaStore as a pass-through video origination and storage service that offers the performance and consistency required for delivering live streaming media combined with the security and durability AWS offers across its services.

Is AWS Elemental MediaStore a standalone service or is it dependent on other services?

AWS Elemental MediaStore can function as a standalone service or within a larger video workflow that includes other AWS Media Services. AWS Media Services are a family of services that form the foundation of cloud-based video workflows, which offer customers the capabilities they need to create, package, and deliver video, all accessible through the AWS Management Console and APIs. Working together with other AWS services, AWS Media Services offer a complete solution for processing and delivery of live or on-demand video content to consumers around the world, cost efficiently and with high quality.

How does AWS Elemental MediaStore work with AWS Elemental MediaLive?

AWS Elemental MediaLive is a video service that allows easy and reliable creation of live outputs for broadcast and streaming delivery at scale. An AWS Elemental MediaStore container can be selected as the output destination for an AWS Elemental MediaLive channel.

How is AWS Elemental MediaStore billed?

With AWS Elemental MediaStore, you are charged a per GB Media Ingest Optimization Fee when content enters the service and charged per GB price for content Storage (per month) for content that you keep in the service.. Request costs are based on the request type, and are charged on the quantity of requests. Visit the AWS Elemental MediaStore Pricing page for more information.

Which media workflow use cases are best suited to AWS Elemental MediaStore?

Serving live adaptive bit-rate video streams that require a HTTP origin are ideal use cases for AWS Elemental MediaStore. With predictable low latency and performance along with immediate read-after-write and read-after-update consistency AWS Elemental MediaStore is optimized to originate fragmented video and ensures the latest versions of manifests for live video streams, which are constantly updated, yet retain the same name, are always the ones that get delivered to players. And requests for video segments are served quickly and reliably providing a better quality of experience with less buffering.

Can on-premises workflows benefit from AWS Elemental MediaStore?
Yes, On-premises encoders can write to AWS Elemental MediaStore and get the benefits of the high performance media origination.

What is the durability of content written to AWS Elemental MediaStore?
Objects ingested into AWS Elemental MediaStore are sent to a replicated write-behind cache that transitions objects to storage backed by Amazon S3 shortly after they are written. While objects are normally transitioned quickly to Amazon S3, there is some chance that this process could be delayed or not occur at all. If your workload requires immediate durability, we recommend using Amazon S3 directly.

AWS Elemental MediaTailor FAQs

What is AWS Elemental MediaTailor?
AWS Elemental MediaTailor is a content personalization and monetization service that allows video providers to serve targeted ads to end users while maintaining broadcast quality-of-service in multiscreen video applications. With AWS Elemental MediaTailor, advertising is inserted upstream before video delivery, on the server side; so a continuous stream arrives at the consumer device, eliminating the possibility of discrimination between content and commercials. Ads are better monetized, maintain video quality that consistently matches the primary content, and are simpler to manage across multiplatform environments. AWS Elemental MediaTailor offers managed transcoding of ad content to provide the best end-user experience while also enabling standards-based client- and server-side ad reporting within a single service. Customers are given full control over the player, origin, and CDN while providing the highest-quality viewer experience.

Who can use AWS Elemental MediaTailor?
AWS Elemental MediaTailor is designed to meet the needs of all types of video providers. Video is increasingly important to companies from small businesses to global enterprises, as well as government agencies, nonprofit organizations, and schools, all of which can use AWS

Elemental MediaTailor to improve the efficiency and effectiveness of their video operations. Companies in the media and entertainment industry, including film and TV studios, broadcast networks, pay TV channels and system operators, programming distributors, internet service providers, online video platforms, and professional sports leagues and teams, can all benefit from AWS Elemental MediaTailor as part of their broadcast, streaming, and over-the-top video offerings.

What challenges does AWS Elemental MediaTailor solve?

Traditional approaches to ad insertion for OTT video can limit the ability of publishers to fully monetize their content in a number of ways. They fail to scale to peaks in live streaming, deliver inconsistent ad reporting metrics, or limit customers to specific CDNs or origin server vendors. Often, by pulling advertising from different networks, client-side solutions fall short of providing the broadcast-grade quality of service that users expect of their viewing experience; moreover, they lose a significant percentage of advertising impressions to ad blocking. AWS Elemental MediaTailor addresses the challenges of personalizing advertising for viewers and accurately reporting ad impressions across devices while mitigating ad blockers, allowing publishers to maximize the monetization potential of their OTT video. It does so without compromising video quality and while scaling to meet the demands of high profile live streaming use cases. AWS Elemental MediaTailor also maintains flexibility for customers as it is agnostic to HTTP CDNs and origin server providers and offers a simple pay-as-you-go pricing model with no up-front costs or commitments.

Why should I use AWS Elemental MediaTailor over other server-side ad insertion solutions?

Other server-side ad insertion solutions typically do not provide detailed client-side viewing metrics. Server-side solutions generally report on CDN server logs of requests to ad servers, which doesn't offer the granularity of client-based viewing metrics that advertisers require. Other solutions may require SDK or specific player integration to handle server-side stitched manifests. In contrast, AWS Elemental MediaTailor does not require specific player or SDK integration to work. In addition, AWS Elemental MediaTailor makes callbacks to a common endpoint for both content and ads rather than known ad serving entities, bypassing ad blocking strategies.

AWS Elemental MediaTailor uses client request information in real-time to communicate with ad decision servers and dynamically generates personalized manifests and ad content. And with AWS Elemental MediaTailor, there is no need for customers to scale origin infrastructure to cope with delivering personalized manifests.

Why should I use AWS Elemental MediaTailor over other client-side ad insertion solutions?

Client-side ad insertion solutions are susceptible to ad blocking and can deliver poor playback quality. As most client-side ad blockers work by blacklisting known ad serving domain names, video clients can end up skipping ad segments entirely, jeopardizing business models that rely on ad revenue from internet-delivered video offerings. AWS Elemental MediaTailor delivers ads in a way that makes ads indistinguishable from content for a seamless viewing experience.

Video providers want viewers to experience the same playback quality as traditional broadcast TV. However, since ads are served by external ad decision services with their own encoders, video processing pipeline, and CDN, it's impossible for client-side ad insertion solutions to ensure ads match the format of the content. This results in increased rebuffer rates and discontinuous transitions from content to ads and back as clients retrieve ads on a best-effort basis.

In other cases, customers wish to ensure that advertisements are compliant with the latest technology regulations such as audio loudness levels. Providers try to solve these issues by preprocessing and preloading ads while working with numerous ad decision services or directly with advertisers, resulting in scaling challenges as the number of ad decision servers and permutations of ad formats increases. AWS Elemental MediaTailor overcomes each of these challenges by pulling down mezzanine-quality assets from ad decision servers and transcoding assets on the fly to the same specifications as the primary content stream. As a result, viewers enjoy the same seamless experience from internet-delivered video as traditional broadcast TV.

How does AWS Elemental MediaTailor simplify my advertising workflow?

Since the same video processing pipeline is used for content and for ads, there is no need to orchestrate complicated ad signaling between the origin server, manifest manipulation service, and ad decision server. In the past, customers had to implement client library changes across all of their supported devices as new video formats, ad insertion specs, or compliance standards emerged. With AWS Elemental MediaTailor, customers can use the same ad insertion workflow to reach all devices, eliminating the need to make custom changes across all client applications to account for new standards.

How does AWS Elemental MediaTailor enable content personalization?

Video content publishers have more information in OTT environments about their end viewer’s demographic profile, viewing habits, and other relevant data than traditional broadcast TV. This allows for increased rates for advertising spots. However, traditional server-side ad insertion solutions don’t typically communicate with and use client device information when stitching together ads and content, for two reasons:

one, they rely on proprietary extensions to pass through client information to ad servers, resulting in communications failures between different implementations and services, and two, ad personalization not only decreases the ability to cache the manifest file, but also greatly increases compute resources required for rewriting manifest files and transcoding unique ad content on-the-fly. Unlike those solutions, AWS Elemental MediaTailor uses the VAST standard when interfacing with external ad servers, while scaling compute resources for manifest manipulation, transcoding, and delivery occurs automatically using the elastic AWS Cloud.

How does AWS Elemental MediaTailor enable accurate reporting of viewing behavior across devices?

Both advertisers and the Interactive Advertising Bureau (IAB) call for granular playback metrics that are measured from end-viewer devices. This requires clients to make numerous HTTP requests to tracking URLs managed by external ad servers as ads are played back. This beacon information allows impressions to be rewarded based on quartile (25 percent, 50 percent, 75 percent) of an ad video that has been played.

By default, AWS Elemental MediaTailor reports these metrics from the server-side without additional integration efforts required. AWS Elemental MediaTailor also provides a client API endpoint to identify when ad content is playing and can be used to implement client-side ad reporting as well as advanced player features to stop scrubbing during ad break, or ad duration countdowns.

How does AWS Elemental MediaTailor improve my viewers' experience when watching my content?

AWS Elemental MediaTailor has a transcode service that works to ensure there are no jarring discontinuities in aspect ratio, resolutions, and video bitrate for transitions between ads and content during playback. AWS Elemental MediaTailor uses standard VAST and VMAP responses from ad servers to pull down a high-quality version of the ad asset and provisions real-time transcoding and packaging resources to format it to the same video and audio parameters as your content.

Is AWS Elemental MediaTailor a standalone service or is it dependent on other services?
AWS Elemental MediaTailor can function as a standalone service or within a larger video workflow that includes other AWS Media Services. AWS Media Services are a family of services that form the foundation of cloud-based video workflows, which offer customers the capabilities they need to create, package, and deliver video, all accessible through the AWS Management Console. Working together with other AWS services, AWS Media Services offer a complete solution for processing and delivery of live or on-demand video content to consumers around the world, cost efficiently and with high quality.

Can I use AWS Elemental MediaTailor with on-premises deployments?
AWS Elemental MediaTailor's manifest manipulation and other other services run in AWS, but can access any origin server hosted on-premises and accesible over HTTP. AWS Elemental MediaTailor itself is an AWS cloud service.

Can I use my own CDN with AWS Elemental MediaTailor?
Yes, AWS Elemental MediaTailor is CDN agnostic.

Can I use my own origin server for AWS Elemental MediaTailor?

Yes, AWS Elemental MediaTailor works with origin servers that are accessible over HTTP and can produce manifests decorated with CUE-IN and CUE-OUT ad markers. Visit AWS Elemental MediaTailor Documentation pages for more information.

How is AWS Elemental MediaTailor billed?

Pricing for AWS Elemental MediaTailor is based on the number of ads inserted. If 1000 people are viewing a stream and there are 5 ads in an ad break, you would be charged for 5000 ad insertions. 10 free ad creative transcodes are included per 1000 ad insertions. Visit the AWS Elemental MediaTailor Pricing page for more information.

How many channels do I need to use AWS Elemental MediaTailor?

There is no minimum or limit to the number of channels supported by AWS Elemental MediaTailor. As pricing is pay as you go, with no minimum commitment and based on ad insertion, there is no penalty of running just one channel. Equally, if there are hundreds of channels or channels with huge expected peaks in concurrent views, AWS Elemental MediaTailor will automatically scale.

AWS Migration Hub FAQs

General

Q: What is AWS Migration Hub? >>

Q: Why should I use AWS Migration Hub? >>

Q: What migration tools integrate with AWS Migration Hub? >>

Q: How does AWS Migration Hub help me track the progress of my application migrations? >>

Q: How does AWS Migration Hub help me understand my IT environment? >>

Q: How much does it cost to use AWS Migration Hub? >>

Getting started

Q: How do I get started with AWS Migration Hub? >>

Q: What is the Migration Hub home region? >>

Q: What regions can I migrate to using AWS Migration Hub? >>

Q: How is access granted to AWS Migration Hub? >>

Discovering servers and grouping applications

Q: How do I view my IT portfolio in AWS Migration Hub? >>

Q: How do I add resources into the Discovery Repository? >>

Q: How do I group servers into an application? >>

Q: How do I view applications? >>

Q: Can I see applications created by other users within the same account? >>

Q: Can I see applications that exist in other AWS accounts? >>

Importing servers and applications

Q: How does the AWS Migration Hub import feature work? >>

Q: What kind of data can I import using the import template? >>

Q: I imported an incorrect file. Can I overwrite or delete it? >>

Q: Is there a limit on the number of import files that I can upload? >>

Q: Do I need to pay for importing data? >>

Q: I do not have data for all the fields in the import template. Can I still import my data? >>

Q: What are the criteria for identifying an incorrect record? >>

Generating EC2 instance recommendations

Q: What is the EC2 instance recommendations feature? >>

Q: Do I need to install AWS Application Discovery Service's Discovery Connector or Discovery Agent to use the EC2 instance recommendations feature? >>

Q: How does the EC2 instance recommendations feature provide a match for a given server? >>

Q: Does the EC2 instance recommendations feature provide recommendations for Burstable Performance Instances? >>

Q: What happens if I have discovery data from multiple sources for the same server in AWS Migration Hub? Which data source is used to calculate the EC2 instance recommendation for that server? >>

Q: Does the EC2 instance recommendations feature recommend current generation instances? >>

Q: When should I use the EC2 instance recommendations feature in AWS Migration Hub versus a more detailed cost assessment with TSO Logic? >>

Tracking migration status

Q: How do I use AWS Migration Hub when migrating applications? >>

Q: Does AWS Migration Hub automatically migrate my applications for me? >>

Q: What do I need to do in order for my application’s migration progress to appear in AWS Migration Hub? >>

Q: What is the experience if I don’t do a strict re-host migration, moving the resources exactly from on-premises to AWS? >>

Q: What if I am using a tool that is not integrated with AWS Migration Hub? >>

Q: How can other tools publish status to AWS Migration Hub? >>

Learn more about product pricing

Review pricing options for AWS Migration Hub.

Learn more »

Sign up for a free account

Instantly get access to the AWS Free Tier.

Sign up »

Start migrating today

Get started migration with AWS Migration Hub on the AWS Management Console.

Sign in »

AWS Application Discovery Service FAQs

General

Q: What is the AWS Application Discovery Service? >>

Q: How does the AWS Application Discovery Service help enterprises migrate to AWS? >>

Q: How does the AWS Application Discovery Service work? >>

Q: How can I get started using the AWS Application Discovery Service? >>

Q: Where is AWS Application Discovery Service available? >>

Q: What is the Migration Hub home region? >>

Q: Should I use agentless or agent-based application discovery? >>

Agent-based discovery

Q: What data does the AWS Application Discovery Agent capture? >>

Q: What operating systems does AWS Application Discovery Service provide agents for? >>

Q: How is the data protected while in transit to AWS? >>

Q: How do I install the AWS Application Discovery Agent in my data center? >>

Q: Will the AWS Application Discovery Service Agent grant AWS remote access to my data center server? >>

Q: Can I run agents in my EC2 instances? >>

Agentless discovery

Q: What does 'agentless' Application Discovery mean? >>

Q: What data does the AWS Application Discovery Agentless Connector capture? >>

Q: What operating systems does the agentless discovery support? >>

Q: How is the data protected while in transit to AWS? >>

Q: How do I install the Application Discovery Agentless Connector in my data center? >>

Q: How can I start the data collection? >>

Q: Will the AWS Application Discovery Agentless Connector grant AWS remote access to my data center servers? >>

Q: Can I run agentless discovery in my EC2 instances? >>

Discovered data

Q: What kind of information is captured by AWS Application Discovery Service? >>

Q: Does this service capture any storage metrics? >>

Q: How often is the information within AWS Application Discovery Service updated? >>

Using the AWS Application Discovery Service data

Q: Can I ingest data into the AWS Application Discovery Service from my existing configuration management database (CMDB)? >>

Q: How do I access the data from this service? >>

Data Exploration in Amazon Athena

Q. What is the Data Exploration feature? >>

Q. What can I see inside Amazon Athena with Data Exploration turned on? >>

Q. Is there cost associated with using Application Discovery Service's Data Exploration feature? >>

Q. Can I enable Data Exploration for some agents only? >>

Q. Do I need to create a special S3 bucket to use this feature? >>

Q. Will the S3 bucket storing my discovery data be secure? >>

AWS Database Migration Service FAQs

Q: Will AWS Database Migration Service help me convert my Oracle PL/SQL and SQL Server T-SQL code to Amazon Aurora or MySQL and PostgreSQL stored procedures? >>

Q: How do I get started with AWS Database Migration Service? >>

Q: In addition to one-time data migration, can I use AWS Database Migration Service for continuous data replication? >>

Q: How are AWS Database Migration Service (DMS) and AWS Schema Conversion Tool (SCT) related? >>

Q: What sources and targets does AWS Database Migration Service support? >>

Q: What sources and targets does AWS Schema Conversion Tool support? >>

Q: Why should I use AWS Database Migration Service instead of my own self-managed replication solution? >>

Q: Can you summarize the database migration steps using AWS Database Migration Service for me? >>

Q: Are these steps different for continuous data replication? >>

Q: Can I monitor the progress of a database migration task? >>

Q: How do I integrate AWS Database Migration Service with other applications? >>

Q: Can I replicate data from encrypted data sources? >>

Q: Does AWS Database Migration Service migrate the database schema for me? >>

Q: Can I use DMS to perform bi-directional replication? >>

Q: How much does DMS cost? >>

AWS DataSync FAQs

General

Q: What is AWS DataSync?

A: AWS DataSync is an online data transfer service that simplifies, automates, and accelerates copying large amounts of data to and from AWS storage services over the internet or AWS Direct Connect. DataSync can copy data between Network File System (NFS) or Server Message Block (SMB) file servers, Amazon Simple Storage Service (Amazon S3) buckets, Amazon Elastic File System (Amazon EFS) file systems, and Amazon FSx for Windows File Server file systems.

Q: Why should I use AWS DataSync?

A: AWS DataSync allows you to move, copy, and synchronize large datasets with millions of files, without having to build custom solutions with open source tools, or license and manage expensive commercial network acceleration software. You can use DataSync to migrate active data to AWS, transfer data to the cloud for analysis and processing, archive data to free up on-premises storage capacity, or replicate data to AWS for business continuity.

Q: What problem does AWS DataSync solve for me?

A: AWS DataSync reduces the complexity and cost of online data transfer, making it simple to transfer datasets between on-premises storage systems, and Amazon S3, Amazon Elastic File System (EFS), and Amazon FSx for Windows File Server. DataSync connects to existing storage systems and data sources with standard storage protocols (NFS or SMB), and uses a purpose-built network protocol and scale-out architecture to accelerate transfer to and from AWS. DataSync automatically scales and handles moving files and objects, scheduling data transfers, monitoring the progress of transfers, encryption, verification of data transfers, and notifying customers of any issues. With

DataSync you pay only for the amount of data copied, with no minimum commitments or upfront fees.

Q: Where can I transfer data to and from?

A: AWS DataSync can transfer data between Network File System (NFS) or Server Message Block (SMB) file servers, Amazon Simple Storage Service (Amazon S3) buckets, Amazon Elastic File System (Amazon EFS) file systems, and Amazon FSx for Windows File Server file systems.

Q: Can I use AWS DataSync to migrate data to AWS?

A: You can use AWS DataSync to migrate from on-premises data to Amazon S3, Amazon EFS, and Amazon FSx for Windows File Server. Configure DataSync to make an initial copy of your entire dataset, and schedule subsequent incremental transfers of changing data until the final cut-over from on-premises to AWS. DataSync includes encryption and integrity validation to help make sure your data arrives securely, intact, and ready to use. To minimize impact on workloads that rely on your network connection, you can schedule your migration to run during off-hours, or limit the amount of network bandwidth that DataSync uses by configuring the built-in bandwidth throttle. Read the storage blog, "Migrating storage with AWS DataSync," to learn more about migration best practices and tips.

Q: How can I use AWS DataSync to archive cold data?

A: You can use AWS DataSync to move cold data from expensive on-premises storage systems directly to durable and secure long-term storage, such as Amazon S3 Glacier or Amazon S3 Glacier Deep Archive. Use DataSync's filtering functionality to exclude copying temporary files and folders, copy only a subset of files from your source location, or split a single file system between multiple destinations. You can select the most cost-effective storage service for your needs: transfer data to any S3 storage class, or use DataSync with EFS Lifecycle Management to store data in Amazon EFS Infrequent Access storage class (EFS IA). Use the built-in task scheduling functionality to regularly archive data that should be retained for compliance or auditing purposes, such as logs, raw footage, or electronic medical records.

Q: How can I use AWS DataSync for recurring transfers between on-premises and AWS for ongoing workflows?

A: You can use AWS DataSync to accelerate and schedule the transfers from on-premises systems into or out of AWS for processing. It can help speed up critical hybrid cloud storage workflows in industries that need to move active files into AWS quickly, including video production in media and entertainment, seismic research in oil and gas, machine learning in life science, and big data analytics in finance. DataSync provides timely delivery to ensure dependent processes are not delayed. You can specify exclude filters, include filters, or both, to determine which files, folders or objects gets transferred each time your task runs.

Q: How can I use AWS DataSync to replicate data to AWS for business continuity?

A: With AWS DataSync, you can periodically replicate files into all Amazon S3 storage classes, or send the data to Amazon EFS or Amazon FSx for Windows File Server for a standby file system. Use the built-in task scheduling functionality to ensure that changes to your dataset are regularly copied to your destination storage.

Q: How can I use AWS DataSync to migrate to Amazon WorkDocs?

A: AWS DataSync is part of the Amazon WorkDocs Migration Service. DataSync makes it easier and faster to migrate home directories and department shares to WorkDocs.

Usage

Q: How do I get started with AWS DataSync?

A: You can transfer data using AWS DataSync with a few clicks in the AWS Management Console or through the AWS Command Line Interface (CLI). To get started, follow these 3 steps:

1. Deploy an agent - Deploy a DataSync agent and associate it to your AWS account via the Management Console or API. The agent will be used to access your NFS server or SMB file share to read data from it or write data to it.

2. Create a data transfer task - Create a task by specifying the location of your data source and destination, and any options you want to use to configure the transfer, such as the desired task schedule.

3. Start the transfer - Start the task and monitor data movement in the console or with Amazon CloudWatch.

Q: How do I deploy an AWS DataSync agent?

A: You deploy an AWS DataSync agent to your VMware ESXi hypervisor or in Amazon EC2. To copy data to or from an on-premises file server, you download the agent virtual machine image (an OVA file) from the AWS Console and deploy to your on-premises VMware ESXi hypervisor. To copy data to or from an in-cloud file server, you create an Amazon EC2 instance from the agent AMI provided in the AWS Console. In both cases the agent must be deployed so that it can access your file server using either the NFS or SMB protocol.

Q: What are the resource requirements for the AWS DataSync agent?

A: You can find the minimum required resources to run the agent here.

Q: How do I start an AWS DataSync data transfer task?

A: AWS DataSync copies data when you initiate a task via the AWS Management Console or AWS Command Line Interface (CLI). Each time a task runs, it scans the source for changes, and performs a copy of any differences between the source to the destination. You can configure which characteristics of the source are used to determine what changed, define filters to include and exclude specific files or folders, and control if files or objects in the destination should be overwritten when changed in the source or deleted when not found in the source.

Q: How does AWS DataSync ensure my data is copied correctly?

A: As AWS DataSync transfers and stores data, it performs integrity checks to ensure the data written to the destination matches the data read from the source. Additionally, an optional verification check can be performed to compare source and destination at the end of the transfer. DataSync will calculate and compare full-file checksums of the data stored in the source and in the destination. You can check either the entire dataset or just the files or objects that DataSync transferred.

Q: How can I monitor the status of data being transferred by AWS DataSync?

A: You can use the AWS Management Console or CLI to monitor the status of data being transferred. Using Amazon CloudWatch Metrics, you can see the number of files and amount of data which has been copied. Amazon CloudWatch Logs are available for detailed error information. In addition, CloudWatch Events are triggered as your tasks transition state, enabling automation of dependent workflows. You can find additional information, such as transfer progress, in the AWS Management Console or CLI.

Q: Can I filter the files and folders that AWS DataSync transfers?

A: Yes. You can specify an exclude filter, an include filter, or both, to limit which files, folders, or objects gets transferred each time a task runs. When creating a task, you configure the file paths or object keys that should always be excluded from being copied. Then, when you start a task, you configure the file paths or object keys that should be included for that execution of the task. If no filters are configured, each time a task runs it will transfer all changes from the source to the destination. Read this AWS storage blog to learn more about using common filters with DataSync.

Q: Can I configure AWS DataSync to transfer on a schedule?

A: Yes. You can schedule your tasks using the AWS DataSync Console or AWS Command Line Interface (CLI), without needing to write and run scripts to manage repeated transfers. Task scheduling automatically runs tasks on the schedule you configure, with hourly, daily, or weekly options provided directly in the Console. This enables you to ensure that changes to your dataset are automatically detected and copied to your destination storage.

Q: Does AWS DataSync preserve the directory structure when copying files?

A: Yes. When transferring files, AWS DataSync creates a director structure on the destination that is similar to the source location's structure.

Q: What happens if an AWS DataSync task is interrupted?

A: If a task is interrupted, for instance, if the network connection goes down or the AWS DataSync agent is restarted, the next run of the task will transfer missing files, and the data will be complete and consistent at the end of this run. Each time a task is started it performs an incremental copy, transferring only the changes from the source to the destination.

Q: Can I use AWS DataSync with AWS Direct Connect?

A: Yes. You can use AWS DataSync with your Direct Connect link to access public service endpoints or private VPC endpoints. When using VPC endpoints, data transferred between the DataSync agent and AWS services does not traverse the public internet or need public IP addresses, increasing the security of data as it is copied over the network.

Q: Does AWS DataSync support VPC endpoints or AWS PrivateLink?

A: Yes. You can use VPC endpoints to ensure data transferred between your AWS DataSync agent, either deployed on-premises or in-cloud, doesn't traverse the public internet or need public IP addresses. Using VPC endpoints increases the security of your data by keeping network traffic within your Amazon Virtual Private Cloud (Amazon VPC). VPC endpoints for DataSync are powered by AWS PrivateLink, a highly available, scalable technology that enables you to privately connect your VPC to supported AWS services.

Q: How do I configure AWS DataSync to use VPC endpoints?

A: To use VPC endpoints with AWS DataSync, you create an AWS PrivateLink interface VPC endpoint for the DataSync service in your chosen VPC, and then choose this endpoint elastic network interface (ENI) when creating your DataSync agent. Your agent will connect to this ENI to activate, and subsequently all data transferred by the agent will remain within your configured VPC. You can use either the AWS DataSync Console, AWS Command Line Interface (CLI), or AWS SDK, to configure VPC endpoints. To learn more, see Using AWS DataSync in a Virtual Private Cloud, and read our AWS storage blog about best practices for migrations using VPC endpoints.

Transferring to and from Amazon S3

Q: Can I copy my data into Amazon S3 Glacier, Amazon S3 Glacier Deep Archive, or other S3 storage classes?

A: Yes. When configuring an S3 bucket for use with AWS DataSync, you can select the S3 storage class that DataSync uses to store objects. DataSync supports storing data directly into S3 Standard, S3 Intelligent-Tiering, S3 Standard-Infrequent Access (S3 Standard-IA), S3 One Zone-Infrequent Access (S3 One Zone-IA), Amazon S3 Glacier (S3 Glacier), and Amazon S3 Glacier Deep Archive (S3 Glacier Deep Archive). More information on Amazon S3 storage classes can be found in the Amazon Simple Storage Service Developer Guide.

Objects smaller than the minimum charge capacity per object will be stored in S3 Standard. For example, folder objects, which are zero-bytes in size and hold only metadata, will be stored in S3 Standard. Read about considerations when working with Amazon S3 storage classes in our documentation, and for more information on minimum charge capacities see Amazon S3 Pricing.

Q: Can I copy data out of S3 Standard-IA and S3 One Zone-IA storage classes?

A: Yes. When using S3 as the source location for an AWS DataSync task, the service will retrieve all objects from the bucket which need to be copied to the destination. Retrieving objects from S3 Standard-IA and S3 One Zone-IA storage will incur a retrieval fee based on the size of the objects. Read about considerations when working with Amazon S3 storage classes in our documentation.

Q: Can I copy data out of S3 Glacier and Amazon S3 Glacier Deep Archive?

A: When using S3 as the source location for an AWS DataSync task, the service will attempt to retrieve all objects from the bucket which need to be copied to the destination. Retrieving objects which are archived in the S3 Glacier or S3 Glacier Deep Archive storage class results in an error. Any errors retrieving archived objects will be logged by DataSync and will result in a failed task completion status. Read about considerations when working with Amazon S3 storage classes in our documentation.

Q: How does AWS DataSync access my Amazon S3 bucket?

A: AWS DataSync assumes an IAM role that you provide. The policy you attach to the role determines which actions the role can perform. DataSync can auto generate this role on your behalf or you can manually configure a role.

Q: How does AWS DataSync convert files and folders to or from objects in Amazon S3?

A: When files or folders are copied to Amazon S3, there is a one-to-one relationship between a file or folder and an object. File and folder metadata timestamps and POSIX permissions, including user ID, group ID, and permissions, are stored in S3 user metadata. File metadata stored in S3 user metadata is interoperable with File Gateway, providing on-premises file-based access to data stored in Amazon S3 by AWS DataSync.

When DataSync copies from an NFS server, the POSIX permissions from the files and folders on the source are stored in the S3 user metadata. When DataSync copies objects that contain this user metadata back to an NFS server, the file metadata is restored. Symbolic links and hard links are also restored when copying back from NFS to S3.

When copying from an SMB file share, default POSIX permissions are stored in S3 user metadata. When copying back to an SMB file share, ownership is set based on the user that was configured in DataSync to access that file share, and default permissions are assigned.

Learn more about how DataSync stores files and metadata in our documentation.

Q: Which Amazon S3 request and storage costs apply when using S3 storage classes with AWS DataSync?

A: Some S3 storage classes have behaviors that can affect your cost, such as data retrieval, minimum storage capacities, and minimum storage durations. DataSync automates management of data to address these factors, and provides settings to minimize data retrieval.

To avoid minimum capacity charge per object, AWS DataSync automatically stores small objects in S3 Standard. To minimize data retrieval fees, you can configure DataSync to verify only files that were transferred by a given task. To avoid minimum storage duration charges, DataSync has controls for overwriting and deleting objects. Read about considerations when working with Amazon S3 storage classes in our documentation.

Transferring to and from Amazon EFS

Q: How does AWS DataSync access my Amazon EFS file system?

A: AWS DataSync accesses your Amazon EFS file system using the NFS protocol. The DataSync service mounts your file system from within your VPC from Elastic Network Interfaces (ENIs) managed by the DataSync service. DataSync fully manages the creation, use, and deletion of these ENIs on your behalf.

Q: Can I use AWS DataSync with all EFS storage classes?

A: Yes. You can use AWS DataSync to copy files into EFS and configure EFS Lifecycle Management to migrate files that have not been accessed for a set period of time to the Infrequent Access (IA) storage class.

Transferring to and from Amazon FSx for Windows File Server

Q: How does DataSync access my Amazon FSx file system?

A: AWS DataSync accesses your Amazon FSx file system using the SMB protocol, authenticating with the username and password you configure in the AWS Console or CLI. The DataSync service mounts your file system from within your VPC from Elastic Network Interfaces (ENIs) managed by the DataSync service. DataSync fully manages the creation, use, and deletion of these ENIs on your behalf.

Q: What Windows metadata is transferred when copying from SMB shares to Amazon FSx?

A: DataSync copies Windows metadata, including file timestamps, file owner, standard file attributes, and NTFS discretionary access lists (DACLs). Copying NTFS system access control lists (SACLs) will be added later in 2020. You can learn more and see the complete list of copied metadata in our documentation.

Performance

Q: How fast can AWS DataSync copy my file system to AWS?

A: The rate at which AWS DataSync can copy a given dataset is a function of amount of data, I/O bandwidth achievable from the source and destination storage, network bandwidth available, and network conditions. A single DataSync agent is capable of saturating a 10 Gbps network link.

Q: Can I control the amount of network bandwidth that an AWS DataSync task uses?

A: Yes. You can control the amount of network bandwidth that AWS DataSync will use by configuring the built-in bandwidth throttle. This can help to minimize impact on other users or applications who rely on the same network connection.

Q: How can I monitor the performance of AWS DataSync?

A: AWS DataSync generates Amazon CloudWatch Metrics to provide granular visibility into the transfer process. Using these metrics, you can see the number of files and amount of data which has been copied, as well as file discovery and verification progress. You can see CloudWatch Graphs with these metrics directly in the DataSync Console.

Q: Will AWS DataSync affect the performance of my source file system?

A: Depending on the capacity of your on-premises file store, and the quantity and size of files to be transferred, AWS DataSync may affect the response time of other clients when accessing the same source data store, because the agent reads or writes data from that storage system. Configuring a bandwidth limit for a task will reduce this impact by limiting the I/O against your storage system.

Security and compliance

Q: Is my data encrypted while being transferred and stored?

A: Yes. All data transferred between the source and destination is encrypted via Transport Layer Security (TLS), which replaced Secure Sockets Layer (SSL). Data is never persisted in AWS DataSync itself. The service supports using default encryption for S3 buckets, Amazon EFS file system encryption of data at rest, and Amazon FSx For Windows File Server encryption at rest and in transit.

Q: How does AWS DataSync access my NFS server or SMB file share?

A: AWS DataSync uses an agent that you deploy into your IT environment or into Amazon EC2 to access your files through the NFS or SMB protocol. This agent connects to DataSync service endpoints within AWS, and is securely managed from the AWS Management Console or CLI.

Q: Does AWS DataSync require setting up a VPN to connect to my destination storage?

A: No. When copying data to or from your premises, there is no need to setup a VPN/tunnel or allow inbound connections. Your AWS DataSync agent can be configured to route through a firewall using standard network ports. You can also deploy DataSync within your Amazon Virtual Private Cloud (Amazon VPC) using VPC endpoints. When using VPC endpoints, data transferred between the DataSync agent and AWS services does not need to traverse the public internet or need public IP addresses.

Q: How do my AWS DataSync agents securely connect to AWS?

A: Your AWS DataSync agent connects to DataSync service endpoints within your chosen AWS Region. You can choose to have the agent connect to public internet facing endpoints, Federal Information Processing Standards (FIPS) validated endpoints, or endpoints within one of your VPCs. Activating your agent securely associates it with your AWS account. To learn more, see Choose a Service Endpoint and Activate Your Agent.

Q: How is my AWS DataSync agent patched and updated?

A: Updates to the agent VM, including both the underlying operating system and the AWS DataSync software packages, are managed by the service once the agent is activated. Updates are applied non-disruptively when the agent is idle and not executing a data transfer task.

Q: Which compliance programs does AWS DataSync support?

A: AWS has the longest-running compliance program in the cloud. AWS is committed to helping customers navigate their requirements. AWS DataSync has been assessed to meet global and industry security standards. DataSync complies with PCI DSS, ISO 9001, 27001, 27017, and 27018; SOC 1, 2, and 3; in addition to being HIPAA eligible. DataSync is also authorized in the AWS US East/West Regions under FedRAMP Moderate and in the AWS GovCloud (US) Regions under FedRamp High. That makes it easier for you to verify our security and meet your own obligations. For more information and resources, visit our compliance pages. You can also go to the Services in Scope by Compliance Program page to see a full list of services and certifications.

Q: Is AWS DataSync PCI compliant?

A: Yes. AWS DataSync is PCI-DSS compliant, which means you can use it to transfer payment information. You can download the PCI Compliance Package in AWS Artifact to learn more about how to achieve PCI Compliance on AWS.

Q: Is AWS DataSync HIPAA eligible?

A: Yes. AWS DataSync is HIPAA eligible, which means if you have a HIPAA BAA in place with AWS, you can use DataSync to transfer protected health information (PHI).

Q: Does AWS DataSync have FedRAMP JAB Moderate Provisional Authorization in the AWS US East/West?

A: AWS DataSync has received a Provisional Authority to Operate (P-ATO) from the Joint Authorization Board (JAB) at the Federal Risk and Authorization Management Program (FedRAMP) Moderate baseline in the US East/West Regions. If you are a federal or commercial customer, you can use AWS DataSync in the AWS East/West Region's authorization boundary with data up to the moderate impact level.

Q: Does AWS DataSync have FedRAMP JAB High Provisional Authorization in the AWS GovCloud (US) Regions?

A: AWS DataSync has received a Provisional Authority to Operate (P-ATO) from the Joint Authorization Board (JAB) at the Federal Risk and Authorization Management Program (FedRAMP) High baseline in the US GovCloud Region. If you are a federal or commercial customer, you can use AWS DataSync in the AWS GovCloud (US) Region’s authorization boundary with data up to the high impact level.

When to choose AWS DataSync

Q: How is AWS DataSync different from using command line tools such as rsync or the Amazon S3 command line interface?

A: AWS DataSync fully automates and accelerates moving large active datasets to AWS, up to 10 times faster than command line tools. It is natively integrated with Amazon S3, Amazon EFS, Amazon FSx for Windows File Server, Amazon CloudWatch, and AWS CloudTrail, which provides seamless and secure access to your storage services, as well as detailed monitoring of the transfer.

DataSync uses a purpose-built network protocol and scale-out architecture to transfer data. A single DataSync agent is capable of saturating a 10 Gbps network link.

DataSync fully automates the data transfer. It comes with retry and network resiliency mechanisms, network optimizations, built-in task scheduling, monitoring via the DataSync API and Console, and CloudWatch metrics, events and logs that provide granular visibility into the transfer process. DataSync performs data integrity verification both during the transfer and at the end of the transfer.

DataSync provides end to end security, and integrates directly with AWS storage services. All data transferred between the source and destination is encrypted via TLS, and access to your AWS storage is enabled via built-in AWS security mechanisms such as IAM roles. DataSync with VPC endpoints are enabled to ensure that data transferred between an organization and AWS does not traverse the public internet, further increasing the security of data as it is copied over the network.

Q: When do I use AWS DataSync and when do I use AWS Snowball Edge?

A: AWS DataSync is ideal for online data transfers. You can use DataSync to migrate active data to AWS, transfer data to the cloud for analysis and processing, archive data to free up on-premises storage capacity, or replicate data to AWS for business continuity.

AWS Snowball Edge is suitable for offline data transfers, for customers who are bandwidth constrained, or transferring data from remote, disconnected, or austere environments.

Q: When do I use AWS DataSync and when do I use AWS Storage Gateway?

A: Use AWS DataSync to migrate existing data to Amazon S3, and then use the File Gateway configuration of AWS Storage Gateway to retain access to the migrated data and for ongoing updates from your on-premises file-based applications.

You can use a combination of DataSync and File Gateway to minimize your on-premises infrastructure while seamlessly connecting on-premises applications to your cloud storage. AWS DataSync enables you to automate and accelerate online data transfers to AWS storage services. File Gateway then provides your on-premises applications with low latency access to the migrated data.

Q: When do I use AWS DataSync, and when do I use Amazon S3 Transfer Acceleration?

A: If your applications are already integrated with the Amazon S3 API, and you want higher throughput for transferring large files to S3, you can use S3 Transfer Acceleration. If you want to transfer data from existing storage systems (e.g. Network Attached Storage), or from instruments that cannot be changed (e.g. DNA sequencers, video cameras), or if you want multiple destinations, you use AWS DataSync. DataSync also automates and simplifies the data transfer by providing additional functionality, such as built-in retry and network resiliency mechanisms, data integrity verification, and flexible configuration to suit your specific needs, including bandwidth throttling, etc.

Q: When do I use AWS DataSync and when do I use AWS Transfer for SFTP?

A: If you currently use SFTP to exchange data with third parties, AWS Transfer for SFTP provides a fully managed SFTP transfer directly into and out of Amazon S3, while reducing your operational burden.

If you want an accelerated and automated data transfer between NFS servers, SMB file shares, Amazon S3, Amazon EFS, and Amazon FSx for Winodws File Server, you can use AWS DataSync. DataSync is ideal for customers who need online migrations for active data sets, timely transfers for continuously generated data, or replication for business continuity.

AWS Server Migration Service FAQs

General

Q: How do I get started with AWS Server Migration Service?

Use the wizard in the AWS Server Migration Service dashboard of the AWS Management Console. Visit the Getting Started Guide for more details, including how to replicate a server.

Q: What is the output of AWS Server Migration Service?

Each server volume replicated is saved as a new Amazon Machine Image (AMI), which can be launched as an EC2 instance (virtual machine) in the AWS cloud. If you are using application groupings, Server Migration Service will launch the servers in a CloudFormation stack using an auto-generated CloudFormation template.

Q: What kind of servers can be migrated to AWS using AWS Server Migration Service?

Currently, you can migrate virtual machines from VMware vSphere, Windows Hyper-V, or Microsoft Azure to AWS using AWS Server Migration Service.

Q: In what Regions is AWS Server Migration Service available?

Refer to the Region Table.

Q: How do I track the status of the migration?

Visit the AWS Server Migration Service dashboard in the AWS Management Console to see the status of the replication.

Q: What operating systems does AWS Server Migration Service support?

AWS Server Migration Service supports migrating Windows Server 2003, 2008, 2012, and 2016, and Windows 7, 8, and 10; Red Hat Enterprise Linux (RHEL), SUSE/SLES, CentOS, Ubuntu, Oracle Linux, Fedora, and Debian Linux operating systems. Learn more.

Q: How long can I replicate my server volumes from on-premises to AWS?

You can replicate your on-premises servers to AWS for up to 90 days (per server). Usage time is calculated from the time a server replication begins until you terminate the replication job. After 90 days, your replication job will be automatically terminated. If you want to increase this limit, please discuss your use case with the AWS Support team.

Q: What is the difference between EC2 VM Import and AWS Server Migration Service?

AWS Server Migration Service is a significant enhancement of EC2 VM Import. The AWS Server Migration Service provides automated, live incremental server replication and AWS Console support. For customers using EC2 VM Import for migration, we recommend using AWS Server Migration Service.

AWS Server Migration Service Connector

Q: What is the AWS Server Migration Service Connector?

The connector appliance is a pre-configured FreeBSD virtual machine (in OVA format). To set up AWS Server Migration Service, you need to first deploy the AWS Server Migration Service Connector virtual appliance on your on-premises VMware vCenter environment.

Q: How many AWS Server Migration Service Connectors do I need to install?

You need to install one AWS Server Migration Service Connector for each VMware vCenter environment.

Q: What permissions does the AWS Server Migration Service Connector require from VMware vCenter?

At minimum, the AWS Server Migration Service Connector requires the ability to create and delete snapshots on VMs that need to be migrated to AWS. Learn more.

Q: Can I use a proxy for communicating with the AWS Server Migration Service?

Yes. The AWS Server Migration Service Connector supports password-based proxy; it does not support NTLM-based proxy.

Security

Q: Are server volumes securely transferred from my data center to AWS?

Yes. Replicated server volumes are encrypted in transit by Transport Layer Security (TLS).

Q: What data does the AWS Server Migration Service Connector capture from VMware vCenter?

The AWS Server Migration Service Connector captures VM inventory information from VMware vCenter and replicates server volumes to AWS.

Q: How do I update the AWS Server Migration Service Connector?

Updates are automatically downloaded and applied when you enable the auto-upgrade option; otherwise, they can be applied on-demand.

Q: Is it secure to deploy the AWS Server Migration Service Connector on my virtualization environment?

Yes, the AWS Server Migration Service Connector only captures basic VM inventory information and snapshots of server volumes from VMware vCenter and does not gather any sensitive information.

Q: Where are my replicated server volumes stored?

Your replicated server volumes are converted to AMIs and stored in your AWS account.

Support

Q: Does AWS Premium Support cover AWS Server Migration Service?

Yes, AWS Premium Support covers issues related to your use of the AWS Server Migration Service. Learn more.

Q: What other support options are available?

Visit the AWS Community discussion forum.

Hyper-V VM migration

Q: What are the key benefits of using AWS Server Migration Service for migrating on premise Hyper-V VMs?

By automating an incremental replication of live server volumes to the AWS cloud as AMIs, SMS allows customers to speed up migration process and reduce manual labor of migration significantly. SMS orchestrates server migrations by allowing customers to schedule replications and track the progress of a group of servers, alleviating the logistical burden of coordinating large-scale server migrations. With the support of incremental replication, customers are able to test server migrations easily. With the support of AWS console-based GUI, customers are able to start and manage migration and track progress easily.

Q: Who can benefit from AWS Server Migration Service for Hyper-V?

Any customer who is looking to migrate their Microsoft Hyper-V virtual machines managed by SCVMM or standalone Hyper-Vs to AWS will benefit from Hyper-V support in SMS. This may include enterprise customers, System integrators, and IT consulting firms who help enterprise customers migrate Hyper-V workloads to AWS.

Q: What components does AWS Server Migration Service have for Hyper-V VM migration?

SMS has an on-premises appliance, the SMS Connector, which talks to the service in AWS. The Connector incrementally transfers volumes of running Hyper-V VMs to the SMS service, and the service creates the AMI incrementally from the transferred volume.

Q: Where can I download the SMS Connector for Hyper-V VM migration?

You can go to the AWS Server Migration Service Console or the SMS public documentation.

Q: What permissions are required in System Center to deploy SMS Connector and start migration?

The SMS connector securely communicates with the SCVMM server using Windows remote management protocol (WinRM). Connector requires a non-admin AD user that is added to “Remote Management Users” group on the SCVMM host, has limited permissions on the CMIV2 and SCVMM WMI objects and is a part of the “Delegated administrator” group within SCVMM. A firewall port needs to be opened on the SCVMM server to allow for secure transfer of remote commands from Connector deployed on a private network within your datacenter. The AD user also needs read permissions on the VM data store on the Hyper-V machine. For additional security, customers may configure WinRM to allow only encrypted traffic over SSL using self-signed certificate and limit access to Connector IP/hostname alone.For more details, please refer to the SMS technical documentation.

Q: How do I set up AWS Server Migration Service for Hyper-V?

AWS console will walk you through a wizard that help you download and install the SMS Connector. You will need to work with your system administrators to create an active directory user for Hyper-V VM migration use. AWS SMS provides an interactive PowerShell script, to be executed with administrative privileges on your Hyper-V host, that automates the permissions setup/configuration. You will provide the AD user credentials and networking/proxy settings as necessary in the SMS Connector. For more details, please refer to the SMS technical documentation.

Q: Do I need to have SCVMM for AWS SMS to work with my Hyper-V hosts?

No. The SMS Connector can be configured to use SCVMM or standalone Hyper-V VMs.

Q: I'm operating both VMware and Hyper-V environments. Can I migrate VMware VMs and Hyper-V VMs simultaneously?

Yes, but you will need two separate SMS Connectors to simultaneously migrate VMs from both VMware and Hyper-V environments.

Q: What versions of Hyper-V, a component of the Windows server, does SMS support?

SMS Supports Hyper-V running on Windows Servers 2012 R2 and above.

Q: What operating systems does AWS Server Migration Service support?

AWS Server Migration Service supports migrating Windows Server 2003, 2008, 2012, and 2016, and Windows 7, 8, and 10; Red Hat Enterprise Linux (RHEL), SUSE/SLES, CentOS, Ubuntu, Oracle Linux, Fedora, and Debian Linux operating systems. Learn more.

Q: How can I view my on-premises Hyper-V VM inventory in AWS console?

After configuring the SMS Connector, customers can import their on-premises Hyper-V inventory by clicking on the "Import server catalog" button in "AWS Console --> Server Migration --> Servers" tab.

Q: I'm going to migrate Hyper-V and VMware VMs simultaneously. Can I view the inventory of Hyper-V and VMware VMs together in a single pane of glass in AWS console?

Yes. SMS enables you to view the VM inventory of both Microsoft's Hyper-V and VMware environments in a single pane of glass. After you import your on-premises Hyper-V VM inventory by clicking on the "Import server catalog" button in AWS SMS console, SMS automatically collects the VM inventory from all the SMS Connectors, deployed across Hyper-V and VMware environments, as a default.

Q: Can I see the VM inventory of Hyper-V and VMware environments separately?

Yes, on the 'servers' page on the AWS SMS console, you can filter by VM manager in the search bar to show only the VMs imported from Hyper-V or VMware environments.

Q: How can I distinguish Hyper-V VMs from VMware VMs during migration?

Yes, on the 'replication jobs' page on the AWS SMS console, you can filter by VM manager in the search bar to show only the progress of SMS replication jobs for VMs from Hyper-V or VMware environments.

Q: an I track the progress of Hyper-V and VMware VMs migration simultaneously?

Yes. Customers can track the progress of Hyper-V and VMware VM migration in a single pane of glass under the 'Replication Jobs' page on the AWS SMS console.

Multi-Server Migration

Q: What is multi-server migration in AWS Server Migration Service?

AWS Server Migration Service now offers multi-server migration support that makes it easier and cost effective to migrate applications from on-premises datacenters to Amazon EC2. Multi-server migration provides you the ability to migrate entire application stacks as opposed to migrating each server individually. You can group servers into applications, replicate the entire application together, and monitor its migration status centrally from the console. You can also easily launch and configure the migrated application with an auto-generated CloudFormation Template.

Q: How does multi-server support work in AWS Server Migration Service?

You can group on-premises servers into an application with one or more sub-groups, specify a replication frequency, provide configuration scripts, and a landing zone where the replicated application should be launched. Server Migration Service will orchestrate migration of all the underlying servers that are part of the application groups, wait for all of the AMIs to be available, and create a Cloud Formation Template that can be launched in the landing zone. Once you launch the application using the auto-generated CloudFormation Template, Server Migration Service will configure it based on your specified configuration scripts.

Q: I already use AWS Server Migration Service. How does this capability benefit me?

With multi-server migration, you no longer have to write custom tooling to coordinate the migrations of multiple servers that make up your application—you can migrate all of your servers together as a single unit. You have the ability to launch all of the servers automatically through a CloudFormation Template and keep it continuously updated with the latest AMIs produced for each replication run. You can start using the multi-server migration without any new setup.

Q: How do I get started with using multi-server migration in Server Migration Service?

Once the on-premises server catalog is imported into Server Migration Service using the SMS Connector, you can get started by configuring an application from the Server Migration Service console, CLI, or APIs. If you are already using Server Migration Service and have configured a SMS role using a managed policy, no action is required. Multi-server migration requires a new role to be able to launch instances using CloudFormation. See technical documentation for more details on how to setup permissions for Server Migration Service.

Q: Can I migrate the applications defined in AWS Application Discovery Service using AWS Server Migration Service?

Currently, the application groupings defined using AWS Application Discovery Service are not available through Server Migration Service. However, support is being added for applications discovered/defined in Application Discovery Service to be automatically available in Server Migration Service for multi-server migration.

Q: Can I customize the CloudFormation Template that AWS Server Migration Service auto-generates?

No, the ability to customize the auto-generated template is not available at this time.

Q: How do I re-configure my application after I launch it in the AWS cloud?

You can provide a custom configuration script per server while defining an application in the Server Migration Service console. Once the server is migrated and launched in the AWS cloud, the service will run the configuration script on your behalf. For example, you can use a configuration script to update database connection strings on your application servers without having to log into each instance and updating it manually.

Q: What type of snapshot consistency does AWS Server Migration Service provide?

AWS Server Migration Service creates crash consistent snapshots for the servers that are part of an application group. These snapshots are triggered around the same time for all the servers within an application. Additionally, for servers running Windows operating system in VMware environments, you can use Volume Shadow Copy Service (VSS) to take application consistent snapshots.

Windows to Linux replatforming assistant for Microsoft SQL Server databases

Q: What is a replatforming assistant for Microsoft SQL Server?

Learn more about the replatforming assistant for Microsoft SQL Server in our latest What's New post.

AWS Snow family FAQs

General

Q: What is the AWS Snow family?

The AWS Snow family are physical devices that help migrate large amounts of data into and out of the cloud without depending on networks. This helps you apply the wide variety of AWS services for analytics, file systems and archives to your data. Snowball is a suitcase-sized device, Snowball Edge is a rack mountable and clusterable suitcase sized device with compute capabilties, and Snowmobile is a shipping container moved with a tractor-trailer. These services can assist with data migration, disaster recovery, data center shutdown and remote data collection projects.

Q: What are some example use cases?

You can use AWS Snow family services for data transfer and occasional pre-processing on location. Some large data transfer examples include cloud migration, disaster recovery, data center relocation, and/or remote data collection projects. These projects typically require you to migrate large amounts of data in the shortest, and most cost-effective, amount of time.

Some example use cases for Snowball Edge computing capabilties include IoT sensor stream capture, on-the-fly media transcoding, image compression, aggregating metrics, and control signaling and alarming.

Q: Why do I need this service?

It can take a long time to transfer large amounts of data over the wire, and some locations don't have any connectivity at all. The Snow family helps expedite data transfers in a more secure and cost-effective way. Each service has a pre-set capacity level to make it easy to choose. Snowball Edge further helps bring computing applications closer to the data source to enhance analysis and deliver real-time results.

Q: Which Snow family service is best for me?

Th Snow family comes in multiple capacities and form factors to fit most data migration and/or remote data collection projects. First determine a capacity point then evaluate your need for on-site computing capabilities.

Please refer to the comparison table for more details.

Q: How quickly can I migrate data?

The Snowball and Snowball Edge services can typically transfer up to 100TBs in about a week. Snowmobile can transfer data at a rate of up to 1TB/s, which means 100PBs can be loaded in just a few weeks. In comparison, a dedicated T3 line at 50Mb/s takes years.

Q: How do I choose between Snow family and other AWS data migration services?

AWS Storage Gateway and Direct Connect services are good choices when network bandwidth limitations do not exist. For the most efficient means of data transfer, whether connected to a network or not, the Snow family of services provides good choices as well. Amazon offers a variety of tools to help you move data via networks, roads, and technology partners. See the pages below to find out which service is best for you.

Learn more about migrating to AWS here »

Learn more about which data migration tool is right for you »

Q: What AWS Regions are supported?

Snow family services are available for use in specific AWS regions. To discuss data transfer needs specific for your region, please follow up with AWS Sales, or see the Regional Service Availability pages, for more information.

AWS Transfer for SFTP FAQs

General

Q: What is AWS Transfer for SFTP?

A: AWS Transfer for SFTP (AWS SFTP) is a fully managed service hosted in AWS that enables the transfer of files over SFTP directly in and out of Amazon S3.

Q: What is SFTP and where is it used?

A: SFTP stands for Secure Shell (SSH) File Transfer Protocol, a network protocol used for secure transfer of data over the internet. The protocol supports the full security and authentication functionality of SSH, and is widely used to exchange data between business partners in a variety of industries including financial services, healthcare, media and entertainment, retail, advertising, and more.

Q: Why should I use AWS SFTP?

A: Today, if you are using SFTP to exchange data with third parties such as vendors, business partners, or customers, and want to manage that data in AWS for processing, analytics, and archival, you have to host and manage your own SFTP service. This requires you to invest in operating and managing infrastructure, patching servers, monitoring for uptime and availability, and building one-off mechanisms to provision users and audit their activity. AWS SFTP solves these challenges by providing a fully managed SFTP service that can reduce your operational burden, while preserving your existing transfer workflows for your end users. The service stores transferred files as objects in your Amazon S3 bucket, so you can extract value from them in your data lake, or for your Customer Relationship Management (CRM) or Enterprise Resource Planning (ERP) workflows, or for archiving in AWS.

Q: What are the benefits of using AWS SFTP?

A: AWS SFTP provides you with a fully managed, highly available SFTP service with auto-scaling capabilities, eliminating the need for you to manage SFTP-related infrastructure. With AWS SFTP, your end users' workflows remain unchanged, while data uploaded and downloaded over SFTP is stored in your Amazon S3 bucket. With the data in Amazon S3, you can now easily use it with the broad array of AWS services for data processing, analytics, machine learning, and archival, in an environment that can meet your compliance requirements.

Q: How do I use AWS SFTP?

A: In 3 simple steps, you can get a persistent and highly available “SFTP server” in AWS. First, you associate your existing SFTP hostname(s) with the SFTP server endpoint. Next, you set up your users by selecting your identity provider for authentication – Service Managed or a directory service like Microsoft AD. Finally, you choose the S3 bucket(s) and assign IAM Roles for access. Once the service endpoint, identity provider, and S3 bucket access policies are enabled, your users can continue to use their existing clients and configurations while the data they access is stored in your S3 bucket.

Q: Can I use CloudFormation to automate deployment of my SFTP servers and users?

A: Yes, you can deploy CloudFormation templates to automate creation of your SFTP servers and users or for integrating an identity provider. Refer to the usage guide for using AWS Transfer for SFTP resources in CloudFormation templates.

Q: Can my users use SCP, FTP, or FTP/S (FTP over SSL) to transfer files using this service?

A: No, your users will need to use SFTP to transfer files. Most file transfer clients offer SFTP as an option that will need to be selected when transferring files using AWS SFTP.

Server endpoint options

Q: Can I use my corporate domain name (sftp.mycompanyname.com) to access my SFTP endpoint?

A: Yes. If you already have a domain name, you can use Amazon Route53 or any DNS service to route your users' traffic from your registered domain to the server endpoint in AWS. Refer to the documentation on how AWS Transfer uses Amazon Route 53 for custom domain names applicable to internet facing endpoints only.

Q: Can I still use the service if I don't have a domain name?

A: Yes, if you don't have a domain name, your users can access your server endpoint using the hostname provided by AWS SFTP. Alternatively, you can register a new domain using the Amazon Route 53 Console or API and route traffic from this new domain to your SFTP server endpoint.

Q: Can I use my domain that already has a public zone?

A: Yes, you will need to CNAME the domain to the SFTP server hostname**.**

Q: Can I set up my SFTP server endpoint to be accessible only within my VPC?

A: Yes. When you create a server or update an existing one, you have the option to specify whether you want the endpoint to be accessible over the public internet or within your VPC. Using a VPC endpoint for your server makes it accessible only to clients within the same VPC, other VPCs you specify, or in on-premises environments using networking technologies that extend your VPC such as AWS Direct Connect, AWS VPN, or VPC peering. You can further restrict access to resources in specific subnets within your VPC using subnet Network Access Control Lists (NACLs) or endpoint Security Groups. Refer to the documentation on creating your server endpoint inside your VPC using AWS PrivateLink for details.

Q: Can my end users use fixed IP addresses to whitelist access to an SFTP server's endpoint in their firewalls?

A: Yes. You can enable fixed IPs for your server endpoint by selecting the VPC endpoint for your server and choosing the internet-facing option. This will allow you to attach Elastic IPs (including BYO IPs) to your server's endpoint, which is assigned as the endpoint's IP address Refer to the section on Creating an internet facing endpoint in the documentation: Creating your server endpoint inside your VPC.

Q: Can I restrict incoming traffic by end users' source IP addresses?

A: Yes. You can attach Security Groups to your server's VPC endpoint which will control inbound traffic to your server. Refer to the section on Creating an internet facing endpoint in the documentation: Creating your server endpoint inside your VPC.

Q: Can my SFTP clients use fixed IP addresses to access my SFTP server whose endpoint type is PUBLIC?

A: No. Fixed IP addresses that are usually used for firewall whitelisting purposes are currently not supported on the PUBLIC Endpoint type.

Q: What IP ranges would my end users need to whitelist to access my SFTP server's endpoint type that is PUBLIC?

A: If you are using the PUBLIC endpoint type, your users will need to whitelist the AWS IP address ranges published here. Refer to the documentation for details on staying up to date with AWS IP Address Ranges.

Q: Will my AWS SFTP server's host key ever change after I create the server?

A: No. The server's host key that is assigned when you create the server remains the same, until you delete and create a new one.

Q: Can I import keys from my current SFTP host so my users do not have to reverify the session information?

A: Yes. You can provide a RSA host key when you create a new server, or update an existing one. This key will be used by your end users' clients to identify your server. Refer to the documentation on using the AWS CLI/SDKs for uploading a Host Key for your server.

User authentication

Q: Can my users continue to use their existing SFTP clients or transfer applications?

A: Yes, any existing SFTP client or SFTP transfer application will continue to work with AWS SFTP. Examples of commonly used SFTP clients include WinSCP, FileZilla, CyberDuck, and OpenSSH clients.

Q: How does the service authenticate my users?

A: The service supports two modes of authentication, using the service to store and access user identities, and using a custom identity provider.

Service managed authentication

Q: How can I authenticate my users using service managed authentication?

A: You can use SSH key based authentication if you are using the service to store and access user identities.

Q: How many SSH keys can I upload per user?

A: You can upload up to 10 SSH keys per user. Note that adding more keys increases login time as the server will need to evaluate each of them until a match is found for successful authentication.

Q: Is key rotation supported for service managed authentication?

A: Yes. Refer to the documentation for details on how to set up key rotation using the service.

Q: Can I lock my service managed users to their designated home directories (“chroot”)?

Yes, when you add a new user, or update an existing one, you can select the “restricted” checkbox. This maps the root of your user’s client to the assigned home directory location in your S3 bucket, “chrooting” them to that location.

Q: Can I use the service managed authentication for password authentication?

A: No. Storing passwords within the service for authentication is currently not supported for SFTP. If you need password authentication for SFTP, visit the blog post on 'Enabling Password Authentication using Secrets Manager.'

Q: Are anonymous users supported?

A: No. Anonymous users are currently not supported.

Custom identity provider

Q: Can I leverage my existing identity provider to manage my SFTP users?

A: AWS SFTP allows you to plug-in your existing identity provider so you can migrate your users whose credentials are stored in your corporate directory. Examples of identity providers include Microsoft Active Directory (AD), Lightweight Directory Access Protocol (LDAP), or custom identity providers.

Q: How do I get started in integrating my existing identity provider for user authentication?

A: To get started, we recommend using the AWS CloudFormation template and provide the necessary information for user authentication and access. Visit the website on custom identity providers to learn more.

Q: When setting up my users via a custom identity provider, what information is used to enable access to my users?

A: Your user will need to provide a username and password (or SSH key) which will be used to authenticate, and access to your bucket is determined by the AWS IAM Role supplied by the API Gateway and Lambda used to query your identity provider. You will also need to provide home directory information, and it is recommended that you lock them down to the designated home folder for an additional layer of security and usability. Refer to this blog post on how to simplify your end users' experience when using a custom identity provider with AWS SFTP.

Q: Why do I need to provide an AWS IAM Role, and how is it used?

A: AWS IAM is used to determine the level of access you want to provide your users. This includes what operations you want to enable on their client and which Amazon S3 buckets they have access to – whether it's the entire bucket or portions of it.

Q: Why do I need to provide home directory information and how is it used?

A: The home directory you set up for your user determines their login directory. As soon as your user logs into the SFTP server, this is the directory path their SFTP client would use as the landing directory. You will need to ensure that the IAM Role supplied provides the user access to the home directory.

Q: I have 100s of users who have similar access settings, but to different portions of my bucket. Can I lock their access to the designated home folder only?

A: Yes. You can use logical directory mappings to specify how you want to make absolute Amazon S3 bucket paths visible to your users. In your identity provider integration's Lambda function, you will need to specify "Entry" as "/" (specifying root) and "Target" as the absolute S3 bucket location that will be their home directory path. You may or may not need to use a scope down policy, as the mappings will be the only S3 bucket locations accessible to your end users. Visit this blog on how to 'Simplify Your AWS SFTP Structure with Chroot and Logical Directories.'

Data uploads and downloads

Q: How are files stored in my Amazon S3 bucket transferred using AWS SFTP?

A: Files transferred over SFTP are stored as objects in your Amazon S3 bucket, and there is a one-to-one mapping between files and objects enabling native access to these objects using AWS services for processing or analytics.

Q: How are Amazon S3 objects stored in my bucket presented to my users?

A: After successful authentication, based on your users' credentials, AWS SFTP presents Amazon S3 objects and folders as files and directories to your users' transfer applications. You can also specify logical directory mappings to customize the way S3 bucket paths are presented to your user.

Q: What file operations are supported by AWS SFTP? What operations are not supported?

A: Common SFTP commands to create, read, update, and delete, files and directories are supported. Files are stored as individual objects in your Amazon S3 bucket. Directories are managed as folder objects in S3, using the same syntax as the S3 console. Directory rename operations, changing ownerships, permissions and timestamps, and use of symbolic links and hard links are currently not supported.

Q: Can I control which operations my users are allowed to perform?

A: Yes, you can enable/disable file operations using the AWS IAM role you have mapped to their username.

Q: Can I provide my SFTP users access to more than one Amazon S3 bucket?

A: Yes. If you include multiple S3 buckets in the IAM policy attached to the AWS IAM Role you assign to your user, they will be able to access it. Additionally, you can present folders from multiple S3 buckets as a single namespace to your users by using Logical Directory Mappings. Refer to this blog for more details.

Q: Can I create a server using AWS Account A and map my SFTP users to Amazon S3 buckets owned by AWS Account B?

A: Yes. You can use the CLI and API to set up cross account access between your server and the buckets you want to use for SFTP. The Console drop down will only list buckets in Account A. Additionally, you'd need to make sure the role being assigned to the user belongs to Account A.

Q: How do I know which SFTP user uploaded a file?

A: You can use Amazon CloudWatch to view your SFTP users' activity. Visit the documentation to learn more on how to enable Amazon CloudWatch logging.

Q: Can I automate processing of a file once it has been uploaded to Amazon S3?

A: Yes, you can use Amazon S3 events to automate processing of the uploaded files using a broad array of AWS services for querying, analysis, machine learning, and more. Visit the documentation to learn more on common examples for post upload processing using Lambda with Amazon S3.

Q: Can I view how much data was uploaded and downloaded using my AWS SFTP server?

A: Yes, data uploaded and downloaded using your server is tracked as metrics in Amazon CloudWatch. Visit the documentation to view the available metrics for tracking and monitoring.

Security and compliance

Q: Is my data secure while in-transit?

A: Yes, the underlying security of the SFTP protocol transfers commands and file data through a secure, encrypted tunnel.

Q: What are my options to encrypt data at rest that was transferred using AWS SFTP?

A: You can choose to encrypt files stored your bucket using Amazon S3 Server-Side Encryption (SSE-S3) or Amazon KMS (SSE-KMS).

Q: Which compliance programs does AWS SFTP support?

A: AWS SFTP is PCI-DSS and GDPR compliant, and HIPAA eligible. AWS SFTP is also SOC 1, 2, and 3 compliant. Learn more about services in scope by compliance programs.

Q: Is AWS SFTP FISMA compliant?

A: AWS East/West and GovCloud (US) Regions are compliant. This compliance is demonstrated through FedRAMP Authorization of these two regions to FedRAMP Moderate and FedRAMP High. We demonstrate compliance through annual assessments and documenting compliance with in-scope NIST SP 800-53 controls within our System Security Plans. Templates are available on Artifact along with our customer responsibility matrix (CRM) which demonstrates, at a detailed level, our responsibility to meet these NIST controls as required by FedRAMP. Artifact is available through the management console accessible by an AWS account for both East/West and GovCloud. If you have any further questions on this topic, please consult the Console.

Q: How does the service ensure integrity of uploaded files?

A: All files uploaded through the SFTP server are verified by comparing the file's pre- and post-upload MD5 checksum.

Q: How can I monitor usage and track my users' activity?

A: You can use Amazon CloudWatch to view your SFTP users' activity. Visit the documentation to learn more on how to enable Amazon CloudWatch logging. Additionally, you can view a record of Amazon S3 API calls made on behalf of your users in AWS CloudTrail.

Billing

Q: What am I paying for when I use AWS SFTP?

A: You pay for the resources you use with AWS SFTP. This includes an hourly charge for the SFTP server endpoint, and charges for SFTP data uploads and downloads. The pricing covers a fully managed, highly available SFTP service that auto-scales in real-time based on your workload demands. Please refer to the AWS SFTP pricing page for more details.

Q: How am I billed for my AWS SFTP server?

A: You are billed on an hourly basis from the time you create and configure your SFTP server, which provisions it for your dedicated use, until the time you delete the server. You are also billed based on the amount of data uploaded and downloaded through your SFTP server. Please refer to the AWS SFTP pricing page for more details.

Q: I have stopped my server. Will I be billed for that server while it is stopped?

A: Yes, stopping the server, by using the console, or by running the “stop-server” CLI command or the “StopServer” API command, does not impact billing. You are billed on an hourly basis from the time you create and configure your SFTP server, which provisions it for your dedicated use, until the time you delete the server.

Frequently asked questions

General

Q: What is AWS Amplify?
AWS Amplify consists of a development framework and developer services that provide the fastest and easiest way to build mobile and web applications on AWS. The open source Amplify Framework provides an opinionated set of libraries, UI components, and a command line interface to build an app backend and integrate it with your iOS, Android, Web, and React Native apps. The Amplify Framework leverages a core set of AWS Cloud Services to offer capabilities including offline data, authentication, analytics, push notifications, bots, and AR/VR at high scale. The AWS Amplify Developer Tools services include the AWS Amplify Console for building, deploying, and hosting web apps and AWS Device Farm for testing mobile apps on real iOS and Android devices.

Q: What are the main components of AWS Amplify?
AWS Amplify contains the open source Amplify Framework, Cloud Services such as AWS AppSync and Amazon Cognito leveraged by the Amplify Framework, and the AWS Amplify Console and AWS Device Farm Developer Tools.

Q: What does it cost to use AWS Amplify?
When you use the Amplify Framework, you pay only for the underlying AWS services you use. There are no additional charges for using the Amplify Framework. To learn about pricing for the AWS Amplify Console, visit the AWS Amplify Console pricing page. To learn about pricing for AWS Device Farm, visit the AWS Device Farm pricing page.

Q: How does the AWS Amplify Console relate to the open source Amplify Framework?
The AWS Amplify Console and the open source Amplify Framework can be used

together or separately. For example, you can use the Amplify Console to deploy and host Single Page App (SPA) frontends and static websites, whether or not they use the Amplify Framework.

If you are using the Amplify Framework's CLI to configure backend resources for your app, the Amplify Console offers additional functionality. On each check-in, the Amplify Console provisions or updates these backend resources prior to deploying your frontend. There is support for a variety of configurations, such as isolated backend deployments per branch or shared backend deployments across branches.

Q: What happened to AWS Mobile Hub?
Existing AWS Mobile Hub customers can continue to use Mobile Hub. For new projects, developers should use AWS Amplify instead.

Q: Where can I find the latest news on AWS Amplify?
Visit our blog and What's New page.

Amplify Framework

Q: What can I do with the open source Amplify Framework?
With the open source Amplify Framework, you can quickly add features such as offline data, multifactor authentication, analytics, and others to your application with a few lines of code. You can configure and integrate the underlying cloud services like AWS AppSync, Amazon Cognito, Amazon Pinpoint, AWS Lambda, Amazon S3, or Amazon Lex directly from your command line minimizing the time required to set-up and manage your back-end services.

Q: What languages and platforms does the Amplify Framework support?
The Amplify Framework supports iOS, Android, Web, and React Native apps. For Web apps, there is deep integration with React, Ionic, Angular, and Vue.js.

Q: Can I use the Amplify Framework libraries even if I do not use the CLI?
Yes. The libraries can be used to access backend resources that were created

without the Amplify CLI.

Q: How is AWS Amplify related to the AWS Mobile SDKs for iOS and Android?
The AWS Mobile SDKs for iOS and Android are part of the Amplify Framework. Get started here.

Q: Where can I find the AWS SDKs for Xamarin and Unity?
To get started with the AWS Mobile SDK for Xamarin, read the AWS Mobile SDK for Xamarin developer guide. To get started with the AWS Mobile SDK for Unity, read the AWS Mobile SDK for Unity developer guide.

Cloud Services

Q: How does the Amplify Framework work with AWS cloud services?
The Amplify Framework is organized based on the features you want to add to your app, such as offline data, multi factor authentication, analytics, and others. When you configure these features using the Amplify CLI, the necessary AWS cloud services are provisioned for you. The configuration is persisted in CloudFormation templates that can be checked into source control and shared with other developers. When you integrate and then use these features into your app via the Amplify library, the library makes the necessary calls to AWS services. For example, 'amplify add analytics' will configure Amazon Pinpoint. Then, when you use the Analytics APIs from Amplify library in your app, the necessary calls will be made to Pinpoint.

Developer Tools

Q: Where can I find the FAQs for the AWS Amplify Console?
Visit AWS Amplify Console FAQs.

Q: Where can I find the FAQs for AWS Device Farm?
Visit the AWS Device Farm FAQs.

Amazon API Gateway FAQs

General

Q: What is Amazon API Gateway? »

Q: Why use Amazon API Gateway? »

Q: What API types are supported by Amazon API Gateway? »

Q. How do I get started with HTTP APIs in API Gateway? »

Q. How do I get started with REST APIs in API Gateway? »

Q: When creating RESTful APIs, when should I use HTTP APIs and when should I use REST APIs? »

Q: Which features come standard with HTTP APIs from API Gateway? »

Q: Can I import an OpenAPI definition to create a HTTP API? »

Q: How can I migrate from my current REST API to a HTTP API? »

Q: How do I know if my current REST API will work as a HTTP API? »

Q: How do I get started with WebSocket APIs in Amazon API Gateway? »

Q: Can I create HTTPS endpoints? »

Q: What data types can I use with Amazon API Gateway? »

Q: With what backends can Amazon API Gateway communicate? »

Q: For which client platforms can Amazon API Gateway generate SDKs? »

Q: In which AWS regions is Amazon API Gateway available? »

Q: What can I manage through the Amazon API Gateway console? »

Q: What is a resource? »

Q: What is a method? »

Q: What is an usage plan? »

Q: What is the Amazon API Gateway API lifecycle? »

Q: What is a stage? »

Q: What are stage variables? »

Q: What is a Resource Policy? »

Q: What if I mistakenly deployed to a stage? »

Q: Can I use my Swagger API definitions? »

Q: How do I monetize my APIs on Amazon API Gateway? »

Q: How do I document my API on Amazon API Gateway? »

Q: How can I avoid creating redundant copies of error messages and other documentation that recurs frequently in my API? »

Q: Can I restrict access to private APIs to a specific Amazon VPC or VPC endpoint? »

back to top >>

Security and Authorization

Q: How do I authorize access to my APIs? »

Q: How does AWS Signature Version 4 work? »

Q: What is a Lambda authorizer? »

Q: Can Amazon API Gateway generate API keys for distribution to third-party developers? »

Q: How can I address or prevent API threats or abuse? »

Q: Can I verify that it is API Gateway calling my backend? »

Q: Can I use AWS CloudTrail with Amazon API Gateway? »

Q: How does Amazon API Gateway work with an Amazon Virtual Private Cloud (Amazon VPC)? »

Q: Can I restrict access to private APIs to a specific Amazon VPC or VPC endpoint? »

Q: Can I configure my REST APIs in API Gateway to use TLS 1.1 or higher? »

back to top >>

Management, Metrics, and Logging

Q: How can I monitor my Amazon API Gateway APIs? »

Q: Can I set up alarms on the Amazon API Gateway metrics? »

Q: How can I set up metrics for Amazon API Gateway? »

Q: Can I determine which version of the API my customers are using? »

Q: Does Amazon API Gateway provide logging support? »

Q: How quickly are logs available? »

back to top >>

Throttling and Caching

Q: How can I protect my backend systems and applications from traffic spikes? »

Q: Can I throttle individual developers calling my APIs? »

Q: How does throttling help me? »

Q: At which levels can Amazon API Gateway throttle inbound API traffic? »

Q: How are throttling rules applied? »

Q: Does Amazon API Gateway provide API result caching? »

Q: What happens if a large number of end users try to invoke my API simultaneously? »

Q: How do APIs scale? »

back to top >>

Billing

Q: How am I charged for using Amazon API Gateway? »

Q: Who pays for Amazon API Gateway API calls generated by third-party developers? »

Q: If an API response is served by cached data, is it still considered an API call for billing purposes? »

back to top >>

WebSocket APIs

Q: What is WebSocket routing in Amazon API Gateway? »

Q: How can I send messages to connected clients from the backend service? »

Q: How can I authorize access to my WebSocket API in Amazon API Gateway? »

Q: How does my backend service know when a client is connected or disconnected from the WebSocket connection in Amazon API Gateway? »

Q: How can my backend service identify if the client is still connected to the WebSocket connection? »

Q: Can I disconnect a client from my backend service? »

Q: What is the maximum message size supported for WebSocket APIs? »

Q: How am I charged for using WebSocket APIs on Amazon API Gateway? »

Q: If messages on the WebSocket connection fail authentication or authorization, do they still count toward my API usage bill? »

back to top >>

Amazon Pinpoint FAQs

General

Q: What is Amazon Pinpoint?

A: Amazon Pinpoint is AWS's Digital User Engagement Service that enables AWS customers to effectively communicate with their end users and measure user engagement across multiple channels including email, Text Messaging (SMS) and Mobile Push Notifications.

Amazon Pinpoint also provides tools that enables audience management and segmentation, campaign management, scheduling, template management, A/B testing, analytics and data integration. It captures data to track deliverability as well as usage and messaging analytics covering a range of dimensions including user, channels and custom attributes.

Amazon Pinpoint is built on a service-based architecture. Developers can extend their applications and backend services in various ways, including: sending messages directly from their applications via the Amazon Pinpoint channels (Email, SMS and Mobile Push), accessing segmentation data to alter their application behavior for segments of users, create and run campaigns directly from their applications, and access deliverability and analytics data to improve the user engagement of their applications. The system empowers customers to send the right message, to the right audience, at the right time and on the most effective channel.

Q: How will developers benefit from Amazon Pinpoint?

A: Amazon Pinpoint offers developers a single API layer, CLI support, and client-side SDK support to be able to extend the communication channels through which their applications engage users. These channels include: email, SMS text messaging, and push notifications, voice messages, and custom channels. Amazon Pinpoint also provides developers with an analytics system that tracks app user behavior and user engagement. With this service, developers can learn how each user prefers to engage and can personalize their end-user's experience to increase the value of the developer's applications.

Amazon Pinpoint also helps developers address multiple messaging use-cases such direct or transactional messaging, targeted or campaign messaging and event-based messaging.

Integrating and enabling all their end-user engagement channels via Amazon Pinpoint, developers can create a 360-degree view of user engagement across all customer touch points.

Q: How will marketers benefit from Amazon Pinpoint?

A: Amazon Pinpoint allows Marketers to create and execute a unified messaging strategy across all engagement channels relevant to their end-users. Pinpoint includes tools and services to let marketers analyze and engage users directly. The console provides marketers with campaign management tools to create, run and manage multi-channel campaigns across their applications, user-base and devices. Campaigns can be scheduled or triggered on user changes and actions. Users and devices can also be grouped through flexibly defined segments which can be used to determine campaign audiences. Marketers can also leverage the multi-channel templating support to personalize end-user messaging and campaign optimization features such as A/B testing, holdout testing and message caps. Marketers can also measure messaging effectiveness using Pinpoint analytics to understand the impact on user behavior.

Q: How will enterprises benefit from Amazon Pinpoint?

A: Enterprises can use Amazon Pinpoint as their Digital User Engagement Service. They can free developers from having to individually integrate different communication channels into their applications and instead focus on leveraging Pinpoint to learn how their end-users and customers are engaging with their applications. It enables them to measure and improve their technology investments by measuring how engaged their digital customers are across all functions of their enterprise.

Q: Why should I use Amazon Pinpoint to run and manage my campaigns?

A: Amazon Pinpoint makes it easy to run targeted campaigns and drive user engagement of applications using different channels: email, SMS and mobile push notifications. Amazon Pinpoint helps you understand user behavior, define which users to target, determine which messages to send, schedule the best time to deliver the messages, and then track the results of your campaign.

Amazon Pinpoint is built to scale, enabling you to collect and process billions of events per day, and send billions of targeted messages to your users.

Marketers can send targeted messages and calls to action when changes occur in their organizations or in a user’s circumstances, like a new product launch to a change in a user’s locale.

Q: If I use another campaign management service how does Amazon Pinpoint help me?

A: Amazon Pinpoint's architecture is services based. Companies can choose which services to use and integrate with their existing systems and processes. Amazon Pinpoint's core services include: engagement analytics, communication channels, deliverability metrics, audience management and segmentation, template management, and campaign management.

The platform also supports data integration services to extend Amazon Pinpoint analytics and segmentation data from external data sources such as S3, as well as data exports to feed external marketing systems via Kinesis Event Streams.

Q: How much does Amazon Pinpoint cost?

A: Amazon Pinpoint has no upfront costs, no minimum charges, and no subscription fees. You pay only for what you use. Visit the Amazon Pinpoint pricing page for more details.

Q: Who owns the data in Amazon Pinpoint?

A: Customers own their data in Amazon Pinpoint. Amazon Web Services does not own or monetize the data customers collect, and does not share it with third parties. We may use the data to improve the service, monitor the health of the service, and provide technical support to you. As with any other AWS service, customers are responsible for how they use the tools we provide; this responsibility includes providing any necessary notice or opt-outs to end users and complying with applicable law.

Product Details

Q: What services and tools does Amazon Pinpoint provide?

A: Amazon Pinpoint includes a console designed for marketers and developers to use. The console provides capabilities to configure communication channels, manage audiences and segmentation, manage and run campaigns, create and manage message templates, create and manage engagement schedules and analyze user engagement. Standard analytics includes: active users, user activities, sessions, user retention, campaign efficacy and user channel engagement metrics. You can create custom analytics to integrate custom attributes and drive analytics for sales conversion, funnel reporting, product adoption by segment and any other metric required to support the business.

Q: I already use Amazon SNS or Amazon SES. What do I gain by switching to Amazon Pinpoint?

A: In typical Amazon SNS and Amazon SES use cases, you have to set up your application to manage each message's audience, content, and delivery schedule. These same features are built in to Amazon Pinpoint. With Amazon Pinpoint, you can create message templates, delivery schedules, highly-targeted segments, and full campaigns.

Q: How does Amazon Pinpoint Voice differ from Amazon Connect?

A: With Amazon Pinpoint Voice, you can engage with your customers by delivering voice messages over the phone. Pinpoint Voice gives customers a great way to deliver transactional messages—such as one-time passwords, appointment reminders, order confirmations, and more. With Pinpoint Voice you can convert a text script to lifelike speech and then deliver the personalized voice message to your customer. Call metrics—such as number of calls completed and number of calls failed—help you to optimize future voice engagements. With both Poinpoint Voice and SMS channels at your disposal, you can send SMS messages to customers who prefer text and deliver voice messages to those who are either unable to receive SMS messages or who would rather interact via a phone call. With the addition of the voice channel, you can now use Pinpoint to seamlessly engage your customers with timely, relevant content through mobile push, email, SMS, and voice calls. To learn more, please see Amazon Pinpoint Voice.

Amazon Connect is a self-service, cloud-based contact center. With only a few clicks in the AWS Management Console agents can begin making or taking calls within minutes. The service makes it possible to design contact flows, similar in concept to Interactive Voice Response (IVR), that adapt the caller experience, changing based on information retrieved by Amazon Connect from AWS services, like Amazon Redshift, or third-party systems, like CRM or analytics solutions. Contact flow integrations with AWS AI services like Amazon Lex enable intelligent conversational bots to turn automated interactions into natural conversations. The self-service graphical interface in Amazon Connect makes it easy for non-technical users to design contact flows, manage agents, and track performance metrics – no specialized skills required. To learn more, please see Amazon Connect.

Q: What data does Amazon Pinpoint store without using SDKs and instrumentation?

A: Amazon Pinpoint can store four different types of data:

Configuration Data from which Amazon Pinpoint services are provided their rules of engagement. This includes:

1. **Communication.** Communication channels can be configured with restrictions per channel or across channels e.g. maximum number of messages a user can receive per day, maximum messages a user can receive for a campaign and quiet times. There are also channel specific configurations e.g. Mobile Push Notification Abbreviated Numbers,

SMS Short Codes, Email Dedicated IP Addresses, permitted message types and lengths configured per channel.

2. **Campaign.** There are different types of campaigns. Scheduled campaigns include schedules, frequency, segments, holdouts, message templates and A/B tests. Event-based campaigns also include trigger or event rules to replace schedules.
3. **Segmentation.** That can be defined through a set of filters driven off user and user engagement data, or they can be defined through data imports and ingested as lists from files extracted from external systems.
4. **Scheduling Configuration.** Scheduling is primarily assigned to campaign-based engagement and defines the frequency and precise time of sending messages.
5. Custom Attributes Configuration. Defines custom attributes and events that applications can capture and return to Amazon Pinpoint Engagement Data.

   User Data which provides Amazon Pinpoint with endpoint information for sending messages across any channel, device or application. User data is extensible, but includes the following per channel:

   - Mobile Push Notifications. This includes userID, appID, appVersion, DeviceID, DeviceModel, DeviceModelVersion, Device OS, OS version, lastTimezone, lastCity, lastCountry, lastLatitude, lastLongitude, lastPostalCode, lastRegion.

   - SMS. Number.

   - Email. PrimaryWorkEmailAddress, SecondaryWorkEmailAddress, PrimaryPersonalEmailAddress, SecondaryPersonalEmailAddress.

   User data can also include:

   - External User Identifiers. Which map users to the IDs in external systems.

   - Custom Attributes. Users can add custom attributes to associate various data that is customized to their utilization of Amazon Pinpoint.

   User Engagement Data which includes default data per channel as well as custom data attributes when configured. Data per channel includes:

   External Data can include any user, segmentation and analytics data.

Q: What are the options for capturing custom application events?

A: You can either use the Mobile SDK within your mobile application to send custom events and attributes for Mobile Push Notifications, or use the Amazon Pinpoint REST API to send events programmatically from any application.

Q: Does Amazon Pinpoint support cross-device/application identity management?

A: Yes. This is captured under User ID.

Q: What OS versions does Amazon Pinpoint support for Mobile Push Notifications?

A: The iOS SDK supports apps running on iOS 7.0 and higher. The Android SDK supports apps running on Android 2.3.3 and higher.

Q: For mobile push notifications is data cached when a user's device is offline?

A: Yes, when using the AWS Mobile SDK, data is cached on the user's device and is uploaded when a network connection is next established.

Q: Are network channels optimized when sending events via the SDK?

A: Yes, the events are batched, and sent once per minute. You can also specify the transport to send the events: cellular and Wi-Fi, or Wi-Fi only.

Analytics

Q: What types of analytics does Amazon Pinpoint provide?

A: Amazon Pinpoint offers several types of standard analytics that provide insight into how customers use your mobile and web applications, how your engagement efforts are performing, and the impact your engagement efforts have on your business outcomes. Standard analytics include metrics for active users, user activities and demographics, sessions, user retention, campaign efficacy, and transactional messages. Using these metrics in combination with the analytics tools on the console, you can perform in-depth analysis by filtering on certain segments, custom attributes, and more. You can also create funnel reports.

Q: Where can I access analytics data?

A: You can view analytics data on the Amazon Pinpoint console. For each of your projects, the console provides detailed charts and metrics that provide insight into areas such as customer demographics, application usage, purchase activity, and delivery and engagement rates for campaigns. You can also access a subset of these metrics programmatically by using the Amazon Pinpoint API.

Q: How long does Amazon Pinpoint store analytics data?

A: Amazon Pinpoint automatically stores your analytics data for 90 days. You can see your data on the console or you can query a subset of data programmatically using the Amazon Pinpoint API. To keep the data for a longer period of time, you can export data from the console to comma-separated values (.csv) files or configure Amazon Pinpoint to stream event data to Amazon Kinesis. Kinesis is an AWS service that can collect, process, and analyze data from other AWS services in real-time. Amazon Pinpoint can send event data to Kinesis Data Firehose, which streams data to AWS data stores such as Amazon S3 or Amazon Redshift. Amazon Pinpoint can also stream data to Kinesis Data Streams, which ingests and stores multiple data streams for processing in analytics applications.

Q: Can Amazon Pinpoint tell if a single user uses the same app on more than one device (for example, on their phone and on a tablet device)?

A: Amazon Pinpoint distinguishes between endpoints and users. An endpoint is a destination that you can send messages to—such as a user's mobile device, email address, or phone number. A user is an individual who has a unique user ID. This ID can be associated with one or more endpoints.

Some of the Amazon Pinpoint analytics charts report on endpoints, and some report on users. To learn more about the individual charts, see Chart Reference for Amazon Pinpoint Analytics in the Amazon Pinpoint User Guide.

Q: How is a "session" defined?

A: A session is one use of an app by the user. A session begins when an app is launched (or brought to the foreground), and ends when the app is terminated (or goes to the background). To accommodate for brief interruptions, like a text message, an inactivity period of up to 5 seconds is not counted as a new session. Total daily sessions shows the number of sessions your app has each day. Average sessions per daily active user shows the mean number of sessions per user per day.

Q: When an app goes to the background does its session end?

A: Yes, the session ends. When the app comes to the foreground, a new session begins.

Q: How are daily and weekly retention defined?

A: Daily retention is measured by determining the number of users that first used your app on a specific day, came back and used your app in the next 7 days (7-day retention), fourteen days (14-day retention), and thirty days (30-day retention).

Q: What is "sticky factor," and how is it calculated?

A: The sticky factor represents the number of monthly users who used the app on a particular day.

Sticky factor is calculated by dividing daily active users (DAU) by monthly active users (MAU). For example, if an app has 100,000 DAU and 300,000 MAU, its sticky factor is .33. A high sticky factor can indicate strong engagement, appeal, and opportunities for monetization.

Q: What are demographics in Amazon Pinpoint?

A: The Demographics charts provide information about the device attributes for your app users. You can also see custom attributes that you define.

Events

Q: What are custom events?

A: Custom events metrics that you define. They help track user actions specific to your app or game. The Events charts provide a view of how often custom events occur. Custom events can be filtered based on attributes and their associated values.

You create custom events by naming them, such as "Item Bought" or "Button Pressed," and then adding context by specifying attributes (for qualitative measures) and metrics (for quantitative measures). For example, if your business goal is to track purchases of items from within the app, you can use "Item Bought" as a custom event, "Item XYZ" as an attribute, and "Item Price" as the metric. The custom events report enables you to search and filter for each attribute or metric. For example, you can find how often "Item XYZ" was purchased or how often “Item Price” was $1.99. You can also review the weighted average of metric values (per session) and track minimum, maximum, or average metric values. As a best practice, we recommend that custom event names be broad and attribute names be specific.

Q: What are the benefits of using custom events?

A: Custom events help you understand the actions that users take when using your app. For example, a game developer might want to understand both how often a level is completed and how much health each player has left at the end of a level. With custom events, you can create an event called "level_complete", with "add_level" as an attribute, and "health" as an attribute value. Each time a level is completed, you can record a "level_complete" event with the name of the level and the player's health. By reviewing the Events charts, you might discover that a level is too easy because players always finish with maximum health. Using this data, you can adjust the level's difficulty to better challenge and engage players, which might improve retention.

You can also use custom events to create event-based campaigns that are sent when your customers take specific actions within your applications. For example, you can set up a campaign to send a message when a customer creates a new account, when they spend a certain dollar amount, or when they add an item to their cart but don't purchase it.

Event-based campaigns help you send messages that are timely, personalized, and relevant to your customers, which ultimately increases their trust in your brand and gives them a reason to return. You can create event-based campaigns by using the Amazon Pinpoint console, or by using the Amazon Pinpoint API.

Q: Are there limits for using custom events in my app?

A: You can have up to 1,500 unique custom event types per app and up to 40 attributes and metrics per custom event. For more information, see Reporting Events in Your Application in the Amazon Pinpoint Developer Guide.

Campaigns

Q: What are campaigns?

A: Campaigns are messages that you send to a targeted segment of users on a predefined schedule. You can use targeted campaigns to increase customer engagement and retention. You can create campaigns for use cases such as welcoming new customers, informing customers of new features in your apps, and promoting special offers and deals.

Q: What's a standard campaign?

A: Standard campaigns include a target segment, a message, and a schedule for sending the message. You can also reuse previously defined segments or define a new segment when you create a campaign. For every scheduled campaign, Amazon Pinpoint recalculates the current audience size based on the criteria associated with the segment.

Q: What's an A/B test campaign?

A: A/B campaigns are campaigns with more than one treatment. Each treatment differs from the other based on the message or the sending schedule. You can compare the response rates for each treatment to determine which one had a bigger impact on your customers.

Q: What are silent push notifications and how do I use them?

A: Silent and in-app notifications are messages that are delivered to customers' devices, but aren't displayed on the devices. You can use these messages to manage the configuration of your app, or to deliver messages to the notification center within your app.

Q: What metrics does Amazon Pinpoint track for standard campaigns?

A: For standard campaigns, you can track messages sent, messages delivered, direct app opens, sessions per user, purchases per user, delivery rate, open rate, user devices messaged, and campaign sessions by time of day.

Q: What are my scheduling options for campaigns?

A: During campaign set up in Amazon Pinpoint, you can choose when the campaign should be sent. You have two options, you can send the campaign at a specific time, or you can send it when an event occurs. Time-based campaigns can be scheduled to run one time immediately or at a time you designate in the future. They can also be scheduled with multiple runs—hourly, daily, weekly, or monthly. To define your recurring campaigns, choose a start date and an end date, and specify whether or not deliver messages based on each recipient's local time zone.

You can also use Amazon Pinpoint to create campaigns that send messages, such as text messages, push notifications, and emails, to your customers when they take specific actions within your apps. You can create event-based campaigns by using the Amazon Pinpoint console, or by using the Amazon Pinpoint API. Event-based campaigns are an effective way to implement both transactional and targeted campaign use cases. Rather than define a time to send your message to customers, you select specific events, attributes, and metric values that you want to use to trigger your campaigns. For more information about event-based campaigns, see our blog post.

Q: What are event-based campaigns?

A: Event-based campaigns send messages, such as text messages, push notifications, and emails, to your customers when they take specific actions within your applications, such as making purchases or watching a video. For example, you can set up a campaign to send a message when a customer creates a new account or when they add an item to their cart but don’t purchase it. You can create event-based campaigns by using the Amazon Pinpoint console, or by using the Amazon Pinpoint API. Event-based campaigns are an effective way to implement both transactional use cases, such as one-time-password and order confirmation messages, and targeted uses cases, such as marketing promotions. Rather than define a time to send your message to customers, you select specific events, attributes, and metric values that you want to use to trigger your campaigns. For more information about event-based campaigns, please view this blog post.

Q: How do I get started with event-based campaigns?

A: The first step in setting up an event-based campaign is to create a new campaign. On step 4 of the campaign creation process, you choose when the campaign should be sent. You can choose to send the campaign at a specific time, or you can send it when an event occurs. Choose **"When an event occurs"**, and then choose the events, attributes, and metrics that trigger your campaign.

Q: What is the cost of event-based campaigns?

A: There are no additional charges associated with creating event-based campaigns. You pay only for the number of endpoints that you target, the number of messages that you send, and the number of analytics events that you send to Amazon Pinpoint. To learn more, see our Pricing page.

Q: How can I learn more about event-based campaigns and best practices?

A: For more information and best practices, see our blog post and the Amazon Pinpoint User Guide.

Q: What are the limitations of event-based campaigns?

A: There are a few limitations to be aware of when creating an event-based campaign. For more information, see the Amazon Pinpoint User Guide.

Q: Can I create and schedule event-based campaigns on the voice channel?

A: No. Currently, you can only send event-based campaigns in the SMS, push notification, and email channels.

Q: Can I use server-side events to trigger my campaigns?

A: Not yet. Amazon Pinpoint only lets you execute campaigns based on events that are associated with individual customers. Specifically, you can only trigger campaigns based on actions that users take in the applications, and that are reported by the AWS Mobile SDK.

Q: Can I create and schedule a voice message campaign?

A: No, the voice channel is only available for transactional messages. You can use the Amazon Pinpoint API to deliver transactional voice messages—such as new account creation notifications or order confirmations—directly to specific recipients.

Q: How do campaign limits work?

A: On the **General Settings** page of the Amazon Pinpoint console, you can configure the maximum number of messages an endpoint can receive for a campaign. This feature is useful when you want to place strict limits on the number of messages that an endpoint can receive from a campaign. For example, if you create a campaign that's automatically sent to all new customers, you can set the limit to 1. This setting ensures that new customers only receive the message once.

It's important to note that this feature is based on the number of messages that *target* an endpoint, as opposed to the number of messages that are actually *delivered* to an endpoint. If a campaign is configured to automatically send a message when a customer creates a new account, but the endpoint isn't able to receive the message for some reason (for example, if the quiet time setting for your campaign applies to the endpoint), then the endpoint is still counted as having been targeted. In this situation, the endpoint is removed from subsequent runs of the campaign.

Journeys

Q: What are journeys?

A: In Amazon Pinpoint, journeys are fully automated, end-to-end messaging solutions for engaging with your customers. Picture a flowchart: actions lead to other actions, sometimes branching into several paths, each with their own unique sets of activities.

You can use journeys to implement a variety of use cases, including customer onboarding scenarios and churn prevention programs. Journeys are flexibile and extensible, giving your teams the control they need to build powerful customer experiences without writing code using an easy-to-use graphical interface.

Q: Are there any prerequisites for creating a journey?

A: You have to have an active AWS account. You also have to set up the email channel and create a project in Amazon Pinpoint.

Q: What is an activity in a journey?

A: Journey activities are the configurable components that make up a journey. These components have different functionality, and you configure them to create the experience you want to build.

For example, an **Email** activity sends an email to every journey participant who arrives on it. A **Wait** activity prevents journey participants from proceeding to the next activity in the journey until a specific date and time, or until a certain amount of time elapses. A **Multivariate split** activity sends journey participants down one of up to five unique paths based on their segment membership, or based on their interactions with messages that you sent earlier in the journey.

To learn more about these activities, see Journeys in the *Amazon Pinpoint User Guide*.

Q: What metrics does Amazon Pinpoint provide for my journeys?

A: You can view metrics for a journey using the same web-based management console that you use to create a journey. From the console, you can quickly determine how many participants entered the journey, as well as the number that arrived on each individual step.

You can also export all of the metrics for your journeys to your preferred destination by using an Amazon Kinesis Data Stream or an Amazon Kinesis Data Firehose stream. This capability makes it possible to perform in-depth, post-journey analyses, or to store your data for an extended period time.

Additional charges apply to exporting data using Amazon Kinesis. For more information, see Amazon Kinesis Data Stream Pricing and Amazon Kinesis Data Firehose Pricing.

Q: Can I schedule my journeys?

A: You can configure each journey to start and end at a specific time. Each journey can run continuously for up to 18 months.

You can also schedule how often new participants enter the journey. When you create a journey, you specify a segment of customers that participate in it. You can set up your journey so that this segment is updated on a regular basis—hourly, daily, monthly, quarterly, annually, or not at all.

Q: What can I do if I make a mistake in my journey?

A: Journeys includes a built-in review process that checks for show-stopping errors, while also providing recommendations and best practices. You have to complete this review process before you launch each journey.

Journeys also includes a test feature, which makes it easy to send a group of test participants through your journey. By testing your journey, you can ensure that it behaves the way that you expect it to behave.

If you encounter issues with your journey while it's running, you can stop it at any time. When you stop a journey, participants halt on the activity they're currently on, and never proceed to the next activity.

Q: How much does it cost to use journeys?

A: There's no additional cost for using journeys. You pay for the customers that you target, and for the messages that you send. To learn more, see Amazon Pinpoint Pricing.

Q: Can I use server-side events to trigger journeys?

A: Currently, you can only insert customers into a journey based on their membership in a segment.

However, you can create dynamic segments that are updated using Lambda functions. When a server-side event occurs, you can use Lambda to update your endpoints, adding them to the target segment for your journey. For more information, see Customizing Segments with AWS Lambda in the *Amazon Pinpoint Developer Guide*.

Q: I'd like to see a feature or capability added to journeys. How can I provide feedback?

A: Our product development roadmap is driven by customer feedback, so we always love to hear your feedback. You can let us know what you think of journeys by clicking the "How do you like journeys?" button in the bottom right corner of the journeys workspace.

We read all of the messages that we receive. If we have any questions about your feedback, we'll contact you directly for more information. However, note that we might not be able to respond to every message that we receive.

Text Messaging

Q: Can I use dedicated phone numbers to send SMS messages?

You can use dedicated short codes or dedicated long codes to send SMS messages to recipients in several countries and regions. The following sections discuss the benefits and disadvantages of each of these options.

Short codes

- **Benefit**: Capable of sending hundreds of SMS messages per second (the actual rate varies based on the country where the short code is based).

- **Benefit**: Easy for recipients to identify and remember.
- **Disadvantage**: Usually more expensive than a long code.
- **Disadvantage**: Often takes several weeks to provision.

Long codes

- **Benefit**: Usually less expensive than a dedicated short code.
- **Benefit**: Typically takes days to provision (actual time varies by country).
- **Benefit**: Can be used to send both SMS messages and voice calls in the US and Canada.
- **Disadvantage**: Allows you to send fewer messages per second (usually between 1 and 15, depending on the country where the long code is based).
- **Disadvantage**: Can be subject to daily sending limits (varies based on the country and mobile carrier of the recipient).
- **Disadvantage**: Harder for recipients to remember and identify.

For more information about obtaining dedicated short and long codes, see the Amazon Pinpoint User Guide.

Q: What is two-way text messaging?

A: Two-way text messaging enables customers of Amazon Pinpoint to receive text messages from their users. When a user sends a text message to a customer's leased number, Amazon Pinpoint passes the text message to the customer, and the customer can use this message to trigger an appropriate response. Depending on the country and the local telecommunication regulation requirements, the customer can use long codes (10-digit phone numbers) and short codes (5 to 6-digit phone numbers).

To receive text messages from their users, the customer enables two-way text messaging in the Amazon Pinpoint console and selects an Amazon SNS topic to receive the text messages. Amazon Pinpoint provides the telephone number of the user and the customer's message ID if the text message was sent to the customer as a reply.

Q: What are the advantages of two-way text messaging?

A: Two-way text messaging enables several customer engagement use cases. For example, a financial services company could use two-way SMS messaging to send confirmation messages when they detect that a customer's account was used to make an unusual purchase. If the customer responds to the message stating that they initiated the purchase, the institution can authorize it.

Appointment confirmations are another common use case for two-way SMS messaging. For example, medical practices, salons, or restaurants can send messages to confirm a customer's appointment or reservation. The customer can respond, indicating whether or not they can keep their appointment. If the customer can't keep their appointment, you can ask if they want to reschedule, and send them the time and date of the next available appointment.

Q: Why do you require a dedicated number for two-way text messaging?

A: Receiving text messages from the same company through multiple numbers makes it hard for users to associate a number with a single business with which they regularly interact. A dedicated number makes it easier for users to participate in two-way text messaging.

Additionally, Amazon Pinpoint delivers text messages using numbers that are shared by multiple customers. Since these numbers are not exclusive to a sender, it is not possible to accurately route the text message to an appropriate customer when a cellular subscriber texts a number that is owned by Amazon Pinpoint. Because of these reasons, Amazon Pinpoint requires companies to lease a dedicated number for two-way communication.

Q: How can I disable two-way text messaging?

A: You can disable two-way text messaging from the Amazon Pinpoint console. When you disable this feature, you stop receiving incoming text messages from your customers.

Q: Can I create automated responses for specific keywords?

A: Yes. If you have a leased phone number (either a long code or a short code), you can use the Amazon Pinpoint console to create responses to specific keywords. When a customer sends a text message that matches the keyword, Amazon Pinpoint sends the corresponding response to the customer. You can also customize the messages that customers receive when they send HELP or STOP messages to your numbers.

Data Privacy

Q. Does Amazon Pinpoint store my customer data?

A: Yes. Amazon Pinpoint stores user, endpoint, and event data. We have to retain this data so that you can create segments, send messages to recipients, and capture application and campaign engagement data.

Q. Who can access the data stored in Amazon Pinpoint?

A: A very limited number of authorized employees have access to the data stored in your Amazon Pinpoint account.

Maintaining your trust is our highest priority. We use sophisticated physical and technical controls to safeguard your privacy and ensure the security of your data.

Your data is encrypted at rest and during transit. Our processes are designed to prevent unauthorized access to or disclosure of your content.

For more information, see the AWS Data Privacy FAQ.

Q: Do I own my content that is processed and stored by Amazon Pinpoint?

A: You always retain ownership of your content. We only use your content with your consent.

Q: How do I delete the data that Amazon Pinpoint stores?

A: You can selectively delete the data stored in your Amazon Pinpoint account. You can also close your entire AWS account, which deletes all of the data stored in Amazon Pinpoint and all other AWS services in every AWS Region. For more information, see Deleting Data from Amazon Pinpoint in the *Amazon Pinpoint Developer Guide*.

Contacting Us

Q: I received spam or other unsolicited email messages from an Amazon Pinpoint user. How do I report these messages?

A: You can report email abuse by sending an email to email-abuse@amazon.com.

To help us handle the issue as quickly and effectively as possible, please include the full headers of the original email. For procedures for obtaining email headers for several common email clients, see How to Get Email Headers on the MxToolbox.com website.

Q: How can I submit feature requests or send other product feedback about Amazon Pinpoint?

A: Your AWS Account Manager can send your feature requests and feedback directly to the appropriate team. If you don't currently have an AWS Account Manager, you can also provide your feedback on the Amazon Pinpoint forum.

Q: How can I get technical support for Amazon Pinpoint?

A: If you have an AWS Support plan, you can create a new support case directly from the web-based AWS management console. AWS Support plans begin at $29 per month. For more information about AWS Support plans, visit https://aws.amazon.com/premiumsupport/.

To open a new technical support case

1. In the console, on the **Support** menu, choose **Support Center**.
2. Next, choose **Create case**.
3. On the **Create case** page, choose **Technical support**.
4. Provide information about the issue you're experiencing, and then submit the ticket.

If you don't have an AWS Support plan, you can also ask questions and get answers on the Amazon Pinpoint forum.

AWS AppSync FAQs

Q. What is AWS AppSync?

AWS AppSync is a new service that enables developers to manage and synchronize mobile app data in real time across devices and users, but still allows the data to be accessed and altered when the mobile device is in an offline state.

The service further allows developers to optimize the user experience by selecting which data is automatically synchronized to each user's device when changes are made, minimizing storage and bandwidth requirements, with a query language called GraphQL.

Using these capabilities, developers can, in minutes, build real time collaborative experiences spanning browsers, mobile apps, Alexa skills, and IoT devices that remain usable when network connectivity is lost.

Q. What types of apps can I build using the features released today?

AWS AppSync can be used to build mobile apps that would benefit from being able to synchronize user and app data across devices, continue functioning when disconnected, and offer real-time collaboration experiences. There are applications across all verticals. Examples include:

- Gaming apps with real-time scoreboards
- News feeds and financial data
- Customer service dashboards
- Shared wallet, travel or itinerary tracking with offline usage
- Social Media with content feeds and search/discovery/messaging
- Dating apps with likes, messaging and geo/proximity awareness

- Field service apps that need to allow for querying and CRUD operations, even when disconnected
- Document collaboration
- 3D collaboration such as shared whiteboards
- AR/VR with multiple actors (doctors in surgery with observers, teachers and students)
- Multi-device (e.g., Alexa, mobile, web, IoT) and multi-modal applications (e.g., task list) that need to work offline yet reflect the same eventually consistent state
- Chat apps, including presence indicators and conversation history

Q. What application developer languages are supported in AWS AppSync?

AWS AppSync SDKs support iOS, Android, and JavaScript. The JavaScript support spans web frameworks such as React and Angular as well as technologies such as React Native and Ionic. You can also use open source clients to connect to the AppSync GraphQL endpoint for using other platform such as generic HTTP libraries or even a simple CURL commands.

Q. What is GraphQL ?

GraphQL is a data language to enable client apps to fetch, change and subscribe to data from servers. In a GraphQL query, the client specifies how the data is to be structured when it is returned by the server. This makes it possible for the client to query only for the data it needs, in the format that it needs it in.

Q. What is a GraphQL Schema?

A GraphQL schema is a definition of what data capabilities are available for the client application to operate on. For example, a schema might say what queries are available or how an app can subscribe to data without needing to know about the underlying data source. Schemas are defined by a type system, which an application's data model can leverage.

Q. Do I need to know GraphQL to get started?

No, AWS AppSync can automatically setup your entire API, schema, and connect data sources with a simple UI builder that allows you to type in your data model in seconds. You can then immediately begin using the endpoint in a client application. The console also provides many sample schema and data sources for fully functioning applications.

Q. Can I use AWS AppSync with my existing AWS resources?

Yes. With AWS AppSync you can use existing tables, functions, and domains from Amazon DynamoDB, AWS Lambda and Amazon Elasticsearch Service with a GraphQL schema. AWS AppSync allows you to create data sources using existing AWS resources and configure the interactions using Mapping Templates.

Q. What is a Mapping Template?

GraphQL requests execute as "resolvers" and need to be converted into the appropriate message format for the different AWS Services that AWS AppSync integrates. For example, a GraphQL query on a field will need to be converted into a unique format for Amazon DynamoDB, AWS Lambda, and Amazon Elasticsearch Service respectively. AWS AppSync provides Mapping Templates for this, which are written in Apache Velocity Template Language (VTL) allowing you to provide custom logic to meet your needs. AWS AppSync also provides built-in templates for the different services and utility functions for enhanced usability.

Q. How is data secured with AWS AppSync?

Application data is stored at rest in your AWS account and not in the AWS AppSync service. You can protect access to this data from applications by using security controls with AWS AppSync including AWS Identity and Access Management (IAM), as well as Amazon Cognito User Pools. Additionally, user context can be passed through for authenticated requests so that you can perform fine-grained access control logic against your resources with Mapping Templates in AWS AppSync.

Q. Can I make my data real-time with AWS AppSync?

Yes. Subscriptions are supported with AWS AppSync against any of the data sources, so that when a mutation occurs, the results can be passed down to clients subscribing to the event stream immediately using either MQTT over WebSockets or pure WebSockets.

Q. How can I do complex queries with AWS AppSync?

The data sources available to AWS AppSync allow you to take full advantage of capabilities provided by Amazon DynamoDB, Amazon Elasticsearch Service, and AWS Lambda when using GraphQL. Features such as indexing and conditional checks, along with Mapping Templates, return comprehensive results from DynamoDB. Use cases such as fuzzy searches, geo searches and more that Amazon Elasticsearch Service offers are available to your application. Finally, Lambda can be used for serial or batched requests to return data from other sources such as Amazon Aurora.

Q. What AWS Regions are available for AWS AppSync?

AWS AppSync is available in US East (N. Virginia), US West (Oregon), US East (Ohio), EU (Ireland), EU (Frankfurt), Asia Pacific (Tokyo), Asia Pacific (Sydney), Asia Pacific (Mumbai) and Asia Pacific (Singapore).

Q. Can I import existing Amazon DynamoDB tables?

AWS AppSync can automatically generate a GraphQL schema from an existing DynamoDB table, including the inference of your table’s key schema and indexes. Once the import is complete GraphQL queries, mutations, and subscriptions can be used with zero coding. AppSync will also “auto-map” non-key attributes from your GraphQL types to DynamoDB attributes.

Q. Can AWS AppSync create a database for me?

Customers can create a GraphQL schema, either by hand or using the console, and AWS AppSync can automatically provision Amazon DynamoDB tables and appropriate indexes for you. Additionally, it will connect the data sources to "GraphQL resolvers" allowing you to just focus on your application code and data structures.

Q. What clients can I use to connect my application to my AppSync API?

You can use any HTTP or GraphQL client to connect to a GraphQL API on AppSync. We do recommend using the Amplify clients which are optimized to connect to the AppSync backend. There are some options depending on your application's use case:

- For DynamoDB data sources, use the DataStore category in the Amplify client. It provides the best developer experience and built-in conflict detection and resolution.
- For non-DynamoDB data sources in scenarios where you have no offline requirements, use the API (GraphQL) category in the Amplify client.
- For non-DynamoDB data sources in scenarios where you have offline requirements, use the AppSync SDK.

AWS Device Farm FAQs

Testing on real mobile devices

General

Q: What is AWS Device Farm? >>

Q: Who should use AWS Device Farm and why? >>

Q: What types of apps does AWS Device Farm support? >>

Q: Does AWS Device Farm use simulators or emulators? >>

Setting up tests & remote access sessions

Q: How do I get started with AWS Device Farm? >>

Q: Which browsers does the AWS Device Farm console support? >>

Q: Which browsers are supported for testing web applications? >>

Q: What is the maximum file size for apps and tests? >>

Q: Do I need to instrument my app or supply source code? >>

Q: Do you store my app, tests, and other files on your servers? For how long? >>

Q: How do you clean up devices after my testing is completed? >>

Q: Do you modify my app? >>

Selecting devices

Q: Which devices are available in AWS Device Farm? How do you select the devices in your fleet? >>

Q: Does AWS Device Farm have international devices from markets like Europe, China, and India? >>

Q: How do I select devices? Can I retest on the same device? >>

Q: Are any apps pre-installed on AWS Device Farm test devices? >>

Q: Are devices able to communicate with other services or systems that are available on the Internet? >>

Q: Can I test different carrier connections and conditions? >>

Q: Can I make phone calls or send SMS from the devices? >>

Q: Can I use the device camera? >>

Testing your app

Q: I don’t have any automated test scripts yet. What do the built-in tests do? >>

Q: What does Fuzz do? >>

Q: I test using an automation framework. Which frameworks do you support? >>

Q: Which test frameworks do you support for web applications? >>

Q: Can you add support for a modified framework or one I designed myself? >>

Q: How does AWS Device Farm decide when to take a screenshot during a test? >>

Q: Android: Is Google Play Services available on your devices? Which version is installed? >>

Q: Android: Is there a default Google account on the devices? >>

Q: Does AWS Device Farm support record and playback automation or do I have to write my scripts? >>

Q: iOS: Do I need to add your UDIDs to my provisioning profile? >>

Q: iOS: My app does not contain debug symbols. Can I supply a dSYM file to AWS Device Farm? >>

Q: Android: My app is obfuscated. Can I still test my app on AWS Device Farm? >>

Q: My app serves ads. Will they be displayed on your devices? Will my ad provider flag this as abuse and ban my account? >>

Q: Can I access the machine hosting the device or access its shell as part of my tests? Can I reach the Internet from it? >>

Q: I'd like to supply media or other data for my app to consume. How do I do that? >>

Q: My app requires dependencies to test all functionality. Can I install other apps? >>

Q: Can I test upgrade flows for my app? How do I install an old version of my app? >>

Q: My app makes use of location services. Can I specify the physical location of the device? >>

Q: Can I run localization tests? How do I change the language of the device? >>

Q: How long does it take before my test starts? >>

Q: What is the maximum test time allowed? >>

Q: Does AWS Device Farm provide a way to run tests and get results through an API? >>

Reviewing results

Q: What’s in an AWS Device Farm test report? >>

Q: Which device logs are included in an AWS Device Farm report? >>

Q: My tests generate and save additional log files. Will I see them in my AWS Device Farm reports? >>

Pricing

Q: How much does AWS Device Farm cost? >>

Q: How does the free trial work? >>

Q: What is a device minute? >>

Q: How does the free trial work? >>

Q: What is the unmetered plan and how do device slots work? >>

Q: What if my testing needs change and I need to add or remove device slots? >>

Q: If I'm on an unmetered plan, can I still make use of metered billing? >>

Q: What is a private device? >>

Q: How do private device subscriptions work and how are they priced? >>

Q: Can I use both private devices and public devices? >>

Testing on desktop browsers

Q: What is Selenium? >>

Q: What is Desktop Browser Testing on AWS Device Farm? >>

Q: How do I get started with Desktop Browser Testing on AWS Device Farm? >>

Q: What operating system are the browsers hosted on? >>

Q: What desktop browsers does AWS Device Farm support? >>

Q: What desired capabilities does AWS Device Farm support? >>

Q: What artifacts are available for troubleshooting test failures? >>

Q: Can I use AWS Device Farm to test my web app on real mobile devices? >>

Q: What are limits of Desktop Browser Testing on AWS Device Farm? >>

Q: How much does Desktop Browser Testing on AWS Device Farm cost? >>

Q: What is instance minute? >>

Amazon VPC FAQs

General Questions

Q. What is Amazon Virtual Private Cloud?

Amazon VPC lets you provision a logically isolated section of the Amazon Web Services (AWS) cloud where you can launch AWS resources in a virtual network that you define. You have complete control over your virtual networking environment, including selection of your own IP address ranges, creation of subnets, and configuration of route tables and network gateways. You can also create a hardware Virtual Private Network (VPN) connection between your corporate datacenter and your VPC and leverage the AWS cloud as an extension of your corporate datacenter.

You can easily customize the network configuration for your Amazon VPC. For example, you can create a public-facing subnet for your web servers that have access to the Internet, and place your backend systems such as databases or application servers in a private-facing subnet with no Internet access. You can leverage multiple layers of security, including security groups and network access control lists, to help control access to Amazon EC2 instances in each subnet.

Q. What are the components of Amazon VPC?

Amazon VPC comprises a variety of objects that will be familiar to customers with existing networks:

- **A Virtual Private Cloud:** A logically isolated virtual network in the AWS cloud. You define a VPC's IP address space from ranges you select.
- **Subnet:** A segment of a VPC's IP address range where you can place groups of isolated resources.
- **Internet Gateway:** The Amazon VPC side of a connection to the public Internet.

- **NAT Gateway:** A highly available, managed Network Address Translation (NAT) service for your resources in a private subnet to access the Internet.
- **Virtual private gateway:** The Amazon VPC side of a VPN connection.
- **Peering Connection:** A peering connection enables you to route traffic via private IP addresses between two peered VPCs.
- **VPC Endpoints:** Enables private connectivity to services hosted in AWS, from within your VPC without using an Internet Gateway, VPN, Network Address Translation (NAT) devices, or firewall proxies.
- **Egress-only Internet Gateway:** A stateful gateway to provide egress only access for IPv6 traffic from the VPC to the Internet.

Q: Why should I use Amazon VPC?

Amazon VPC enables you to build a virtual network in the AWS cloud - no VPNs, hardware, or physical datacenters required. You can define your own network space, and control how your network and the Amazon EC2 resources inside your network are exposed to the Internet. You can also leverage the enhanced security options in Amazon VPC to provide more granular access to and from the Amazon EC2 instances in your virtual network.

Q. How do I get started with Amazon VPC?

Your AWS resources are automatically provisioned in a ready-to-use default VPC. You can choose to create additional VPCs by going to the Amazon VPC page in the AWS Management Console and selecting "Start VPC Wizard".

You'll be presented with four basic options for network architectures. After selecting an option, you can modify the size and IP address range of the VPC and its subnets. If you select an option with Hardware VPN Access, you will need to specify the IP address of the VPN hardware on your network. You can modify the VPC to add or remove secondary IP ranges and gateways, or add more subnets to IP ranges.

The four options are:

1. Amazon VPC with a single public subnet only

2. Amazon VPC with public and private subnets
3. Amazon VPC with public and private subnets and AWS Site-to-Site VPN access
4. Amazon VPC with a private subnet only and AWS Site-to-Site VPN access

Q. What are the different types of VPC endpoints available on Amazon VPC?

VPC endpoints enable you to privately connect your VPC to services hosted on AWS without requiring an Internet gateway, a NAT device, VPN, or firewall proxies. Endpoints are horizontally scalable and highly available virtual devices that allow communication between instances in your VPC and AWS services. Amazon VPC offers two different types of endpoints: gateway type endpoints and interface type endpoints.

Gateway type endpoints are available only for AWS services including S3 and DynamoDB. These endpoints will add an entry to your route table you selected and route the traffic to the supported services through Amazon’s private network.

Interface type endpoints provide private connectivity to services powered by PrivateLink, being AWS services, your own services or SaaS solutions, and supports connectivity over Direct Connect. More AWS and SaaS solutions will be supported by these endpoints in the future. Please refer to VPC Pricing for the price of interface type endpoints.

Billing

Q. How will I be charged and billed for my use of Amazon VPC?

There are no additional charges for creating and using the VPC itself. Usage charges for other Amazon Web Services, including Amazon EC2, still apply at published rates for those resources, including data transfer charges. If you connect your VPC to your corporate datacenter using the optional hardware VPN connection, pricing is per VPN connection-hour (the amount of time you have a VPN connection in the "available" state.) Partial hours are billed as full hours. Data transferred over VPN connections will be charged at standard AWS Data Transfer rates. For VPC-VPN pricing information, please visit the pricing section of the Amazon VPC product page.

Q. What usage charges will I incur if I use other AWS services, such as Amazon S3, from Amazon EC2 instances in my VPC?

Usage charges for other Amazon Web Services, including Amazon EC2, still apply at published rates for those resources. Data transfer charges are not incurred when accessing Amazon Web Services, such as Amazon S3, via your VPC's Internet gateway.

If you access AWS resources via your VPN connection, you will incur Internet data transfer charges.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Connectivity

Q. What are the connectivity options for my Amazon VPC?

You may connect your Amazon VPC to:

- The internet (via an internet gateway)
- Your corporate data center using an AWS Site-to-Site VPN connection (via the virtual private gateway)
- Both the internet and your corporate data center (utilizing both an internet gateway and a virtual private gateway)
- Other AWS services (via internet gateway, NAT, virtual private gateway, or VPC endpoints)
- Other Amazon VPCs (via VPC peering connections)

Q. How do I connect my VPC to the Internet?

Amazon VPC supports the creation of an Internet gateway. This gateway enables Amazon EC2 instances in the VPC to directly access the Internet.

Q. Are there any bandwidth limitations for Internet gateways? Do I need to be concerned about its availability? Can it be a single point of failure?

No. An Internet gateway is horizontally-scaled, redundant, and highly available. It imposes no bandwidth constraints.

Q. How do instances in a VPC access the Internet?

You can use public IP addresses, including Elastic IP addresses (EIPs), to give instances in the VPC the ability to both directly communicate outbound to the Internet and to receive unsolicited inbound traffic from the Internet (e.g., web servers). You can also use the solutions in the next question.

Q. How do instances without public IP addresses access the Internet

Instances without public IP addresses can access the Internet in one of two ways:

1. Instances without public IP addresses can route their traffic through a NAT gateway or a NAT instance to access the Internet. These instances use the public IP address of the NAT gateway or NAT instance to traverse the Internet. The NAT gateway or NAT instance allows outbound communication but doesn’t allow machines on the Internet to initiate a connection to the privately addressed instances.
2. For VPCs with a hardware VPN connection or Direct Connect connection, instances can route their Internet traffic down the virtual private gateway to your existing datacenter. From there, it can access the Internet via your existing egress points and network security/monitoring devices.

Q. Can I connect to my VPC using a software VPN?

Yes. You may use a third-party software VPN to create a site to site or remote access VPN connection with your VPC via the Internet gateway.

Q. Does traffic go over the internet when two instances communicate using public IP addresses?

Traffic between two EC2 instances in the same AWS Region stays within the AWS network, even when it goes over public IP addresses.
Traffic between EC2 instances in different AWS Regions stays within the AWS network, if there is an Inter-Region VPC Peering connection between the VPCs where the two instances reside.
Traffic between EC2 instances in different AWS Regions where there is no Inter-Region VPC Peering connection between the VPCs where these instances reside, is not guaranteed to stay within the AWS network.

Q. How does an AWS Site-to-Site VPN connection work with Amazon VPC?

An AWS Site-to-Site VPN connection connects your VPC to your datacenter. Amazon supports Internet Protocol Security (IPSec) VPN connections. Data transferred between your VPC and datacenter routes over an encrypted VPN connection to help maintain the confidentiality and integrity of data in transit. An internet gateway is not required to establish an AWS Site-to-Site VPN connection.

IP Addressing

Q. What IP address ranges can I use within my Amazon VPC?

You can use any IPv4 address range, including RFC 1918 or publicly routable IP ranges, for the primary CIDR block. For the secondary CIDR blocks, certain restrictions apply. Publicly routable IP blocks are only reachable via the Virtual Private Gateway and cannot be accessed over the Internet through the Internet gateway. AWS does not advertise customer-owned IP address blocks to the Internet. You can allocate an Amazon-provided IPv6 CIDR block to a VPC by calling the relevant API or via the AWS Management Console.

Q. How do I assign IP address ranges to Amazon VPCs?

You assign a single Classless Internet Domain Routing (CIDR) IP address range as the primary CIDR block when you create a VPC and can add up to four (4) secondary CIDR blocks after creation of the VPC. Subnets within a VPC are addressed from these CIDR ranges by you. Please note that while you can create multiple VPCs with overlapping IP address ranges, doing so will prohibit you from connecting these VPCs to a common home network via the hardware VPN connection. For this reason we recommend using non-overlapping IP address ranges. You can allocate an Amazon-provided IPv6 CIDR block to your VPC.

Q. What IP address ranges are assigned to a default Amazon VPC?

Default VPCs are assigned a CIDR range of 172.31.0.0/16. Default subnets within a default VPC are assigned /20 netblocks within the VPC CIDR range.

Q. Can I advertise my VPC public IP address range to the internet and route the traffic through my datacenter, via the AWS Site-to-Site VPN, and to my Amazon VPC?

Yes, you can route traffic via the AWS Site-to-Site VPN connection and advertise the address range from your home network.

Q. Can I use my public IPv4 addresses in VPC and access them over the Internet?

Yes, you can bring your public IPv4 addresses into AWS VPC and statically allocate them to subnets and EC2 instances. To access these addresses over the Internet, you will have to advertise them to the Internet from your on-premises network. You will also have to route the traffic over these addresses between your VPC and on-premises network using AWS DX or AWS VPN connection. You can route the traffic from your VPC using the Virtual Private Gateway. Similarly, you can route the traffic from your on-premises network back to your VPC using your routers.

Q. How large of a VPC can I create?

Currently, Amazon VPC supports five (5) IP address ranges, one (1) primary and four (4) secondary for IPv4. Each of these ranges can be between /28 (in CIDR notation) and /16 in size. The IP address ranges of your VPC should not overlap with the IP address ranges of your existing network.

For IPv6, the VPC is a fixed size of /56 (in CIDR notation). A VPC can have both IPv4 and IPv6 CIDR blocks associated to it.

Q. Can I change the size of a VPC?

Yes. You can expand your existing VPC by adding four (4) secondary IPv4 IP ranges (CIDRs) to your VPC. You can shrink your VPC by deleting the secondary CIDR blocks you have added to your VPC. You cannot however change the size of the IPv6 address range of your VPC.

Q. How many subnets can I create per VPC?

Currently you can create 200 subnets per VPC. If you would like to create more, please submit a case at the support center.

Q. Is there a limit on how large or small a subnet can be?

The minimum size of a subnet is a /28 (or 14 IP addresses.) for IPv4. Subnets cannot be larger than the VPC in which they are created.

For IPv6, the subnet size is fixed to be a /64. Only one IPv6 CIDR block can be allocated to a subnet.

Q. Can I use all the IP addresses that I assign to a subnet?

No. Amazon reserves the first four (4) IP addresses and the last one (1) IP address of every subnet for IP networking purposes.

Q. How do I assign private IP addresses to Amazon EC2 instances within a VPC?

When you launch an Amazon EC2 instance within a VPC, you may optionally specify the primary private IP address for the instance. If you do not specify the primary private IP address, AWS automatically addresses it from the IP address range you assign to that subnet. You can assign secondary private IP addresses when you launch an instance, when you create an Elastic Network Interface, or any time after the instance has been launched or the interface has been created.

Q. Can I change the private IP addresses of an Amazon EC2 instance while it is running and/or stopped within a VPC?

Primary private IP addresses are retained for the instance's or interface's lifetime. Secondary private IP addresses can be assigned, unassigned, or moved between interfaces or instances at any time.

Q. If an Amazon EC2 instance is stopped within a VPC, can I launch another instance with the same IP address in the same VPC?

No. An IP address assigned to a running instance can only be used again by another instance once that original running instance is in a “terminated” state.

Q. Can I assign IP addresses for multiple instances simultaneously?

No. You can specify the IP address of one instance at a time when launching the instance.

Q. Can I assign any IP address to an instance?

You can assign any IP address to your instance as long as it is:

- Part of the associated subnet's IP address range
- Not reserved by Amazon for IP networking purposes
- Not currently assigned to another interface

Q. Can I assign multiple IP addresses to an instance?

Yes. You can assign one or more secondary private IP addresses to an Elastic Network Interface or an EC2 instance in Amazon VPC. The number of secondary private IP addresses you can assign depends on the instance type. See EC2 User Guide for more information on the number of secondary private IP addresses that can be assigned per instance type.

Q. Can I assign one or more Elastic IP (EIP) addresses to VPC-based Amazon EC2 instances?

Yes, however, the EIP addresses will only be reachable from the Internet (not over the VPN connection). Each EIP address must be associated with a unique private IP address on the instance. EIP addresses should only be used on instances in subnets configured to route their traffic directly to the Internet gateway. EIPs cannot be used on instances in subnets configured to use a NAT gateway or a NAT instance to access the Internet. This is applicable only for IPv4. Amazon VPCs do not support EIPs for IPv6 at this time.

Topology

Q. Can I specify which subnet will use which gateway as its default?

Yes. You may create a default route for each subnet. The default route can direct traffic to egress the VPC via the Internet gateway, the virtual private gateway, or the NAT gateway.

Security and Filtering

Q. How do I secure Amazon EC2 instances running within my VPC?

Amazon EC2 security groups can be used to help secure instances within an Amazon VPC. Security groups in a VPC enable you to specify both inbound and outbound network traffic that is allowed to or from each Amazon EC2 instance. Traffic which is not explicitly allowed to or from an instance is automatically denied.

In addition to security groups, network traffic entering and exiting each subnet can be allowed or denied via network Access Control Lists (ACLs).

Q. What are the differences between security groups in a VPC and network ACLs in a VPC?

Security groups in a VPC specify which traffic is allowed to or from an Amazon EC2 instance. Network ACLs operate at the subnet level and evaluate traffic entering and exiting a subnet. Network ACLs can be used to set both Allow and Deny rules. Network ACLs do not filter traffic between instances in the same subnet. In addition, network ACLs perform stateless filtering while security groups perform stateful filtering.

Q. What is the difference between stateful and stateless filtering?

Stateful filtering tracks the origin of a request and can automatically allow the reply to the request to be returned to the originating computer. For example, a stateful filter that allows inbound traffic to TCP port 80 on a webserver will allow the return traffic, usually on a high numbered port (e.g., destination TCP port 63, 912) to pass through the stateful filter between the client and the webserver. The filtering device maintains a state table that tracks the origin and destination port numbers and IP addresses. Only one rule is required on the filtering device: Allow traffic inbound to the web server on TCP port 80.

Stateless filtering, on the other hand, only examines the source or destination IP address and the destination port, ignoring whether the traffic is a new request or a reply to a request. In the above example, two rules would need to be implemented on the filtering device: one rule to allow traffic inbound to the web server on TCP port 80, and another rule to allow outbound traffic from the webserver (TCP port range 49, 152 through 65, 535).

Q. Within Amazon VPC, can I use SSH key pairs created for instances within Amazon EC2, and vice versa?

Yes.

Q. Can Amazon EC2 instances within a VPC communicate with Amazon EC2 instances not within a VPC?

Yes. If an Internet gateway has been configured, Amazon VPC traffic bound for Amazon EC2 instances not within a VPC traverses the Internet gateway and then enters the public AWS network to reach the EC2 instance. If an Internet gateway has not been configured, or if the instance is in a subnet configured to route through the virtual private gateway, the traffic traverses the VPN connection, egresses from your datacenter, and then re-enters the public AWS network.

Q. Can Amazon EC2 instances within a VPC in one region communicate with Amazon EC2 instances within a VPC in another region?

Yes. Instances in one region can communicate with each other using Inter-Region VPC Peering, public IP addresses, NAT gateway, NAT instances, VPN Connections or Direct Connect connections.

Q. Can Amazon EC2 instances within a VPC communicate with Amazon S3?

Yes. There are multiple options for your resources within a VPC to communicate with Amazon S3. You can use VPC Endpoint for S3, which makes sure all traffic remains within Amazon's network and enables you to apply additional access policies to your Amazon S3 traffic. You can use an Internet gateway to enable Internet access from your VPC and instances in the VPC can communicate with Amazon S3. You can also make all traffic to Amazon S3 traverse the Direct Connect or VPN connection, egress from your datacenter, and then re-enter the public AWS network.

Q. Can I monitor the network traffic in my VPC?

Yes. You can use Amazon VPC traffic mirroring and Amazon VPC flow logs features to monitor the network traffic in your Amazon VPC.

VPC Traffic Mirroring

Q. What is Amazon VPC traffic mirroring?

Amazon VPC traffic mirroring makes it easy for customers to replicate network traffic to and from an Amazon EC2 instance and forward it to out-of-band security and monitoring appliances for use-cases such as content inspection, threat monitoring, and troubleshooting. These appliances can be deployed on an individual EC2 instance or a fleet of instances behind a Network Load Balancer (NLB) with User Datagram Protocol (UDP) listener.

Q. How does Amazon VPC traffic mirroring work?

The traffic mirroring feature copies network traffic from Elastic Network Interface (ENI) of EC2 instances in your Amazon VPC. The mirrored traffic can be sent to another EC2 instance or to an NLB with a UDP listener. Traffic mirroring encapsulates all copied traffic with VXLAN headers. The mirror source and destination (monitoring appliances) can be in the same VPC or in a different VPC, connected via VPC peering or AWS Transit Gateway.

Q. Which resources can be monitored with Amazon VPC traffic mirroring ?

Traffic mirroring supports network packet captures at the Elastic Network Interface (ENI) level for EC2 instances. This feature is currently supported on all virtualized Nitro based EC2 instances.

Q. What type of appliances are supported with Amazon VPC traffic mirroring?

Customers can either use open source tools or choose from a wide-range of monitoring solution available on AWS Marketplace. Traffic mirroring allows customers to stream replicated traffic to any network packet collector/broker or analytics tool, without requiring them to install vendor-specific agents.

Q. How is Amazon VPC traffic mirroring different from Amazon VPC flow logs?

Amazon VPC flow logs allow customers to collect, store, and analyze network flow logs. The information captured in flow logs includes information about allowed and denied traffic, source and destination IP addresses, ports, protocol number, packet and byte counts, and an action (accept or reject). You can use this feature to troubleshoot connectivity and security issues and to make sure that the network access rules are working as expected.

Amazon VPC traffic mirroring, provides deeper insight into network traffic by allowing you to analyze actual traffic content, including payload, and is targeted for use-cases when you need to analyze the actual packets to determine the root cause a performance issue, reverse-engineer a sophisticated network attack, or detect and stop insider abuse or compromised workloads.

Amazon VPC and EC2

Q. Within which Amazon EC2 region(s) is Amazon VPC available?

Amazon VPC is currently available in multiple Availability Zones in all Amazon EC2 regions.

Q. Can a VPC span multiple Availability Zones?

Yes.

Q. Can a subnet span Availability Zones?

No. A subnet must reside within a single Availability Zone.

Q. How do I specify which Availability Zone my Amazon EC2 instances are launched in?

When you launch an Amazon EC2 instance, you must specify the subnet in which to launch the instance. The instance will be launched in the Availability Zone associated with the specified subnet.

Q. How do I determine which Availability Zone my subnets are located in?

When you create a subnet you must specify the Availability Zone in which to place the subnet. When using the VPC Wizard, you can select the subnet's Availability Zone in the wizard confirmation screen. When using the API or the CLI you can specify the Availability Zone for the subnet as you create the subnet. If you don't specify an Availability Zone, the default "No Preference" option will be selected and the subnet will be created in an available Availability Zone in the region.

Q. Am I charged for network bandwidth between instances in different subnets?

If the instances reside in subnets in different Availability Zones, you will be charged $0.01 per GB for data transfer.

Q. When I call DescribeInstances(), do I see all of my Amazon EC2 instances, including those in EC2-Classic and EC2-VPC?

Yes. DescribeInstances() will return all running Amazon EC2 instances. You can differentiate EC2-Classic instances from EC2-VPC instances by an entry in the subnet field. If there is a subnet ID listed, the instance is within a VPC.

Q. When I call DescribeVolumes(), do I see all of my Amazon EBS volumes, including those in EC2-Classic and EC2-VPC?

Yes. DescribeVolumes() will return all your EBS volumes.

Q. How many Amazon EC2 instances can I use within a VPC?

You can run any number of Amazon EC2 instances within a VPC, so long as your VPC is appropriately sized to have an IP address assigned to each instance. You are initially limited to launching 20 Amazon EC2 instances at any one time and a maximum VPC size of /16 (65,536 IPs). If you would like to increase these limits, please complete the following form.

Q. Can I use my existing AMIs in Amazon VPC?

You can use AMIs in Amazon VPC that are registered within the same region as your VPC. For example, you can use AMIs registered in us-east-1 with a VPC in us-east-1. More information is available in the Amazon EC2 Region and Availability Zone FAQ.

Q. Can I use my existing Amazon EBS snapshots?

Yes, you may use Amazon EBS snapshots if they are located in the same region as your VPC. More details are available in the Amazon EC2 Region and Availability Zone FAQ.

Q: Can I boot an Amazon EC2 instance from an Amazon EBS volume within Amazon VPC?

Yes, however, an instance launched in a VPC using an Amazon EBS-backed AMI maintains the same IP address when stopped and restarted. This is in contrast to similar instances launched outside a VPC, which get a new IP address. The IP addresses for any stopped instances in a subnet are considered unavailable.

Q. Can I use Amazon EC2 Reserved Instances with Amazon VPC?

Yes. You can reserve an instance in Amazon VPC when you purchase Reserved Instances. When computing your bill, AWS does not distinguish whether your instance runs in Amazon VPC or standard Amazon EC2. AWS automatically optimizes which instances are charged at the lower Reserved Instance rate to ensure you always pay the lowest amount. However, your instance reservation will be specific to Amazon VPC. Please see the Reserved Instances page for further details.

Q. Can I employ Amazon CloudWatch within Amazon VPC?

Yes.

Q. Can I employ Auto Scaling within Amazon VPC?

Yes.

Q. Can I launch Amazon EC2 Cluster Instances in a VPC?

Yes. Cluster instances are supported in Amazon VPC, however, not all instance types are available in all regions and Availability Zones.

Default VPCs

Q. What is a default VPC?

A default VPC is a logically isolated virtual network in the AWS cloud that is automatically created for your AWS account the first time you provision Amazon EC2 resources. When you launch an instance without specifying a subnet-ID, your instance will be launched in your default VPC.

Q. What are the benefits of a default VPC?

When you launch resources in a default VPC, you can benefit from the advanced networking functionalities of Amazon VPC (EC2-VPC) with the ease of use of Amazon EC2 (EC2-Classic). You can enjoy features such as changing security group membership on the fly, security group egress filtering, multiple IP addresses, and multiple network interfaces without having to explicitly create a VPC and launch instances in the VPC.

Q. What accounts are enabled for default VPC?

If your AWS account was created after March 18, 2013 your account may be able to launch resources in a default VPC. See this Forum Announcement to determine which regions have been enabled for the default VPC feature set. Also, accounts created prior to the listed dates may utilize default VPCs in any default VPC enabled region in which you've not previously launched EC2 instances or provisioned Amazon Elastic Load Balancing, Amazon RDS, Amazon ElastiCache, or Amazon Redshift resources.

Q. How can I tell if my account is configured to use a default VPC?

The Amazon EC2 console indicates which platforms you can launch instances in for the selected region, and whether you have a default VPC in that region. Verify that the region you'll use is selected in the navigation bar. On the Amazon EC2 console dashboard, look for "Supported Platforms" under "Account Attributes". If there are two values, EC2-Classic and EC2-VPC, you can launch instances into either platform. If there is one value, EC2-VPC, you can launch instances only into EC2-VPC. Your default VPC ID will be listed under "Account Attributes" if your account is configured to use a default VPC. You can also use the EC2 DescribeAccountAttributes API or CLI to describe your supported platforms.

Q. Will I need to know anything about Amazon VPC in order to use a default VPC?

No. You can use the AWS Management Console, AWS EC2 CLI, or the Amazon EC2 API to launch and manage EC2 instances and other AWS resources in a default VPC. AWS will automatically create a default VPC for you and will create a default subnet in each Availability Zone in the AWS region. Your default VPC will be connected to an Internet gateway and your instances will automatically receive public IP addresses, just like EC2-Classic.

Q. What are the differences between instances launched in EC2-Classic and EC2-VPC?

See Differences between EC2-Classic and EC2-VPC in the EC2 User Guide.

Q. Do I need to have a VPN connection to use a default VPC?

No. Default VPCs are attached to the Internet and all instances launched in default subnets in the default VPC automatically receive public IP addresses. You can add a VPN connection to your default VPC if you choose.

Q. Can I create other VPCs and use them in addition to my default VPC?

Yes. To launch an instance into nondefault VPCs you must specify a subnet-ID during instance launch.

Q. Can I create additional subnets in my default VPC, such as private subnets?

Yes. To launch into nondefault subnets, you can target your launches using the console or the --subnet option from the CLI, API, or SDK.

Q. How many default VPCs can I have?

You can have one default VPC in each AWS region where your Supported Platforms attribute is set to "EC2-VPC".

Q. What is the IP range of a default VPC?

The default VPC CIDR is 172.31.0.0/16. Default subnets use /20 CIDRs within the default VPC CIDR.

Q. How many default subnets are in a default VPC?

One default subnet is created for each Availability Zone in your default VPC.

Q. Can I specify which VPC is my default VPC?

Not at this time.

Q. Can I specify which subnets are my default subnets?

Not at this time.

Q. Can I delete a default VPC?

Yes, you can delete a default VPC. Once deleted, you can create a new default VPC directly from the VPC Console or by using the CLI. This will create a new default VPC in the region. This does not restore the previous VPC that was deleted.

Q. Can I delete a default subnet?

Yes, you can delete a default subnet. Once deleted, you can create a new default subnet in the availability zone by using the CLI or SDK. This will create a new default subnet in the availability zone specified. This does not restore the previous subnet that was deleted.

Q. I have an existing EC2-Classic account. Can I get a default VPC?

The simplest way to get a default VPC is to create a new account in a region that is enabled for default VPCs, or use an existing account in a region you've never been to before, as long as the Supported Platforms attribute for that account in that region is set to "EC2-VPC".

Q. I really want a default VPC for my existing EC2 account. Is that possible?

Yes, however, we can only enable an existing account for a default VPC if you have no EC2-Classic resources for that account in that region. Additionally, you must terminate all non-VPC provisioned Elastic Load Balancers, Amazon RDS, Amazon ElastiCache, and Amazon Redshift resources in that region. After your account has been configured for a default VPC, all future resource launches, including instances launched via Auto Scaling, will be placed in your default VPC. To request your existing account be setup with a default VPC, please go to *Account and Billing* -> *Service: Account* -> *Category: Convert EC2 Classic to VPC* and raise a request. We will review your request, your existing AWS services and EC2-Classic presence and guide you through the next steps.

Q. How are IAM accounts impacted by default VPC?

If your AWS account has a default VPC, any IAM accounts associated with your AWS account use the same default VPC as your AWS account.

Elastic Network Interfaces

Q. Can I attach or detach one or more network interfaces to an EC2 instance while it’s running?

Yes.

Q. Can I have more than two network interfaces attached to my EC2 instance?

The total number of network interfaces that can be attached to an EC2 instance depends on the instance type. See the EC2 User Guide for more information on the number of allowed network interfaces per instance type.

Q. Can I attach a network interface in one Availability Zone to an instance in another Availability Zone?

Network interfaces can only be attached to instances residing in the same Availability Zone.

Q. Can I attach a network interface in one VPC to an instance in another VPC?

Network interfaces can only be attached to instances in the same VPC as the interface.

Q. Can I use Elastic Network Interfaces as a way to host multiple websites requiring separate IP addresses on a single instance?

Yes, however, this is not a use case best suited for multiple interfaces. Instead, assign additional private IP addresses to the instance and then associate EIPs to the private IPs as needed.

Q. Will I get charged for an Elastic IP Address that is associated to a network interface but the network interface isn't attached to a running instance?

Yes.

Q. Can I detach the primary interface (eth0) on my EC2 instance?

No. You can attach and detach secondary interfaces (eth1-ethn) on an EC2 instance, but you can't detach the eth0 interface.

Peering Connections

Q. Can I create a peering connection to a VPC in a different region?

Yes. Peering connections can be created with VPCs in different regions. Inter-region VPC peering is available globally in all commercial regions (excluding China).

Q. Can I peer my VPC with a VPC belonging to another AWS account?

Yes, assuming the owner of the other VPC accepts your peering connection request.

Q. Can I peer two VPCs with matching IP address ranges?

No. Peered VPCs must have non-overlapping IP ranges.

Q. How much do VPC peering connections cost?

There is no charge for creating VPC peering connections, however, data transfer across peering connections is charged. See the Data Transfer section of the EC2 Pricing page for data transfer rates.

Q. Can I use AWS Direct Connect or hardware VPN connections to access VPCs I'm peered with?

No. "Edge to Edge routing" isn't supported in Amazon VPC. Refer to the VPC Peering Guide for additional information.

Q. Do I need an Internet Gateway to use peering connections?

No. VPC peering connections do not require an Internet Gateway.

Q. Is VPC peering traffic within the region encrypted?

No. Traffic between instances in peered VPCs remains private and isolated – similar to how traffic between two instances in the same VPC is private and isolated.

Q. If I delete my side of a peering connection, will the other side still have access to my VPC?

No. Either side of the peering connection can terminate the peering connection at any time. Terminating a peering connection means traffic won't flow between the two VPCs.

Q. If I peer VPC A to VPC B and I peer VPC B to VPC C, does that mean VPCs A and C are peered?

No. Transitive peering relationships are not supported.

Q. What if my peering connection goes down?

AWS uses the existing infrastructure of a VPC to create a VPC peering connection; it is neither a gateway nor a VPN connection, and does not rely on a separate piece of physical hardware. There is no single point of failure for communication or a bandwidth bottleneck.

Inter-Region VPC Peering operates on the same horizontally scaled, redundant, and highly available technology that powers VPC today. Inter-Region VPC Peering traffic goes over the AWS backbone that has in-built redundancy and dynamic bandwidth allocation. There is no single point of failure for communication.

If an Inter-Region peering connection does go down, the traffic will not be routed over the internet.

Q. Are there any bandwidth limitations for peering connections?

Bandwidth between instances in peered VPCs is no different than bandwidth between instances in the same VPC. **Note:** A placement group can span peered VPCs; however, you will not get full-bisection bandwidth between instances in peered VPCs. Read more about Placement Groups.

Q. Is Inter-Region VPC Peering traffic encrypted?

Traffic is encrypted using modern AEAD (Authenticated Encryption with Associated Data) algorithms. Key agreement and key management is handled by AWS.

Q. How do DNS translations work with Inter-Region VPC Peering?

By default, a query for a public hostname of an instance in a peered VPC in a different region will resolve to a public IP address. Route 53 private DNS can be used to resolve to a private IP address with Inter-Region VPC Peering.

Q. Can I reference security groups across an Inter-Region VPC Peering connection?

No. Security groups cannot be referenced across an Inter-Region VPC Peering connection.

Q. Does Inter-Region VPC Peering support with IPv6?

No. Inter-Region VPC Peering does not support IPv6.

Q. Can Inter-Region VPC Peering be used with EC2-Classic Link?

No. Inter-Region VPC Peering cannot be used with EC2-ClassicLink.

Q. Are there AWS Services that cannot be used over Inter-Region VPC Peering?

Network Load Balancers, AWS PrivateLink and Elastic File System cannot be used over Inter-Region VPC Peering.

ClassicLink

Q. What is ClassicLink?

Amazon Virtual Private Cloud (VPC) ClassicLink allows EC2 instances in the EC2-Classic platform to communicate with instances in a VPC using private IP addresses. To use ClassicLink, enable it for a VPC in your account, and associate a Security Group from that VPC with an instance in EC2-Classic. All the rules of your VPC Security Group will apply to communications between instances in EC2-Classic and instances in the VPC.

Q. What does ClassicLink cost?

There is no additional charge for using ClassicLink; however, existing cross Availability Zone data transfer charges will apply. For more information, consult the EC2 pricing page.

Q. How do I use ClassicLink?

In order to use ClassicLink, you first need to enable at least one VPC in your account for ClassicLink. Then you associate a Security Group from the VPC with the desired EC2-Classic instance. The EC2-Classic instance is now linked to the VPC and is a member of the selected Security Group in the VPC. Your EC2-Classic instance cannot be linked to more than one VPC at the same time.

Q. Does the EC2-Classic instance become a member of the VPC?

The EC2-Classic instance does not become a member of the VPC. It becomes a member of the VPC Security Group that was associated with the instance. All the rules and references to the VPC Security Group apply to communication between instances in EC2-Classic instance and resources within the VPC.

Q. Can I use EC2 public DNS hostnames from my EC2-Classic and EC2-VPC instances to address each other, in order to communicate using private IP?

No. The EC2 public DNS hostname will not resolve to the private IP address of the EC2-VPC instance when queried from an EC2-Classic instance, and vice-versa.

Q. Are there any VPCs for which I cannot enable ClassicLink?

Yes. ClassicLink cannot be enabled for a VPC that has a Classless Inter-Domain Routing (CIDR) that is within the 10.0.0.0/8 range, with the exception of 10.0.0.0/16 and 10.1.0.0/16. In addition, ClassicLink cannot be enabled for any VPC that has a route table entry pointing to the 10.0.0.0/8 CIDR space to a target other than "local".

Q. Can traffic from an EC2-Classic instance travel through the Amazon VPC and egress through the Internet gateway, virtual private gateway, or to peered VPCs?

Traffic from an EC2-Classic instance can only be routed to private IP addresses within the VPC. They will not be routed to any destinations outside the VPC, including Internet gateway, virtual private gateway, or peered VPC destinations.

Q. Does ClassicLink affect the access control between the EC2-Classic instance, and other instances that are in the EC2-Classic platform?

ClassicLink does not change the access control defined for an EC2-Classic instance through its existing Security Groups from the EC2-Classic platform.

Q. Will ClassicLink settings on my EC2-Classic instance persist through stop/start cycles?

The ClassicLink connection will not persist through stop/start cycles of the EC2-Classic instance. The EC2-Classic instance will need to be linked back to a VPC after it is stopped and started. However, the ClassicLink connection will persist through instance reboot cycles.

Q. Will my EC2-Classic instance be assigned a new, private IP address after I enable ClassicLink?

There is no new private IP address assigned to the EC2-Classic instance. When you enable ClassicLink on an EC2-Classic instance, the instance retains and uses its existing private IP address to communication with resources in a VPC.

Q: Does ClassicLink allow EC2-Classic Security Group rules to reference VPC Security Groups, or vice versa?

ClassicLink does not allow EC2-Classic Security Group rules to reference VPC Security Groups, or vice versa.

AWS PrivateLink

Q. What is AWS PrivateLink?

AWS PrivateLink enables customers to access services hosted on AWS in a highly available and scalable manner, while keeping all the network traffic within the AWS network. Service users can use this to privately access services powered by PrivateLink from their Amazon Virtual Private Cloud (VPC) or their on-premises, without using public IPs, and without requiring the traffic to traverse across the Internet. Service owners can register their Network Load Balancers to PrivateLink services and provide the services to other AWS customers.

Q. How can I use AWS PrivateLink?

As a service user, you will need to create interface type VPC endpoints for services that are powered by PrivateLink. These service endpoints will appear as Elastic Network Interfaces (ENIs) with private IPs in your VPCs. Once these endpoints are created, any traffic destined to these IPs will get privately routed to the corresponding AWS services.

As a service owner, you can onboard your service to AWS PrivateLink by establishing a Network Load Balancer (NLB) to front your service and create a PrivateLink service to register with the NLB. Your customers will be able to establish endpoints within their VPC to connect to your service after you whitelisted their accounts and IAM roles.

Q. Which services are currently available on AWS PrivateLink?

The following AWS services support this feature: Amazon Elastic Compute Cloud (EC2), Elastic Load Balancing (ELB), Kinesis Streams, Service Catalog, EC2 Systems Manager, Amazon SNS, and AWS DataSync. Many SaaS solutions support this feature as well. Please visit AWS Marketplace for more SaaS products powered by AWS PrivateLink.

Q. Can I privately access services powered by AWS PrivateLink over AWS Direct Connect?

Yes. The application in your on-premises can connect to the service endpoints in Amazon VPC over AWS Direct Connect. The service endpoints will automatically direct the traffic to AWS services powered by AWS PrivateLink.

Q. What CloudWatch metrics are available for the interface-based VPC endpoint?

Currently, no CloudWatch metric is available for the interface-based VPC endpoint.

Q. Who pays the data transfer costs for the traffic going via the interface-based VPC endpoint?

The concept of data transfer costs is similar to that of data transfer costs for EC2 instances. Since an interface-based VPC endpoint is an ENI in the subnet, data transfer charges depend on the source of the traffic. If the traffic to this interface is coming from a resource across AZ, EC2 cross-AZ data transfer charges apply to the consumer end. Customers in the consumer VPC can use AZ-specific DNS endpoint to make sure the traffic stays within the same AZ if they have provisioned each AZ available in their account.

Bring Your Own IP

Q. What is the Bring Your Own IP feature?

Bring Your Own IP (BYOIP) enables customers to move all or part of their existing publicly routable IPv4 address space to AWS for use with their AWS resources. Customers will continue to own the IP range, however, AWS will take over its advertisement on the internet. Customers can create Elastic IPs from the IP space they bring to AWS and use them with EC2 instances, NAT Gateways, and Network Load Balancers. Customers will continue to have access to Amazon-supplied IPs and can choose to use BYOIP Elastic IPs, Amazon-supplied IPs, or both.

Q. Why should I use BYOIP?

You may want to bring your own IP addresses to AWS for the following reasons:

IP Reputation: Many customers consider the reputation of their IP addresses to be a strategic asset and want to use those IPs on AWS with their resources. For example, customers who maintain services such as outbound e-mail MTA and have high reputation IPs, can now bring over their IP space and successfully maintain their existing sending success rate.

Customer whitelisting: BYOIP also enables customers to move workloads that rely on IP address whitelisting to AWS without the need to re-establish the whitelists with new IP addresses.

Hardcoded dependencies: Several customers have IPs hardcoded in devices or have taken architectural dependencies on their IPs. BYOIP enables such customers hassle free migration to AWS.

Regulation and compliance: Many customers are required to use certain IPs because of regulation and compliance reasons. They too are unlocked by BYOIP.

Q. How can I use IP addresses from a BYOIP prefix with AWS resources?

Your BYOIP prefix will show as an IP pool in your account. You can create Elastic IPs (EIPs) from the IP pool and use them like regular Elastic IPs (EIPs) with any AWS resource that supports EIPs. Currently, EC2 instances, NAT Gateways, and Network Load Balancers support EIPs.

Q. What happens if I release a BYOIP Elastic IP?

When you release a BYOIP Elastic IP it goes back to the BYOIP IP pool from which it was allocated.

Q. In which AWS Regions is BYOIP available?

The feature is currently available in the US-East (N.Virginia), US-East (Ohio), US-West (Oregon), EU (Dublin), EU (London), EU (Frankfurt), Canada (Central), Asia Pacific (Mumbai), Asia Pacific (Sydney), Asia Pacific (Tokyo), Asia Pacific (Singapore) and South America (Sao Paulo) AWS Regions.

Q. Can a BYOIP prefix be shared with multiple VPCs in the same account?

Yes. You can use the BYOIP prefix with any number of VPCs in the same account.

Q. How many IP ranges can I bring via BYOIP?

You can bring a maximum of five IP ranges to your account.

Q. What is the most specific prefix that I can bring via BYOIP?

The most specific prefix you can bring via BYOIP is a /24 IPv4 prefix.

Q. Which RIR prefixes can I use for BYOIP?

You can use ARIN, RIPE, and APNIC registered prefixes.

Q. Can I bring a reassigned or reallocated prefix?

We are not accepting reassigned or reallocated prefixes at this time. IP ranges should be a net type of direct allocation or direct assignment.

Q. Can I move a BYOIP prefix from one AWS Region to another?

Yes. You can do that by de-provisioning the BYOIP prefix from the current region and then provisioning it to the new region.

Additional Questions

Q. Can I use the AWS Management Console to control and manage Amazon VPC?

Yes. You can use the AWS Management Console to manage Amazon VPC objects such as VPCs, subnets, route tables, Internet gateways, and IPSec VPN connections. Additionally, you can use a simple wizard to create a VPC.

Q. How many VPCs, subnets, Elastic IP addresses, and internet gateways can I create?

You can have:

- Five Amazon VPCs per AWS account per region
- Two hundred subnets per Amazon VPC
- Five Amazon VPC Elastic IP addresses per AWS account per region
- One internet gateway per Amazon VPC

See the Amazon VPC user guide for more information on VPC limits.

Q. Can I obtain AWS support with Amazon VPC?

Yes. Click here for more information on AWS support.

Q. Can I use ElasticFox with Amazon VPC?

ElasticFox is no longer officially supported for managing your Amazon VPC. Amazon VPC support is available via the AWS APIs, command line tools, and the AWS Management Console, as well as a variety of third-party utilities.

Amazon API Gateway FAQs

General

Q: What is Amazon API Gateway? »

Q: Why use Amazon API Gateway? »

Q: What API types are supported by Amazon API Gateway? »

Q. How do I get started with HTTP APIs in API Gateway? »

Q. How do I get started with REST APIs in API Gateway? »

Q: When creating RESTful APIs, when should I use HTTP APIs and when should I use REST APIs? »

Q: Which features come standard with HTTP APIs from API Gateway? »

Q: Can I import an OpenAPI definition to create a HTTP API? »

Q: How can I migrate from my current REST API to a HTTP API? »

Q: How do I know if my current REST API will work as a HTTP API? »

Q: How do I get started with WebSocket APIs in Amazon API Gateway? »

Q: Can I create HTTPS endpoints? »

Q: What data types can I use with Amazon API Gateway? »

Q: With what backends can Amazon API Gateway communicate? »

Q: For which client platforms can Amazon API Gateway generate SDKs? »

Q: In which AWS regions is Amazon API Gateway available? »

Q: What can I manage through the Amazon API Gateway console? »

Q: What is a resource? »

Q: What is a method? »

Q: What is an usage plan? »

Q: What is the Amazon API Gateway API lifecycle? »

Q: What is a stage? »

Q: What are stage variables? »

Q: What is a Resource Policy? »

Q: What if I mistakenly deployed to a stage? »

Q: Can I use my Swagger API definitions? »

Q: How do I monetize my APIs on Amazon API Gateway? »

Q: How do I document my API on Amazon API Gateway? »

Q: How can I avoid creating redundant copies of error messages and other documentation that recurs frequently in my API? »

Q: Can I restrict access to private APIs to a specific Amazon VPC or VPC endpoint? »

back to top >>

Security and Authorization

Q: How do I authorize access to my APIs? »

Q: How does AWS Signature Version 4 work? »

Q: What is a Lambda authorizer? »

Q: Can Amazon API Gateway generate API keys for distribution to third-party developers? »

Q: How can I address or prevent API threats or abuse? »

Q: Can I verify that it is API Gateway calling my backend? »

Q: Can I use AWS CloudTrail with Amazon API Gateway? »

Q: How does Amazon API Gateway work with an Amazon Virtual Private Cloud (Amazon VPC)? »

Q: Can I restrict access to private APIs to a specific Amazon VPC or VPC endpoint? »

Q: Can I configure my REST APIs in API Gateway to use TLS 1.1 or higher? »

back to top >>

Management, Metrics, and Logging

Q: How can I monitor my Amazon API Gateway APIs? »

Q: Can I set up alarms on the Amazon API Gateway metrics? »

Q: How can I set up metrics for Amazon API Gateway? »

Q: Can I determine which version of the API my customers are using? »

Q: Does Amazon API Gateway provide logging support? »

Q: How quickly are logs available? »

back to top >>

Throttling and Caching

Q: How can I protect my backend systems and applications from traffic spikes? »

Q: Can I throttle individual developers calling my APIs? »

Q: How does throttling help me? »

Q: At which levels can Amazon API Gateway throttle inbound API traffic? »

Q: How are throttling rules applied? »

Q: Does Amazon API Gateway provide API result caching? »

Q: What happens if a large number of end users try to invoke my API simultaneously? »

Q: How do APIs scale? »

back to top >>

Billing

Q: How am I charged for using Amazon API Gateway? »

Q: Who pays for Amazon API Gateway API calls generated by third-party developers? »

Q: If an API response is served by cached data, is it still considered an API call for billing purposes? »

back to top >>

WebSocket APIs

Q: What is WebSocket routing in Amazon API Gateway? »

Q: How can I send messages to connected clients from the backend service? »

Q: How can I authorize access to my WebSocket API in Amazon API Gateway? »

Q: How does my backend service know when a client is connected or disconnected from the WebSocket connection in Amazon API Gateway? »

Q: How can my backend service identify if the client is still connected to the WebSocket connection? »

Q: Can I disconnect a client from my backend service? »

Q: What is the maximum message size supported for WebSocket APIs? »

Q: How am I charged for using WebSocket APIs on Amazon API Gateway? »

Q: If messages on the WebSocket connection fail authentication or authorization, do they still count toward my API usage bill? »

back to top >>

Amazon CloudFront FAQs

General

Q. What is Amazon CloudFront?

Amazon CloudFront is a web service that gives businesses and web application developers an easy and cost effective way to distribute content with low latency and high data transfer speeds. Like other AWS services, Amazon CloudFront is a self-service, pay-per-use offering, requiring no long term commitments or minimum fees. With CloudFront, your files are delivered to end-users using a global network of edge locations.

Q. What can I do with Amazon CloudFront?

Amazon CloudFront provides a simple API that lets you:

- Distribute content with low latency and high data transfer rates by serving requests using a network of edge locations around the world.
- Get started without negotiating contracts and minimum commitments.

Q. How do I get started with Amazon CloudFront?

Click the “Create Free Account” button on the Amazon CloudFront detail page. If you choose to use another AWS service as the origin for the files served through Amazon CloudFront, you must sign up for that service before creating CloudFront distributions.

Q. How do I use Amazon CloudFront?

To use Amazon CloudFront, you:

- For static files, store the definitive versions of your files in one or more origin servers. These could be Amazon S3 buckets. For your dynamically generated content that is personalized or customized, you can use Amazon EC2 – or any other web server – as the origin server. These origin servers will store or generate your content that will be distributed through Amazon CloudFront.

- Register your origin servers with Amazon CloudFront through a simple API call. This call will return a CloudFront.net domain name that you can use to distribute content from your origin servers via the Amazon CloudFront service. For instance, you can register the Amazon S3 bucket “bucketname.s3.amazonaws.com” as the origin for all your static content and an Amazon EC2 instance “dynamic.myoriginserver.com” for all your dynamic content. Then, using the API or the AWS Management Console, you can create an Amazon CloudFront distribution that might return “abc123.cloudfront.net” as the distribution domain name.
- Include the cloudfront.net domain name, or a CNAME alias that you create, in your web application, media player, or website. Each request made using the cloudfront.net domain name (or the CNAME you set-up) is routed to the edge location best suited to deliver the content with the highest performance. The edge location will attempt to serve the request with a local copy of the file. If a local copy is not available, Amazon CloudFront will get a copy from the origin. This copy is then available at that edge location for future requests.

Q. How does Amazon CloudFront provide higher performance?

Amazon CloudFront employs a global network of edge locations and regional edge caches that cache copies of your content close to your viewers. Amazon CloudFront ensures that end-user requests are served by the closest edge location. As a result, viewer requests travel a short distance, improving performance for your viewers. For files not cached at the edge locations and the regional edge caches, Amazon CloudFront keeps persistent connections with your origin servers so that those files can be fetched from the origin servers as quickly as possible. Finally, Amazon CloudFront uses additional optimizations – e.g. wider TCP initial congestion window – to provide higher performance while delivering your content to viewers.

Q. How does Amazon CloudFront lower my costs to distribute content over the Internet?

Like other AWS services, Amazon CloudFront has no minimum commitments and charges you only for what you use. Compared to self-hosting, Amazon CloudFront spares you from the expense and complexity of operating a network of cache servers in multiple sites across the internet and eliminates the need to over-provision capacity in order to serve potential spikes in traffic. Amazon CloudFront also uses techniques such as collapsing simultaneous viewer requests at an edge location for the same file into a single request to your origin server. This reduces the load on your origin servers reducing the need to scale your origin infrastructure, which can bring you further cost savings.

Additionally, if you are using an AWS origin (e.g., Amazon S3, Amazon EC2, etc.), effective December 1, 2014, **we are no longer charging for AWS data transfer out to Amazon CloudFront**. This applies to data transfer from all AWS regions to all global CloudFront edge locations.

Q. How does Amazon CloudFront speed up my entire website?

Amazon CloudFront uses standard cache control headers you set on your files to identify static and dynamic content. Delivering all your content using a single Amazon CloudFront distribution helps you make sure that performance optimizations are applied to your entire website or web application. When using AWS origins, you benefit from improved performance, reliability, and ease of use as a result of AWS's ability to track and adjust origin routes, monitor system health, respond quickly when any issues occur, and the integration of Amazon CloudFront with other AWS services. You also benefit from using different origins for different types of content on a single site – e.g. Amazon S3 for static objects, Amazon EC2 for dynamic content, and custom origins for third-party content – paying only for what you use.

Q. How is Amazon CloudFront different from Amazon S3?

Amazon CloudFront is a good choice for distribution of frequently accessed static content that benefits from edge delivery—like popular website images, videos, media files or software downloads.

Q. How is Amazon CloudFront different from traditional content delivery solutions?

Amazon CloudFront lets you quickly obtain the benefits of high performance content delivery without negotiated contracts or high prices. Amazon CloudFront gives all developers access to inexpensive, pay-as-you-go pricing – with a self-service model. Developers also benefit from tight integration with other Amazon Web Services. The solution is simple to use with Amazon S3, Amazon EC2, and Elastic Load Balancing as origin servers, giving developers a powerful combination of durable storage and high performance delivery. Amazon CloudFront also integrates with Amazon Route 53 and AWS CloudFormation for further performance benefits and ease of configuration.

Q. What types of content does Amazon CloudFront support?

Amazon CloudFront supports content that can be sent using the HTTP or WebSocket protocols. This includes dynamic web pages and applications, such as HTML or PHP pages or WebSocket-based applications, and any popular static files that are a part of your web application, such as website images, audio, video, media files or software downloads. Amazon CloudFront also supports delivery of live or on-demand media streaming over HTTP.

Q. Does Amazon CloudFront work with non-AWS origin servers?

Yes. Amazon CloudFront works with any origin server that holds the original, definitive versions of your content, both static and dynamic. There is no additional charge to use a custom origin.

Q. How does Amazon CloudFront enable origin redundancy?

For every origin that you add to a CloudFront distribution, you can assign a backup origin that can be used to automatically serve your traffic if the primary origin is unavailable. You can choose a combination of HTTP 4xx/5xx status codes that, when returned from the primary origin, trigger the failover to the backup origin. The two origins can be any combination of AWS and non-AWS origins.

Q: Does Amazon CloudFront offer a Service Level Agreement (SLA)?

Yes. The Amazon CloudFront SLA provides for a service credit if a customer’s monthly uptime percentage is below our service commitment in any billing cycle. More information can be found here.

Q: Can I use the AWS Management Console with Amazon CloudFront?

Yes. You can use the AWS Management Console to configure and manage Amazon CloudFront though a simple, point-and-click web interface. The AWS Management Console supports most of Amazon CloudFront's features, letting you get Amazon CloudFront's low latency delivery without writing any code or installing any software. Access to the AWS Management Console is provided free of charge at https://console.aws.amazon.com.

Q: What tools and libraries work with Amazon CloudFront?

There are a variety of tools for managing your Amazon CloudFront distribution and libraries for various programming languages available in our resource center.

Q. Can I point my zone apex (example.com versus www.example.com) at my Amazon CloudFront distribution?

Yes. By using Amazon Route 53, AWS's authoritative DNS service, you can configure an 'Alias' record that lets you map the apex or root (example.com) of your DNS name to your Amazon CloudFront distribution. Amazon Route 53 will then respond to each request for an Alias record with the right IP address(es) for your CloudFront distribution. Route 53 doesn't charge for queries to Alias records that are mapped to a CloudFront distribution. These queries are listed as "Intra-AWS-DNS-Queries" on the Amazon Route 53 usage report.

Edge locations

Q. What is CloudFront Regional Edge Cache?

CloudFront delivers your content through a worldwide network of data centers called edge locations. The regional edge caches are located between your origin web server and the global edge locations that serve content directly to your viewers. This helps improve performance for your viewers while lowering the operational burden and cost of scaling your origin resources.

Q. How does regional edge cache work?

Amazon CloudFront has added several regional edge cache locations globally, at close proximity to your viewers. They are located between your origin webserver and the global edge locations that serve content directly to your viewers. As objects become less popular, individual edge locations may remove those objects to make room for more popular content. Regional Edge Caches have a larger cache width than any individual edge location, so objects remain in the cache longer at the nearest regional edge caches. This helps keep more of your content closer to your viewers, reducing the need for CloudFront to go back to your origin webserver and improving overall performance for viewers. For example, CloudFront edge locations in Europe now go to the regional edge cache in Frankfurt to fetch an object before going back to your origin webserver. Regional edge cache locations are currently used only for requests that need to go back to a custom origin; i.e. requests to S3 origins will skip regional edge cache locations.

Q. Is regional edge cache feature enabled by default?

Yes. You do not need to make any changes to your CloudFront distributions; this feature is enabled by default for all new and existing CloudFront distributions. There are no additional charges to use this feature.

Q. Where are the edge network locations used by Amazon CloudFront located?

Amazon CloudFront uses a global network of edge locations and regional edge caches for content delivery. You can see a full list of Amazon CloudFront locations here.

Q. Can I choose to serve content (or not serve content) to specified countries?

Yes, the Geo Restriction feature lets you specify a list of countries in which your users can access your content. Alternatively, you can specify the countries in which your users cannot access your content. In both cases, CloudFront responds to a request from a viewer in a restricted country with an HTTP status code 403 (Forbidden).

Q. How accurate is your GeoIP database?

The accuracy of the IP Address to country lookup database varies by region. Based on recent tests, our overall accuracy for the IP address to country mapping is 99.8%.

Q. Can I serve a custom error message to my end users?

Yes, you can create custom error messages (for example, an HTML file or a .jpg graphic) with your own branding and content for a variety of HTTP 4xx and 5xx error responses. Then you can configure Amazon CloudFront to return your custom error messages to the viewer when your origin returns one of the specified errors to CloudFront.

Q. How long will Amazon CloudFront keep my files at the edge locations?

By default, if no cache control header is set, each edge location checks for an updated version of your file whenever it receives a request more than 24 hours after the previous time it checked the origin for changes to that file. This is called the “expiration period.” You can set this expiration period as short as 0 seconds, or as long as you’d like, by setting the cache control headers on your files in your origin. Amazon CloudFront uses these cache control headers to determine how frequently it needs to check the origin for an updated version of that file. For expiration period set to 0 seconds, Amazon CloudFront will revalidate every request with the origin server. If your files don’t change very often, it is best practice to set a long expiration period and implement a versioning system to manage updates to your files.

Q. How do I remove an item from Amazon CloudFront edge locations?

There are multiple options for removing a file from the edge locations. You can simply delete the file from your origin and as content in the edge locations reaches the expiration period defined in each object’s HTTP header, it will be removed. In the event that offensive or potentially harmful material needs to be removed before the specified expiration time, you can use the Invalidation API to remove the object from all Amazon CloudFront edge locations. You can see the charge for making invalidation requests here.

Q. Is there a limit to the number of invalidation requests I can make?

If you're invalidating objects individually, you can have invalidation requests for up to 3,000 objects per distribution in progress at one time. This can be one invalidation request for up to 3,000 objects, up to 3,000 requests for one object each, or any other combination that doesn't exceed 3,000 objects.

If you're using the * wildcard, you can have requests for up to 15 invalidation paths in progress at one time. You can also have invalidation requests for up to 3,000 individual objects per distribution in progress at the same time; the limit on wildcard invalidation requests is independent of the limit on invalidating objects individually. If you exceed this limit, further invalidation requests will receive an error response until one of the earlier request completes.

You should use invalidation only in unexpected circumstances; if you know beforehand that your files will need to be removed from cache frequently, it is recommended that you either implement a versioning system for your files and/or set a short expiration period.

Compliance

Q. Is Amazon CloudFront PCI compliant?

Yes, Amazon CloudFront is included in the set of services that are compliant with the Payment Card Industry Data Security Standard (PCI DSS) Merchant Level 1, the highest level of compliance for service providers. Please see our developer's guide for more information.

Q: Is Amazon CloudFront HIPAA eligible?

Yes, AWS has expanded its HIPAA compliance program to include Amazon CloudFront as a HIPAA eligible service. If you have an executed Business Associate Agreement (BAA) with AWS, you can use Amazon CloudFront to accelerate the delivery of protected health information (PHI). For more information, see HIPAA Compliance and our developer's guide.

Q: Is Amazon CloudFront SOC compliant?

Yes, Amazon CloudFront is compliant with SOC (System & Organization Control) measures. SOC Reports are independent third-party examination reports that demonstrate how AWS achieves key compliance controls and objectives. For more information see, AWS SOC Compliance and our developer's guide.

Q: How do I request an AWS SOC1, SOC 2, or SOC 3 Report?

The AWS SOC 1 and SOC 2 reports are available to customers by using AWS Artifact, a self-service portal for on-demand access to AWS compliance reports. Sign in to AWS Artifact in the AWS Management Console, or learn more at Getting Started with AWS Artifact. The latest AWS SOC 3 Report is publicly available on the AWS website.

HTTP and HTTP/2

Q. What types of HTTP requests are supported by Amazon CloudFront?

Amazon CloudFront currently supports GET, HEAD, POST, PUT, PATCH, DELETE and OPTIONS requests.

Q. Does Amazon CloudFront cache POST responses?

Amazon CloudFront does not cache the responses to POST, PUT, DELETE, and PATCH requests – these requests are proxied back to the origin server. You may enable caching for the responses to OPTIONS requests.

Q. How do I use HTTP/2?

If you have an existing Amazon CloudFront distribution, you can turn on HTTP/2 using the API or the Management Console. In the Console, go to the “Distribution Configuration” page and navigate to the section “Supported HTTP Versions.” There, you can select "HTTP/2, HTTP/1.1, or HTTP/1.0". HTTP/2 is automatically enabled for all new CloudFront distributions.

Q. What if my origin does not support HTTP/2?

Amazon CloudFront currently supports HTTP/2 for delivering content to your viewers’ clients and browsers. For communication between the edge location and your origin servers, Amazon CloudFront will continue to use HTTP/1.1.

Q. Does Amazon CloudFront support HTTP/2 without TLS?

Not currently. However, most of the modern browsers support HTTP/2 only over an encrypted connection. You can learn more about using SSL with Amazon CloudFront here.

WebSocket

Q. What are WebSockets?

WebSocket is a real-time communication protocol that provides bidirectional communication between a client and a server over a long-held TCP connection. By using a persistent open connection, the client and the server can send real-time data to each other without the client having to frequently reinitiate connections checking for new data to exchange. WebSocket connections are often used in chat applications, collaboration platforms, multiplayer games, and financial trading platforms. Refer to our documentation to learn more about using the WebSocket protocol with Amazon CloudFront.

Q. How do I enable my Amazon CloudFront distribution to support the WebSocket protocol?

You can use WebSockets globally, and no additional configuration is needed to enable the WebSocket protocol within your CloudFront resource as it is now supported by default.

Q. When is a WebSocket connection established through Amazon CloudFront?

Amazon CloudFront establishes WebSocket connections only when the client includes the 'Upgrade: websocket' header and the server responds with the HTTP status code 101 confirming that it can switch to the WebSocket protocol.

Q. Does Amazon CloudFront support secured WebSockets over TLS?

Yes. Amazon CloudFront supports encrypted WebSocket connections (WSS) using the SSL/TLS protocol.

Security

Q. Can I configure my CloudFront distribution to deliver content over HTTPS using my own domain name?

By default, you can deliver your content to viewers over HTTPS by using your CloudFront distribution domain name in your URLs, for example, https://dxxxxx.cloudfront.net/image.jpg. If you want to deliver your content over HTTPS using your own domain name and your own SSL certificate, you can use one of our Custom SSL certificate support features. Learn more.

Q. What is Field-Level Encryption?

Field-Level Encryption is a feature of CloudFront that allows you to securely upload user-submitted data such as credit card numbers to your origin servers. Using this functionality, you can further encrypt sensitive data in an HTTPS form using field-specific encryption keys (which you supply) before a PUT/ POST request is forwarded to your origin. This ensures that sensitive data can only be decrypted and viewed by certain components or services in your application stack. To learn more about field-level encryption, see Field-Level Encryption in our documentation.

Q. I am already using SSL/ TLS encryption with CloudFront, do I still need Field-Level Encryption?

Many web applications collect sensitive data such as credit card numbers from users that is then processed by application services running on the origin infrastructure. All these web applications use SSL/TLS encryption between the end user and CloudFront, and between CloudFront and your origin. Now, your origin could have multiple micro-services that perform critical operations based on user input. However, typically sensitive information only needs to be used by a small subset of these micro-services, which means most components have direct access to these data for no reason. A simple programming mistake, such as logging the wrong variable could lead to a customer's credit card number being written to a file.

With field-level encryption, CloudFront's edge locations can encrypt the credit card data. From that point on, only applications that have the private keys can decrypt the sensitive fields. So the order fulfillment service can only view encrypted credit card numbers, but the payment services can decrypt credit card data. This ensures a higher level of security since even if one of the application services leaks cipher text, the data remains cryptographically protected.

Q. What is the difference between SNI Custom SSL and Dedicated IP Custom SSL of Amazon CloudFront?

Dedicated IP Custom SSL allocates dedicated IP addresses to serve your SSL content at each CloudFront edge location. Because there is a one to one mapping between IP addresses and SSL certificates, Dedicated IP Custom SSL works with browsers and other clients that do not support SNI. Due to the current IP address costs, Dedicated IP Custom SSL is $600/month prorated by the hour.

SNI Custom SSL relies on the SNI extension of the Transport Layer Security protocol, which allows multiple domains to serve SSL traffic over the same IP address by including the hostname viewers are trying to connect to. As with Dedicated IP Custom SSL, CloudFront delivers content from each Amazon CloudFront edge location and with the same security as the Dedicated IP Custom SSL feature. SNI Custom SSL works with most modern browsers, including Chrome version 6 and later (running on Windows XP and later or OS X 10.5.7 and later), Safari version 3 and later (running on Windows Vista and later or Mac OS X 10.5.6. and later), Firefox 2.0 and later, and Internet Explorer 7 and later (running on Windows Vista and later). Older browsers that do not support SNI cannot establish a connection with CloudFront to load the HTTPS version of your content. SNI Custom SSL is available at no additional cost beyond standard CloudFront data transfer and request fees.

Q. What is Server Name Indication?

Server Name Indication (SNI) is an extension of the Transport Layer Security (TLS) protocol. This mechanism identifies the domain (server name) of the associated SSL request so the proper certificate can be used in the SSL handshake. This allows a single IP address to be used across multiple servers. SNI requires browser support to add the server name, and while most modern browsers support it, there are a few legacy browsers that do not. For more details see the SNI section of the CloudFront Developer Guide or the SNI Wikipedia article.

Q. Does CloudFront Integrate with AWS Certificate Manager?

Yes, you can now provision SSL/TLS certificates and associate them with CloudFront distributions within minutes. Simply provision a certificate using the new AWS Certificate Manager (ACM) and deploy it to your CloudFront distribution with a couple of clicks, and let ACM manage certificate renewals for you. ACM allows you to provision, deploy, and manage the certificate with no additional charges.

Note that CloudFront still supports using certificates that you obtained from a third-party certificate authority and uploaded to the IAM certificate store.

Q. Does Amazon CloudFront support access controls for paid or private content?

Yes, Amazon CloudFront has an optional private content feature. When this option is enabled, Amazon CloudFront will only deliver files when you say it is okay to do so by securely signing your requests. Learn more about this feature by reading the CloudFront Developer Guide.

Q. How can I safeguard my web applications delivered via CloudFront from DDoS attacks?

As an AWS customer, you get AWS Shield Standard at no additional cost. AWS Shield is a managed service that provides protection against DDoS attacks for web applications running on AWS. AWS Shield Standard provides protection for all AWS customers against common and most frequently occurring Infrastructure (layer 3 and 4) attacks like SYN/UDP Floods, Reflection attacks, and others to support high availability of your applications on AWS.

AWS Shield Advanced is an optional paid service available to AWS Business Support and AWS Enterprise Support customers. AWS Shield Advanced provides additional protections against larger and more sophisticated attacks for your applications running on Elastic Load Balancing (ELB), Amazon CloudFront and Route 53.

Q. How can I protect my web applications delivered via CloudFront?

You can integrate your CloudFront distribution with AWS WAF, a web application firewall that helps protect web applications from attacks by allowing you to configure rules based on IP addresses, HTTP headers, and custom URI strings. Using these rules, AWS WAF can block, allow, or monitor (count) web requests for your web application. Please see AWS WAF Developer Guide for more information.

Caching

Q. Can I add or modify request headers forwarded to the origin?

Yes, you can configure Amazon CloudFront to add custom headers, or override the value of existing headers, to requests forwarded to your origin. You can use these headers to help validate that requests made to your origin were sent from CloudFront; you can even configure your origin to only allow requests that contain the custom header values you specify. Additionally, if you use multiple CloudFront distributions with the same origin, you can use custom headers to distinguish origin request made by each different distribution. Finally, custom headers can be used to help determine the right CORS headers returned for your requests. You can configure custom headers via the CloudFront API and the AWS Management Console. There are no additional charges for this feature. For more details on how to set your custom headers, you can read more here.

Q. How does Amazon CloudFront handle HTTP cookies?

Amazon CloudFront supports delivery of dynamic content that is customized or personalized using HTTP cookies. To use this feature, you specify whether you want Amazon CloudFront to forward some or all of your cookies to your custom origin server. Amazon CloudFront then considers the forwarded cookie values when identifying a unique object in its cache. This way, your end users get both the benefit of content that is personalized just for them with a cookie and the performance benefits of Amazon CloudFront. You can also optionally choose to log the cookie values in Amazon CloudFront access logs.

Q. How does Amazon CloudFront handle query string parameters in the URL?

A query string may be optionally configured to be part of the cache key for identifying objects in the Amazon CloudFront cache. This helps you build dynamic web pages (e.g. search results) that may be cached at the edge for some amount of time.

Q. Can I specify which query parameters are used in the cache key?

Yes, the query string whitelisting feature allows you to easily configure Amazon CloudFront to only use certain parameters in the cache key, while still forwarding all of the parameters to the origin.

Q. Is there a limit to the number of query parameters that can be whitelisted?

Yes, you can configure Amazon CloudFront to whitelist up to 10 query parameters.

Q. What parameter types are supported?

Amazon CloudFront supports URI query parameters as defined in section 3.4 of RFC3986. Specifically, it supports query parameters embedded in an HTTP GET string after the '?' character, and delimited by the '&' character.

Q. Does CloudFront support gzip compression?

Yes, CloudFront can automatically compress your text or binary data. To use the feature, simply specify in your cache behavior settings that you would like CloudFront to compress objects automatically and ensure that your client adds Accept-Encoding: gzip in the request header (most modern web browsers do this by default). For more information on this feature, please see our developer guide.

Streaming

Q. What is streaming? Why would I want to stream?

Generally, streaming refers to delivering audio and video to end users over the Internet without having to download the media file prior to playback. The protocols used for streaming include those that use HTTP for delivery such as Apple's HTTP Live Streaming (HLS), MPEG Dynamic Adaptive Streaming over HTTP (MPEG-DASH), Adobe's HTTP Dynamic Streaming (HDS) and Microsoft's Smooth Streaming. These protocols are different than the delivery of web pages and other online content because streaming protocols deliver media in real time – viewers watch the bytes as they are delivered. Streaming content has several potential benefits for you and your end-users:

- Streaming can give viewers more control over their viewing experience. For instance, it is easier for a viewer to seek forward and backward in a video using streaming than using traditional download delivery.
- Streaming can give you more control over your content, as no file remains on the viewer's client or local drive when they finish watching a video.
- Streaming can help reduce your costs, as it only delivers the portions of a media file that viewers actually watch. In contrast, with traditional downloads, frequently the whole media file will be delivered to viewers, even if they only watch a portion of the file.

Q. Does Amazon CloudFront support video-on-demand (VOD) streaming protocols?

Yes, Amazon CloudFront provides you with multiple options to deliver on-demand video content. If you have media files that have been converted to HLS, MPEG-DASH, or Microsoft Smooth Streaming, for example using AWS Elemental MediaConvert, prior to being stored in Amazon S3 (or a custom origin), you can use an Amazon CloudFront web distribution to stream in that format without having to run any media servers.

Alternatively, you can also run a third party streaming server (e.g. Wowza Media Server available on AWS Marketplace) on Amazon EC2, which can convert a media file to the required HTTP streaming format. This server can then be designated as the origin for an Amazon CloudFront web distribution.

Visit the Video on Demand (VOD) on AWS page to learn more.

Q. Does Amazon CloudFront support live streaming to multiple platforms?

Yes. You can use Amazon CloudFront live streaming with any live video origination service that outputs HTTP-based streams, such as AWS Elemental MediaPackage or AWS Elemental MediaStore. MediaPackage is a video origination and just-in-time packaging service that allows video distributors to securely and reliably deliver streaming content at scale using multiple delivery and content protection standards. MediaStore is an HTTP origination and storage service that offers the high performance, immediate consistency, and predictable low latency required for live media combined with the security and durability of Amazon storage.

Visit the AWS Live Video Streaming page to learn more.

Limits

Q. Can I use Amazon CloudFront if I expect usage peaks higher than 10 Gbps or 15,000 RPS?

Yes. Complete our request for higher limits here, and we will add more capacity to your account within two business days.

Q: Is there a limit to the number of distributions my Amazon CloudFront account may deliver?

For the current limit on the number of distributions that you can create for each AWS account, see Amazon CloudFront Limits in the Amazon Web Services General Reference. To request a higher limit, please go to the CloudFront Limit Increase Form.

Q: What is the maximum size of a file that can be delivered through Amazon CloudFront?

The maximum size of a single file that can be delivered through Amazon CloudFront is 20 GB. This limit applies to all Amazon CloudFront distributions.

Logging and reporting

Q: Can I get access to request logs for content delivered through Amazon CloudFront?

Yes. When you create or modify a CloudFront distribution, you can enable access logging. When enabled, this feature will automatically write detailed log information in a W3C extended format into an Amazon S3 bucket that you specify. Access logs contain detailed information about each request for your content, including the object requested, the date and time of the request, the edge location serving the request, the client IP address, the referrer, the user agent, the cookie header, and the result type (for example, cache hit/miss/error).

Q: Does Amazon CloudFront offer ready-to-use reports so I can learn more about my usage, viewers, and content being served?

Yes. Whether it's receiving detailed cache statistics reports, monitoring your CloudFront usage, seeing where your customers are viewing your content from, or setting near real-time alarms on operational metrics, Amazon CloudFront offers a variety of solutions for your reporting needs. You can access all our reporting options by visiting the Amazon CloudFront Reporting & Analytics dashboard in the AWS Management Console. You can also learn more about our various reporting options by viewing Amazon CloudFront's Reports & Analytics page.

Q: Can I tag my distributions?

Yes. Amazon CloudFront supports cost allocation tagging. Tags make it easier for you to allocate costs and optimize spending by categorizing and grouping

AWS resources. For example, you can use tags to group resources by administrator, application name, cost center, or a specific project. To learn more about cost allocation tagging, see Using Cost Allocation Tags. If you are ready to add tags to you CloudFront distributions, see Amazon CloudFront Add Tags page.

Q: Can I get a history of all Amazon CloudFront API calls made on my account for security, operational or compliance auditing?

Yes. To receive a history of all Amazon CloudFront API calls made on your account, you simply turn on AWS CloudTrail in the CloudTrail's AWS Management Console. For more information, visit AWS CloudTrail home page.

Q: Do you have options for monitoring and alarming metrics in real time?

You can monitor, alarm and receive notifications on the operational performance of your Amazon CloudFront distributions within just a few minutes of the viewer request using Amazon CloudWatch. CloudFront automatically publishes six operational metrics, each at 1-minute granularity, into Amazon CloudWatch. You can then use CloudWatch to set alarms on any abnormal patterns in your CloudFront traffic. To learn how to get started monitoring CloudFront activity and setting alarms via CloudWatch, please view our walkthrough in the Amazon CloudFront Developer Guide or simply navigate to the Amazon CloudFront Management Console and select Monitoring & Alarming in the navigation pane.

Lambda@Edge

Q: What is Lambda@Edge?

Lambda@Edge allows you to run code at global AWS edge locations without provisioning or managing servers, responding to end users at the lowest network latency. You just upload your Node.js/Python code to AWS Lambda and configure your function to be triggered in response to Amazon CloudFront requests (i.e., when a viewer request lands, when a request is forwarded to or received back from the origin, and right before responding back to the end user). The code is then ready to execute at every AWS edge location when a request for content is received, and scales with the volume of requests across CloudFront edge locations. Learn more in our documentation.

Q. How do I customize content with Lambda@Edge?

Once you have identified a content delivery decision you would like to make at the CloudFront edge, identify which cache behaviors, and what point in the request flow the logic applies to (i.e., when a viewer request lands, when a request is forwarded to or received back from the origin, or right before responding back to the end viewer). Next, write a Node.js/Python Lambda function using the Lambda console or API, and associate it with the selected CloudFront trigger event for your distribution. Once saved, the next time an applicable request is made to your distribution, the function is propagated to the CloudFront edge, and will scale and execute as needed. Learn more in our documentation.

Q: What events can be triggered with Amazon CloudFront?

Your functions will automatically trigger in response to the following Amazon CloudFront events:

- **Viewer Request** - This event occurs when an end user or a device on the Internet makes an HTTP(S) request to CloudFront, and the request arrives at the edge location closest to that user.
- **Viewer Response** - This event occurs when the CloudFront server at the edge is ready to respond to the end user or the device that made the request.
- **Origin Request** - This event occurs when the CloudFront edge server does not already have the requested object in its cache, and the viewer request is ready to be sent to your backend origin webserver (e.g. Amazon EC2, or Application Load Balancer, or Amazon S3).
- **Origin Response** - This event occurs when the CloudFront server at the edge receives a response from your backend origin webserver.

IPv6

Q. What is IPv6?

Every server and device connected to the Internet must have a numeric Internet Protocol (IP) address. As the Internet and the number of people using it grows exponentially, so does the need for IP addresses. IPv6 is a new version of the Internet Protocol that uses a larger address space than its predecessor IPv4. Under IPv4, every IP address is 32 bits long, which allows 4.3 billion unique addresses. An example IPv4 address is 192.0.2.1. In comparison, IPv6 addresses are 128 bits, which allow for approximately three hundred and forty trillion, trillion unique IP addresses. An example IPv6 address is: 2001:0db8:85a3:0:0:8a2e:0370:7334

Q. What can I do with IPv6?

Using IPv6 support for Amazon CloudFront, your applications can connect to Amazon CloudFront edge locations without needing any IPv6 to IPv4 translation software or systems. You can meet the requirements for IPv6 adoption set by governments - including the U.S. Federal government – and benefit from IPv6 extensibility, simplicity in network management, and additional built-in support for security.

Q. Should I expect a change in Amazon CloudFront performance when using IPv6?

No, you will see the same performance when using either IPv4 or IPv6 with Amazon CloudFront.

Q: Are there any Amazon CloudFront features that will not work with IPv6?

All existing features of Amazon CloudFront will continue to work on IPv6, though there are two changes you may need for internal IPv6 address processing before you turn on IPv6 for your distributions.

1. If you have turned on the Amazon CloudFront Access Logs feature, you will start seeing your viewer's IPv6 address in the "c-ip" field and may need to verify that your log processing systems continue to work for IPv6.
2. When you enable IPv6 for your Amazon CloudFront distribution, you will get IPv6 addresses in the 'X-Forwarded-For' header that is sent to your origins. If

your origin systems are only able to process IPv4 addresses, you may need to verify that your origin systems continue to work for IPv6.

Additionally, if you use IP whitelists for Trusted Signers, you should use an IPv4-only distribution for your Trusted Signer URLs with IP whitelists and an IPv4 / IPv6 distribution for all other content. This model sidesteps an issue that would arise if the signing request arrived over an IPv4 address and was signed as such, only to have the request for the content arrive via a different IPv6 address that is not on the whitelist.

To learn more about IPv6 support in Amazon CloudFront, see "IPv6 support on Amazon CloudFront" in the Amazon CloudFront Developer Guide.

Q: Does that mean if I want to use IPv6 at all I cannot use Trusted Signer URLs with IP whitelist?

No. If you want to use IPv6 and Trusted Signer URLs with IP whitelist you should use two separate distributions. You should dedicate a distribution exclusively to your Trusted Signer URLs with IP whitelist and disable IPv6 for that distribution. You would then use another distribution for all other content, which will work with both IPv4 and IPv6.

Q. If I enable IPv6, will the IPv6 address appear in the Access Log?

Yes, your viewer's IPv6 addresses will now be shown in the "c-ip" field of the access logs, if you have the Amazon CloudFront Access Logs feature enabled. You may need to verify that your log processing systems continue to work for IPv6 addresses before you turn on IPv6 for your distributions. Please contact Developer Support if you have any issues with IPv6 traffic impacting your tool or software's ability to handle IPv6 addresses in access logs. For more details, please refer to the Amazon CloudFront Access Logs documentation.

Q: Can I disable IPv6 for all my new distributions?

Yes, for both new and existing distributions, you can use the Amazon CloudFront console or API to enable / disable IPv6 per distribution.

Q: Are there any reasons why I would want to disable IPv6?

In discussions with customers, the only common case we heard about was internal IP address processing. When you enable IPv6 for your Amazon CloudFront distribution, in addition to getting an IPv6 address in your detailed access logs, you will get IPv6 addresses in the 'X-Forwarded-For' header that is sent to your origins. If your origin systems are only able to process IPv4 addresses, you may need to verify that your origin systems continue to work for IPv6 addresses before you turn on IPv6 for your distributions.

Q: I enabled IPv6 for my distribution but a DNS lookup doesn't return any IPv6 addresses. What is happening?

Amazon CloudFront has very diverse connectivity around the globe, but there are still certain networks that do not have ubiquitous IPv6 connectivity. While the long term future of the Internet is obviously IPv6, for the foreseeable future every endpoint on the Internet will have IPv4 connectivity. When we find parts of the Internet that have better IPv4 connectivity than IPv6, we will prefer the former.

Q: If I use Route 53 to handle my DNS needs and I created an alias record pointing to an Amazon CloudFront distribution, do I need to update my alias records to enable IPv6?

Yes, you can create Route 53 alias records pointing to your Amazon CloudFront distribution to support both IPv4 and IPv6 by using "A" and "AAAA" record type respectively. If you want to enable IPv4 only, you need only one alias record with type "A". For details on alias resource record sets, please refer to the Amazon Route 53 Developer Guide.

Billing

Q. How will I be charged for my use of Amazon CloudFront?

Amazon CloudFront charges are based on actual usage of the service in four areas: Data Transfer Out, HTTP/HTTPS Requests, Invalidation Requests, and Dedicated IP Custom SSL certificates associated with a CloudFront distribution.

With the AWS Free Usage Tier, you can get started with Amazon CloudFront for free. Upon sign-up, new AWS customers receive 50 GB Data Transfer Out and 2,000,000 HTTP and HTTPS Requests for Amazon CloudFront each month for one year.

- **Data Transfer Out to Internet**
  You are charged for the volume of data transferred out from Amazon CloudFront edge locations, measured in GB. You can see the rates for Amazon CloudFront data transfer to the internet here. Note that your data transfer usage is totaled separately for specific geographic regions, and then cost is calculated based on pricing tiers for each area. If you use other AWS services as the origins of your files, you are charged separately for your use of those services, including for storage and compute hours. If you use an AWS origin (such as Amazon S3, Amazon EC2, and so on), effective December 1, 2014, **we do not charge for AWS data transfer out to Amazon CloudFront**. This applies to data transfer from all AWS Regions to all global CloudFront edge locations.
- Data Transfer Out to Origin
  You will be charged for the volume of data transferred out, measured in GB, from the Amazon CloudFront edge locations to your origin (both AWS origins and other origin servers). You can see the rates for Amazon CloudFront data transfer to Origin here.
- **HTTP/HTTPS Requests**
  You will be charged for number of HTTP/HTTPS requests made to Amazon CloudFront for your content. You can see the rates for HTTP/HTTPS requests here.
- **Invalidation Requests**
  You are charged per path in your invalidation request. A path listed in your invalidation request represents the URL (or multiple URLs if the path contains a wildcard character) of the object you want to invalidate from CloudFront cache. You can request up to 1,000 paths each month from Amazon CloudFront at no additional charge. Beyond the first 1,000 paths, you will be charged per path listed in your invalidation requests. You can see the rates for invalidation requests here.

- **Dedicated IP Custom SSL**
  You pay $600 per month for each custom SSL certificate associated with one or more CloudFront distributions using the Dedicated IP version of custom SSL certificate support. This monthly fee is pro-rated by the hour. For example, if you had your custom SSL certificate associated with at least one CloudFront distribution for just 24 hours (i.e. 1 day) in the month of June, your total charge for using the custom SSL certificate feature in June will be (1 day / 30 days) * $600 = $20. To use Dedicated IP Custom SSL certificate support, upload a SSL certificate and use the AWS Management Console to associate it with your CloudFront distributions. If you need to associate more than two custom SSL certificates with your CloudFront distribution, please include details about your use case and the number of custom SSL certificates you intend to use in the CloudFront Limit Increase Form.

Usage tiers for data transfer are measured separately for each geographic region. The prices above are exclusive of applicable taxes, fees, or similar governmental charges, if any exist, except as otherwise noted.

Q: Does your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Q: How am I charged for 304 responses?

A 304 is a response to a conditional GET request and will result in a charge for the HTTP/HTTPS request and the Data Transfer Out to Internet. A 304 response does not contain a message-body; however, the HTTP headers will consume some bandwidth for which you would be charged standard CloudFront data transfer fees. The amount of data transfer depends on the headers associated with your object.

Q. Can I choose to only serve content from less expensive Amazon CloudFront regions?

Yes, "Price Classes" provides you an option to lower the prices you pay to deliver content out of Amazon CloudFront. By default, Amazon CloudFront minimizes end user latency by delivering content from its entire global network of edge locations. However, because we charge more where our costs are higher, this means that you pay more to deliver your content with low latency to end-users in some locations. Price Classes let you reduce your delivery prices by excluding Amazon CloudFront's more expensive edge locations from your Amazon CloudFront distribution. In these cases, Amazon CloudFront will deliver your content from edge locations within the locations in the price class you selected and charge you the data transfer and request pricing from the actual location where the content was delivered.

If performance is most important to you, you don't need to do anything; your content will be delivered by our whole network of locations. However, if you wish to use another Price Class, you can configure your distribution through the AWS Management Console or via the Amazon CloudFront API. If you select a price class that does not include all locations, some of your viewers, especially those in geographic locations that are not in your price class, may experience higher latency than if your content were being served from all Amazon CloudFront locations.

Note that Amazon CloudFront may still occasionally serve requests for your content from an edge location in a location that is not included in your price class. When this occurs, you will only be charged the rates for the least expensive location in your price class.

You can see the list of locations making up each price class here.

Amazon Route 53 FAQs

Getting Started

Q. What is a Domain Name System (DNS) Service?

DNS is a globally distributed service that translates human readable names like www.example.com into the numeric IP addresses like 192.0.2.1 that computers use to connect to each other. The Internet's DNS system works much like a phone book by managing the mapping between names and numbers. For DNS, the names are domain names (www.example.com) that are easy for people to remember and the numbers are IP addresses (192.0.2.1) that specify the location of computers on the Internet. DNS servers translate requests for names into IP addresses, controlling which server an end user will reach when they type a domain name into their web browser. These requests are called "queries."

Q. What is Amazon Route 53?

Amazon Route 53 provides highly available and scalable Domain Name System (DNS), domain name registration, and health-checking web services. It is designed to give developers and businesses an extremely reliable and cost effective way to route end users to Internet applications by translating names like example.com into the numeric IP addresses, such as 192.0.2.1, that computers use to connect to each other. You can combine your DNS with health-checking services to route traffic to healthy endpoints or to independently monitor and/or alarm on endpoints. You can also purchase and manage domain names such as example.com and automatically configure DNS settings for your domains. Route 53 effectively connects user requests to infrastructure running in AWS – such as Amazon EC2 instances, Elastic Load Balancing load balancers, or Amazon S3 buckets – and can also be used to route users to infrastructure outside of AWS.

Q. What can I do with Amazon Route 53?

With Amazon Route 53, you can create and manage your public DNS records. Like a phone book, Route 53 lets you manage the IP addresses listed for your domain names in the Internet's DNS phone book. Route 53 also answers requests to translate specific domain names like into their corresponding IP addresses like 192.0.2.1. You can use Route 53 to create DNS records for a new domain or transfer DNS records for an existing domain. The simple, standards-based REST API for Route 53 allows you to easily create, update and manage DNS records. Route 53 additionally offers health checks to monitor the health and performance of your application as well as your web servers and other resources. You can also register new domain names or transfer in existing domain names to be managed by Route 53.

Q. How do I get started with Amazon Route 53?

Amazon Route 53 has a simple web service interface that lets you get started in minutes. Your DNS records are organized into “hosted zones” that you configure with the AWS Management Console or Route 53’s API. To use Route 53, you simply:

- Subscribe to the service by clicking on the sign-up button on the service page.
- If you already have a domain name:
    - Use the AWS Management Console or the CreateHostedZone API to create a hosted zone that can store DNS records for your domain. Upon creating the hosted zone, you receive four Route 53 name servers across four different Top-Level Domains (TLDs) to help ensure a high level of availability.
    - Additionally, you can transfer your domain name to Route 53’s management via either the AWS Management Console or the API.
- If you don't already have a domain name:
    - Use the AWS Management Console or the API to register your new domain name.
    - Route 53 automatically creates a hosted zone that stores DNS records for your domain. You also receive four Route 53 name servers across four different Top-Level Domains (TLDs) to help ensure a high level of availability.
- Your hosted zone will be initially populated with a basic set of DNS records, including four virtual name servers that will answer queries for your domain. You can add, delete or change records in this set by using the AWS Management Console or by calling the ChangeResourceRecordSet API. A list of supported DNS records is available here.
- If your domain name is not managed by Route 53, you will need to inform the registrar with whom you registered your domain name to update the name servers for your domain to the ones associated with your hosted zone. If your domain name is managed by Route 53 already, your domain name will be automatically associated with the name servers hosting your zone.

Q. How does Amazon Route 53 provide high availability and low latency?

Route 53 is built using AWS’s highly available and reliable infrastructure. The globally distributed nature of our DNS servers helps ensure a consistent ability to route your end users to your application by circumventing any internet or network related issues. Route 53 is designed to provide the level of dependability required by important applications. Using a global anycast network of DNS servers around the world, Route 53 is designed to automatically answer queries from the optimal location depending on network conditions. As a result, the service offers low query latency for your end users.

Q. What are the DNS server names for the Amazon Route 53 service?

To provide you with a highly available service, each Amazon Route 53 hosted zone is served by its own set of virtual DNS servers. The DNS server names for each hosted zone are thus assigned by the system when that hosted zone is created.

Q. What is the difference between a Domain and a Hosted Zone?

A domain is a general DNS concept. Domain names are easily recognizable names for numerically addressed Internet resources. For example, amazon.com is a domain. A hosted zone is an Amazon Route 53 concept. A hosted zone is analogous to a traditional DNS zone file; it represents a collection of records that can be managed together, belonging to a single parent domain name. All resource record sets within a hosted zone must have the hosted zone's domain name as a suffix. For example, the amazon.com hosted zone may contain records named www.amazon.com, and www.aws.amazon.com, but not a record named www.amazon.ca. You can use the Route 53 Management Console or API to create, inspect, modify, and delete hosted zones. You can also use the Management Console or API to register new domain names and transfer existing domain names into Route 53's management.

Q. What is the price of Amazon Route 53?

Amazon Route 53 charges are based on actual usage of the service for Hosted Zones, Queries, Health Checks, and Domain Names. For full details, see the Amazon Route 53 pricing page.

You pay only for what you use. There are no minimum fees, no minimum usage commitments, and no overage charges. You can estimate your monthly bill using the AWS Simple Monthly Calculator.

Q. What types of access controls can I set for the management of my Domains on Amazon Route 53?

You can control management access to your Amazon Route 53 hosted zone by using the AWS Identity and Access Management (IAM) service. AWS IAM allows you to control who in your organization can make changes to your DNS records by creating multiple users and managing the permissions for each of these users within your AWS Account. Learn more about AWS IAM here.

Q. I have subscribed for Amazon Route 53 but when I try to use the service it says "The AWS Access Key ID needs a subscription for the service."

When you sign up for a new AWS service, it can take up to 24 hours in some cases to complete activation, during which time you cannot sign up for the service again. If you've been waiting longer than 24 hours without receiving an email confirming activation, this could indicate a problem with your account or the authorization of your payment details. Please contact AWS Customer Service for help.

Q. Does Amazon Route 53 offer a Service Level Agreement (SLA)?

Yes. The Amazon Route 53 SLA provides for a service credit if a customer’s monthly uptime percentage is below our service commitment in any billing cycle. More information can be found here.

Q. When is my hosted zone charged?

Hosted zones are billed once when they are created and then on the first day of each month.

Q. Why do I see two charges for the same hosted zone in the same month?

Hosted zones have a grace period of 12 hours--if you delete a hosted zone within 12 hours after you create it, we don't charge you for the hosted zone. After the grace period ends, we immediately charge the standard monthly fee for a hosted zone. If you create a hosted zone on the last day of the month (for example, January 31st), the charge for January might appear on the February invoice, along with the charge for February.

Q. Does Amazon Route 53 provide query logging capability?

You can configure Amazon Route 53 to log information about the queries that Amazon Route 53 receives including date-time stamp, domain name, query type, location etc. When you configure query logging, Amazon Route 53 starts to send logs to CloudWatch Logs. You use CloudWatch Logs tools to access the query logs. For more information please see our documentation.

Domain Name Systems (DNS)

Q. Does Amazon Route 53 use an anycast network?

Yes. Anycast is a networking and routing technology that helps your end users' DNS queries get answered from the optimal Route 53 location given network conditions. As a result, your users get high availability and improved performance with Route 53.

Q. Is there a limit to the number of hosted zones I can manage using Amazon Route 53?

Each Amazon Route 53 account is limited to a maximum of 500 hosted zones and 10,000 resource record sets per hosted zone. Complete our request for a higher limit and we will respond to your request within two business days.

Q. How can I import a zone into Route 53?

Route 53 supports importing standard DNS zone files which can be exported from many DNS providers as well as standard DNS server software such as BIND. For newly-created hosted zones, as well as existing hosted zones that are empty except for the default NS and SOA records, you can paste your zone file directly into the Route 53 console, and Route 53 automatically creates the records in your hosted zone. To get started with zone file import, read our walkthrough in the Amazon Route 53 Developer Guide.

Q. Can I create multiple hosted zones for the same domain name?

Yes. Creating multiple hosted zones allows you to verify your DNS setting in a "test" environment, and then replicate those settings on a "production" hosted zone. For example, hosted zone Z1234 might be your test version of example.com, hosted on name servers ns-1, ns-2, ns-3, and ns-4. Similarly, hosted zone Z5678 might be your production version of example.com, hosted on ns-5, ns-6, ns-7, and ns-8. Since each hosted zone has a virtual set of name servers associated with that zone, Route 53 will answer DNS queries for example.com differently depending on which name server you send the DNS query to.

Q. Does Amazon Route 53 also provide website hosting?

No. Amazon Route 53 is an authoritative DNS service and does not provide website hosting. However, you can use Amazon Simple Storage Service (Amazon S3) to host a static website. To host a dynamic website or other web applications, you can use Amazon Elastic Compute Cloud (Amazon EC2), which provides flexibility, control, and significant cost savings over traditional web hosting solutions. Learn more about Amazon EC2 here. For both static and dynamic websites, you can provide low latency delivery to your global end users with Amazon CloudFront. Learn more about Amazon CloudFront here.

Q. Which DNS record types does Amazon Route 53 support?

Amazon Route 53 currently supports the following DNS record types:

- A (address record)
- AAAA (IPv6 address record)
- CNAME (canonical name record)
- CAA (certification authority authorization)
- MX (mail exchange record)
- NAPTR (name authority pointer record)
- NS (name server record)
- PTR (pointer record)
- SOA (start of authority record)
- SPF (sender policy framework)
- SRV (service locator)
- TXT (text record)
- Amazon Route 53 also offers alias records, which are an Amazon Route 53-specific extension to DNS. You can create alias records to route traffic to selected AWS resources, including Amazon Elastic Load Balancing load balancers, Amazon CloudFront distributions, AWS Elastic Beanstalk environments, API Gateways, VPC interface endpoints, and Amazon S3 buckets that are configured as websites. Alias record typically have a type of A or AAAA, but they work like a CNAME record. Using an alias record, you can map your record name (example.com) to the DNS name for an AWS resource(elb1234.elb.amazonaws.com). Resolvers see the A or AAAA record and the IP address of the AWS resource.

We anticipate adding additional record types in the future.

Q. Does Amazon Route 53 support wildcard entries? If so, what record types support them?

Yes. To make it even easier for you to configure DNS settings for your domain, Amazon Route 53 supports wildcard entries for all record types, except NS records. A wildcard entry is a record in a DNS zone that will match requests for any domain name based on the configuration you set. For example, a wildcard DNS record such as *.example.com will match queries for www.example.com and subdomain.example.com.

Q. What is the default TTL for the various record types and can I change these values?

The time for which a DNS resolver caches a response is set by a value called the time to live (TTL) associated with every record. Amazon Route 53 does not have a default TTL for any record type. You must always specify a TTL for each record so that caching DNS resolvers can cache your DNS records to the length of time specified through the TTL.

Q. Can I use 'Alias' records with my sub-domains?

Yes. You can also use Alias records to map your sub-domains (www.example.com, pictures.example.com, etc.) to your ELB load balancers, CloudFront distributions, or S3 website buckets.

Q. Are changes to resource record sets transactional?

Yes. A transactional change helps ensure that the change is consistent, reliable, and independent of other changes. Amazon Route 53 has been designed so that changes complete entirely on any individual DNS server, or not at all. This helps ensure your DNS queries are always answered consistently, which is important when making changes such as flipping between destination servers. When using the API, each call to ChangeResourceRecordSets returns an identifier that can be used to track the status of the change. Once the status is reported as INSYNC, your change has been performed on all of the Route 53 DNS servers.

Q. Can I associate multiple IP addresses with a single record?

Yes. Associating multiple IP addresses with a single record is often used for balancing the load of geographically-distributed web servers. Amazon Route 53 allows you to list multiple IP addresses for an A record and responds to DNS requests with the list of all configured IP addresses.

Q. How quickly will changes I make to my DNS settings on Amazon Route 53 propagate globally?

Amazon Route 53 is designed to propagate updates you make to your DNS records to its world-wide network of authoritative DNS servers within 60 seconds under normal conditions. A change is successfully propagated world-wide when the API call returns an INSYNC status listing.

Note that caching DNS resolvers are outside the control of the Amazon Route 53 service and will cache your resource record sets according to their time to live (TTL). The INSYNC or PENDING status of a change refers only to the state of Route 53’s authoritative DNS servers.

Q. Can I see a history of my changes and other operations on my Route 53 resources?

Yes, via AWS CloudTrail you can record and log the API call history for Route 53. Please reference the CloudTrail product page to get started.

Q. Can I use AWS CloudTrail logs to roll back changes to my hosted zones?

No. We recommend that you do not use CloudTrail logs to roll back changes to your hosted zones, because reconstruction of your zone change history using your CloudTrail logs may be incomplete.

Your AWS CloudTrail logs can be used for the purposes of security analysis, resource change tracking, and compliance auditing.

Q. Does Amazon Route 53 support DNSSEC?

Amazon Route 53 does not support DNSSEC for DNS at this time. But Amazon Route 53 allows DNSSEC on domain registration.

Q. Does Amazon Route 53 support IPv6?

Yes. Amazon Route 53 supports both forward (AAAA) and reverse (PTR) IPv6 records. The Amazon Route 53 service itself is also available over IPv6. Recursive DNS resolvers on IPv6 networks can use either IPv4 or IPv6 transport in order to submit DNS queries to Amazon Route 53. Amazon Route 53 health checks also support monitoring of endpoints using the IPv6 protocol.

Q. Can I point my zone apex (example.com versus www.example.com) at my Elastic Load Balancer?

Yes. Amazon Route 53 offers a special type of record called an 'Alias' record that lets you map your zone apex (example.com) DNS name to the DNS name for your ELB load balancer (such as my-loadbalancer-1234567890.us-west-2.elb.amazonaws.com). IP addresses associated with load balancers can change at any time due to scaling up, scaling down, or software updates. Route 53 responds to each request for an Alias record with one or more IP addresses for the load balancer. Route 53 supports alias records for three types of load balancers: Application Load Balancers, Network Load Balancers, and Classic Load Balancers. There is no additional charge for queries to Alias records that are mapped to AWS ELB load balancers. These queries are listed as “Intra-AWS-DNS-Queries” on the Amazon Route 53 usage report.

Q. Can I point my zone apex (example.com versus www.example.com) at my website hosted on Amazon S3?

Yes. Amazon Route 53 offers a special type of record called an ‘Alias’ record that lets you map your zone apex (example.com) DNS name to your Amazon S3 website bucket (i.e. example.com.s3-website-us-west-2.amazonaws.com). IP addresses associated with Amazon

S3 website endpoints can change at any time due to scaling up, scaling down, or software updates. Route 53 responds to each request for an Alias record with one IP address for the bucket. Route 53 doesn't charge for queries to Alias records that are mapped to an S3 bucket that is configured as a website. These queries are listed as “Intra-AWS-DNS-Queries” on the Amazon Route 53 usage report.

Q. Can I point my zone apex (example.com versus www.example.com) at my Amazon CloudFront distribution?

Yes. Amazon Route 53 offers a special type of record called an ‘Alias’ record that lets you map your zone apex (example.com) DNS name to your Amazon CloudFront distribution (for example, d123.cloudfront.net). IP addresses associated with Amazon CloudFront endpoints vary based on your end user’s location (in order to direct the end user to the nearest CloudFront edge location) and can change at any time due to scaling up, scaling down, or software updates. Route 53 responds to each request for an Alias record with the IP address(es) for the distribution. Route 53 doesn't charge for queries to Alias records that are mapped to a CloudFront distribution. These queries are listed as “Intra-AWS-DNS-Queries” on the Amazon Route 53 usage report.

Q. Can I point my zone apex (example.com versus www.example.com) at my AWS Elastic Beanstalk environment?

Yes. Amazon Route 53 offers a special type of record called an ‘Alias’ record that lets you map your zone apex (example.com) DNS name to your AWS Elastic Beanstalk DNS name (i.e. example.elasticbeanstalk.com). IP addresses associated with AWS Elastic Beanstalk environments can change at any time due to scaling up, scaling down, or software updates. Route 53 responds to each request for an Alias record with one or more IP addresses for the environment. Queries to Alias records that are mapped to AWS Elastic Beanstalk environments are free. These queries are listed as “Intra-AWS-DNS-Queries” on the Amazon Route 53 usage report.

Q. Can I point my zone apex (example.com versus www.example.com) at my Amazon API Gateway?

Yes. Amazon Route 53 offers a special type of record called an ‘Alias’ record that lets you map your zone apex (example.com) DNS name to your Amazon API Gateway DNS name (i.e. api-id.execute-api.region.amazonaws.com/stage). IP addresses associated with Amazon API Gateway can change at any time due to scaling up, scaling down, or software updates. Route 53 responds to each request for an Alias record with one or more IP addresses for the API Gateway. There is no additional charge for queries to Alias records that are mapped to Amazon API Gateways. These queries are listed as “Intra-AWS-DNS-Queries” on the Route 53 usage report.

Q. Can I point my zone apex (example.com versus www.example.com) at my Amazon VPC endpoint?

Yes. Amazon Route 53 offers a special type of record called an ‘Alias’ record that lets you map your zone apex (example.com) DNS name to your Amazon VPC Endpoint DNS name (i.e. vpce-svc-03d5ebb7d9579a2b3.us-east-1.vpce.amazonaws.com). IP addresses associated with Amazon VPC Endpoints can change at any time due to scaling up, scaling down, or software updates. Route 53 responds to each request for an Alias record with one or more IP addresses for the VPC endpoint. There is no additional charge for queries to Alias records that are mapped to Amazon VPC endpoints. These queries are listed as “Intra-AWS-DNS-Queries” on the Amazon Route 53 usage report.

Q. How can I use Amazon Route 53 with Amazon Simple Storage Service (Amazon S3) and Amazon CloudFront?

For websites delivered via Amazon CloudFront or static websites hosted on Amazon S3, you can use the Amazon Route 53 service to create an Alias record for your domain which points to the CloudFront distribution or S3 website bucket. For S3 buckets not configured to host static websites, you can create a CNAME record for your domain and the S3 bucket name. In all cases, note that you will also need to configure your S3 bucket or your CloudFront distribution respectively with the alternate domain name entry to completely establish the alias between your domain name and the AWS domain name for your bucket or distribution.

For CloudFront distributions and S3 buckets configured to host static websites, we recommend creating an ‘Alias’ record that maps to your CloudFront distribution or S3 website bucket, instead of using CNAMEs. Alias records have two advantages: first, unlike CNAMEs, you can create an Alias record for your zone apex (e.g. example.com, instead of www.example.com), and second, queries to Alias records are free of charge.

Q. Why does the DNS Query Test Tool return a response different than the dig or nslookup commands?

When resource record sets are changed in Amazon Route 53, the service propagates updates you make to your DNS records to its world-wide network of authoritative DNS servers. If you test the record before propagation is complete, you may see an old value when you use the dig or nslookup utilities. Additionally, DNS resolvers on the internet are outside the control of the Amazon Route 53 service and will cache your resource record sets according to their time to live (TTL), which means a dig/nslookup command might return a cached value. You should also make sure that your domain name registrar is using the name servers in your Amazon Route 53 hosted zone. If not, Amazon Route 53 will not be authoritative for queries to your domain.

DNS Routing Policies

Q. Does Amazon Route 53 support Weighted Round Robin (WRR)?

Yes. Weighted Round Robin allows you to assign weights to resource record sets in order to specify the frequency with which different responses are served. You may want to use this capability to do A/B testing, sending a small portion of traffic to a server on which you've made a software change. For instance, suppose you have two record sets associated with one DNS name—one with weight 3 and one with weight 1. In this case, 75% of the time Route 53 will return the record set with weight 3 and 25% of the time Route 53 will return the record set with weight 1. Weights can be any number between 0 and 255.

Q. What is Amazon Route 53's Latency Based Routing (LBR) feature?

LBR (Latency Based Routing) is a new feature for Amazon Route 53 that helps you improve your application's performance for a global audience. You can run applications in multiple AWS regions and Amazon Route 53, using dozens of edge locations worldwide, will route end users to the AWS region that provides the lowest latency.

Q. How do I get started using Amazon Route 53's Latency Based Routing (LBR) feature?

You can start using Amazon Route 53's new LBR feature quickly and easily by using either the AWS Management Console or a simple API. You simply create a record set that includes the IP addresses or ELB names of various AWS endpoints and mark that record set as an LBR-enabled Record Set, much like you mark a record set as a Weighted Record Set. Amazon Route 53 takes care of the rest - determining the best endpoint for each request and routing end users accordingly, much like Amazon CloudFront, Amazon's global content delivery service, does. You can learn more about how to use Latency Based Routing in the Amazon Route 53 Developer Guide.

Q. What is the price for Amazon Route 53's Latency Based Routing (LBR) feature?

Like all AWS services, there are no upfront fees or long term commitments to use Amazon Route 53 and LBR. Customers simply pay for the hosted zones and queries they actually use. Please visit the Amazon Route 53 pricing page for details on pricing for Latency Based Routing queries.

Q. What is Amazon Route 53's Geo DNS feature?

Route 53 Geo DNS lets you balance load by directing requests to specific endpoints based on the geographic location from which the request originates. Geo DNS makes it possible to customize localized content, such as presenting detail pages in the right language or restricting distribution of content to only the markets you have licensed. Geo DNS also lets you balance load across endpoints in a predictable, easy-to-manage way, ensuring that each end-user location is consistently routed to the same endpoint. Geo DNS provides three levels of geographic granularity: continent, country, and state, and Geo DNS also provides a global record which is served in cases where an end user's location doesn't match any of the specific Geo DNS records you have created. You can also combine Geo DNS with other routing types, such as Latency Based Routing and DNS Failover, to enable a variety of low-latency and fault-tolerant architectures. For information on how to configure various routing types, please see the Amazon Route 53 documentation.

Q. How do I get started using Amazon Route 53's Geo DNS feature?

You can start using Amazon Route 53's Geo DNS feature quickly and easily by using either the AWS Management Console or the Route 53 API. You simply create a record set and specify the applicable values for that type of record set, mark that record set as a Geo DNS-enabled Record Set, and select the geographic region (global, continent, country, or state) that you want the record to apply to. You can learn more about how to use Geo DNS in the Amazon Route 53 Developer Guide.

Q. When using Geo DNS, do I need a "global" record? When would Route 53 return this record?

Yes, we strongly recommend that you configure a global record, to ensure that Route 53 can provide a response to DNS queries from all possible locations—even if you have created specific records for each continent, country, or state where you expect your end users will be located. Route 53 will return the value contained in your global record in the following cases:

The DNS query comes from an IP address not recognized by Route 53's Geo IP database.

The DNS query comes from a location not included in any of the specific Geo DNS records you have created.

Q. Can I have a Geo DNS record for a continent and different Geo DNS records for countries within that continent? Or a Geo DNS record for a country and Geo DNS records for states within that country?

Yes, you can have Geo DNS records for overlapping geographic regions (e.g., a continent and countries within that continent, or a country and states within that country). For each end user's location, Route 53 will return the most specific Geo DNS record that includes that location. In other words, for a given end user's location, Route 53 will first return a state record; if no state record is found, Route 53 will return a country record; if no country record is found, Route 53 will return a continent record; and finally, if no continent record is found, Route 53 will return the global record.

Q. What is the price for Route 53's Geo DNS feature?

Like all AWS services, there are no upfront fees or long term commitments to use Amazon Route 53 and Geo DNS. Customers simply pay for the hosted zones and queries they actually use. Please visit the Amazon Route 53 pricing page for details on pricing for Geo DNS queries.

Q. What is the difference between Latency Based Routing and Geo DNS?

Geo DNS bases routing decisions on the geographic location of the requests. In some cases, geography is a good proxy for latency; but there are certainly situations where it is not. LatencyBased Routing utilizes latency measurements between viewer networks and AWS datacenters. These measurements are used to determine which endpoint to direct users toward.

If your goal is to minimize end-user latency, we recommend using Latency Based Routing. If you have compliance, localization requirements, or other use cases that require stable routing from a specific geography to a specific endpoint, we recommend using Geo DNS.

Q. Does Amazon Route 53 support multiple values in response to DNS queries?

Route 53 now supports multivalue answers in response to DNS queries. While not a substitute for a load balancer, the ability to return multiple health-checkable IP addresses in response to DNS queries is a way to use DNS to improve availability and load balancing. If you want to route traffic randomly to multiple resources, such as web servers, you can create one multivalue answer record for each resource and, optionally, associate an Amazon Route 53 health check with each record. Amazon Route 53 supports up to eight healthy records in response to each DNS query.

Traffic Flow

Q. What is Amazon Route 53 Traffic Flow?

Amazon Route 53 Traffic Flow is an easy-to-use and cost-effective global traffic management service. With Amazon Route 53 Traffic Flow, you can improve the performance and availability of your application for your end users by running multiple endpoints around the world, using Amazon Route 53 Traffic Flow to connect your users to the best endpoint based on latency, geography, and endpoint health. Amazon Route 53 Traffic Flow makes it easy for developers to create policies that route traffic based on the constraints they care most about, including latency, endpoint health, load, geoproximity and geography. Customers can customize these templates or build policies from scratch using a simple visual policy builder in the AWS Management Console.

Q. What is the difference between a traffic policy and a policy record?

A **traffic policy** is the set of rules that you define to route end users' requests to one of your application's endpoints. You can create a traffic policy using the visual policy builder in the Amazon Route 53 Traffic Flow section of the Amazon Route 53 console. You can also create traffic policies as JSON-formatted text files and upload these policies using the Route 53 API, the AWS CLI, or the various AWS SDKs.

By itself, a traffic policy doesn't affect how end users are routed to your application because it isn't yet associated with your application's DNS name (such as www.example.com). To start using Amazon Route 53 Traffic Flow to route traffic to your application using the traffic policy you've created, you create a **policy record** which associates the traffic policy with the appropriate DNS name within an Amazon Route 53 hosted zone that you own. For example, if you want to use a traffic policy that you've named my-first-traffic-policy to manage traffic for your application at www.example.com, you will create a policy record for www.example.com within your hosted zone example.com and choose my-first-traffic-policy as the traffic policy.

Policy records are visible in both the Amazon Route 53 Traffic Flow and Amazon Route 53 Hosted Zone sections of the Amazon Route 53 console.

Q. Can I use the same policy to manage routing for more than one DNS name?

Yes. You can reuse a policy to manage more than one DNS name in one of two ways. First, you can create additional policy records using the policy. Note that there is an additional charge for using this method because you are billed for each policy record that you create.

The second method is to create one policy record using the policy, and then for each additional DNS name that you want to manage using the policy, you create a standard CNAME record pointing at the DNS name of the policy record that you created. For example, if you create a policy record for example.com, you can then create DNS records for www.example.com, blog.example.com, and www.example.net with a CNAME value of example.com for each record. Note that this method is not possible for records at the zone apex, such as example.net, example.org, or example.co.uk (without www or another subdomain in front of the domain name). For records at the zone apex, you must create a policy record using your traffic policy.

Q. Can I create an Alias record pointing to a DNS name that is managed by a traffic policy?

Yes, it is possible to create an Alias record pointing to a DNS name that is being managed by a traffic policy.

Q. Is there a charge for traffic policies that don’t have a policy record?

No. We only charge for policy records; there is no charge for creating the traffic policy itself.

Q. How am I billed for using Amazon Route 53 Traffic Flow?

You are billed per policy record. A policy record represents the application of a Traffic Flow policy to a specific DNS name (such as www.example.com) in order to use the traffic policy to manage how requests for that DNS name are answered. Billing is monthly and is prorated for partial months. There is no charge for traffic policies that are not associated with a DNS name via a policy record. For details on pricing, see the Amazon Route 53 pricing page.

Q. What are the advanced query types supported in Amazon Route 53 Traffic Flow?

Traffic Flow supports all Amazon Route 53 DNS Routing policies including latency, endpoint health, multivalue; answers, weighted round robin, and geo. In addition to these, Traffic Flow also supports geoproximity based routing with traffic biasing.

Q. How does a traffic policy using geoproximity rule route DNS traffic?

When you create a traffic flow policy, you can specify either an AWS region (if you're using AWS resources) or the latitude and longitude for each endpoint. For example, suppose you have EC2 instances in the AWS US East (Ohio) region and in the US West (Oregon) region. When an user in Seattle visits your website, geoproximity routing will route the DNS query to the EC2 instances in the US West (Oregon) region because it's closer geographically. For more information please see the documentation on geoproximity routing.

Q. How does the geoproximity bias value of an endpoint affect DNS traffic routing to other endpoints?

Changing the geoproximity bias value on an endpoint either expands or shrinks the area from which Route 53 routes traffic to a resource. The geoproximity bias can't accurately predict the load factor, though, because a small shift in the size of geographic areas might include or exclude major metropolitan areas that generate large numbers of queries. For more information please refer to our documentation.

Q. Can I use bias for other Traffic Flow rules?

As of today, bias can only be applied to geoproximity rules.

Private DNS

Q. What is Private DNS?

Private DNS is a Route 53 feature that lets you have authoritative DNS within your VPCs without exposing your DNS records (including the name of the resource and its IP address(es) to the Internet.

Q. Can I use Amazon Route 53 to manage my organization's private IP addresses?

Yes, you can manage private IP addresses within Virtual Private Clouds (VPCs) using Amazon Route 53's Private DNS feature. With Private DNS, you can create a private hosted zone, and Route 53 will only return these records when queried from within the VPC(s) that you have associated with your private hosted zone. For more details, see the Amazon Route 53 Documentation.

Q. How do I set up Private DNS?

You can set up Private DNS by creating a hosted zone in Route 53, selecting the option to make the hosted zone "private", and associating the hosted zone with one of your VPCs. After creating the hosted zone, you can associate it with additional VPCs. See the Amazon Route 53 Documentation for full details on how to configure Private DNS.

Q. Do I need connectivity to the outside Internet in order to use Private DNS?

You can resolve internal DNS names from resources within your VPC that do not have Internet connectivity. However, to update the configuration for your Private DNS hosted zone, you need Internet connectivity to access the Route 53 API endpoint, which is outside of VPC.

Q. Can I still use Private DNS if I'm not using VPC?

No. Route 53 Private DNS uses VPC to manage visibility and provide DNS resolution for private DNS hosted zones. To take advantage of Route 53 Private DNS, you must configure a VPC and migrate your resources into it.

Q. Can I use the same private Route 53 hosted zone for multiple VPCs?

Yes, you can associate multiple VPCs with a single hosted zone.

Q. Can I associate VPCs and private hosted zones that I created under different AWS accounts?

Yes, you can associate VPCs belonging to different accounts with a single hosted zone. You can see more details here.

Q. Will Private DNS work across AWS regions?

Yes. DNS answers will be available within every VPC that you associate with the private hosted zone. Note that you will need to ensure that the VPCs in each region have connectivity with each other in order for resources in one region to be able to reach resources in another region. Route 53 Private DNS is supported today in the US East (Northern Virginia), US West (Northern California), US West (Oregon), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), EU (Frankfurt), EU (Ireland), and South America (Sao Paulo) regions.

Q. Can I configure DNS Failover for Private DNS hosted zones?

Yes, it is possible to configure DNS Failover by associating health checks with resource record sets within a Private DNS hosted zone. If your endpoints are within a Virtual Private Cloud (VPC), you have several options to configure health checks against these endpoints. If the endpoints have public IP addresses, then you can create a standard health check against the public IP address of each endpoint. If your endpoints only have private IP addresses, then you cannot create standard health checks against these endpoints. However, you can create metric based health checks, which function like standard Amazon Route 53 health checks except that they use an existing Amazon CloudWatch metric as the source of endpoint health information instead of making requests against the endpoint from external locations.

Q. Can I use Private DNS to block domains and DNS names that I don’t want to be reached from within my VPC?

Yes, you can block domains and specific DNS names by creating these names in one or more Private DNS hosted zones and pointing these names to your own server (or another location that you manage).

Health Checks & DNS Failover

Q. What is DNS Failover?

DNS Failover consists of two components: health checks and failover. Health checks are automated requests sent over the Internet to your application to verify that your application is reachable, available, and functional. You can configure the health checks to be similar to the typical requests made by your users, such as requesting a web page from a specific URL. With DNS failover, Route 53 only returns answers for resources that are healthy and reachable from the outside world, so that your end users are routed away from a failed or unhealthy part of your application.

Q. How do I get started with DNS Failover?

Visit the Amazon Route 53 Developer Guide for details on getting started. You can also configure DNS Failover from within the Route 53 Console.

Q. Does DNS Failover support Elastic Load Balancers (ELBs) as endpoints?

Yes, you can configure DNS Failover for Elastic Load Balancers (ELBs). To enable DNS Failover for an ELB endpoint, create an Alias record pointing to the ELB and set the “Evaluate Target Health” parameter to true. Route 53 creates and manages the health checks for your ELB automatically. You do not need to create your own Route 53 health check of the ELB. You also do not need to associate your resource record set for the ELB with your own health check, because Route 53 automatically associates it with the health checks that Route 53 manages on your behalf. The ELB health check will also inherit the health of your backend instances behind that ELB. For more details on using DNS Failover with ELB endpoints, please consult the Route 53 Developer Guide.

Q. Can I configure a backup site to be used only when a health check fails?

Yes, you can use DNS Failover to maintain a backup site (for example, a static site running on an Amazon S3 website bucket) and fail over to this site in the event that your primary site becomes unreachable.

Q. What DNS record types can I associate with Route 53 health checks?

You can associate any record type supported by Route 53 except SOA and NS records.

Q. Can I health check an endpoint if I don’t know its IP address?

Yes. You can configure DNS Failover for Elastic Load Balancers and Amazon S3 website buckets via the Amazon Route 53 Console without needing to create a health check of your own. For these endpoint types, Route 53 automatically creates and manages health checks on your behalf which are used when you create an Alias record pointing to the ELB or S3 website bucket and enable the "Evaluate Target Health" parameter on the Alias record.

For all other endpoints, you can specify either the DNS name (e.g. www.example.com) or the IP address of the endpoint when you create a health check for that endpoint.

Q. One of my endpoints is outside AWS. Can I set up DNS Failover on this endpoint?

Yes. Just like you can create a Route 53 resource record that points to an address outside AWS, you can set up health checks for parts of your application running outside AWS, and you can fail over to any endpoint that you choose, regardless of location. For example, you may have a legacy application running in a datacenter outside AWS and a backup instance of that application running within AWS. You can set up health checks of your legacy application running outside AWS, and if the application fails the health checks, you can fail over automatically to the backup instance in AWS.

Q. If failover occurs and I have multiple healthy endpoints remaining, will Route 53 consider the load on my healthy endpoints when determining where to send traffic from the failed endpoint?

No, Route 53 does not make routing decisions based on the load or available traffic capacity of your endpoints. You will need to ensure that you have available capacity at your other endpoints, or the ability to scale at those endpoints, in order to handle the traffic that had been flowing to your failed endpoint.

Q. How many consecutive health check observations does an endpoint need to fail to be considered "failed"?

The default is a threshold of three health check observations: when an endpoint has failed three consecutive observations, Route 53 will consider it failed. However, Route 53 will continue to perform health check observations on the endpoint and will resume sending traffic to it once it passes three consecutive observations. You can change this threshold to any value between 1 and 10 observations. For more details, see the Amazon Route 53 Developer Guide.

Q. When my failed endpoint becomes healthy again, how is the DNS failover reversed?

After a failed endpoint passes the number of consecutive health check observations that you specify when creating the health check (the default threshold is three observations), Route 53 will restore its DNS records automatically, and traffic to that endpoint will resume with no action required on your part.

Q. What is the interval between health check observations?

By default, health check observations are conducted at an interval of 30 seconds. You can optionally select a fast interval of 10 seconds between observations.

By checking three times more often, fast interval health checks enable Route 53 to confirm more quickly that an endpoint has failed, shortening the time required for DNS failover to redirect traffic in response to the endpoint's failure.

Fast interval health checks also generate three times the number of requests to your endpoint, which may be a consideration if your endpoint has a limited capacity to serve web traffic. Visit the Route 53 pricing page for details on pricing for fast interval health checks and other optional health check features. For more details, see the Amazon Route 53 Developer Guide.

Q. How much load should I expect a health check to generate on my endpoint (for example, a web server)?

Each health check is conducted from multiple locations around the world. The number and set of locations is configurable; you can modify the number of locations from which each of your health checks is conducted using the Amazon Route 53 console or API. Each location checks the endpoint independently at the interval that you select: the default interval of 30 seconds, or an optional fast interval of 10 seconds. Based on the current default number of health checking locations, you should expect your endpoint to receive one request every 2-3 seconds on average for standard interval health checks and one or more requests per second for fast-interval health checks.

Q. Do Route 53 health checks follow HTTP redirects?

No. Route 53 health checks consider an HTTP 3xx code to be a successful response, so they don't follow the redirect. This may cause unexpected results for string-matching health checks. The health check searches for the specified string in the body of the redirect. Because the health check doesn't follow the redirect, it never sends a request to the location that the redirect points to and never gets a response from that location. For string matching health checks, we recommend that you avoid pointing the health check at a location that returns an HTTP redirect.

Q. What is the sequence of events when failover happens?

In simplest terms, the following events will take place if a health check fails and failover occurs:

Route 53 conducts a health check of your application. In this example, your application fails three consecutive health checks, triggering the following events.

Route 53 disables the resource records for the failed endpoint and no longer serves these records. This is the failover step, which causes traffic to begin being routed to your healthy endpoint(s) instead of your failed endpoint.

Q. Do I need to adjust the TTL for my records in order to use DNS Failover?

The time for which a DNS resolver caches a response is set by a value called the time to live (TTL) associated with every record. We recommend a TTL of 60 seconds or less when using DNS Failover, to minimize the amount of time it takes for traffic to stop being routed to your failed endpoint. In order to configure DNS Failover for ELB and S3 Website endpoints, you need to use Alias records which have fixed TTL of 60 seconds; for these endpoint types, you do not need to adjust TTLs in order to use DNS Failover.

Q. What happens if all of my endpoints are unhealthy?

Route 53 can only fail over to an endpoint that is healthy. If there are no healthy endpoints remaining in a resource record set, Route 53 will behave as if all health checks are passing.

Q. Can I use DNS Failover without using Latency Based Routing (LBR)?

Yes. You can configure DNS Failover without using LBR. In particular, you can use DNS failover to configure a simple failover scenario where Route 53 monitors your primary website and fails over to a backup site in the event that your primary site is unavailable.

Q. Can I configure a health check on a site accessible only via HTTPS?

Yes. Route 53 supports health checks over HTTPS, HTTP or TCP.

Q. Do HTTPS health checks validate the endpoint’s SSL certificate?

No, HTTPS health checks test whether it’s possible to connect with the endpoint over SSL and whether the endpoint returns a valid HTTP response code. However, they do not validate the SSL certificate returned by the endpoint.

Q. Do HTTPS health checks support Server Name Indication (SNI)?

Yes, HTTPS health checks support SNI.

Q. How can I use health checks to verify that my web server is returning the correct content?

You can use Route 53 health checks to check for the presence of a designated string in a server response by selecting the “Enable String Matching” option. This option can be used to check a web server to verify that the HTML it serves contains an expected string. Or, you can create a dedicated status page and use it to check the health of the server from an internal or operational perspective. For more details, see the Amazon Route 53 Developer Guide.

Q. How do I see the status of a health check that I've created?

You can view the current status of a health check, as well as details on why it has failed, in the Amazon Route 53 console and via the Route 53 API.

Additionally, each health check's results are published as Amazon CloudWatch metrics showing the endpoint's health and, optionally, the latency of the endpoint's response. You can view a graph of the Amazon CloudWatch metric in the health checks tab of the Amazon Route 53 console to see the current and historical status of the health check. You can also create Amazon CloudWatch alarms on the metric in order to send notifications if the status of the health check changes.

The Amazon CloudWatch metrics for all of your Amazon Route 53 health checks are also visible in the Amazon CloudWatch console. Each Amazon CloudWatch metric contains the Health Check ID (for example, 01beb6a3-e1c2-4a2b-a0b7-7031e9060a6a) which you can use to identify which health check the metric is tracking.

Q. How can I measure the performance of my application's endpoints using Amazon Route 53?

Amazon Route 53 health checks include an optional latency measurement feature which provides data on how long it takes your endpoint to respond to a request. When you enable the latency measurement feature, the Amazon Route 53 health check will generate additional Amazon CloudWatch metrics showing the time required for Amazon Route 53's health checkers to establish a connection and to begin receiving data. Amazon Route 53 provides a separate set of latency metrics for each AWS region where Amazon Route 53 health checks are conducted.

Q. How can I be notified if one of my endpoints starts failing its health check?

Because each Route 53 health check publishes its results as a CloudWatch metric, you can configure the full range of CloudWatch notifications and automated actions which can be triggered when the health check value changes beyond a threshold that you specify. First, in either the Route 53 or CloudWatch console, configure a CloudWatch alarm on the health check metric. Then add a notification action and specify the email or SNS topic that you want to publish your notification to. Please consult the Route 53 Developer Guide for full details.

Q: I created an alarm for my health check, but I need to re-send the confirmation email for the alarm's SNS topic. How can I re-send this email?

Confirmation emails can be re-sent from the SNS console. To find the name of the SNS topic associated with the alarm, click the alarm name within the Route 53 console and looking in the box labeled "Send notification to."

Within the SNS console, expand the list of topics, and select the topic from your alarm. Open the "Create Subscription" box and select Email for protocol and enter the desired email address. Clicking "Subscribe" will re-send the confirmation email.

Q. I'm using DNS Failover with Elastic Load Balancers (ELBs) as endpoints. How can I see the status of these endpoints?

The recommended method for setting up DNS Failover with ELB endpoints is to use Alias records with the "Evaluate Target Health" option. Because you don't create your own health checks for ELB endpoints when using this option, there are no specific CloudWatch metrics generated by Route 53 for these endpoints.

You can get metrics on the health of your load balancer in two ways. First, Elastic Load Balancing publishes metrics that indicate the health of the load balancer and the number of healthy instances behind it. For details on configuring CloudWatch metrics for ELB, consult the ELB developer guide. Second, you can create your own health check against the CNAME provided by the ELB, e.g. elb-example-123456678.us-west-2.elb.amazonaws.com. You won't use this health check for DNS Failover itself (because the "Evaluate Target Health" option provides DNS Failover for you), but you can view the CloudWatch metrics for this health check and create alarms to be notified if the health check fails.

For complete details on using DNS Failover with ELB endpoints, please consult the Route 53 Developer Guide.

Q. For Alias records pointing to Amazon S3 Website buckets, what is being health checked when I set Evaluate Target Health to "true"?

Amazon Route 53 performs health checks of the Amazon S3 service itself in each AWS region. When you enable Evaluate Target Health on an Alias record pointing to an Amazon S3 Website bucket, Amazon Route 53 will take into account the health of the Amazon S3 service in the AWS region where your bucket is located. Amazon Route 53 does not check whether a specific bucket exists or contains valid website content; Amazon Route 53 will only fail over to another location if the Amazon S3 service itself is unavailable in the AWS region where your bucket is located.

Q. What is the cost to use CloudWatch metrics for my Route 53 health checks?

CloudWatch metrics for Route 53 health checks are available free of charge.

Q. Can I configure DNS Failover based on internal health metrics, such as CPU load, network, or memory?

Yes. Amazon Route 53’s metric based health checks let you perform DNS failover based on any metric that is available within Amazon CloudWatch, including AWS-provided metrics and custom metrics from your own application. When you create a metric based health check within Amazon Route 53, the health check becomes unhealthy whenever its associated Amazon CloudWatch metric enters an alarm state.

Metric based health checks are useful to enable DNS failover for endpoints that cannot be reached by a standard Amazon Route 53 health check, such as instances within a Virtual Private Cloud (VPC) that only have private IP addresses. Using Amazon Route 53’s calculated health check feature, you can also accomplish more sophisticated failover scenarios by combining the results of metric based health checks with the results of standard Amazon Route 53 health checks, which make requests against an endpoint from a network of checkers around the world. For example, you can create a configuration which fails away from an endpoint if either its public-facing web page is unavailable, or if internal metrics such as CPU load, network in/out, or disk reads show that the server itself is unhealthy.

Q. My web server is receiving requests from a Route 53 health check that I did not create. How can I stop these requests?

Occasionally, Amazon Route 53 customers create health checks that specify an IP address or domain name that does not belong to them. If your web server is getting unwanted HTTP(s) requests that you have traced to Amazon Route 53 health checks, please provide information on the unwanted health check using this form, and we will work with our customer to fix the problem.

Q. If I specify a domain name as my health check target, will Amazon Route 53 check over IPv4 or IPv6?

If you specify a domain name as the endpoint of an Amazon Route 53 health check, Amazon Route 53 will look up the IPv4 address of that domain name and will connect to the endpoint using IPv4. Amazon Route 53 will not attempt to look up the IPv6 address for an endpoint that is specified by domain name. If you want to perform a health check over IPv6 instead of IPv4, select "IP address" instead of "domain name" as your endpoint type, and enter the IPv6 address in the “IP address” field.

Q. Where can I find the IPv6 address ranges for Amazon Route 53's DNS servers and health checkers?

AWS now publishes its current IP address ranges in JSON format. To view the current ranges, download the .json file using the following link. If you access this file programmatically, ensure that the application downloads the file only after successfully verifying the TLS certificate that is returned by the AWS server.

Download: ip-ranges.json

To find IP ranges for Route 53 servers, search for the following values in the "service" field:

Route 53 DNS servers: Search for "ROUTE53"

Route 53 health checkers: Search for "ROUTE53_HEALTHCHECKS"

For more information, see AWS IP Address Ranges in the Amazon Web Services General Reference.

Please note that the IPv6 ranges may not yet appear in this file. For reference, the IPv6 ranges for Amazon Route 53 health checkers are as follows:

2600:1f1c:7ff:f800::/53
2a05:d018:fff:f800::/53
2600:1f1e:7ff:f800::/53
2600:1f1c:fff:f800::/53
2600:1f18:3fff:f800::/53
2600:1f14:7ff:f800::/53
2600:1f14:fff:f800::/53
2406:da14:7ff:f800::/53
2406:da14:fff:f800::/53
2406:da18:7ff:f800::/53
2406:da1c:7ff:f800::/53
2406:da1c:fff:f800::/53
2406:da18:fff:f800::/53
2600:1f18:7fff:f800::/53
2a05:d018:7ff:f800::/53
2600:1f1e:fff:f800::/53
2620:107:300f::36b7:ff80/122
2a01:578:3::36e4:1000/122
2804:800:ff00::36e8:2840/122
2620:107:300f::36f1:2040/122

2406:da00:ff00::36f3:1fc0/122
2620:108:700f::36f4:34c0/122
2620:108:700f::36f5:a800/122
2400:6700:ff00::36f8:dc00/122
2400:6700:ff00::36fa:fdc0/122
2400:6500:ff00::36fb:1f80/122
2403:b300:ff00::36fc:4f80/122
2403:b300:ff00::36fc:fec0/122
2400:6500:ff00::36ff:fec0/122
2406:da00:ff00::6b17:ff00/122
2a01:578:3::b022:9fc0/122
2804:800:ff00::b147:cf80/122

Domain Name Registration

Q. Can I register domain names with Amazon Route 53?

Yes. You can use the AWS Management Console or API to register new domain names with Route 53. You can also request to transfer in existing domain names from other registrars to be managed by Route 53. Domain name registration services are provided under our Domain Name Registration Agreement.

Q. What Top Level Domains ("TLDs") do you offer?

Route 53 offers a wide selection of both generic Top Level Domains ("gTLDs": for example, .com and .net) and country-code Top Level Domains ("ccTLDs": for example, .de and .fr). For the complete list, please see the Route 53 Domain Registration Price List.

Q. How can I register a domain name with Route 53?

To get started, log into your account and click on "Domains". Then, click the big blue "Register Domain" button and complete the registration process.

Q. How long does it take to register a domain name?

Depending on the TLD you've selected, registration can take from a few minutes to several hours. Once the domain is successfully registered, it will show up in your account.

Q. How long is my domain name registered for?

The initial registration period is typically one year, although the registries for some top-level domains (TLDs) have longer registration periods. When you register a domain with Amazon Route 53 or you transfer domain registration to Amazon Route 53, we configure the domain to renew automatically. For more information, see Renewing Registration for a Domain in the Amazon Route 53 Developer Guide.

Q. What information do I need to provide to register a domain name?

In order to register a domain name, you need to provide contact information for the registrant of the domain, including name, address, phone number, and email address. If the administrative and technical contacts are different, you need to provide that contact information, too.

Q. Why do I need to provide personal information to register a domain?

ICANN, the governing body for domain registration, requires that registrars provide contact information, including name, address, and phone number, for every domain name registration, and that registrars make this information publicly available via a Whois database. For domain names that you register as an individual (i.e., not as a company or organization), Route 53 provides privacy protection, which hides your personal phone number, email address, and physical address, free of charge. Instead, the Whois contains the registrar's name and mailing address, along with a registrar-generated forwarding email address that third parties may use if they wish to contact you.

Q. Does Route 53 offer privacy protection for domain names I have registered?

Yes, Route 53 provides privacy protection at no additional charge. The privacy protection hides your phone number, email address, and physical address. Your first and last name will be hidden if the TLD registry and registrar allow it. When you enable privacy protection, a Whois query for the domain will contain the registrar's mailing address in place of your physical address, and the registrar's name in place of your name (if allowed). Your email address will be a registrar-generated forwarding email address that third parties may use if they wish to contact you. Domain names registered by companies or organizations are eligible for privacy protection if the TLD registry and registrar allow it.

Q. Where can I find the requirements for specific TLDs?

For a list of TLDs please see the price list and for the specific registration requirements for each, please see the Amazon Route 53 Developer Guide and our Domain Name Registration Agreement.

Q. What name servers are used to register my domain name?

When your domain name is created we automatically associate your domain with four unique Route 53 name servers, known as a delegation set. You can view the delegation set for your domain in the Amazon Route 53 console. They're listed in the hosted zone that we create for you automatically when you register a domain.

By default, Route 53 will assign a new, unique delegation set for each hosted zone you create. However, you can also use the Route 53 API to create a “reusable delegation set”, which you can then apply to multiple hosted zones that you create. For customers with large numbers of domain names, reusable delegation sets make migration to Route 53 simple, because you can instruct your domain name registrar to use the same delegation set for all your domains managed by Route 53. This feature also makes it possible for you to create “white label” name server addresses such as ns1.example.com, ns2.example.com, etc., which you can point to your Route 53 name servers. You can then use your “white label” name server addresses as the authoritative name servers for as many of your domain names as desired. For more details, see the Amazon Route 53 documentation.

Q. Will I be charged for my name servers?

You will be charged for the hosted zone that Route 53 creates for your domain name, as well as for the DNS queries against this hosted zone that Route 53 serves on your behalf. If you do not wish to be charged for Route 53’s DNS service, you can delete your Route 53 hosted zone. Please note that some TLDs require you to have valid name servers as part of your domain name registration. For a domain name under one of these TLDs, you will need to procure DNS service from another provider and enter that provider’s name server addresses before you can safely delete your Route 53 hosted zone for that domain name.

Q. What is Amazon Registrar, Inc. and what is a registrar of record?

AWS resells domain names that are registered with ICANN-accredited registrars. Amazon Registrar, Inc. is an Amazon company that is accredited by ICANN to register domains. The registrar of record is the “Sponsoring Registrar” listed in the WHOIS record for your domain to indicate which registrar your domain is registered with.

Q. Who is Gandi?

Amazon is a reseller of the registrar Gandi. As the registrar of record, Gandi is required by ICANN to contact the registrant to verify their contact information at the time of initial registration. You MUST verify your contact information if requested by Gandi within the first 15 days of registration in order to prevent your domain name from being suspended. Gandi also sends out reminder notices before the domain comes up for renewal.

Q. Which top-level domains does Amazon Route 53 register through Amazon Registrar and which ones does it register through Gandi?

See our documentation for a list of the domains that you can currently register using Amazon Route 53. This list includes information about which registrar is the current registrar of record for each TLD that we sell.

Q. Can I transfer my .com and .net domain registrations from Gandi to Amazon?

No. We plan to add this functionality soon.

Q. What is Whois? Why is my information shown in Whois?

Whois is a publicly available database for domain names that lists the contact information and the name servers that are associated with a domain name. Anyone can access the Whois database by using the WHOIS command, which is widely available. It's included in many operating systems, and it's also available as a web application on many websites. The Internet Corporation for Assigned Names and Numbers (ICANN) requires that all domain names have publicly available contact information in case someone needs to get in contact with the domain name holder.

Q. How do I transfer my domain name to Route 53?

To get started, log into your account and click on “Domains”. Then, click the “Transfer Domain” button at the top of the screen and complete the transfer process. Please make sure before you start the transfer process, (1) your domain name is unlocked at your current registrar, (2) you have disabled privacy protection on your domain name (if applicable), and (3) that you have obtained the valid Authorization Code, or “authcode”, from your current registrar which you will need to enter as part of the transfer process.

Q. How do I transfer my existing domain name registration to Amazon Route 53 without disrupting my existing web traffic?

First, you need to get a list of the DNS record data for your domain name, generally available in the form of a “zone file” that you can get from your existing DNS provider. With the DNS record data in hand, you can use Route 53’s Management Console or simple web-services interface to create a hosted zone that can store the DNS records for your domain name and follow its transfer process, which will include such steps as updating the name servers for your domain name to the ones associated with your hosted zone. To complete the domain name transfer process, contact the registrar with whom you registered your domain name and follow its transfer process, which will include steps such as updating the name servers for your domain name to the ones associated with your hosted zone. As soon as your registrar propagates the new name server delegations, the DNS queries from your end users will start to get answered by the Route 53 DNS servers.

Q. How do I check on the status of my transfer request?

You can view the status of domain name transfers in the "Alerts" section on the homepage of the Route 53 console.

Q. What do I do if my transfer wasn't successful?

You will need to contact your current registrar in order to determine why your transfer failed. Once they have resolved the issue, you can resubmit your transfer request.

Q. How do I transfer my domain name to a different registrar?

In order to move your domain name away from Route 53, you need to initiate a transfer request with your new registrar. They will request the domain name be moved to their management.

Q. Is there a limit to the number of domains I can manage using Amazon Route 53?

Each new Amazon Route 53 account is limited to a maximum of 50 domains. Complete our request form for a higher limit and we will respond to your request within two business days.

Q. Does Amazon Route 53 DNS support DNSSEC?

Amazon Route 53's DNS services does NOT support DNSSEC at this time. However, our domain name registration service supports configuration of signed DNSSEC keys for domains when DNS service is configured at another provider. More information on configuring DNSSEC for your domain name registration can be found here.

Q. How do I transfer a domain registration that has DNSSEC enabled to Amazon Route 53?

See our documentation for a step-by-step guide on transferring your DNSSEC-enabled domain to Amazon Route 53.

Route 53 Resolver

Q. What is Amazon Route 53 Resolver?

Route 53 Resolver is a regional DNS service that provides recursive DNS lookups for names hosted in EC2 as well as public names on the internet. This functionality is available by default in every Amazon Virtual Private Cloud (VPC). For hybrid cloud scenarios you can configure conditional forwarding rules and DNS endpoints to enable DNS resolution across AWS Direct Connect and AWS Managed VPN.

Q. What is recursive DNS?

Amazon Route 53 is both an Authoritative DNS service and Recursive DNS service. Authoritative DNS contains the final answer to a DNS query, generally an IP address. Clients (such as mobile devices, applications running in the cloud, or servers in your datacenter) don’t actually talk directly to authoritative DNS services, except in very rare cases. Instead, clients talk to recursive DNS services (also known as DNS resolvers) which find the correct authoritative answer for any DNS query. Route 53 Resolver is a recursive DNS service.

When receiving a query, a recursive DNS service like Route 53 Resolver may either be configured to automatically forward the query directly to a specific recursive DNS server, or it may recursively search beginning with the root of the domain and continuing until it finds the final answer. In either case, once an answer is found, the recursive DNS server may cache the answer for a period of time so it can answer subsequent queries for the same name more quickly in the future.

Q. What are conditional forwarding rules?

Conditional forwarding rules allow Resolver to forward queries for specified domains to the target IP address of your choice, typically an on-premises DNS resolver. Rules are applied at the VPC level and can be managed from one account and shared across multiple accounts.

Q. What are DNS endpoints?

A DNS endpoint includes one or more elastic network interfaces (ENI) that attach to your Amazon Virtual Private Cloud (VPC). Each ENI is assigned an IP address from the subnet space of the VPC where it is located. This IP address can then serve as a forwarding target for on-premises DNS servers to forward queries. Endpoints are required both for DNS query traffic that you're forwarding from VPCs to your network and from your network to your VPCs over AWS Direct Connect and Managed VPN.

Q. How do I share rules across accounts?

Route 53 Resolver is integrated with AWS Resource Access Manager (RAM) which provides customers with a simple way to share their resources across AWS accounts or within their AWS Organization. Rules can be created in one primary account and then shared across multiple accounts using RAM. Once shared, the rules still need to be applied to VPCs in those accounts before they can take effect. For more information, see the AWS RAM documentation.

Q. What happens if I decide to stop sharing rules with other accounts?

Those rules will no longer be usable by the accounts you previously shared them with. This means that if those rules were associated to VPCs in those accounts, they will be disassociated from those VPCs.

Q. What regions are available for Route 53 Resolver?

Visit our AWS Region Table to see which regions Route 53 Resolver has launched in.

Q. Does regional support for Route 53 Resolver mean that all of Amazon Route 53 is now regional?

No. Amazon Route 53 public and private DNS, traffic flow, health checks, and domain name registration are all global services.

Q. How do I get started with Route 53 Resolver?

Visit the Amazon Route 53 developer guide for details on getting started. You can also configure Resolver from within the Amazon Route 53 console.

Learn more about Amazon Route 53 pricing

Simple pricing to only pay for what you need.

Learn more »

Sign up for a free account

Instantly get access to the AWS Free Tier.

Sign up »

Start building in the console

Get started with Amazon Route 53 in the AWS Console.

Sign in »

AWS App Mesh FAQs

General

Q: What is AWS App Mesh?

A: AWS App Mesh is a new technology that makes it easy to monitor, control, and debug the communications between services. App Mesh uses Envoy, an open source service mesh proxy which is deployed alongside your microservice containers. App Mesh is integrated with AWS services for monitoring and tracing, and it works with many popular third-party tools. App Mesh can be used with microservice containers managed by Amazon ECS, Amazon EKS, AWS Fargate, Kubernetes running on AWS, and services running on Amazon EC2.

Q: Why should I use App Mesh?

A: App Mesh makes it easy to get visibility and control over the communications between your services without writing new code or running additional AWS infrastructure. Using App Mesh, you can standardize how services communicate, implement rules for communications between services, and capture metrics, logs, and traces directly into AWS services and third-party tools of your choice.

Q: How does App Mesh work?

A: App Mesh sets up and manages a service mesh for your services. To do this, you run the open source Envoy proxy alongside each service, and App Mesh configures the proxy to handle all communications into and out of each container. App Mesh collects metrics, such as error rates, and connections per second, which can be exported to Amazon CloudWatch using a statsd collector. Using App Mesh APIs, you can route traffic based on path or weights to specific service versions.

Q: What is a service mesh?

A: A service mesh is a new software layer that handles all of the communications between services. It provides new features to connect and manage connections between services and is independent of each service's code, allowing it to work across network boundaries and with multiple service management systems.

Integrations

Q: How does App Mesh work with Amazon Elastic Container Services (ECS) and AWS Fargate?

A: App Mesh provides new communication, observation, and management capabilities to applications managed by Amazon ECS and AWS Fargate. You add the Envoy proxy image to the task definition. App Mesh manages Envoy configuration to provide service mesh capabilities. App Mesh exports metrics, logs, and traces to the endpoints specified in the Envoy bootstrap configuration provided. App Mesh provides an API to configure traffic routes and other controls between microservices that are mesh-enabled.

Q: How does App Mesh work with Amazon Elastic Container Service for Kubernetes (EKS)?

A: Use the open source AWS App Mesh controller and mutating webhook admission controller. These controllers connect your Kubernetes services to App Mesh and ensure that the Envoy proxy is injected into your pods. App Mesh exports metrics, logs, and traces to the endpoints specified in the Envoy bootstrap configuration provided. App Mesh provides an API to configure traffic routes and other controls between microservices that are mesh-enabled.

Q: How does App mesh work with services running on Amazon EC2?

A: Run the Envoy proxy as a container or process on your EC2 instance. Use the AWS-provided container proxy init container, or run your own script, to redirect network traffic on the instance through the proxy. App Mesh manages Envoy configuration to provide service mesh capabilities. App Mesh exports metrics, logs, and traces to the endpoints specified in the Envoy bootstrap configuration provided. App Mesh provides an API to configure traffic routes and other controls between microservices that are mesh-enabled.

Q: Why should I use App Mesh instead of AWS Elastic Load Balancers?

A: We recommend using AWS Elastic Load Balancing to handle all internet traffic and traffic from clients that are not within your trust boundary. For internal services that connect to other services within an AWS region, App Mesh provides flexibility, consistency, and a greater degree of control and monitoring for services communications.

Monitoring, logging, and tracing

Q: What type of monitoring capabilities does App Mesh provide?

A: With App Mesh, you get consistent metrics and logs for every hop between services. These logs and metrics include metadata such as service-names and request identifiers. With these, you can aggregate, filter, a see graphical dashboards of service-to-service communications using tools like Amazon CloudWatch. Common dashboards might include error rates and error codes between your service and dependent services. App Mesh automatically collects traces for each service and makes it easy to visualize a service map with details of all service API calls. These capabilities make it easier to debug and identify the root cause of communication issues between your microservices.

Q: Can I use non-AWS tools for monitoring, logging, or tracing with App Mesh? Yes.

A: Yes. App Mesh supports any third-party tool that works with Envoy. This includes Splunk, Prometheus, and Grafana, as well as open-tracing solutions like Zipkin and LightStep.

Traffic control

Q: What type of traffic controls does App Mesh provide?

A: App Mesh gives you a set of client-side controls for traffic routing. App Mesh provides APIs to route traffic between applications based on service names and versions. These capabilities make it easier to deploy new versions of your microservices.

Getting started

Q: How much does App Mesh cost?

A: There is no additional charge for using AWS App Mesh. You pay only for the AWS resources (i.e. EC2 instances or requested Fargate CPU and memory) consumed by the App Mesh proxy that runs alongside your containers. You pay only for what you use, as you use it; there are no minimum fees and no upfront commitments.

Q: How do I start using App Mesh?

A: App Mesh is generally available today. You can start using App Mesh from the AWS CLI or SDK. Learn more on the Getting Started page.

AWS Cloud Map FAQs

What is AWS Cloud Map?

AWS Cloud Map is a cloud resource discovery service. Cloud Map enables you to name your application resources with custom names, and it automatically updates the locations of these dynamically changing resources. This increases your application availability because your applications always discover the most up-to-date locations of its resources.

Why should I use Cloud Map?

Modern applications are typically composed of multiple services that are accessible over an API and perform a specific function. Each service interacts with a variety of other resources such as databases, queues, object stores, and customer-defined microservices, and they also need to be able to find the location of all the infrastructure resources on which it depends, in order to function.

Cloud Map allows you to register any application resources such as databases, queues, microservices, and other cloud resources with custom names. Cloud Map then constantly checks the health of resources to make sure the location is up-to-date. The resources in your application can then query the registry for the location of the exact resources needed based on the application version and deployment environment.

What is the difference between Amazon Route 53 Auto Naming and AWS Cloud Map?

Amazon Route 53 Auto Naming, which was released on December 05, 2017, automates service name management in DNS and supported IP-based resources only. AWS Cloud Map extends the capabilities of the Auto Naming APIs by providing a service registry for resources, represented by IPs, URLs, and ARNs and offering an API-based service discovery mechanism with a faster change propagation and the ability to use attributes to narrow down the set of discovered resources. All the existing Amazon Route 53 Auto Naming resources are automatically upgraded to AWS Cloud Map and are available for API-based discovery.

What is a namespace?

A *namespace* is a logical entity in AWS Cloud Map that allows for grouping your services and enforcing a common level of visibility—either public (accessible from the public internet) or private (visible in a specific VPC only). You define how your applications should discover services at the namespace level by allowing the discovery via AWS SDK and API, or enabling optional discovery via DNS.

What is the difference between a “service” and a “resource” in AWS Cloud Map?

In AWS Cloud Map, a service is an application component that serves a particular purpose, such as generating bills or resizing thumbnail images. When a service is deployed, it runs on some type of infrastructure, for example, EC2 instances, ECS tasks, DynamoDB tables, SQS queues, or Lambda functions. In AWS Cloud Map, these are resources. Your service may require only one resource, or it could be running on thousands of resources that dynamically come and go as it scales.

What is the difference between the answers AWS Cloud Map returns over DNS vs. HTTPS?

Over DNS, AWS Cloud Map provides resource locations that consist of IP addresses or IP:port combinations (using either IPv4 or IPv6). Using API-based discovery, AWS Cloud Map can return all of those types of locations, as well as URLs or ARNs. For resources that have IP or IP:port locations, you can specify whether AWS Cloud Map should return resources when queried over DNS, API, or both. For resources that have URL or ARN locations, applications must query over API. When clients query AWS Cloud Map over API, they can narrow down the results by specifying attributes.

How does AWS Cloud Map perform health checking of my IP-based resources?

When you register a resource in AWS Cloud Map, you can optionally specify settings for health checks, including: health-checking type (regular or path-based), an optional path to check, and the number of retries to deem the resource unhealthy. As soon as you register IP-based resources, AWS Cloud Map automatically provisions health checks for these resources based on the settings in the service. AWS Cloud Map also makes resource health information available via the /describeHealthStatus API.

Will AWS Cloud Map work with resources in my VPC?

Yes, AWS Cloud Map allows you to register resources in your VPC and make them discoverable. AWS Cloud Map also includes a regional API that you can use to register and deregister the resources inside or outside of VPC.

How does integration with Amazon Elastic Container Service (ECS) work?

When you create an ECS service, you can choose to enable service discovery by specifying a custom name, attributes, and optional health-checking settings for the service. As ECS launches tasks for your service, it registers them as resources in the AWS Cloud Map service registry, which ensures that the tasks become discoverable via API calls and DNS queries.

How does integration with Amazon Elastic Container Service for Kubernetes (EKS) work?

We created a Cloud Map provider in an open-source Kubernetes connector ExternalDNS that automatically propagates internal service locations to the AWS Cloud Map service registry as Amazon EKS services launch and removes them on termination. All EKS services then become discoverable via AWS Cloud Map, which provides a unified service registry for all container workloads on EKS.

Ready to get started?

Build your first Cloud Map

Sign into the Cloud Map console »

Sign up for a free account

Instantly get access to the AWS Free Tier

Sign up »

Start building in the console

Get started building with Cloud Map in the AWS Console

Sign in »

AWS Direct Connect FAQs

General Questions

Q. What is AWS Direct Connect?

AWS Direct Connect is a network service that provides an alternative to using the Internet to connect customer's on premise sites to AWS.

Q. What can I do with AWS Direct Connect?

Using AWS Direct Connect, data that would have previously been transported over the Internet can now be delivered through a private network connection between AWS and your datacenter or corporate network.

Q. What are the benefits of using AWS Direct Connect and private network connections?

In many circumstances, private network connections can reduce costs, increase bandwidth, and provide a more consistent network experience than Internet-based connections.

Q. Which AWS services can be used with AWS Direct Connect?

All AWS services, including Amazon Elastic Compute Cloud (EC2), Amazon Virtual Private Cloud (VPC), Amazon Simple Storage Service (S3), and Amazon DynamoDB can be used with AWS Direct Connect.

Q. Can I use the same private network connection with Amazon Virtual Private Cloud (VPC) and other AWS services simultaneously?

Yes. Each AWS Direct Connect connection can be configured with one or more virtual interfaces. Virtual interfaces may be configured to access AWS services such as Amazon EC2 and Amazon S3 using public IP space, or resources in a VPC using private IP space.

Q. If I’m using Amazon CloudFront and my origin is in my own data center, can I use AWS Direct Connect to transfer the objects stored in my own data center?

Yes. Amazon CloudFront supports custom origins including origins you run outside of AWS. The access to the CloudFront edge locations will be restricted to the geographically nearest AWS region, with the exception of the North America regions which currently allow access to all North American region's on-net CloudFront origins. With AWS Direct Connect, you will pay AWS Direct Connect data transfer rates for origin transfer.

Through Direct Connect, customer traffic will remain in Amazon's backbone network after it enters it. Therefore, prefixes of CloudFront locations that are not on the Amazon backbone network will not be advertised through Direct Connect. You can also find more details about IP prefixes advertised on AWS Direct Connect public virtual interfaces here. You can also refer to this link to know more about Direct Connect routing policy.

Q. Where is AWS Direct Connect available?

You can find the complete list of Direct Connect locations on the Product Details page.

Q. Can I use AWS Direct Connect if my network is not present at an AWS Direct Connect location?

Yes. AWS Direct Connect Partners can help you extend your preexisting data center or office network to an AWS Direct Connect location. Please see AWS Direct Connect Partners for more information.

Q. How can I get started with AWS Direct Connect?

Use the AWS Direct Connect tab on the AWS Management Console to create a new connection. Then you will change the region to the region you wish to use. When requesting a connection, you will be asked to select the AWS Direct Connect location you wish to use, the number of ports, and the port speed. You will also have the opportunity to request to have an AWS Direct Connect Partner contact you if you need assistance extending your office or data center network to the AWS Direct Connect location.

Q. Can I order a port for AWS GovCloud (US) in the AWS Management Console?

If you wish to order a port to connect to AWS GovCloud (US) you will need to use the AWS GovCloud (US) management console. Details about getting started in the AWS GovCloud (US) region can be found here.

Billing

Q. Are there any setup charges or a minimum service term commitment required to use AWS Direct Connect?

There are no setup charges, and you may cancel at any time. Services provided by AWS Direct Connect Partners may have other terms or restrictions that apply.

Q. How will I be charged and billed for my use of AWS Direct Connect?

AWS Direct Connect has two separate charges: port-hours and Data Transfer. Pricing is per port-hour consumed for each port type. Partial port-hours consumed are billed as full hours. The account that owns the port will be charged the port-hour charges.

Data Transfer via AWS Direct Connect will be billed in the same month in which the usage occurred. See additional Q & A below to understand how Data Transfer will be billed.

Q. Will regional data transfer be billed at the AWS Direct Connect rate?

No, data transfer between Availability Zones in a region will be billed at the regular regional data transfer rate in the same month in which the usage occurred.

Q. What defines billable port-hours for Dedicated Connections?

Port-hours are billed once the connection between the AWS router and your router is established, or 90 days after you ordered the port, whichever comes first. Port charges will continue to be billed anytime the AWS Direct Connect port is provisioned for your use. If you no longer wish to be charged for your port, please follow the cancellation process detailed in How do I cancel the AWS Direct Connect service?

Q. What defines billable port-hours for Hosted Connections?

Port-hours are billed once you have accepted the Hosted Connection. Port charges will continue to be billed as long as the Hosted Connection is provisioned for your use. If you no longer wish to be charged for your Hosted Connection, please work with the AWS Direct Connect Partner to cancel the Hosted Connection.

Q. What is the format for Hosted Connection port-hour charges?

All Hosted Connection port-hour charges at a Direct Connect location are grouped by capacity.

For example, consider the bill for a customer with two separate 200Mbps Hosted Connections at a Direct Connect location, and no other Hosted Connections at that location. The port-hour charges for the two separate 200Mbps Hosted Connections will be summarized under a single item with a label ending in “HCPortUsage:200M”. For a month with 720 total hours, the port-hour total for this item will be 1,440, or the total number of hours in the month multiplied by the total number of 200Mbps Hosted Connections at this location.

The Hosted Connection capacity identifiers which may appear on your bill are as follows:

HCPortUsage:50M
HCPortUsage:100M
HCPortUsage:200M
HCPortUsage:300M
HCPortUsage:400M
HCPortUsage:500M
HCPortUsage:1G
HCPortUsage:2G
HCPortUsage:5G
HCPortUsage:10G

Note that these capacity identifiers will appear by location depending on which Hosted Connection capacities you have at each location.

Q. Which AWS account gets charged for the Data Transfer Out performed over a public virtual interface?

For publicly addressable AWS resources (for example, Amazon S3 buckets, Classic EC2 instances, or EC2 traffic that goes through an internet gateway), if the outbound traffic is destined for public prefixes owned by the same AWS payer account and actively advertised to AWS through an AWS Direct Connect public virtual Interface, the Data Transfer Out (DTO) usage is metered toward the resource owner at AWS Direct Connect data transfer rate.

For AWS Direct Connect pricing information, please see AWS Direct Connect pricing. If using an AWS Direct Connect Partner to facilitate a Direct Connect connection, contact the AWS Direct Connect Partner regarding any fees they may charge.

Q. Which AWS account gets charged for the Data Transfer Out performed over a transit/private virtual interface?

With the introduction of the granular Data Transfer Out allocation feature, the AWS account responsible for the Data Transfer Out will be charged for the Data Transfer Out performed over a transit/private virtual interface. The AWS account responsible for the Data Transfer Out will be determined based on the customer’s use of the private/transit virtual interface as follows:

- Private virtual interface(s) is used to interface with Amazon Virtual Private Cloud(s) with or without Direct Connect gateway(s). In the case of the private virtual interface, the AWS account owning the AWS resources responsible for the Data Transfer Out will be charged.
- Transit virtual interface(s) is used to interface with AWS Transit Gateway(s). In the case of the transit virtual interface, the AWS account owning the Amazon Virtual Private Cloud(s) attached to the AWS Transit Gateway associated with the Direct Connect gateway attached to the transit virtual interface will be charged. Please note that all applicable AWS Transit Gateway specific charges (Data Processing and Attachment) will be in addition to the AWS Direct Connect Data Transfer Out.

Q. How does AWS Direct Connect work with consolidated billing?

AWS Direct Connect data transfer usage will be aggregated to your master account.

Q. How do I cancel the AWS Direct Connect service?

You can cancel AWS Direct Connect service by deleting your ports from the AWS management console. You should also cancel any service(s) offered by a third party. For example, contact the colocation provider to disconnect any cross-connects to AWS Direct Connect, and/or a network service provider who may be providing network connectivity from your remote locations to the AWS Direct Connect location.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Technical

Q. What connection capacities are supported by AWS Direct Connect?

For Dedicated Connections, 1Gbps and 10Gbps ports are available. For Hosted Connections, capacities of 50Mbps, 100Mbps, 200Mbps, 300Mbps, 400Mbps, 500Mbps, 1Gbps, 2Gbps, 5Gbps and 10Gbps may be ordered from approved AWS Direct Connect Partners. See AWS Direct Connect Partners for more information.

Q. Are there limits on the amount of data that I can transfer using AWS Direct Connect?

No. You may transfer any amount of data up to the limit of your selected port capacity.

Q. Are there limits on the number of routes I can advertise towards AWS using AWS Direct Connect?

Yes, you can advertise up to 100 routes over each Border Gateway Protocol session using AWS Direct Connect.

Q. What happens if I advertise more than 100 routes over a Border Gateway Protocol session?

Your Border Gateway Protocol session will go down if you advertise more than 100 routes over a Border Gateway Protocol session. This will prevent all network traffic flowing over that virtual interface until you reduce the number of routes to less than 100.

No. You may transfer any amount of data up to the limit of your selected port capacity.Q. What are the technical requirements for the connection?

AWS Direct Connect supports 1000BASE-LX or 10GBASE-LR connections over singlemode fiber using Ethernet transport. Your device must support 802.1Q VLANs. See the AWS Direct Connect User Guide for more detailed requirements information.

Q. What AWS region(s) can I connect to via this connection?

Using direct connect gateway, you can connect to VPCs deployed in any AWS Region from this location. See the Direct Connect Gateway page to get more details.

Direct connect locations can also access the public resources in any AWS Region using a public virtual interface.

Q. What Availability Zone(s) can I connect to via this connection?

Using direct connect gateway, you can connect to VPCs deployed in any AWS Region Availability Zone(s) from this location. See the direct connect gateway page to get more details.

Q. Are connections to AWS Direct Connect redundant?

Each connection consists of a single dedicated connection between ports on your router and an Amazon router. We recommend establishing a second connection if redundancy is required. When you request multiple ports at the same AWS Direct Connect location, they will be provisioned on redundant Amazon routers. To achieve high availability, we recommend you to have connections at multiple AWS Direct Connect locations. You can refer to this page to learn more about achieving highly available network connectivity.

Q. Will I lose connectivity if my AWS Direct Connect link fails?

If you have established a second AWS Direct Connect connection, traffic will failover to the second link automatically. We recommend enabling Bidirectional Forwarding Detection (BFD) when configuring your connections to ensure fast detection and failover.

To achieve high availability, we recommend you to have connections at multiple AWS Direct Connect locations. You can refer to this page to learn more about achieving highly available network connectivity.

If you have configured a back-up IPsec VPN connection instead, all VPC traffic will failover to the VPN connection automatically. Traffic to/from public resources such as Amazon S3 will be routed over the Internet. If you do not have a backup AWS Direct Connect link or a IPsec VPN link, then Amazon VPC traffic will be dropped in the event of a failure. Traffic to/from public resources will be routed over the Internet.

Q. Can I extend one of my VLANs to the AWS Cloud using AWS Direct Connect?

No, VLANs are utilized in AWS Direct Connect only to separate traffic between virtual interfaces.

Q. Does AWS Direct Connect offer a Service Level Agreement (SLA)?

Yes, AWS Direct Connect offers SLA. Please see here for more details.

Q: What are the technical requirements for virtual interfaces to public AWS services such as Amazon EC2 and Amazon S3?

This connection requires the use of the Border Gateway Protocol (BGP) with an Autonomous System Number (ASN) and IP Prefixes. You will need the following information to complete the connection:

- A public or private ASN. If you are using a public ASN, you must own it. If you are using a private ASN, it must be in the 64512 to 65535 range.
- A new unused VLAN tag that you select
- Public IPs (/30) allocated by you for the BGP session

By default, Amazon will advertise global public IP prefixes via BGP. You must advertise public IP prefixes (/30 or smaller) that you own via BGP. For more details, consult the AWS Direct Connect User Guide.

Q: What is an Autonomous System Number (ASN) and do I need one to use AWS Direct Connect?

Autonomous System numbers are used to identify networks that present a clearly defined external routing policy to the Internet. AWS Direct Connect requires an ASN to create a public or private virtual interface. You may use a public ASN which you own, or you can pick any private ASN number between 64512 to 65535 range.

Q. What IP address will be assigned to each end of a virtual interface?

If you are configuring a virtual interface to the public AWS cloud, the IP addresses for both ends of the connection must be allocated from public IP space that you own. If the virtual interface is to a VPC and you choose to have AWS auto-generate the peer IP CIDR, the IP address space for both ends of the connection will be allocated by AWS in the 169.254.0.0/16 range.

Q: Can I connect to the Internet via this connection?

No.

Q: If I have more than one virtual interface attached, can I exchange traffic between the two ports?

Not for public Direct Connect virtual interfaces; but you can exchange traffic between the two ports in the same region if they are connecting to the same VGW.

Q. Can I locate my hardware next to the equipment that powers AWS Direct Connect?

You can procure rack space within the facility housing the AWS Direct Connect location and deploy your equipment nearby. However, AWS customer equipment cannot be placed within AWS Direct Connect racks or cage areas for security reasons. For more information, contact the operator for the particular facility. Once deployed, you can connect this equipment to AWS Direct Connect using a cross-connect.

Q. How do I enable BFD on my Direct Connect connection?

Asynchronous BFD is automatically enabled for each Direct Connect virtual interface, but will not take effect until it's configured on your router. AWS has set the BFD liveness detection minimum interval to 300, and the BFD liveness detection multiplier to 3.

Q. How do I set up Direct Connect for the AWS GovCloud (US) Region?

See the AWS GovCloud (US) User Guide for detailed instructions on how to set up a Direct Connect connection for the AWS GovCloud (US) region.

Link Aggregation Group

Q. What is this feature?

Link Aggregation Groups (LAG) are a way for customers to order and manage multiple direct connect ports as a single larger connection instead of as separate discrete connections.

Q. What’s the max number of links I can have in a LAG group?
The maximum number of links will be 4x in a LAG group.

Q. What does the LOA look like?

You will receive a single LOA document with dedicated page for each connection.

Q. What are you using for Link Aggregation Groups?

We are using the industry standard of LACP.

Q. Are these LAGs Static or Dynamic LACP?

We are configuring Dynamic LACP bundles. Static LACP bundles are not supported.

Q. Are these in Active/Active or Active/Passive mode?

They will be in Active/Active. That means, that AWS ports will always be sending Link Aggregation Control Protocol Data Units (LACPDUs).

Q. Does the MTU change at all?

The MTU of the LAG can be changed, please refer to Jumbo Frame documentation here to know more.

Q. Can I have my ports configured for Active/Passive instead of Active/Active?

You could configure LAG at your endpoint with LACP active or passive mode, AWS side is always configured as Active mode LACP.

Q. Can I mix interface types and have a few 1G ports and a few 10G ports in the same LAG?

No, you can create LAG using the same type of ports (either 1G or 10G).

Q. What ports types will this be available on?

It will be available for 1G and 10G Dedicated Connection ports.

Q. Can I LAG Hosted Connections as well?

No. It will only be available for 1G and 10G Dedicated Connections. It will not be available for Hosted Connections.

Q. Can I create a LAG out of my existing ports?

Yes, if your ports are on the same chassis. Please note this will cause your ports to go down for a moment while they are reconfigured as a LAG. They will not come back up until LAG is configured on your side as well.

Q. Can I have a LAG that spans multiple AWS routers?

LAG will only include ports on the same AWS device. We don't support multi-chassis LAG.

Q. How do I add links to my LAG once it's set up?

You can request another port for your LAG, but if we do not have ports available in the same chassis you will need to order a new LAG and migrate your connections. For example, if you have 3x 1G links, and would like to add a fourth but we do not have a port available on that chassis, you will need to order a new LAG of 4x 1G ports.

Q. What does the new LOA look like when I order additional connection to add to the LAG?

You will receive a separate LOA for each the new members of the LAG group.

Q. You're out of ports and I have to order a new LAG, but I have VIFs configured! How do I move those?

You can have multiple VIFs attached to a VGW at once, and you can configure VIFs on a connection even when it's down. We suggest you create the new VIFs on your new bundle, and then move the connections over to the new bundle once you've created all of your VIFS. Remember to delete the old connections so we stop billing you for them.

Q. Can I delete a single port from my LAG?

Yes, but only if your min links is set to lower than the ports you'll have left. Ex: You have 4 ports and Min links set to 4 – you won't be able to delete a port from the bundle. If min links is set to 3, you can then delete a port from the bundle. We will return a notification with the specific panel/port you've deleted and a reminder to disconnect the cross connect and circuit from Amazon.

Q. Can I delete my LAG all at once?

Yes, but just like a regular connection you won't be able to delete it if you have VIFs configured.

Q. If I have only 2 ports in my LAG can I still delete one?

Yes, you can have a single port in a LAG.

Q. Can I order a LAG with only one port?

Yes you can. Please note we can't guarantee there will be more ports available on the same
chassis in the future if you wish to add more ports.

Q. Can I convert a LAG back to individual ports?
Yes. This can be done with the DisassociateConnectionWithLag API call. See the API section.

Q. Can you just create a tool to move my VIFs for me?

You can use AssociateVirtualInterface API or console to do this operation.

Q. Does the LAG show as a single connection or a collection of connections?

It will show as a single dxlag and we'll list the connection id's under it.

Q. What does Min Links mean, and why do I have a check box for it when I order my bundle?

Min links is a feature in LACP where you can set the minimum number of links needed to be active in a bundle for that bundle to be active and pass traffic. If, for example, you have 4 ports, your min links is set to 3, and you only have 2 active ports, your bundle will not be active. If you have 3 or more then the bundle is active and will pass traffic if you have a VIF configured.

Q. What's the behavior if I don't click the Min Links?

We'll set Min Links to 0 by default.

Q. Can I change the Min Links after I’ve set up my LAG?

Yes. You can change the min links value after you’ve set up the bundle, either via console or via API.

Q. When I associate my existing DirectConnect connection with a LAG what happens with existing Virtual Interfaces already created with DirectConnect connection?

When a DirectConnect connection with existing Virtual Interfaces (VIFs) is associated to a LAG, Virtual Interfaces are migrated to the LAG; Please note that certain parameters associated with VIFs needs to be unique like VLAN numbers to be moved to LAG.

Q. If I have multiple LAGs, can I still use BFD to improve fail over time between paths?

BFD is still supported.

Q. Can I set link priority on a specific link?

We’ll treat all links as equal, so we won’t set “link priority” on any specific link.

Q. Does having a LAG make my connection more resilient?

No, LAG does not make your connectivity to AWS more resilient. If you have more than one link in your LAG, and if your min links is set to one, your LAG will let you protect against single link failure. It won’t protect against a single device failure at AWS where your LAG is terminating.

To achieve high availability connectivity to AWS we recommend you to have connections at multiple AWS Direct Connect locations. You can refer to this page to learn more about achieving highly available network connectivity.

Q. Can I have VIFs on two different LAG connected to the same VGW?

Yes. This behavior is exactly like creating VIFs on single ports.

Q. Can I have a 40GE interface on my side that connects to 4x 10GE on the AWS side?

You will need 4x 10GE interfaces on your router to connect to AWS. A single 40GE interface connecting to a 4x 10GE LACP is not supported.

Q. Is there a charge for LAG?

There is no extra charge for LAG.

IPv6

Q. Can I run IPv4 and IPv6 on the same virtual interface (VIF)?

AWS Direct Connect supports both single and dual stack configurations on public and private VIFs. You will be able to add an IPv6 peering session to an existing VIF with IPv4 peering session (or vice versa). You can also create 2 separate VIFs – one for IPv4 and another one for IPv6

Q. I need a public IPv6 range, can Amazon assign me a range?

Yes. Addressing for both public and private VIFs is provided by default and with a netmask of /125.

Q. What IP address will Amazon assign my private VIF if I select "assign an IP" in the console?

For a private IPv4 VIF, Amazon will provide you a /30 CIDR. For a private IPv6 VIF, Amazon will provide you a /125 CIDR.

Q. Will I still need to run BGP on my VIFs?

Yes. Both private and public Direct Connect require a native peering from IPv4 or IPv6. Multiprotocol BGP is not supported at this time.

Q. Are there any changes to VLAN assignment?

No. Layer 2 functionality remains the same for IPv4 and IPv6.

Q. Will I still be able to use BFD for faster BGP failover times?

Yes. BFD is supported for IPv6 BGP peerings.

Q. Are there any changes in the length of CIDR you can advertise to AWS?

Yes, for IPv6 we will limit the length of CIDR you can advertise to AWS to /64 (or shorter) for public Direct Connect Virtual Interface. For IPv4, prefix limits will remain the same.

Q. What routes will AWS announce to me over a public VIF?

All public routes.

Q. Will you support multicast or anycast over IPv6 VIFs?

We will not support multicast or anycast on Direct Connect.

Q. What routes will I learn from AWS over a public VIF?

AWS Public Direct Connect will advertise IPv6 prefixes for all IPv6 enabled services.

Q. Can I create a Hosted Virtual Interface for someone that is IPv6 enabled?

Yes you can.

Q. Will this impact policers associated with Hosted Connections?

It will not.

Q. Will cloudhub still work in my VGW? (note also impacts VPN)

It will only work for like-for-like traffic; that is, you can send IPv4 traffic out an IPv4 interface. Translation between IPv4 and IPv6 is not supported. Translation between IPv4 and IPv6 is not supported.

Using AWS Direct Connect with Amazon Virtual Private Cloud

Q. What are the technical requirements for virtual interfaces to VPCs?

This connection requires the use of Border Gateway Protocol (BGP). You will need the following information to complete the connection:

- A public or private ASN. If you are using a public ASN you must own it. If you are using a private ASN, it must be in the 64512 to 65535 range.
- A new unused VLAN tag that you select
- The VPC Virtual Private Gateway (VGW) ID

AWS will allocate private IPs (/30) in the 169.x.x.x range for the BGP session and will advertise the VPC CIDR block over BGP. You can advertise the default route via BGP.

Q. How does AWS Direct Connect differ from an IPSec VPN Connection?

A VPC VPN Connection utilizes IPSec to establish encrypted network connectivity between your intranet and Amazon VPC over the Internet. VPN Connections can be configured in minutes and are a good solution if you have an immediate need, have low to modest bandwidth requirements, and can tolerate the inherent variability in Internet-based connectivity. AWS Direct Connect does not involve the Internet; instead, it uses dedicated, private network connections between your intranet and Amazon VPC.

Q. Can I use AWS Direct Connect and a VPN Connection to the same VPC simultaneously?

Yes. However, only in fail-over scenarios. The Direct Connect path will always be preferred, when established, regardless of AS path prepending.

Q. Can I establish a Layer 2 connection between VPC and my network?

No, Layer 2 connections are not supported.

Direct Connect Gateway

Q. What is Direct Connect gateway?

Direct Connect gateway is a grouping of virtual private gateways (VGWs) and private virtual interfaces (VIFs).

Q. What is the multi-account support for Direct Connect gateway?

Multi-account support for Direct Connect gateway will allow you to associate up to 10 Amazon Virtual Private Clouds (Amazon VPCs) or up to 3 AWS Transit Gateways from multiple AWS accounts with a Direct Connect gateway.

Q. Can I associate Amazon Virtual Private Clouds (Amazon VPCs) owned by any AWS account with a Direct Connect gateway owned by any AWS account?

Yes, you can associate Amazon Virtual Private Clouds (Amazon VPCs) owned by any AWS account with a Direct Connect gateway owned by any AWS account.

Q. Can I associate AWS Transit Gateway owned by any AWS account with a Direct Connect gateway owned by any AWS account?

Yes, you can associate AWS Transit Gateway owned by any AWS account with a Direct Connect gateway owned by any AWS account.

Q. How do I use multi-account support for Direct Connect gateway?

You can use the AWS Management Console or AWS Direct Connect APIs to use multi-account support for the Direct Connect gateway. If you are the owner of the Direct Connect gateway, follow the steps outline in the AWS Direct Connect user guide.

Q. Why is Direct Connect gateway needed?

It provides three main functions. First; Direct Connect gateway will enable you to interface with VPCs in any AWS Region (except AWS China Region), enabling you to use your AWS Direct Connect connections to interface with more than one AWS Regions.

Second; you can share a private virtual interface to interface with up to ten Virtual Private Clouds (VPCs), enabling you to reduce the number of Border gateway Protocol sessions between your on-premises network and AWS deployments.

Third: By attaching transit virtual interface(s) to a Direct Connect gateway and associating Transit Gateway(s) with the Direct Connect gateway, you can share transit virtual interface(s) to interface with up to three Transit Gateways, enabling you to reduce the number of Border Gateway Protocol sessions between your on-premises network and AWS deployments.

Q. If I use Direct Connect gateway, does my traffic to the desired AWS Region go via the associated home AWS Region?

No. When using Direct Connect gateway, your traffic will take the shortest path from your Direct Connect location to the destination AWS Region and vice versa regardless of the associated home AWS Region of the Direct Connect location that you are connected at.

Q. Are there additional fees when using Direct Connect gateway and working with remote regions?

There are no charges for using a Direct Connect gateway. You will pay applicable egress data charges based on the source remote AWS Region and port hour charges as per AWS Direct Connect pricing.

Q. Do the private/transit virtual interfaces(s), Direct Connect gateway, Virtual Private Gateway or AWS Transit Gateways need to be in the same account to use Direct Connect gateway functionality?

Yes, private virtual interface and Direct Connect gateway must be in the same AWS account to use Direct Connect gateway functionality. Similarly, transit virtual interface and Direct Connect gateway must be in the same AWS account to use Direct Connect gateway functionality

Virtual private gateway(s) or AWS Transit Gateway(s) can be in different AWS accounts then the account owning the Direct Connect gateway.

Q. Can I continue to use all my VPC features if I associate VGW (associated with Amazon VPC) to Direct Connect Gateway?

Yes, Networking features such as Elastic File System, Elastic Load Balancer, Application Load Balancer, Security Groups, Access Control List, AWS PrivateLink will still work with Direct Connect gateway.

Direct Connect gateway will not support CloudHub functionality, but if you are using AWS Classic VPN or AWS VPN connection to VGW that is associated with your Direct Connect gateway, you will be able to use your VPN connection to failover.

Features that are currently not supported by Direct Connect, AWS Classic VPN, or AWS VPN, such as edge-to-edge routing, VPC peering, VPC endpoint, will not be supported by Direct Connect gateway.

Q. I am working with one of the AWS Direct Connect Partners to get private virtual interface provisioned for my account, can I use Direct Connect gateway?

Yes, you can associate a provisioned private virtual interface with your Direct Connect gateway when you confirm your provisioned Private in your AWS account.

Q. What if I just want to connect to VPCs in my local region?

You can continue to use the current practice of attaching your VIF to VGW; you will continue to have intra-region VPC connectivity, and will be charged egress rate that is applicable based on geographical regions.

Q. What are the limits associated with Direct Connect gateway usage?

Please refer to AWS Direct Connect Limits to get limits associated with the Direct Connect gateway feature.

Q. Can a VGW (associated with a VPC) be part of more than one Direct Connect gateway?

No, a VGW-VPC pair cannot be part of more than one Direct Connect gateway.

Q. Can a private virtual interface be attached to more than one Direct Connect gateway?

No, one private virtual interface can only attach to a single Direct Connect gateway OR a single Virtual Private Gateway. We recommend that you follow AWS Direct Connect resiliency recommendations and attach more than one private virtual interface.

Q. Can I associate multiple VGWs (each associated with a VPC) to a Direct Connect gateway?

Yes, as long as the IP CIDR blocks of the Amazon VPC associated with the Virtual Private Gateway do not overlap.

Q. Does Direct Connect gateway break existing CloudHub functionality for customers?

No, Direct Connect gateway does not break existing CloudHub for customers. Direct Connect gateway enables connectivity between on-premise networks and any AWS region's VPC. CloudHub enables connectivity between on-premise network using Direct Connect or VPN within the same region the VIF is associated with the VGW directly. Existing CloudHub functionality will continue to be supported.

Q. What type of traffic is supported, and not supported by Direct Connect gateway?

Please refer to AWS Direct Connect User Guide to review supported and not supported traffic patterns.

Q. Will intra-region CloudHub continue to be supported?

Yes, customers will still be able to attach a Direct Connect VIF directly to a VGW to support CloudHub

Q. I currently have a VPN in us-east-1 attached to a VGW. I want to enable CloudHub in us-east-1 between that VPN and a new VIF. Can I do this with Direct Connect gateway?

No, you cannot do this with a Direct Connect gateway, but the option to attach a VIF directly to a VGW is available to enable the VPN <-> Direct Connect CloudHub use case.

Q. I have an existing private virtual interface associated with VGW, can I associate my existing private virtual interface with Direct Connect gateway?

No, an existing private virtual interface associated with VGW cannot be associated with the Direct Connect gateway. Please create a new private virtual interface, and at the time of creation, associate with your Direct Connect gateway.

Q. Does Direct Connect gateway deprecate CloudHub functionality?

No. You can continue using your already created CloudHub.

Q. Can I create a new CloudHub between my VPN connection and Direct Connect VIF?

Yes, you can create a new CloudHub between your VPN and Direct Connect VIF by using a VGW attachment instead of a Direct Connect gateway attachement.

Q. If I have a VGW attached to a VPN and a Direct Connect gateway and my Direct Connect circuit goes down, will my VPC traffic route out the VPN?

Yes, as long as the VPC route table still has routes to the VGW towards the VPN.

Q. Can I attach a VGW that is not attached to a VPC to a Direct Connect gateway?

No, you cannot associate an unattached VGW to Direct Connect gateway.

Q. I have created a Direct Connect gateway with one Direct Connect Private , and three non-overlapping VGWs (each associated with a VPC), what happens if I detach one of the VGW from the VPC?

Traffic from your on-premise network to the detached VPC will stop, and VGW's association with the Direct Connect gateway will be deleted.

Q. I have created a Direct Connect gateway with one Direct Connect VIF, and three non-overlapping VGW-VPC pairs, what happens if I detach one of the VGW from the Direct Connect gateway?

Traffic from your on-premise network to the detached VGW (associated with a VPC) will stop.

Q. Can I send traffic from one VPC associated with a Direct Connect gateway to another VPC associated to the same Direct Connect gateway?

No, Direct Connect gateway only supports routing traffic from Direct Connect VIFs to VGW (associated with VPC). In order to send traffic between 2 VPCs, you would configure a VPC peering connection, the same as you do today.

Q. I currently have a VPN in us-east-1 attached to a VGW. If I associate this VGW to a Direct Connect gateway, can I send traffic from that VPN to a VIF attached to the Direct Connect gateway in a different region?

No, a Direct Connect gateway will not route traffic between a VPN and a Direct Connect VIF. To enable this use case, you would create a VPN in the region of the VIF and attach the VIF and the VPN to the same VGW.

Q. How do I detach my VGW-VPC pair from a Direct Connect gateway?

You can detach a VGW-VPC pair from a Direct Connect gateway using the AWS Console or API.

Q. Do you provide any SLA for Direct Connect gateway?

Please see here to review AWS Direct Connect SLA.

Q. Can I resize my Amazon VPC associated with a Direct Connect gateway?

Yes, you can resize the Amazon VPC. If you re-size your Amazon VPC, you must re-send the proposal with the re-sized VPC CIDR to the Direct Connect gateway owner. Once the Direct Connect gateway owner approves the new proposal, the re-sized VPC CIDR will be advertised towards your on-premise network.

Q. Is there a way to configure Direct Connect gateway to selectively propagate prefixes to/from Amazon VPCs?

Yes, Direct Connect gateway offers a way for you to selectively announce prefixes towards your on-premise networks. As the owner of the Direct Connect gateway, you can override the prefixes being advertised towards the on-premises network before you accept the association proposal OR when you can update the association request with allowed prefixes. Please see this documentation to get more information.

For prefixes getting advertised from your on-premise networks, each VPC associated with a Direct Connect gateway will receive all prefixes announced from your on-premises networks.

If you want to limit traffic to and from any specific Amazon VPC, you should consider using Access Control Lists (ACLs) for each VPC.

Q. Can I associate multiple AWS Transit Gateways to a Direct Connect gateway?

Yes, you can associate up to three AWS Transit Gateways to a Direct Connect gateway as long as the IP CIDR blocks announced from your AWS Transit Gateways do not overlap.

Direct Connect Gateway - Bring your own Private ASN

Q. What is this feature?

Configurable Private Autonomous System Number (ASN). This allows customers to set the ASN on the Amazon side of the BGP session for private VIFs on any newly created Direct Connect Gateway.

Q. Where are these features available?

All commercial AWS Regions (except AWS China Region) and GovCloud (US).

Q. How can I configure/assign my ASN to be advertised as the Amazon side ASN?

You can configure/assign an ASN to be advertised as the Amazon side ASN during creation of the new Direct Connect Gateway. You can create a Direct Connect Gateway using the AWS Direct Connect console or a CreateDirectConnectGateway API call.

Q. Can I use any ASN - public and private?

You can assign any private ASN to the Amazon side. You cannot assign any other public ASN.

Q. Why can't I assign a public ASN for the Amazon half of the BGP session?

Amazon is not validating ownership of the ASNs, therefore we're limiting the Amazon-side ASN to private ASNs. We want to protect customers from BGP spoofing.

Q. What ASN can I choose?

You can choose any private ASN. Ranges for 16-bit private ASNs include 64512 to 65534. You can also provide 32-bit ASNs between 4200000000 and 4294967294.

Q. What will happen if I try to assign a public ASN to the Amazon half of the BGP session?

We will ask you to re-enter a private ASN once you attempt to create the Direct Connect Gateway.

Q. If I don't provide an ASN for the Amazon half of the BGP session, what ASN can I expect Amazon to assign to me?

Amazon will provide an ASN of 64512 for the Direct Connect Gateway if you don't choose one.

Q. Where can I view the Amazon side ASN?

You can view the Amazon side ASN in the AWS Direct Connect console and in the response of the DescribeDirectConnectGateways or using DescribeVirtualInterfaces API.

Q. If I have a public ASN, will it work with a private ASN on the AWS side?

Yes, you can configure the Amazon side of the BGP session with a private ASN and your side with a public ASN.

Q. I have private VIFs already configured and want to set a different Amazon side ASN for the BGP session on an existing VIF. How can I make this change?

You will need to create a new Direct Connect Gateway with desired ASN, and create a new VIF with the newly created Direct Connect Gateway. Your device configuration also needs to change appropriately.

Q. I'm attaching multiple private VIFs to a single Direct Connect Gateway. Can each VIF have a separate Amazon side ASN?

No, you can assign/configure separate Amazon side ASN for each Direct Connect Gateway, not each VIF. Amazon side ASN for VIF is inherited from the Amazon side ASN of the attached Direct Connect Gateway.

Q. Can I use different private ASNs for my Direct Connect Gateway and Virtual Private Gateway?

Yes, you can use different private ASNs for your Direct Connect Gateway and Virtual Private Gateway. Please note, the Amazon side ASN you will receive depends on your private virtual interface association.

Q. Can I use same private ASNs for my Direct Connect Gateway and Virtual Private Gateway?

Yes, you can use same private ASNs for your Direct Connect Gateway and Virtual Private Gateway. Please note, the Amazon side ASN you will receive depends on your private virtual interface association.

Q. I'm attaching multiple Virtual Private Gateways with their own private ASN to a single Direct Connect Gateway configured with its own private ASN. Which private ASN takes precedence, VGW or Direct Connect Gateway?

Direct Connect Gateway private ASN will be used as the Amazon side ASN for the Border Gateway Protocol (BGP) session between your network and AWS.

Q. Where can I select my own private ASN?

When creating a Direct Connect Gateway in the AWS Direct Connect Gateway console. Once Direct Connect Gateway is configured with Amazon side ASN, the private virtual interfaces associated with the Direct Connect Gateway will use your configured ASN as the Amazon side ASN.

Q. I use CloudHub today. Will I have to adjust my configuration in the future?

You will not have to make any changes.

Q. I want to select a 32-bit ASN. What is the range of 32-bit private ASNs?

We will support 32-bit ASNs from 4200000000 to 4294967294.

Q. Once the Direct Connect Gateway is created, can I change or modify the Amazon side ASN?

No, you cannot modify the Amazon side ASN after creation. You can delete the Direct Connect Gateway and recreate a new Direct Connect Gateway with the desired private ASN.

Using Public Virtual Interfaces

Q. When creating a virtual interface to work with AWS services using public IP space, what IP prefixes will I receive via BGP?

You will receive all Amazon IP prefixes for the region that you are connecting to in supported AWS Regions, and on-net prefixes from other AWS non-regional point of presence (PoP) as available such as CloudFront you can refer to this link for more information. This includes prefixes necessary to reach AWS services, and may include prefixes for other Amazon affiliates, including those of www.amazon.com. For the current list of prefixes advertised by AWS, please download the JSON of AWS IP Address Ranges. Note, that AWS may advertise a prefix in more specific (or-longer) ranges. For example, prefix 96.127.0.0/17 in the file may be advertised as 96.127.0.0/21, 96.127.8.0/21, 96.127.32.0/19, and 96.127.64.0/18 so please ensure you're matching more specific ranges.

When customers use AWS Direct Connect, customers' traffic will remain in AWS global network backbone, after it enters AWS global network backbone. Therefore, prefixes of services such as Route53 or certain CloudFront locations that are not on the Amazon backbone network will not be advertised through Direct Connect.

For the newly created public VIF, Direct Connect customers will receive all Amazon public IP prefixes in supported AWS regions and on-net prefixes from other AWS non-region points of presence (POP) as available such as CloudFront. Standard AWS Direct Connect data transfer out rates apply for all traffic routed through your AWS Direct Connect connection. Please see the AWS Direct Connect community forum for the additional details in the routing policy of the public virtual interface.

Note: Currently AWS Global Accelerator prefixes are not advertised over AWS Direct Connect public virtual interface.

Q. What IP prefixes should I advertise over BGP for virtual interfaces to public AWS services?

You should advertise appropriate public IP prefixes that you own over BGP. Traffic from AWS services destined for these prefixes will be routed over your AWS Direct Connect connection.

Q. I am going to create a new public virtual interface. Do I need to take any action to get Amazon's global public IP address prefixes?

No, by default, AWS Direct Connect will advertise local and remote AWS Region prefixes where available and includes on-net prefixes from other AWS non-Region points of presence (PoP) where available; for example, CloudFront and Route53. Note: Currently AWS Global Accelerator prefixes are not advertised over AWS Direct Connect public virtual interface.

Q. How do I receive IP address prefixes for AWS Global Accelerator over my public virtual interface?

At this time, AWS Direct Connect does not advertise IP address prefixes for AWS Global Accelerator over public virtual interface.

Q. I see asymmetric traffic with AWS Global Accelerator. My traffic that goes to AWS Global Accelerator traverses the internet, but the return traffic that comes to my on-premises network traverses my AWS Direct Connect public virtual interface. How can I make sure that I get symmetric traffic between my on-premises network and AWS Global Accelerator?

Currently, we recommend that you do not advertise IP addresses that you use to communicate with AWS Global Accelerator over your AWS Direct Connect public virtual interface.

Q. I am okay with asymmetric traffic for AWS Global Accelerator. My traffic that goes to AWS Global Accelerator traverses the internet, but the return traffic that come to my on-premises network traverses my AWS Direct Connect public virtual interface. What Data Transfer Out charges do I pay for the AWS Global Accelerator traffic over AWS Direct Connect?

You pay internet Data Transfer Out rates for your AWS Global Accelerator traffic that traverses the AWS Direct Connect public virtual interface.

Q. Can I enable Jumbo MTU on Public Virtual Interface?

No, Jumbo MTU is not supported for Public Virtual Interface

Q. Will this new capability affect my existing public virtual interfaces?

No, your existing public virtual interfaces will not get affected.

Q. How many prefixes will you advertise over my newly created public virtual interface?

You should receive approximately 2,000 prefixes, and it will continue to increase.

Q. I do not want global public IP prefixes, can I opt out?

Yes, you can opt out using scoping communities. Please refer to this link to learn more about scoping communities supported by AWS Direct Connect.

Q. I want to migrate my existing public virtual interface to receive global prefixes; how can I do this migration?

You have two options to do such a migration. First, create a new public virtual interface, migrate traffic from your existing public virtual interface to the newly created public virtual interface; delete your old public virtual interface. Second, open a support case to request scope change for your existing public virtual interface, you will experience a Border Gateway Protocol flap during the scope change.

Jumbo Frames

Q. What is the Maximum Transmission Unit (MTU) supported by AWS Direct Connect?

AWS Direct Connect and Direct Connect Gateway support both 1500 and 9001 Maximum Transmission Unit (MTU). MTU is a configurable option on the AWS Direct Connect Private Virtual Interface.

Q. How do I change the MTU of a Private Virtual Interface?

- If you own both the AWS Direct Connect port and the Virtual Private Interface created on the AWS Direct Connect port, you can modify the MTU of an existing Virtual Interface or you can create a new Virtual Interface with 9001 MTU using API, CLI or Console.

- If the AWS Direct Connect port is owned by another AWS account, the port owner will need to enable Jumbo Frames on the port by modifying the MTU on an existing Virtual Interface or creating a new Virtual Interface with 9001 MTU.
- If your AWS Direct Connect connection is provided by an AWS Direct Connect Partner, then you need to check if the port is Jumbo Frames capable using API, CLI or console. If the port is Jumbo Frames capable, you can modify the MTU setting on the Virtual Interface. Otherwise, the AWS Direct Connect Partner will need to contact AWS support to make the port Jumbo Frames capable.

Q. Can I use Jumbo Frames over AWS Direct Connect with both propagated and static routes?

Jumbo Frames only apply to propagated routes from Direct Connect. If you add static routes pointing to your Virtual Private Gateway to the route table, traffic routed through static routes will be sent using a 1500 MTU.

Q. If I have two Private Virtual Interfaces that advertise the same route and both Interfaces have different MTUs, which MTU will be used?

If two virtual interfaces advertise the same route, but use different MTUs, 1500 MTU will be used for both virtual interfaces.

Q. Will Jumbo Frames work with AWS Direct Connect and AWS Managed VPN when both advertise the same routes?

AWS Managed VPN service does not support Jumbo Frames. If the same route is advertised over AWS Direct Connect and AWS Managed VPN, the 1500 MTU will be used.

Q. Do you support moving of a Jumbo Frame enabled Virtual Private Interface from one Direct Connect port to another?

If the destination port is not Jumbo Frames capable then you cannot move the Jumbo Frames enabled Virtual Interface to it. You will need to disable Jumbo Frames on the Virtual Interface and move it then re-enable Jumbo Frames.

Alternatively, you can enable Jumbo on any Virtual Interface on the destination connection before moving the Jumbo Frames enabled Virtual Interface.

Q. Is the downtime expected when enabling Jumbo Frames on a Private Virtual Interface?

Yes, if the owner of an AWS Direct Connect port (with Jumbo Frames not enabled on any virtual interface) creates a Jumbo Frame enabled Private Virtual Interface for the first time, there will be expected downtime of 5 to 30 seconds on the physical port. Other virtual interfaces on this port, regardless of their account, will also observe this downtime. If the physical port has at least one Virtual Interface that is Jumbo Frames enabled, then there will be no downtime observed on the physical interface.

Q. How do I enable Jumbo Frames on a Link Aggregation Group (LAG) Private Virtual Interface?

You will need to enable Jumbo Frames for at least one Private Virtual Interface in the LAG to enable Jumbo Frames on the LAG.

Local preference communities for private virtual interface

Q. What is this feature?

This feature provides support for local preference communities for private virtual interfaces. With communities, customers can influence the return path for traffic sourced from VPC address space.

Q. Can I use this feature for my existing EBGP sessions?

Yes, all existing BGP sessions on private virtual interfaces support the use of local preference communities.

Q. Do you charge additionally for this feature?

There is no additional charge for using this feature.

Q. Will this feature be available on both Public and Private Virtual Interfaces?

No, this feature is currently available for private virtual interfaces only.

Q. Will this feature work with Direct Connect Gateway?

Yes, this feature will work with private virtual interfaces attached with Direct Connect Gateway.

Q. Can I verify communities being received by AWS?

No, at this time we do not provide such monitoring features.

Q. What are the supported local preference communities for Direct Connect private virtual interface?

The following communities are supported for private virtual interface and are evaluated in order of lowest to highest preference. Communities are mutually exclusive. Prefixes marked with the same communities, and bearing identical MED*, AS_PATH attributes are candidates for multi-pathing.

- 7224:7100 – Low Preference
- 7224:7200 – Medium Preference
- 7224:7300 – High Preference

Q. What is the default behavior in case I do not use the supported communities?

If you do not specify Local Preference communities for your private VIF, the default local preference is based on the distance to the Direct Connect Locations from the local region. In such situation, egress behavior across multiple VIFs from multiple Direct Connect Locations may be arbitrary.

Q. I have two private VIFs on a physical connection at a Direct Connect location; can I use supported communities to influence egress behavior across these two private VIFs?

Yes, you can use this feature to influence egress traffic behavior between two VIFs on the same physical connection.

Q. I have two Direct Connect connections, both 1G, I want all incoming traffic into my network load balanced across these two connections, can I use community based routing to achieve such load balancing across the locations?

Yes, you can use community based routing to enable load balancing across Direct Connect locations. To do so, any prefixes requiring load-balancing must be marked with the same communities.

Q. Will the local preference communities feature support failover?

Yes. This can be accomplished by advertising prefixes over the primary/active virtual interface with a community for higher local preference than prefixes advertised over the backup/passive virtual interface. This feature is backwards compatible with pre-existing methods for achieving failover; if your Direct Connect is currently configured for failover, no additional changes are necessary.

Q. I have already configured my routers with AS_PATH, do I need to change the configuration to use community tags and disrupt my network?

No, we will continue to respect AS_PATH attribute. This feature is an additional knob you can use to get better control over the incoming traffic from AWS. Direct Connect follows the standard approach for path selection. Bear in mind that local preference is evaluated before the AS_PATH attribute.

Q. I have two Direct Connect connections, one is 1G and another is 10G, and both are advertising the same prefix. I would like to receive all traffic for this destination across the 10G Direct Connect connection, but still be capable of failing over to the 1G connection. Can local preference communities be used to balance traffic in this scenario?

Yes. By marking the prefix advertised over the 10G Direct Connection with a community of a higher local preference, it will be the preferred path. In the event that the 10G fails or the prefix withdrawn, the 1G interface will become the return path.

Q. How wide will you multipath traffic to my network?

We will multipath per prefix at up to 16 next-hops wide, where each next-hop is a unique AWS endpoint.

Q. Can I have v4 and v6 BGP sessions running over a single VPN tunnel?

At this time, we will only allow v4 BGP session running single VPN tunnel with IPv4 address. In future, we will allow v6 BGP sessions running over the single VPN tunnel with IPv4 endpoint address.

Q. Is there any difference to the BGP configuration/setup details outlined for DX?

VPN BGP will work the same as DX

Q. Can I terminate my tunnel to an endpoint with an IPv6 address?

At this time, we will only support IPv4 endpoint address for VPN. In future, we will support VPN endpoint with IPv6 address.

Q. Can I terminate my tunnel to an IPv4 address and run IPv6 BGP sessions over the tunnel?

At this time, we will only allow v4 BGP session running single VPN tunnel with IPv4 address. In future, we will allow v6 BGP sessions running over the single VPN tunnel with IPv4 endpoint address.

AWS Transit Gateway Support

Q: Which AWS Regions offer AWS Direct Connect support for AWS Transit Gateway?

AWS Direct Connect support for AWS Transit Gateway is now available in AWS GovCloud (US-East), AWS GovCloud (US-West), Canada (Central), US East (N. Virginia), US East (Ohio), US West (N. California), US West (Oregon), EU (Ireland), EU (London), EU (Frankfurt), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Sydney), Asia Pacific (Singapore), Asia Pacific (Tokyo), EU (Paris), South America (Sao Paulo), and Asia Pacific (Hong Kong) AWS Regions.

Q: What is transit virtual interface?

Transit virtual interface is a type of virtual interface you can create on any AWS Direct Connect 1/2/5/10 Gbps connection. Transit virtual interface can only be attached to a Direct Connect gateway. You can use the AWS Direct Connect gateway attached with one or more transit virtual interface to interface with up to three AWS Transit Gateways in any supported AWS Regions.

Similar to the private virtual interface, you can establish one IPv4 BGP session and one IPv6 BGP session over a single transit virtual interface, and if you have 1G/10G connections at a location enabled with logical redundancy, you can create redundant layer 3 connection over a transit virtual interface.

Q: How do I create transit virtual interface?

You can use the AWS Management console or APIs to create transit virtual interface.

Q: Can I allocate transit virtual interface in another AWS account?

Yes, you can allocate transit virtual interface in any AWS account.

Q: Can I attach transit virtual interface to my Virtual Private Gateway?

No, you cannot attach transit virtual interface to your Virtual Private Gateway

Q: Can I attach private virtual interface to my AWS Transit Gateway?

No, you cannot attach private virtual interface to your AWS Transit Gateway.

Q: Who will pay the data transfer out charges for the data transfer done on the transit virtual interface?

The AWS account owning the transit virtual interface will pay for the AWS Direct Connect Data Transfer Out.

Q: What are the limits associated with transit virtual interface?

Please refer to AWS Direct Connect limits page to know more about the limits associated with transit virtual interface.

Q: Can I add more transit virtual interfaces to the connection?

No, you can create only one transit virtual interface for any AWS Direct Connect 1/2/5/10 Gbps connection.

Q: I have an existing Direct Connect gateway attached to a private virtual interface, can I attach a transit virtual interface to this Direct Connect gateway?

No, a Direct Connect Gateway can only have one type of virtual interface attached.

Q: Can I associate my AWS Transit Gateway to the Direct Connect gateway attached to private virtual interface?

No, an AWS Transit Gateway can only be associated with the Direct Connect gateway attached to transit virtual interface.

Q: How long does it take to establish an association between AWS Transit Gateway and AWS Direct Connect gateway?

It can take up to 40 minutes to establish an association between AWS Transit Gateway and AWS Direct Connect gateway.

Q: How many total virtual interfaces can I create per 1 Gbps or 10 Gbps dedicated connection?

You can create up to 51 virtual interfaces per 1 Gbps or 10Gbps dedicated connection inclusive of the transit virtual interface.

Q: Does a transit virtual interface support jumbo frames?

Yes, transit virtual interface will support jumbo frames. Maximum transmission unit (MTU) size will be limited to 8,500.

Q: I have 4x10G LAG, how many transit virtual interfaces can I create on this link aggregation group (LAG)?

You can create one transit virtual interface on the 4x10G LAG.

Q: Do you support all the border gateway protocol (BGP) attributes that you support on the Private virtual interface for the transit virtual interface?

Yes, you can continue to use supported BGP attributes (AS_PATH, Local Pref, NO_EXPORT) on the transit virtual interface.

Q: Can I create transit virtual interface on 1/2/5/10 Gbps hosted connection?

Yes, you can create one transit virtual interface on a 1/2/5/10 Gbps hosted connection.

Q: I want to associate my Transit Gateway to a Direct Connect gateway, can I use the same Autonomous System Number (ASN) for the Direct Connect gateway and the Transit Gateway?

No, you cannot use the same ASN for the Transit Gateway and the Direct Connect gateway.

Virtual Private Network (VPN)

Q. Can I have v4 and v6 BGP sessions running over a single VPN tunnel?

At this time, we will only allow v4 BGP session running single VPN tunnel with IPv4 address. In future, we will allow v6 BGP sessions running over the single VPN tunnel with IPv4 endpoint address.

Q. Is there any difference to the BGP configuration/setup details outlined for DX?

VPN BGP will work the same as DX

Q. Can I terminate my tunnel to an endpoint with an IPv6 address?

At this time, we will only support IPv4 endpoint address for VPN. In future, we will support VPN endpoint with IPv6 address.

Q. Can I terminate my tunnel to an IPv4 address and run IPv6 BGP sessions over the tunnel?

At this time, we will only allow v4 BGP session running single VPN tunnel with IPv4 address. In future, we will allow v6 BGP sessions running over the single VPN tunnel with IPv4 endpoint address.

AWS Global Accelerator FAQs

General

Q: What is AWS Global Accelerator?

A: AWS Global Accelerator is a networking service that helps you improve the availability and performance of the applications that you offer to your global users. AWS Global Accelerator is easy to set up, configure, and manage. It provides static IP addresses that provide a fixed entry point to your applications and eliminate the complexity of managing specific IP addresses for different AWS Regions and Availability Zones. AWS Global Accelerator always routes user traffic to the optimal endpoint based on performance, reacting instantly to changes in application health, your user's location, and policies that you configure. Like other AWS services, AWS Global Accelerator is a self-service, pay-per-use offering, requiring no long term commitments or minimum fees.

Q: What can I do with AWS Global Accelerator?

A: By using AWS Global Accelerator, you can:

- Associate the static IP addresses provided by AWS Global Accelerator to regional AWS resources or endpoints, such as Network Load Balancers, Application Load Balancers, EC2 Instances, and Elastic IP addresses. The IP addresses are anycast from AWS edge locations so they provide onboarding to the AWS global network close to your users.
- Easily move endpoints between Availability Zones or AWS Regions without needing to update your DNS configuration or change client-facing applications.
- Dial traffic up or down for a specific AWS Region by configuring a traffic dial percentage for your endpoint groups. This is especially useful for testing performance and releasing updates.
- Control the proportion of traffic directed to each endpoint within an endpoint group by assigning weights across the endpoints.

Q: How do I get started with AWS Global Accelerator?

A: You can get started with setting up AWS Global Accelerator by using the API or through the AWS Management Console. Because AWS Global Accelerator is a global service, it's not tied to any specific AWS Region. Here are three easy steps to set up AWS Global Accelerator for your application:

1. Create an accelerator: When you create your accelerator, AWS Global Accelerator provisions two static IP addresses for it. Then you configure one or more listeners to process inbound connections from end clients to your accelerator, based on the protocol and port that you specify.
2. Configure endpoint groups: You choose one or more regional endpoint groups to associate to your accelerator's listener by specifying the AWS Regions to which you want to distribute traffic. Your listener routes requests to the registered endpoints in this endpoint group. AWS Global Accelerator monitors the health of endpoints within the group using the health check settings defined for each endpoint. You can configure a traffic dial percentage for each endpoint group, which controls the amount of traffic that an endpoint group accepts. By default, the traffic dial is set to 100% for all regional endpoint groups.
3. Register endpoints for endpoint groups: You register one or more regional resources, such as Application Load Balancers, Network Load Balancers, EC2 Instances, or Elastic IP addresses, in each endpoint group. Then you can set weights to choose how much traffic is routed to each endpoint.

Q: How does AWS Global Accelerator work together with Elastic Load Balancing (ELB)?

A: Both of these services solve the challenge of routing user requests to healthy application endpoints. AWS Global Accelerator relies on ELB to provide the traditional load balancing features such as support for internal and non-AWS endpoints, pre-warming, and Layer 7 routing. However, while ELB provides load balancing within one Region, AWS Global Accelerator provides traffic management across multiple Regions.

A regional ELB load balancer is an ideal target for AWS Global Accelerator. By using a regional ELB load balancer, you can precisely distribute incoming application traffic across backends, such as Amazon EC2 instances or Amazon ECS tasks, within an AWS Region. AWS Global Accelerator complements ELB by extending these capabilities beyond a single AWS Region, allowing you to provision a global interface for your applications in any number of Regions. If you have workloads that cater to a global client base, we recommend that you use AWS Global Accelerator. If you have workloads hosted in a single AWS Region and used by clients in and around the same Region, you can use an Application Load Balancer or Network Load Balancer to manage your resources.

Q: How is AWS Global Accelerator different from Amazon CloudFront?

A: AWS Global Accelerator and Amazon CloudFront are separate services that use the AWS global network and its edge locations around the world. CloudFront improves performance for both cacheable content (such as images and videos) and dynamic content (such as API acceleration and dynamic site delivery). Global Accelerator improves performance for a wide range of applications over TCP or UDP by proxying packets at the edge to applications running in one or more AWS Regions. Global Accelerator is a good fit for non-HTTP use cases, such as gaming (UDP), IoT (MQTT), or Voice over IP, as well as for HTTP use cases that specifically require static IP addresses or deterministic, fast regional failover. Both services integrate with AWS Shield for DDoS protection.

Q: Can I use AWS Global Accelerator for my on-premises services?

A: You can’t directly configure on-premises resources as endpoints for your static IP addresses, but you can configure a Network Load Balancer (NLB) in each AWS Region to address your on-premises endpoints. Then you can register the NLBs as endpoints in your AWS Global Accelerator configuration.

Benefits

Q: What benefits does AWS Global Accelerator provide?

A: AWS Global Accelerator includes the following benefits:

Instant regional failover: AWS Global Accelerator automatically checks the health of your applications and routes user traffic only to healthy application endpoints. If the health status changes or you make configuration updates, AWS Global Accelerator reacts instantaneously to route your users to the next available endpoint.

High availability: AWS Global Accelerator has a fault-isolating design that increases the availability of your application. When you create an accelerator, you are allocated two IPv4 static IP addresses that are serviced by independent network zones. Similar to Availability Zones, these network zones are isolated units with their own physical infrastructure and serve static IP addresses from a unique IP subnet. If one static IP address becomes unavailable due to IP address blocking or unreachable networks, AWS Global Accelerator provides fault tolerance to client applications by rerouting to a healthy static IP address from the other isolated network zone.

No variability around clients that cache IP addresses: Some client devices and internet resolvers cache DNS answers for long periods of time. So when you make a configuration update, or there’s an application failure or change in your routing preference, you don’t know how long it will take before all of your users receive updated IP addresses. With AWS Global Accelerator, you don’t have to rely on the IP address caching settings of client devices. Change propagation takes a matter of seconds, which reduces your application downtime.

Improved performance: AWS Global Accelerator ingresses traffic from the edge location that is closest to your end clients through anycast static IP addresses. Then traffic traverses the congestion-free and redundant AWS global network, which optimizes the path to your application that is running in an AWS Region. AWS Global Accelerator chooses the optimal AWS Region based on the geography of end clients, which reduces first-byte latency and improves performance.

Easy manageability: The static IP addresses provided by AWS Global Accelerator are fixed and provide a single entry point to your applications. This lets you easily move your endpoints between Availability Zones or between AWS Regions, without having to update your DNS configuration or client-facing applications. Use cases include A/B testing, application updates, and failover simulations. Corporate proxies can also whitelist your application’s static IP addresses in their firewalls.

Fine-grained control: AWS Global Accelerator lets you set a traffic dial for your regional endpoint groups, to dial traffic up or down for a specific AWS Region when you conduct performance testing or application updates. In addition, if you have stateful applications, you can choose to direct all requests from a user to the same endpoint, regardless of the source port and protocol, to maintain client affinity. These features give you fine-grained control.

Q: I operate only in a single AWS Region. Can I get any benefit from AWS Global Accelerator?

A: Yes. While you might not want to use the intelligent traffic routing capabilities of AWS Global Accelerator, there are a number of advantages to using static IP addresses. First, by using these addresses, you increase the Quality of Service (QoS) for your users by onboarding their traffic onto the AWS global network as close to them as possible. Typically, traffic must take multiple hops through the public internet, over potentially congested and non-redundant network paths, to reach your destination AWS Region. With AWS Global Accelerator, you get to leverage the AWS globally redundant network to help improve your application availability and performance. Second, you have the freedom to easily move your application between AWS Regions without changing your public interface. This means that you can plan for the future, knowing that if your needs change, you can easily migrate or add additional AWS Regions without worrying about how your users will connect to your applications.

High availability

Q: How does AWS Global Accelerator make it easy to move to a multi-Region setup?

A: You may want to run your applications in multiple AWS Regions for regional redundancy and to improve performance by running your applications closer to your users. By providing a network layer between your application and clients, AWS Global Accelerator can perform health checks, and then automatically route traffic around failed endpoints, without disrupting clients. This graceful shutdown and startup of new endpoints improves availability and performance for your users while ensuring that internet traffic is routed to the closest available endpoint.

Q: How does AWS Global Accelerator help support multi-Region failover?

A: AWS Global Accelerator provides you with a set of static IP addresses that can map to multiple application endpoints across AWS Regions, to improve redundancy. If your application experiences failure in a specific AWS Region, AWS Global Accelerator automatically detects the unhealthy endpoints and redirects traffic to the next optimal AWS Region, ensuring high availability and disaster recovery.

Q: How fast will my application failover between AWS Regions?

A: AWS Global Accelerator can detect an unhealthy endpoint and take it out of service in less than one minute.

Compliance

Q: What compliance certifications does AWS Global Accelerator support?

A: AWS Global Accelerator certifications make it easier for you to verify our high security standards and meet your own regulatory and compliance obligations. It has been assessed to comply with PCI DSS, ISO 9001, 27001, 27017, 27018, 27018, and SOC (System & Organization Control), in addition to being HIPAA-eligible.

Additional questions

Q: Can I use my own IP addresses with Global Accelerator?

A: You can Bring Your Own IP address (BYOIP) to AWS Global Accelerator, which enables you to use your own IP addresses as a fixed entry point to your application endpoints. This allows you to move your on-premises applications that have hardcoded IP address dependencies to AWS, without making any client-facing changes. This is helpful for example in regulated environments that require whitelisting of IP address ranges. The accelerators that use your own IP addresses work exactly the same as your accelerators which use Amazon-provided IP addresses.

Q: Does AWS Global Accelerator support IPv4 and IPv6?

A: The service currently supports IPv4 addresses.

Q: What protocols does AWS Global Accelerator support?

A: AWS Global Accelerator supports both TCP and UDP protocols.

Q: How is AWS Global Accelerator different from a DNS-based traffic management solution?

A: First, some client devices and internet resolvers cache DNS answers for long periods of time. So when you make a configuration update, or there's an application failure or change in your routing preference, you don't know how long it will take before all of your users receive updated IP addresses. With AWS Global Accelerator, you don't have to rely on the IP address caching settings of client devices. Change propagation takes a matter of seconds, which reduces your application downtime. Second, with Global Accelerator, you get static IP addresses that provide a fixed entry point to your applications. This lets you easily move your endpoints between Availability Zones or between AWS Regions, without having to update the DNS configuration or client-facing applications.

Q: Can I use AWS Global Accelerator with AWS Direct Connect?

A: We recommend that you don't advertise IP addresses that you use to communicate with AWS Global Accelerator over your AWS Direct Connect public virtual interface. Direct Connect does not advertise IP address prefixes for

Global Accelerator over a public virtual network. For more information about public virtual interfaces and Direct Connect, see Using Public Virtual Interfaces.

AWS Transit Gateway FAQs

General

Q: In which AWS Regions is AWS Transit Gateway available?

A: AWS Transit Gateway is available in US East (Virginia), US East (Ohio), US West (Oregon), US West (Northern California), AWS GovCloud (US-East), AWS GovCloud (US-West), Canada (Central), South America (São Paulo), EU (Ireland), EU (London), EU (Frankfurt), EU (Paris), EU (Stockholm), Asia Pacific (Hong Kong), Asia Pacific (Mumbai), Asia Pacific (Tokyo), Asia Pacific (Singapore), Asia Pacific (Seoul), and Asia Pacific (Sydney) AWS Regions with support for other regions coming soon.

Transit Gateway inter-region peering support is available for gateways in US East (Virginia), US East (Ohio), US West (Oregon), EU (Ireland), and EU (Frankfurt) AWS Regions with support for other regions coming soon.

Q: How do I control which Amazon VPCs can communicate with each other?

A: You can segment your network by creating multiple route tables in an AWS Transit Gateway and associate Amazon VPCs and VPNs to them. This will allow you to create isolated networks inside an AWS Transit Gateway similar to virtual routing and forwarding (VRFs) in traditional networks. The AWS Transit Gateway will have a default route table. The use of multiple route tables is optional.

Q: How does routing work in AWS Transit Gateway?

A: AWS Transit Gateway supports dynamic and static routing between attached Amazon VPCs and VPNs. By default, Amazon VPCs, VPNs, Direct Connect gateways, and peered Transit Gateways are associated to the default route table. You can create additional route tables and associate Amazon VPCs, Direct Connect gateways, and VPNs with it.

The routes decide the next hop depending on the destination IP address of the packet. Routes can point to an Amazon VPC or a VPN connection, a Direct Connect gateway, or a peered Transit Gateway.

Q: How do routes get propagated into the AWS Transit Gateway?

A: There are 2 ways where routes get propagated in the AWS Transit Gateway:

1. Routes propagated to/from on-premises networks: When you connect VPN, routes will propagate between the AWS Transit Gateway and your on-premises router using Border Gateway Protocol (BGP).
2. Routes Propagated to/from Amazon VPCs: When you attach an Amazon VPC to an AWS Transit Gateway or resize an attached Amazon VPC, the Amazon VPC Classless Inter-Domain Routing (CIDR) will propagate into the AWS Transit Gateway route table using internal APIs (not BGP). CIDR is a method for allocating IP addresses and IP routing to slow the growth of routing tables on routers across the Internet, and to help slow the rapid exhaustion of IPv4 addresses. Routes in the AWS Transit Gateway route table will not be propagated to the Amazon VPC's route table. Amazon VPC owner need to create static route to send Traffic to the AWS Transit Gateway.

Inter-region peering attachments between Transit Gateways do not support route propagation.

Q: Can I connect Amazon VPCs with overlapping CIDRs?

A: AWS Transit Gateway doesn't support routing between Amazon VPCs with overlapping CIDRs. If you attach a new Amazon VPC that has a CIDR which overlaps with an already attached Amazon VPC, AWS Transit Gateway will not propagate the new Amazon VPC route into the AWS Transit Gateway route table.

Performance and limits

Q: What are the service limits that I need to keep in mind while using AWS Transit Gateways?

A: The table below list the different service limits:

Limit	Default
Number of AWS Transit Gateway attachments	5,000
Maximum bandwidth per VPN connection*	1.25 Gbps
Maximum bandwidth (burst) per VPC, Direct Connect gateway, or peered Transit Gateway connection	50 Gbps
Number of AWS Transit Gateways per account	5
Number of AWS Transit Gateway attachments per VPC	5
Number of routes	10,000
Number of Direct Connect gateways per AWS Transit Gateway	20

*You can use equal-cost multi-path routing (ECMP) to get higher VPN bandwidth by aggregating multiple VPN connections.

Security and compliance

Q: With which compliance programs does AWS Transit Gateway conform?

A: AWS Transit Gateway inherits compliance from Amazon Virtual Private Cloud (Amazon VPC) and meets the standards for PCI DSS Level 1, ISO 9001, ISO 27001, ISO 27017, ISO 27018, SOC 1, SOC 2, SOC 3, FedRAMP Moderate, FedRAMP High and HIPAA eligibility.

For more information, visit our compliance page.

Feature interoperability

Q: Does AWS Transit Gateway support IPv6?

A: Yes, AWS Transit Gateway supports attaching Amazon VPCs with IPv6 CIDRs.

Q: Which Amazon VPC features are not supported in the first release?

A: Security Group Referencing on Amazon VPC is not supported at launch. Spoke Amazon VPCs cannot reference security groups in other spokes connected to the same AWS Transit Gateway.

Q: Can I associate my AWS Transit Gateway with a Direct Connect gateway in a different account?

A: Yes, you can associate your AWS Transit Gateway with an AWS Direct Connect gateway from a different AWS account, if both of their AWS accounts have the same AWS payer account ID. Only the owner of the AWS Transit Gateway can create association to a Direct Connect gateway. You cannot use Resource Access Manager to associate your AWS Transit Gateway with Direct Connect gateway. For more information, please review the AWS Transit Gateway Support section in the Direct Connect FAQs.

Q: I want to associate my Transit Gateway to a Direct Connect gateway, can I use the same Autonomous System Number (ASN) for the Direct Connect gateway and the Transit Gateway?

A: No, you cannot use the same ASN for the Transit Gateway and the Direct Connect gateway.

Q: Which attachment types can I use to route multicast traffic?

A: You can route multicast traffic within and between VPC attachments to a Transit Gateway. Multicast routing is not supported over AWS Direct Connect, AWS Site-to-Site VPN, and peering attachments.

Network manager

Q: What is AWS Transit Gateway network manager?

A: AWS Transit Gateway network manager is a feature of AWS Transit Gateway. It centralizes management and monitoring of networking resources and connections to remote branch locations.

Q: How do I setup AWS Transit Gateway network manager?

A: Use the following steps to setup and manage Transit Gateway network manager:

- Create a new 'global network', initially an empty object.
- Register your AWS Transit Gateways from any AWS Region.
- Add on-premises resources: Input information about your on-premises devices, sites, links, and the Site-to-Site VPN connections with which they are associated.
- Monitor your global network: through Network Manager's visualizations, events, and metrics.

Q: Which AWS partners are supporting AWS Transit Gateway network manager?

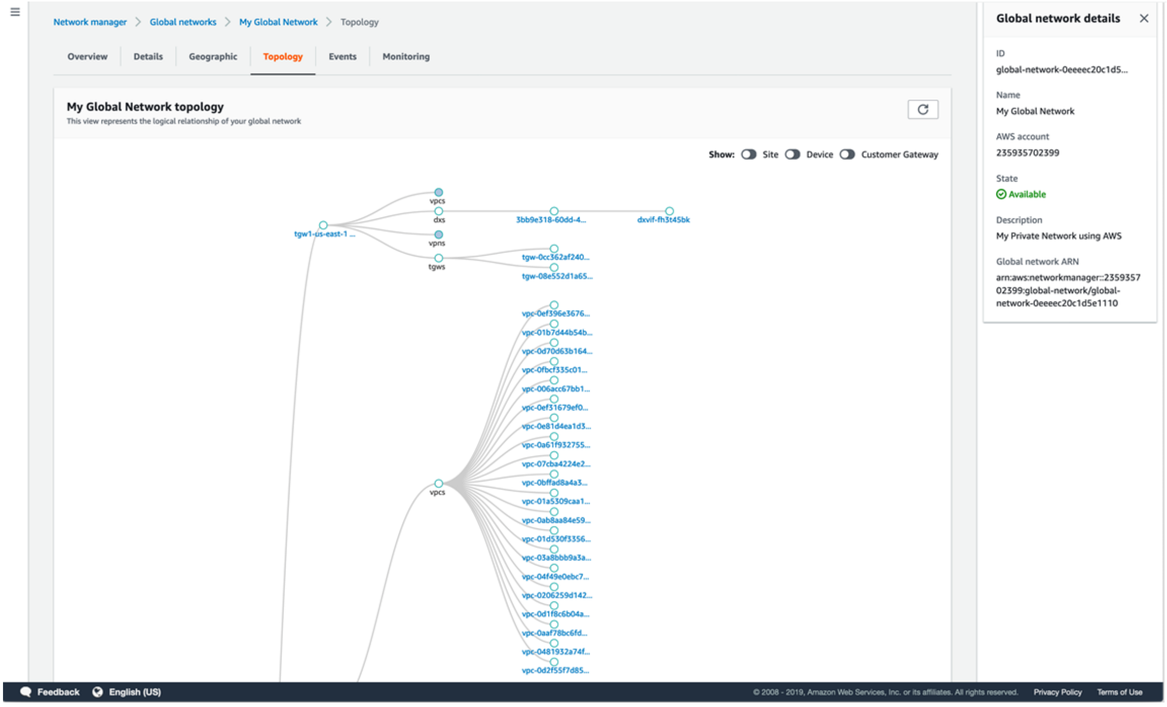
A: Currently, Cisco, Aruba, Silver Peak, and Aviatrix are supporting AWS Transit Gateway network manager. Their integration of network manager into their SD-WAN solutions enables to automate the branch-cloud connectivity and provides end-to-end monitoring of the global network from a single dashboard.

Q: What is a global network?

A: A 'Global Network' is an object in the AWS Transit Gateway network manager service that represents your private global network in AWS. It includes your AWS Transit Gateway hubs, their attachments, and on-premises devices, sites, and links.

Q: What resources are automatically included in the global network when I register an AWS Transit Gateway?

A: For registered AWS Transit Gateways, all attachments are automatically included. Attachments include VPCs, VPNs, Direct Connect gateways, and AWS Transit Gateway-AWS Transit Gateway peering.



My Global Network topology

This view represents the logical relationship of your global network

Show: Site Device Customer Gateway

Global network details

ID
global-network-0eeeec20c1d5...

Name
My Global Network

AWS account
235935702399

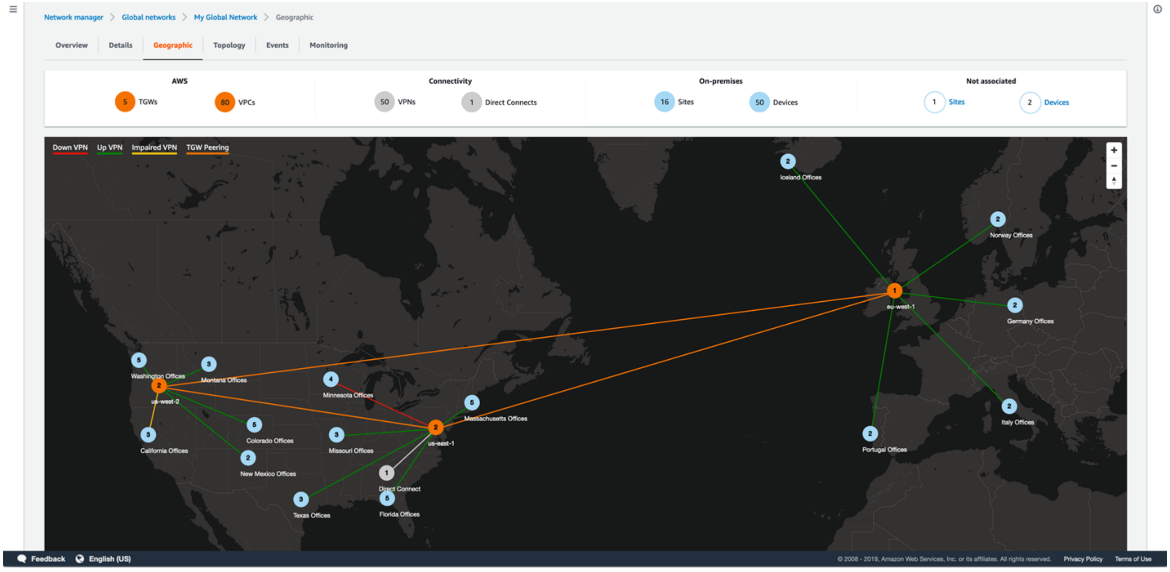
State
Available

Description
My Private Network using AWS

Global network ARN
arn:aws:networkmanager::235935702399:global-network/global-network-0eeeec20c1d5e1110

Q: How can I visualize the resources and connections in my global network?

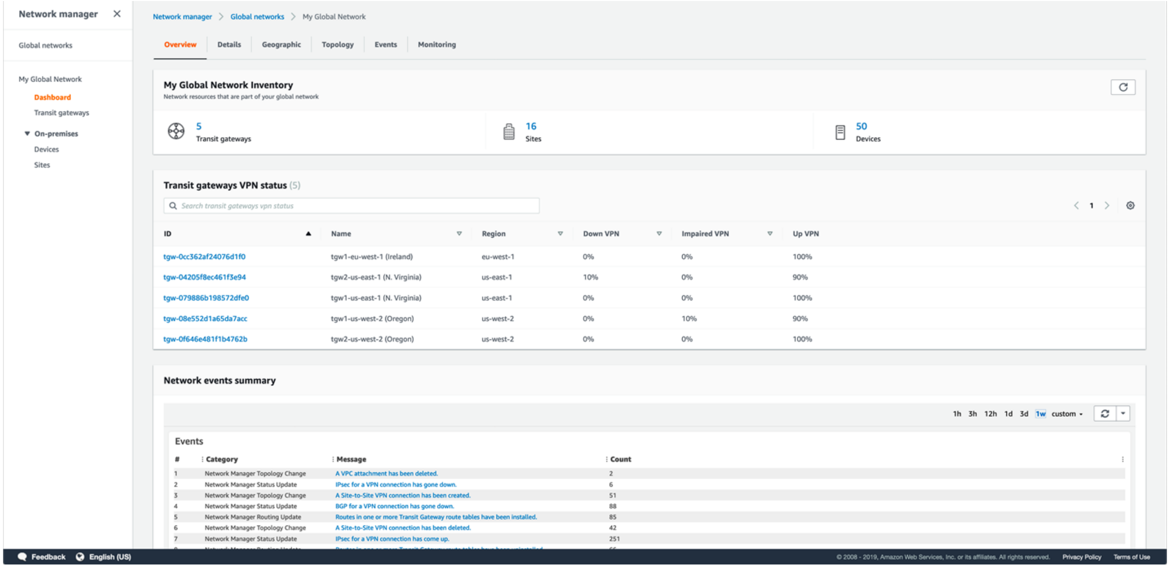
A: The AWS Transit Gateway network manager dashboard shows your AWS Transit Gateways across all AWS Regions and on-premises. It offers a logical view and a geographic view of your network resources and connections, along with connection status.



AWS		Connectivity		On-premises		Not associated	
5 TGWs	80 VPCs	50 VPNs	1 Direct Connects	16 Sites	50 Devices	1 Sites	2 Devices

Q: How does AWS Transit Gateway network manager help me monitor my global network?

A: The dashboard of AWS Transit Gateway network manager also shows you these events and metrics, such as bytes in/out, packets in/out, and packets dropped. Connection status is embedded into the the topology and goegraphic views of your global network. AWS Transit Gateway network manager also offers real-time network events and metrics for your global network through AWS CloudWatch. These events, metrics, and visualizations help you monitor your network and take actions as needed.



My Global Network Inventory

Network resources that are part of your global network

5 Transit gateways

16 Sites

50 Devices

Transit gateways VPN status (5)

Search transit gateways vpn status

ID	Name	Region	Down VPN	Impaired VPN	Up VPN
tgw-0cc362af24076d1f0	tgw1-eu-west-1 (Ireland)	eu-west-1	0%	0%	100%
tgw-04205f8ec461f3e94	tgw2-us-east-1 (N. Virginia)	us-east-1	10%	0%	90%
tgw-079886b198572dfe0	tgw1-us-east-1 (N. Virginia)	us-east-1	0%	0%	100%
tgw-08e552d1a65da7acc	tgw1-us-west-2 (Oregon)	us-west-2	0%	10%	90%
tgw-0f646e481f1b4762b	tgw2-us-west-2 (Oregon)	us-west-2	0%	0%	100%

Network events summary

1h 3h 12h 1d 3d 1w custom

Events

#	Category	Message	Count
1	Network Manager Topology Change	A VPC attachment has been deleted.	2
2	Network Manager Status Update	IPsec for a VPN connection has gone down.	6
3	Network Manager Topology Change	A Site-to-Site VPN connection has been created.	51
4	Network Manager Status Update	BGP for a VPN connection has gone down.	88
5	Network Manager Routing Update	Routes in one or more Transit Gateway route tables have been installed.	85
6	Network Manager Topology Change	A Site-to-Site VPN connection has been deleted.	42
7	Network Manager Status Update	IPsec for a VPN connection has come up.	251
8	Network Manager Routing Update	Routes in one or more Transit Gateway route tables have been uninstalled.	66

© 2008 - 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Q: What metrics are available in AWS Transit Gateway network manager?

A: From the dashboard of network manager, you can view Transit Gateway availability and performance metrics, such as bytes in/out, packets in/out, and packets dropped. AWS Site-to-Site VPN up/down metrics are also available to view for your on-premises devices and links.

Q: What network events are available in AWS Transit Gateway network manager?

A: AWS Transit Gateway network manager offers built-in event notifications for network topology changes, routing updates, and connection status updates. These events are delivered through CloudWatch Events.

Q: How do AWS partners support AWS Transit Gateway network manager?

A: SD-WAN providers offer integration with AWS Transit Gateway network manager. Their integration of network manager into their SD-WAN solutions enables them to automate the branch-cloud connectivity and provides end-to-end monitoring of the global network from a single pane of glass, the dashboard of the network manager.

Q: How do I automatically connect using a partner SD-WAN device?

A: Your SD-WAN solution from the partner uses AWS application programming interfaces (APIs) on your behalf to automatically register the branch device, create a VPN connection, and then applies the VPN configurations to the branch device to establish the connection.

AWS VPN FAQs

AWS VPN is comprised of two services: AWS Site-to-Site VPN and AWS Client VPN. AWS Site-to-Site VPN enables you to securely connect your on-premises network or branch office site to your Amazon Virtual Private Cloud (Amazon VPC). AWS Client VPN enables you to securely connect users to AWS or on-premises networks.

General questions

Q: What is a Client VPN endpoint?

A: The Client VPN endpoint is a regional construct that you configure to use the service. The VPN sessions of the end users terminate at the Client VPN endpoint. As part of configuring the Client VPN endpoint, you specify the authentication details, server certificate information, client IP address allocation, logging, and VPN options.

Q: What is a target network?

A: A target network, is a network that you associate to the Client VPN endpoint that enables secure access to your AWS resources as well as access to on-premises. Currently, the target network is a subnet in your Amazon VPC.

Billing

Q: What defines billable VPN connection-hours?

A: VPN connection-hours are billed for any time your VPN connections are in the "available" state. You can determine the state of a VPN connection via the AWS Management Console, CLI, or API. If you no longer wish to use your VPN connection, you simply terminate the VPN connection to avoid being billed for additional VPN connection-hours.

Q: Do your prices include taxes?

A: Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

AWS Site-to-Site VPN setup and management

Q: Can I use the AWS Management Console to control and manage AWS Site-to-Site VPN?

A: Yes. You can use the AWS Management Console to manage IPSec VPN connections, such as AWS Site-to-Site VPN.

Q: How many customer gateways, virtual private gateways, and AWS Site-to-Site VPN connections can I create?

A: You can have:

- One internet gateway per VPC
- Five virtual private gateways per AWS account per AWS Region
- Fifty customer gateways per AWS account per AWS Region
- Ten IPsec VPN Connections per virtual private gateway

See the VPC User Guide for more information on VPC limits.

AWS Site-to-Site VPN connectivity

Q: What are the VPN connectivity options for my VPC?

A: You may connect your VPC to your corporate data center using a Hardware VPN connection via the virtual private gateway.

Q: How do instances without public IP addresses access the Internet?

A: Instances without public IP addresses can access the Internet in one of two ways:

Instances without public IP addresses can route their traffic through a network address translation (NAT) gateway or a NAT instance to access the internet. These instances use the public IP address of the NAT gateway or NAT instance to traverse the internet. The NAT gateway or NAT instance allows outbound communication but doesn’t allow machines on the internet to initiate a connection to the privately addressed instances.

For VPCs with a hardware VPN connection or Direct Connect connection, instances can route their Internet traffic down the virtual private gateway to your existing datacenter. From there, it can access the Internet via your existing egress points and network security/monitoring devices.

Q: How does an AWS Site-to-Site VPN connection work with Amazon VPC?

A: An AWS Site-to-Site VPN connection connects your VPC to your datacenter. Amazon supports Internet Protocol security (IPsec) VPN connections. Data transferred between your VPC and datacenter routes over an encrypted VPN connection to help maintain the confidentiality and integrity of data in transit. An Internet gateway is not required to establish a Site-to-Site VPN connection.

Q: What is IPSec?

A: IPsec is a protocol suite for securing Internet Protocol (IP) communications by authenticating and encrypting each IP packet of a data stream.

Q: Which customer gateway devices can I use to connect to Amazon VPC?

A: You can create two types of AWS Site-to-Site VPN connections: statically routed VPN connections and dynamically-routed VPN connections. Customer gateway devices supporting statically-routed VPN connections must be able to:

Establish IKE Security Association using Pre-Shared Keys

Establish IPsec Security Associations in Tunnel mode

Utilize the AES 128-bit or 256-bit encryption function

Utilize the SHA-1 or SHA-2 (256) hashing function

Utilize Diffie-Hellman (DH) Perfect Forward Secrecy in "Group 2" mode, or one of the additional DH groups we support

Perform packet fragmentation prior to encryption

In addition to the above capabilities, devices supporting dynamically-routed Site-to-Site VPN connections must be able to:

Establish Border Gateway Protocol (BGP) peering

Bind tunnels to logical interfaces (route-based VPN)

Utilize IPsec Dead Peer Detection

Q: Which Diffie-Hellman groups do you support?

A; We support the following Diffie-Hellman (DH) groups in Phase 1 and Phase 2.

Phase 1 DH groups 2, 14-18, 22, 23, 24

Phase 2 DH groups 2, 5, 14-18, 22, 23, 24

Q: What customer gateway devices are known to work with Amazon VPC?

A: In The network administrator guide, you will find a list of the devices meeting the aforementioned requirements, that are known to work with hardware VPN connections, and that will support in the command line tools for automatic generation of configuration files appropriate for your device.

Q: If my device is not listed, where can I go for more information about using it with Amazon VPC?

A: We recommend checking the Amazon VPC forum as other customers may be already using your device.

Q: What is the approximate maximum throughput of a Site-to-Site VPN connection?

A: Virtual gateway supports IPSEC VPN throughput up to 1.25 Gbps. Multiple VPN connections to the same VPC are cumulatively bound by the virtual gateway throughput of 1.25 Gbps.

Q: What factors affect the throughput of my VPN connection?

A: VPN connection throughput can depend on multiple factors, such as the capability of your customer gateway, the capacity of your connection, average packet size, the protocol being used, TCP vs. UDP, and the network latency between your customer gateway and the virtual private gateway.

Q: What tools are available to me to help troubleshoot my Site-to-Site VPN configuration?

A: The DescribeVPNConnection API displays the status of the VPN connection, including the state ("up"/"down") of each VPN tunnel and corresponding error messages if either tunnel is "down". This information is also displayed in the AWS Management Console.

Q: How do I connect a VPC to my corporate datacenter?

A: Establishing a hardware VPN connection between your existing network and Amazon VPC allows you to interact with Amazon EC2 instances within a VPC as if they were within your existing network. AWS does not perform network address translation (NAT) on Amazon EC2 instances within a VPC accessed via a hardware VPN connection.

Q: Can I NAT my customer gateway behind a router or firewall?

A: You will use the public IP address of your NAT device.

Q: What IP address do I use for my customer gateway address?

A: You will use the public IP address of your NAT device.

Q: How do I disable NAT-T on my connection?

A: You will need to disable NAT-T on your device. If you don't plan on using NAT-T and it is not disabled on your device, we will attempt to establish a tunnel over UDP port 4500. If that port is not open the tunnel will not establish.

Q: I would like to have multiple customer gateways behind a NAT, what do I need to do to configure that?

A: You will need to disable NAT-T on your device. If you don't plan on using NAT-T and it is not disabled on your device, we will attempt to establish a tunnel over UDP port 4500. If that port is not open the tunnel will not establish.

Q: How many IPsec security associations can be established concurrently per tunnel?

A: The AWS VPN service is a route-based solution, so when using a route-based configuration you will not run into SA limitations. If, however, you are using a policy-based solution you will need to limit to a single SA, as the service is a route-based solution.

Q: Can I advertise my VPC public IP address range to the internet and route the traffic through my datacenter, via the Site-to-Site VPN, and to my VPC?

A: Yes, you can route traffic via the hardware VPN connection and advertise the address range from your home network.

AWS Accelerated Site-to-Site VPN

Q: Why should I use Accelerated Site-to-Site VPN?

A: VPN connections face inconsistent availability and performance as traffic traverses through multiple public networks on the internet before reaching the VPN endpoint in AWS. These public networks can be congested. Each hop can introduce availability and performance risks. Accelerated Site-to-Site VPN makes user experience more consistent by using the highly available and congestion-free AWS global network.

Q: How can I create an Accelerated Site-to-Site VPN?

A: When creating a VPN connection, set the option “Enable Acceleration” to ‘true’.

Q: How do I find out whether my existing VPN connection is an Accelerated Site-to-Site VPN?

A: In the description of your VPN connection, the value for “Enable Acceleration” should be set to ‘true’.

Q: How can I convert my existing Site-to-Site VPN to an Accelerated Site-to-Site VPN?

A: Create a new Accelerated Site-to-Site VPN, update your customer gateway device to connect to this new VPN connection, and then delete your existing VPN connection. You will get new tunnel endpoint internet protocol (IP) addresses since accelerated VPNs use separate IP address ranges from non-accelerated VPN connections.

Q: Is Accelerated Site-to-Site VPN supported for both virtual gateway and AWS Transit Gateway?

A: Only Transit Gateway supports Accelerated Site-to-Site VPN. A Transit Gateway should be specified when creating a VPN connection. The VPN endpoint on the AWS side is created on the Transit Gateway.

Q: Does an Accelerated Site-to-Site VPN connection offer two tunnels for high availability?

A: Yes, each VPN connection offers two tunnels for high availability.

Q: Are there any protocol differences between Accelerated and non-Accelerated Site-to-Site VPN tunnels?

A: NAT-T is required and is enabled by default for Accelerated Site-to-Site VPN connections. Other that that, Accelerated and non-Accelerated VPN tunnels support the same IP security (IPSec) and internet key exchange (IKE) protocols, and also offer the same bandwidth, tunnel options, routing options, and authentication types.

Q: Does Accelerated Site-to-Site VPN offer two network zones for high availability?

A: Yes, we select AWS Global Accelerator global internet protocol addresses (IPs) from independent network zones for the two tunnel endpoints.

Q: Is Accelerated Site-to-Site VPN an option in AWS Global Accelerator?

A: No, Accelerated Site-to-Site VPN can only by created through AWS Site-to-Site VPN. Accelerated Site-to-Site VPNs cannot be created through the AWS Global Accelerator console or API.

Q: Can I use Accelerated VPN over public AWS Direct Connect virtual interfaces?

A: No, Accelerated Site-to-Site VPN over public Direct Connect virtual interfaces is not available. In most cases there is no acceleration benefit of Accelerated Site-to-Site VPN when used over public Direct Connect.

Q: In which AWS Regions is Accelerated Site-to-Site VPN available?

A: Accelerated Site-to-Site VPN available is currently available in these AWS Regions: US East (N. Virginia), US East (Ohio), US West (Oregon), US West (N. California), EU (Ireland), EU (Frankfurt), EU (London), EU (Paris), Asia Pacific (Singapore), Asia Pacific (Tokyo), Asia Pacific (Sydney), Asia Pacific (Seoul), Asia Pacific (Mumbai), and Canada (Central).

AWS Client VPN setup and management

Q: How do I setup AWS Client VPN?

A: The IT administrator creates a Client VPN endpoint, associates a target network to that endpoint and sets up the access policies to allow end user connectivity. The IT administrator distributes the client VPN configuration file to the end users. End users will need to download an OpenVPN client and use the client VPN configuration file to create their VPN session.

Q: What should an end user do to setup a connection?

A: The end user should download an OpenVPN client to their device. Next, the user will import the AWS Client VPN configuration file to the OpenVPN client and initiate a VPN connection.

AWS Client VPN connectivity

Q: How do I enable connectivity to other networks?

A: You can enable connectivity to other networks like peered Amazon VPCs, on-premises networks via virtual gateway or AWS services, such as S3, via endpoints, networks via AWS PrivateLink or other resources via internet gateway. To enable connectivity, add a route to the specific network in the Client VPN route table, and add authorization rule enabling access to the specific network.

Q: Can the Client VPN endpoint belong to a different account from the associated subnet?

A: No, the subnet being associated has to be in the same account as Client VPN endpoint.

Q: Can I access resources in a VPC within a different region different from the region in which I setup the TLS session, using a Private IP address?

A: You can achieve this by following the two steps: First, set up a cross-region peering connection between your destination VPC (in the different region) and the Client VPN associated VPC. Second, you should add a route and access rule for the destination VPC in the Client VPN endpoint. Your users can now access the resources in the destination VPC that is in a different region from your Client VPN endpoint.

Q: What transport protocols are supported by Client VPN?

A: You can choose either TCP or UDP for the VPN session.

Q: Does AWS Client VPN support split tunnel?

A: Yes. You may choose to create an endpoint with split tunnel enabled or disabled. If you've previously created an endpoint with split tunnel disabled, you may choose to modify it it to enable split tunnel. If split tunnel is enabled, traffic destined for routes configured on the endpoint will be routed via the VPN tunnel. All other traffic will be routed via your local network interface. If split tunnel is disabled, all the traffic from the device will traverse through the VPN tunnel.

AWS Client VPN authentication and authorization

Q: What authentication mechanisms does AWS Client VPN support?

A: Client VPN supports authentication with Active Directory using AWS Directory Services and Certificate-based authentication.

Q: Can I use an on-premises Active Directory service to authenticate users?

A: Yes. AWS Client VPN integrates with AWS Directory Service that will allow you to connect to on-premises Active Directory.

Q: Does AWS Client VPN support mutual authentication?

A: Yes, AWS Client VPN supports mutual authentication. When mutual authentication is enabled, customer have to upload the root certificate used to issue the client certificate on the server.

Q: Can I blacklist client certificates?

A: Yes, AWS Client VPN supports statically-configured Certificate Revocation List (CRL).

Q: Does AWS Client VPN support the ability for a customer to bring their own certificate?

A: Yes. You should upload the certificate, root certification authority (CA) certificate, and the private key of the server. These are uploaded to AWS Certificate Manager.

Q: Does AWS Client VPN integrate with AWS Certificate Manager (ACM) to generate server certificates?

A: Yes. You can use ACM as a subordinate CA chained to an external root CA. ACM then generates the server certificate. In this scenario, ACM also does the server certificate rotation.

Q: Does AWS Client VPN support posture assessment?

A: No. AWS Client VPN does not support posture assessment. Other AWS services, such as Amazon Inspectors, support posture assessment.

Q: Does AWS Client VPN support Multi-Factor Authentication (MFA)?

A: Yes, AWS Client VPN supports MFA through Active Directory using AWS Directory Services.

Q: How does AWS Client VPN support authorization?

A: You configure authorization rules which limit the users who can access a network. For a specified network, you configure the Active Directory group that is allowed access. Only users belonging to this Active Directory group can access the specified network.

Q: Does AWS Client VPN support security group?

A: Client VPN supports security group. You can specify security group for the group of associations. When a subnet is associated, we will automatically apply the default security group of the VPC of the subnet.

Q: How do I use security group to restrict access to my applications for only Client VPN connections?

A: For your application, you can specify to allow access only from the security groups that were applied to the associated subnet. Now you limit access to only users connected via Client VPN.

AWS Client VPN visibility and monitoring

Q: What logs are supported for AWS Client VPN?

A: Client VPN exports the connection log as a best effort to CloudWatch logs. These logs are exported periodically at 15 minute intervals. The connection logs include details on created and terminated connection requests.

Q: Does Client VPN support Amazon VPC Flow Logs in the endpoint?

A: No. You can use Amazon VPC Flow Logs in the associated VPC.

Q: Can I monitor active connections?

A: Yes, using the CLI or console, you can view the current active connections for an endpoint and terminate active connections.

Q: Can I monitor by endpoint using CloudWatch?

A: Yes. Using CloudWatch monitor you can see Ingress and Egress bytes and Active connections for each Client VPN Endpoint.

VPN clients

Q: How do I deploy the free software client for AWS Client VPN?

A: The software client for AWS Client VPN is compatible with existing AWS Client VPN configurations. The client supports adding profiles using the OpenVPN configuration file generated by the AWS Client VPN service. Once the profile is created, the client will connect to your endpoint based on your settings.

Q: What is the additional price to use the software client of AWS Client VPN?

A: The software client is provided free of charge. You will only be billed for AWS Client VPN service usage.

Q: What type of devices and operating system versions are supported?

A: We currently support Windows 7 (and above) and macOS (64 bit versions of macOS Sierra and above) desktop devices. These include Windows 10 and macOS devices. As new operating systems are released, we will add support for them quickly.

Q: Do my connection profiles synchronize between all of my devices?

A: No, but IT administrators can provide configuration files for their software client deployment to pre-configure settings.

Q: Do I need admin permission on my device to run the software client of AWS Client VPN?

A: Yes. You need admin access to install the app on both Windows and Mac. After that point, admin access is not required.

Q: What VPN protocol is used by the client of AWS Client VPN?

A: AWS Client VPN, including the software client, supports the OpenVPN protocol.

Q: Will all the features supported by AWS Client VPN service be supported using the software client?

A: Yes. The client supports all the features provided by the AWS Client VPN service.

Q: Does the software client of AWS Client VPN allow LAN access when connected?

A: Yes, you can access your local area network when connected to AWS VPN Client.

Q: What authentication capabilities can I support using the software client?

A: The software client of AWS Client VPN supports Active Directory using AWS Directory Services (with and without MFA) and mutual authentication.

Q: What type of client logging will be supported by AWS Client VPN?

A: When a user attempts to connect, the details of the connection setup are logged. Connection attempts are saved up to 30 days with a maximum file size of 90 MB.

Q: Can I mix the software client of AWS Client VPN and standards based OpenVPN clients connecting to AWS Client VPN endpoint?

A: Yes, assuming that the authentication type defined on the AWS Client VPN endpoint is supported by the standards-based OpenVPN client.

Q: Where can I download the software client of AWS Client VPN?

A: You can download the generic client without any customizations from the AWS Client VPN product page. IT administrators may choose to host the download within their own system.

Q: Can I run multiple types of VPN clients on one device?

A: We do not recommend running multiple VPN clients on a device. This can cause conflicts or the VPN clients can interfere with each other and cause unsuccessful connections. That said, the AWS Client VPN can be installed alongside another VPN client.

Virtual private gateway

Q: What is this feature?

A: For any new virtual gateways, configurable Private Autonomous System Number (ASN) allows customers to set the ASN on the Amazon side of the BGP session for VPNs and AWS Direct Connect private VIFs.

Q: What is the cost of using this feature?

A: There is no additional charge for this feature.

Q: How can I configure/assign my ASN to be advertised as Amazon side ASN?

A: You can configure/assign an ASN to be advertised as the Amazon side ASN during creation of the new Virtual Private Gateway (virtual gateway). You can create a virtual gateway using the VPC console or a EC2/CreateVpnGateway API call.

Q: What ASN did Amazon assign prior to this feature?

A: Amazon assigned the following ASNs: EU West (Dublin) 9059; Asia Pacific (Singapore) 17493 and Asia Pacific (Tokyo) 10124. All other regions were assigned an ASN of 7224; these ASNs are referred as “legacy public ASN” of the region.

Q: Can I use any ASN – public and private?

A: You can assign any private ASN to the Amazon side. You can assign the "legacy public ASN" of the region until June 30th 2018, you cannot assign any other public ASN. After June 30th 2018, Amazon will provide an ASN of 64512.

Q: Why can’t I assign a public ASN for the Amazon half of the BGP session?

A: Amazon is not validating ownership of the ASNs, therefore, we’re limiting the Amazon-side ASN to private ASNs. We want to protect customers from BGP spoofing.

Q: What ASN can I choose?

A: You can choose any private ASN. Ranges for 16-bit private ASNs include 64512 to 65534. You can also provide 32-bit ASNs between 4200000000 and 4294967294.

Amazon will provide a default ASN for the virtual gateway if you don’t choose one. Until June 30th 2018, Amazon will continue to provide the “legacy public ASN” of the region. After June 30th 2018, Amazon will provide an ASN of 64512.

Q: What will happen if I try to assign a public ASN to the Amazon half of the BGP session?

A: We will ask you to re-enter a private ASN once you attempt to create the virtual gateway, unless it is the "legacy public ASN" of the region.

Q: If I don’t provide an ASN for the Amazon half of the BGP session, what ASN can I expect Amazon to assign to me?

A: Amazon will provide an ASN for the virtual gateway if you don’t choose one. Until June 30th 2018, Amazon will continue to provide the “legacy public ASN” of the region. After June 30th 2018, Amazon will provide an ASN of 64512.

Q: Where can I view the Amazon side ASN?

A: You can view the Amazon side ASN in the virtual gateway page of VPC console and in the response of EC2/DescribeVpnGateways API.

Q: If I have a public ASN, will it work with a private ASN on the AWS side?

A: Yes, you can configure the Amazon side of the BGP session with a private ASN and your side with a public ASN.

Q: I have private VIFs already configured and want to set a different Amazon side ASN for the BGP session on an existing VIF. How can I make this change?

A: You will need to create a new virtual gateway with desired ASN, and create a new VIF with the newly created virtual gateway. Your device configuration also needs to change appropriately.

Q: I have VPN connections already configured and want to modify the Amazon side ASN for the BGP session of these VPNs. How can I make this change?

A: You will need to create a new virtual gateway with the desired ASN, and recreate your VPN connections between your Customer Gateways and the newly created virtual gateway.

Q: I already have a virtual gateway and a private VIF/VPN connection configured using an Amazon assigned public ASN of 7224. If Amazon automatically generates the ASN for the new private virtual gateway, what Amazon side ASN will I be assigned?

A: Amazon will assign 64512 to the Amazon side ASN for the new virtual gateway.

Q: I have a virtual gateway and a private VIF/VPN connection configured using an Amazon assigned public ASN. I want to use the same Amazon assigned public ASN for the new private VIF/VPN connection I’m creating. How do I do this?

A: You can configure/assign an ASN to be advertised as the Amazon side ASN during creation of the new Virtual Private Gateway (virtual gateway). You can create virtual gateway using console or EC2/CreateVpnGateway API call. As noted earlier, until June 30th 2018, Amazon will continue to provide the “legacy public ASN” of the region. After June 30th 2018, Amazon will provide an ASN of 64512.

Q: I have a virtual gateway and a private VIF/VPN connection configured using an Amazon assigned public ASN of 7224. If Amazon auto generates the ASN for the new private VIF/VPN connection using the same virtual gateway, what Amazon side ASN will I be assigned?

A: Amazon will assign 7224 to the Amazon side ASN for the new VIF/VPN connection. The Amazon side ASN for your new private VIF/VPN connection is inherited from your existing virtual gateway and defaults to that ASN.

Q: I’m attaching multiple private VIFs to a single virtual gateway. Can each VIF have a separate Amazon side ASN?

A: No, you can assign/configure separate Amazon side ASN for each virtual gateway, not each VIF. Amazon side ASN for VIF is inherited from the Amazon side ASN of the attached virtual gateway.

Q: I’m creating multiple VPN connections to a single virtual gateway. Can each VPN connection have a separate Amazon side ASN?

A: No, you can assign/configure separate Amazon side ASN for each virtual gateway, not each VPN connection. Amazon side ASN for VPN connection is inherited from the Amazon side ASN of the virtual gateway.

Q: Where can I select my own ASN?

A: When creating a virtual gateway in the VPC console, uncheck the box asking if you want an auto-generated Amazon BGP ASN and provide your own private ASN for the Amazon half of the BGP session. Once virtual gateway is configured with Amazon side ASN, the private VIFs or VPN connections created using the virtual gateway will use your Amazon side ASN.

Q. I use CloudHub today. Will I have to adjust my configurations in the future?

A: You will not have to make any changes.

Q: I want to select a 32-bit ASN. What is the range of 32-bit private ASNs?

A: We will support 32-bit ASNs from 4200000000 to 4294967294.

Q: Once the virtual gateway is created, can I change or modify the Amazon side ASN?

A: No, you cannot modify the Amazon side ASN after creation. You can delete the virtual gateway and recreate a new virtual gateway with the desired ASN.

Q: Is there a new API to configure/assign the Amazon side ASN?

A: No. You can do this with the same API as before (EC2/CreateVpnGateway). We just added a new parameter (amazonSideAsn) to this API.

Q: Is there a new API to view the Amazon side ASN?

A: No. You can view the Amazon side ASN with the same EC2/DescribeVpnGateways API. We just added a new parameter (amazonSideAsn) to this API.

Elastic Load Balancing FAQs

General

Q: How do I decide which load balancer to select for my application?

A: Elastic Load Balancing supports three types of load balancers. You can select the appropriate load balancer based on your application needs. If you need to load balance HTTP requests, we recommend you to use Application Load Balancer. For network/transport protocols (layer4 – TCP, UDP) load balancing, and for extreme performance/low latency applications we recommend using Network Load Balancer. If your application is built within the EC2 Classic network then you should use Classic Load Balancer.

Q: Can I privately access Elastic Load Balancing APIs from my Amazon Virtual Private Cloud (VPC) without using public IPs?

A: Yes, you can privately access Elastic Load Balancing APIs from your Amazon Virtual Private Cloud (VPC) by creating VPC endpoints. With VPC endpoints, the routing between the VPC and Elastic Load Balancing APIs is handled by the AWS network without the need for an Internet gateway, NAT gateway, or VPN connection. The latest generation of VPC Endpoints used by Elastic Load Balancing are powered by AWS PrivateLink, an AWS technology enabling the private connectivity between AWS services using Elastic Network Interfaces (ENI) with private IPs in your VPCs. To learn more about AWS PrivateLink, visit the AWS PrivateLink documentation.

Q: Is there an SLA for load balancers?

A: Yes, Elastic Load Balancing guarantees a monthly availability of at least 99.99% for your load balancers (Classic, Application or Network). To learn more about the SLA and know if you are qualified for a credit, visit https://aws.amazon.com/elasticloadbalancing/sla/.

Application Load Balancer

Q: Which operating systems does an Application Load Balancer support?

A: An Application Load Balancer supports targets with any operating system currently supported by the Amazon EC2 service.

Q: Which protocols does an Application Load Balancer support?

A: An Application Load Balancer supports load balancing of applications using HTTP and HTTPS (Secure HTTP) protocols.

Q: Is HTTP/2 Supported on an Application Load Balancer?

A: Yes. HTTP/2 support is enabled natively on an Application Load Balancer. Clients that support HTTP/2 can connect to an Application Load Balancer over TLS.

Q: What TCP ports can I use to load balance?

A: You can perform load balancing for the following TCP ports: 1-65535

Q: Is WebSockets supported on an Application Load Balancer?

A: Yes. WebSockets and Secure WebSockets support is available natively and ready for use on an Application Load Balancer.

Q: Is Request tracing supported on an Application Load Balancer?

A: Yes. Request tracing is enabled by default on your Application Load Balancer.

Q: Does a Classic Load Balancer have the same features and benefits as an Application Load Balancer?

A: While there is some overlap, there is no feature parity between the two types of load balancers. Application Load Balancers are the foundation of our application layer load-balancing platform for the future.

Q: Can I configure my Amazon EC2 instances to accept traffic only from my Application Load Balancers?

A: Yes.

Q: Can I configure a security group for the front-end of an Application Load Balancer?

A: Yes.

Q: Can I use the existing APIs that I use with my Classic Load Balancer with an Application Load Balancer?

A: No. Application Load Balancers require a new set of APIs.

Q: How do I manage both Application and Classic Load Balancers simultaneously?

A: The ELB Console will allow you to manage Application and Classic Load Balancers from the same interface. If you are using the CLI or an SDK, you will use a different 'service' for Application Load Balancers. For example, in the CLI you will describe your Classic Load Balancers using `aws elb describe-load-balancers` and your Application Load Balancers using `aws elbv2 describe-load-balancers`.

Q: Can I convert my Classic Load Balancer to an Application Load Balancer (and vice versa)?

A: No, you cannot convert one load balancer type into another.

Q: Can I migrate to Application Load Balancer from Classic Load Balancer?

A: Yes. You can migrate to Application Load Balancer from Classic Load Balancer using one of the options listed in this document.

Q: Can I use an Application Load Balancer as a Layer-4 load balancer?

A: No. If you need Layer-4 features, you should use Network Load Balancer.

Q: Can I use a single Application Load Balancer for handling HTTP and HTTPS requests?

A: Yes, you can add listeners for HTTP port 80 and HTTPS port 443 to a single Application Load Balancer.

Q: Can I get a history of Application Load Balancing API calls made on my account for security analysis and operational troubleshooting purposes?

A: Yes. To receive a history of Application Load Balancing API calls made on your account, use AWS CloudTrail.

Q: Does an Application Load Balancer support HTTPS termination?

A: Yes, you can terminate HTTPS connection on the Application Load Balancer. You must install an SSL certificate on your load balancer. The load balancer uses this certificate to terminate the connection and then decrypt requests from clients before sending them to targets.

Q: What are the steps to get a SSL certificate?

A: You can either use AWS Certificate Manager to provision an SSL/TLS certificate or you can obtain the certificate from other sources by creating the certificate request, getting the certificate request signed by a CA, and then uploading the certificate either using AWS Certification Manager or the AWS Identity and Access Management (IAM) service.

Q: How does an Application Load Balancer integrate with AWS Certificate Manager (ACM)?

A: An Application Load Balancer is integrated with AWS Certificate Management (ACM). Integration with ACM makes it very simple to bind a certificate to the load balancer thereby making the entire SSL offload process very easy. Purchasing, uploading, and renewing SSL/TLS certificates is a time-consuming manual and complex process. With ACM integration with Application Load Balancer, this whole process has been shortened to simply requesting a trusted SSL/TLS certificate and selecting the ACM certificate to provision it with the load balancer.

Q: Is back-end server authentication supported with an Application Load Balancer?

A: No, only encryption is supported to the back-ends with an Application Load Balancer.

Q: How can I enable Server Name Indication (SNI) for my Application Load Balancer?

A: SNI is automatically enabled when you associate more than one TLS certificate with the same secure listener on a load balancer. Similarly, SNI mode for a secure listener is automatically disabled when you have only one certificate associated to a secure listener.

Q: Can I associate multiple certificates for the same domain to a secure listener?

A: Yes, you can associate multiple certificates for the same domain to a secure listener. For example, you can associate:

ECDSA and RSA certificates

Certificates with different key sizes (e.g. 2K and 4K) for SSL/TLS certificates

Single-Domain, Multi-Domain (SAN) and Wildcard certificates

Q: Is IPv6 supported with an Application Load Balancer?

A: Yes, IPv6 is supported with an Application Load Balancer.

Q: How do you set up rules on an Application Load Balancer?

A: You can configure rules for each of the listeners that you have on the load balancer. The rules include conditions and corresponding actions if the conditions are satisfied. The supported conditions are Host header, path, HTTP headers, methods, query parameters, and source IP CIDRs. The supported actions are redirect, fixed response, authenticate, and forward. Once you have set this up, the load balancer will use the rules to determine how a particular HTTP request should be routed. You can use multiple conditions and actions in a rule and in each condition can specify a match on multiple values.

Q: Are there limits on the resources for an Application Load Balancer?

A: Your AWS account has these limits for an Application Load Balancer.

Q. How can I protect my web applications behind a load balancer from web attacks?

A: You can integrate your Application Load Balancer with AWS WAF, a web application firewall that helps protect web applications from attacks by allowing you to configure rules based on IP addresses, HTTP headers, and custom URI strings. Using these rules, AWS WAF can block, allow, or monitor (count) web requests for your web application. Please see AWS WAF developer guide for more information.

Q: Can I load balance to any arbitrary IP address?

A: You can use any IP address from the load balancer's VPC CIDR for targets within load balancer's VPC and any IP address from RFC 1918 ranges (10.0.0.0/8, 172.16.0.0/12, and 192.168.0.0/16) or RFC 6598 range (100.64.0.0/10) for targets located outside the load balancer's VPC (for example, targets in Peered VPC, EC2-Classic and on-premises locations reachable over AWS Direct Connect or VPN connection).

Q: How can I load balance applications distributed across a VPC and on-premises location?

A: There are various ways to achieve hybrid load balancing. If an application runs on targets distributed between a VPC and an on-premises location, you can add them to the same target group using their IP addresses. To migrate to AWS without impacting your application, gradually add VPC targets to the target group and remove on-premises targets from the target group. If you have two different applications such that the targets for one application are in a VPC and the targets for other applications are in on-premises location, you can put the VPC targets in one target group and the on-premises targets in another target group and use content based routing to route traffic to each target group. You can also use separate load balancers for VPC and on-premises targets and use DNS weighting to achieve weighted load balancing between VPC and on-premises targets.

Q: How can I load balance to EC2-Classic instances?

A: You cannot load balance to EC2-Classic Instances when registering their Instance IDs as targets. However if you link these EC2-Classic instances to the load balancer's VPC using ClassicLink and use the private IPs of these EC2-Classic instances as targets, then you can load balance to the EC2-Classic instances. If you are using EC2 Classic instances today with a Classic Load Balancer, you can easily migrate to an Application Load Balancer.

Q: How do I enable cross-zone load balancing in Application Load Balancer?

A: Cross-zone load balancing is already enabled by default in Application Load Balancer.

Q: When should I authenticate users using the Application Load Balancer's integration with Amazon Cognito vs. the Application Load Balancers' native support for OpenID Connect (IODC) identity providers (IdPs)?

A: You should use authentication through Amazon Cognito if:

- You want to provide flexibility to your users to authenticate via social network identities (Google, Facebook, and Amazon) or enterprise identities (SAML) or via your own user directories provided by Amazon Cognito's User Pool.
- You are managing multiple identity providers including OpenID Connect and want to create a single authentication rule in Application Load Balancer (ALB), that can use Amazon Cognito to federate your multiple identity providers.
- You have a need to actively manage user profiles with one or more social or OpenID Connect identity providers from one central place. For example, you can put users in groups and add custom attributes to represent user status and control access for paid users.

Alternatively, if you have invested in developing custom IdP solutions and simply want to authenticate with a single identity provider that is OpenID Connect-compatible, you may prefer using Application Load Balancer's native OIDC solution.

Q: What type of redirects does Application Load Balancer support ?

A: The following three types of redirects are supported.

Types of redirects	Examples
HTTP to HTTP	http://hostA to http://hostB
HTTP to HTTPS	http://hostA to https://hostB https://hostA:portA/pathA to https://hostB:portB/pathB
HTTPS to HTTPS	https://hostA to https://hostB

Q: What content types does ALB support for the message body of fixed-response action?

A: The following content types are supported: text/plain, text/css, text/html, application/javascript, application/json.

Q: How does Lambda invocation via Application Load Balancer work?

A: HTTP(S) requests received by a load balancer are processed by the content-based routing rules. If the request content matches the rule with an action to forward it to a target group with a Lambda function as a target then the corresponding Lambda function is invoked. The content of the request (including headers and body) is passed on to the Lambda function in JSON format. The response from the Lambda function should be in JSON format. The response from the Lambda function is transformed into an HTTP response and sent to the client. The load balancer invokes your Lambda function using the AWS Lambda Invoke API and requires that you have provided invoke permissions for your Lambda function to Elastic Load Balancing service.

Q: Does Lambda invocation via Application Load Balancer support requests over both HTTP and HTTPS protocol?

A: Yes. Application Load Balancer supports Lambda invocation for requests over both HTTP and HTTPS protocol.

Q: In which AWS Regions can I use Lambda functions as targets with the Application Load Balancer?

A: You can use Lambda as a target with the Application Load Balancer in US East (N. Virginia), US East (Ohio), US West (Northern California), US West (Oregon), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), EU (Frankfurt), EU (Ireland), EU (London), EU (Paris), South America (São Paulo), and GovCloud (US-West) AWS Regions.

Q: Is the Application Load Balancer available in Local Zones?

A: Yes, Application Load Balancer is available in the Local Zone in Los Angeles. Within the Los Angeles Local Zone, Application Load Balancer will operate in a single subnet and scale automatically to meet varying levels of application load without manual intervention.

Application Load Balancer Pricing FAQs

Q: How does Application Load Balancer pricing work?

A: You are charged for each hour or partial hour that an Application Load Balancer is running and the number of Load Balancer Capacity Units (LCU) used per hour.

Q: What is a Load Balancer Capacity Unit (LCU)?

A: An LCU is a new metric for determining how you pay for an Application Load Balancer. An LCU defines the maximum resource consumed in any one of the dimensions (new connections, active connections, bandwidth and rule evaluations) the Application Load Balancer processes your traffic.

Q: Will I be billed on Classic Load Balancers by LCU?

A: No, Classic Load Balancers will continue to be billed for bandwidth and hourly usage.

Q: How do I know the number of LCUs an Application Load Balancer is using?

A: We expose the usage of all four dimensions that constitute an LCU via CloudWatch.

Q: Will I be billed on all the dimensions in an LCU?

A: No. The number of LCUs per hour will be determined based on maximum resource consumed amongst the four dimensions that constitutes a LCU.

Q: Will I be billed on partial LCUs?

A: Yes.

Q: Is a free tier offered on an Application Load Balancer for new AWS accounts?

A: Yes. For new AWS accounts, a free tier for an Application Load Balancer offers 750 hours and 15 LCUs. This free tier offer is only available to new AWS customers, and is available for 12 months following your AWS sign-up date

Q: Can I use a combination of Application Load Balancer and Classic Load Balancer as part of my free tier?

A: Yes. You can use both Classic and Application Load Balancers for 15GB and 15 LCUs respectively. The 750 load balancer hours are shared between both Classic and Application Load Balancers.

Q: What are rule evaluations?

A: Rule evaluations are defined as the product of number of rules processed and the request rate averaged over an hour.

Q: How does the LCU billing work with different certificate types and key sizes?

A: Certificate key size affects only the number of new connections per second in the LCU computation for billing. The following table lists the value of this dimension for different key sizes for RSA and ECDSA certificates.

RSA certificates

Key Size	**<=2K**	**<=4K**	**<=8K**	**>8K**
New connections/sec	25	5	1	0.25

ECDSA Certificates

Key Size	**<=256**	**<=384**	**<=521**	**>521**
New connections/sec	25	5	1	0.25

Q: Am I charged for regional AWS data-transfer for cross-zone load balancing in Application Load Balancer?

A: No. Since cross-zone load balancing is always on with Application Load Balancer, you are not charged for this type of regional data transfer.

Q: Is user authentication in Application Load Balancer charged separately?

A: No. There is no separate charge for enabling the authentication functionality in Application Load Balancer. When using Amazon Cognito with Application Load Balancer, Amazon Cognito pricing will apply.

Q: How do you charge for Application Load Balancer usage with Lambda targets?

A: You are charged as usual for each hour or partial hour that an Application Load Balancer is running and the number of Load Balancer Capacity Units (LCU) used per hour. For Lambda targets, each LCU offers 0.4 GB processed bytes per hour, 25 new connections per second, 3,000 active connections per minute and 1000 rule evaluations per second. For the processed bytes dimension, each LCU provides 0.4 GB per hour for Lambda targets versus 1GB per hour for all other target types like EC2 instances, containers and IP addresses. Please note that usual AWS Lambda charges apply to Lambda invocations by Application Load Balancer.

Q: How can I know the bytes processed by Lambda targets versus bytes processed by other targets (EC2, containers, and on-premises servers)?

A: Applications Load Balancers emit two new CloudWatch metrics. LambdaTargetProcessedBytes metric indicates the bytes processed by Lambda targets and the StandardProcessedBytes metric indicates bytes processed by all other target types.

Network Load Balancer

Q: Can I create a TCP or UDP (Layer 4) listener for my Network Load Balancer?

A: Yes. Network Load Balancers support both TCP, UDP, and TCP+UDP (Layer 4) listeners, as well as TLS listeners.

Q: What are the key features available with the Network Load Balancer?

A: Network Load Balancer provides both TCP and UDP (Layer 4) load balancing. It is architected to handle millions of requests/sec, sudden volatile traffic patterns and provides extremely low latencies. In addition Network Load Balancer also supports TLS termination, preserves the source IP of the clients, and provides stable IP support and Zonal isolation. It also supports long-running connections that are very useful for WebSocket type applications.

Q: Can Network Load Balancer process both TCP and UDP protocol traffic on the same port?

A: Yes. To achieve this, you can use a TCP+UDP listener. For example, for a DNS services using both TCP and UDP you can create a TCP+UDP listener on port 53, and the load balancer will process traffic for both UDP and TCP requests on that port. You must associate a TCP+UDP listener with a TCP+UDP target group.

Q: How does Network Load Balancer compare to what I get with the TCP listener on a Classic Load Balancer?

A: Network Load Balancer preserves the source IP of the client which in the Classic Load Balancer is not preserved. Customers can use proxy protocol with Classic Load Balancer to get the source IP. Network Load Balancer automatically provides a static IP per Availability Zone to the load balancer and also enables assigning an Elastic IP to the load balancer per Availability Zone. This is not supported with Classic Load Balancer.

Q: Can I migrate to Network Load Balancer from Classic Load Balancer?

A: Yes. You can migrate to Network Load Balancer from Classic Load Balancer using one of the options listed in this document.

Q: Are there limits on the resources for my Network Load Balancer?

A: Yes, please refer to Network Load Balancer limits documentation for more information.

Q: Can I use the AWS Management Console to set up my Network Load Balancer?

A: Yes, you can use the AWS Management Console, AWS CLI, or the API to set up a Network Load Balancer.

Q: Can I use the existing API for Classic Load Balancers for my Network Load Balancers?

A: No. To create a Classic Load Balancer, use the 2012-06-01 API. To create a Network Load Balancer or an Application Load Balancer, use the 2015-12-01 API.

Q: Can I create my Network Load Balancer in a single Availability Zone?

A: Yes, you can create your Network Load Balancer in a single availability zone by providing a single subnet when you create the load balancer.

Q: Does Network Load Balancer support DNS regional and zonal fail-over?

A: Yes, you can use Amazon Route 53 health checking and DNS failover features to enhance the availability of the applications running behind Network Load Balancers. Using Route 53 DNS failover, you can run applications in multiple AWS Availability zones and designate alternate load balancers for failover across regions. In the event that you have your Network Load Balancer configured for multi-AZ, if there are no healthy EC2 instances registered with the load balancer for that Availability Zone or if the load balancer nodes in a given zone are unhealthy, then R-53 will fail away to alternate load balancer nodes in other healthy availability zones.

Q: Can I have a Network Load Balancer with a mix of ELB-provided IPs and Elastic IPs or assigned private IPs?

A: No. A Network Load Balancer's addresses must be completely controlled by you, or completely controlled by ELB. This is to ensure that when using Elastic IPs with a Network Load Balancer, all addresses known to your clients do not change.

Q: Can I assign more than one EIP to my Network Load Balancer in each subnet?

A: No. For each associated subnet that a Network Load Balancer is in, the Network Load Balancer can only support a single public/internet facing IP address.

Q: If I remove/delete a Network Load Balancer what will happen to the Elastic IP addresses that were associated with it?

A: The Elastic IP Addresses that were associated with your load balancer will be returned to your allocated pool and made available for future use.

Q: Does Network Load Balancer support internal load balancers?

A: Network Load Balancer can be set-up as an internet-facing load balancer or an internal load balancer similar to what is possible with Application Load Balancer and Classic Load Balancer.

Q: Can the internal Network Load balancer support more than one private IP in each subnet?

A: No. For each associated subnet that a load balancer is in, the Network Load Balancer can only support a single private IP.

Q: Can I set up Websockets with my Network Load Balancer?

A: Yes, configure TCP listeners that route the traffic to the targets that implement WebSockets protocol (https://tools.ietf.org/html/rfc6455). Because WebSockets is a layer 7 protocol and Network Load Balancer is operating at layer 4, no special handling exists in Network Load Balancer for WebSockets or other higher level protocols.

Q: Can I load balance to any arbitrary IP address?

A: Yes. You can use any IP address from the load balancer's VPC CIDR for targets within load balancer's VPC and any IP address from RFC 1918 ranges (10.0.0.0/8, 172.16.0.0/12, and 192.168.0.0/16) or RFC 6598 range (100.64.0.0/10) for targets located outside the load balancer's VPC (EC2-Classic and on-premises locations reachable over AWS Direct Connect). Load balancing to IP address target type is supported for TCP listeners only, and is currently not supported for UDP listeners.

Q: Can I use Network Load Balancer to setup PrivateLink?

A: Yes, Network Load Balancers with TCP and TLS Listeners can be used to setup PrivateLink. You cannot setup PrivateLink with UDP listeners on Network Load Balancers.

Q: What is a UDP flow?

A: While UDP is connectionless, the load balancer maintains UDP flow state based on 5-tuple hash, making sure that packets sent in the same context are consistently forwarded to the same target. The flow is considered active as long as traffic is flowing and until the idle timeout is reached. Once the timeout threshold is reached, the load balancer will forget the affinity, and incoming UDP packet will be considered as a new flow and load-balanced to a new target.

Q: What is the idle timeout supported by Network Load Balancer?

A: Network Load Balancer idle timeout for TCP connections is 350 seconds. The idle timeout for UDP flows is 120 seconds.

Q: What benefit will I get by targeting containers behind a load balancer with IP addresses instead of instance IDs?

A: Each container on an instance can now have its own security group and does not need to share security rules with other containers. You can attach security groups to an ENI and each ENI on an instance can have a different security group. You can map a container to the IP address of a particular ENI to associate security group(s) per container. Load balancing using IP addresses also allows multiple containers running on an instance use the same port (say port 80). The ability to use the same port across containers allows containers on an instance to communicate with each other through well-known ports instead of random ports.

Q: How can I load balance applications distributed across a VPC and on-premises location?

A: There are various ways to achieve hybrid load balancing. If an application runs on targets distributed between a VPC and an on-premises location, you can add them to the same target group using their IP addresses. To migrate to AWS without impacting your application, gradually add VPC targets to the target group and remove on-premises targets from the target group. You can also use separate load balancers for VPC and on-premises targets and use DNS weighting to achieve weighted load balancing between VPC and on-premises targets.

Q: How can I load balance to EC2-Classic instances?

A: You cannot load balance to EC2-Classic Instances when registering their Instance IDs as targets. However if you link these EC2-Classic instances to the load balancer's VPC using ClassicLink and use the private IPs of these EC2-Classic instances as targets, then you can load balance to the EC2-Classic instances. If you are using EC2 Classic instances today with a Classic Load Balancer, you can easily migrate to a Network Load Balancer.

Q: How do I enable cross-zone load balancing in Network Load Balancer?

A: You can enable cross-zone loading balancing only after creating your Network Load Balancer. You achieve this by editing the load balancing attributes section and then by selecting the cross-zone load balancing support checkbox.

Q: Am I charged for regional AWS data-transfer when I enable cross-zone load balancing in Network Load Balancer?

A: Yes, you will be charged for regional data transfer between Availability Zones with Network Load Balancer when cross-zone load balancing is enabled. Check the charges in the data-transfer section at Amazon EC2 On-Demand Pricing page.

Q: Is there any impact of cross-zone load balancing on Network Load Balancer limits?

A: Yes. Network Load Balancer currently supports 200 targets per Availability Zone. For example, if you are in 2 Availability-Zones, you can have up to 400 targets registered with Network Load Balancer. If cross-zone load balancing is on, then the maximum targets reduces from 200 per Availability Zone to 200 per load balancer. So, in the example above when cross-zone load balancing is on, even though your load balancer is in 2 Availability Zones, you are limited to 200 targets that can be registered to the load balancer.

Q: Does Network Load Balancer support TLS termination?

A: Yes, you can terminate TLS connections on the Network Load Balancer. You must install an SSL certificate on your load balancer. The load balancer uses this certificate to terminate the connection and then decrypt requests from clients before sending them to targets.

Q: Is source IP is preserved when terminating TLS on Network Load Balancer?

A: Source IP continues to be preserved even if you terminate TLS on the Network Load Balancer.

Q: What are the steps to get a SSL certificate?

A: You can either use AWS Certificate Manager to provision an SSL/TLS certificate or you can obtain the certificate from other sources by creating the certificate request, getting the certificate request signed by a CA, and then uploading the certificate either using AWS Certification Manager (ACM) or the AWS Identity and Access Management (IAM) service.

Q: How can I enable Server Name Indication (SNI) for my Network Load Balancer?

A: SNI is automatically enabled when you associate more than one TLS certificate with the same secure listener on a load balancer. Similarly, SNI mode for a secure listener is automatically disabled when you have only one certificate associated to a secure listener.

Q: How does the Network Load Balancer integrate with AWS Certificate Manager (ACM) or Identity Access Manager (IAM)?

A: Network Load Balancer is integrated with AWS Certificate Management (ACM). Integration with ACM makes it very simple to bind a certificate to the load balancer thereby making the entire SSL offload process very easy. Purchasing, uploading, and renewing SSL/TLS certificates is a time-consuming manual and complex process. With ACM integration with Network Load Balancer, this whole process has been shortened to simply requesting a trusted SSL/TLS certificate and selecting the ACM certificate to provision it with the load balancer. Once you create a Network Load balancer, you can now configure a TLS listener and then you have an option to select a certificate from either ACM or Identity Access Manager (IAM). This experience is similar to what you have in Application Load Balancer or Classic Load Balancer.

Q: Is back-end server authentication supported with Network Load Balancer?

A: No, only encryption is supported to the back-ends with Network Load Balancer.

Q: What are the certificate types supported by Network Load Balancer?

A: Network Load Balancer only supports RSA certificates with 2K key size. We currently do not support RSA certificate key sizes greater than 2K or ECDSA certificates on the Network Load Balancer.

Q: In which AWS Regions is TLS Termination on Network Load Balancer supported?

A: You can use TLS Termination on Network Load Balancer in US East (N. Virginia), US East (Ohio), US West (Northern California), US West (Oregon), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), EU (Frankfurt), EU (Ireland), EU (London), EU (Paris), South America (São Paulo), and GovCloud (US-West) AWS Regions.

Network Load Balancer Pricing FAQs

Q: How does Network Load Balancer pricing work?

A: You are charged for each hour or partial hour that a Network Load Balancer is running and the number of Load Balancer Capacity Units (LCU) used by Network Load Balancer per hour.

Q: What is a Load Balancer Capacity Unit (LCU)?

A: An LCU is a new metric for determining how you pay for a Network Load Balancer. An LCU defines the maximum resource consumed in any one of the dimensions (new connections/flows, active connections/flows, and bandwidth) the Network Load Balancer processes your traffic.

Q: What is the LCU metrics for TCP traffic on Network Load Balancer?

A: The LCU metrics for the TCP traffic is as follows:

- 800 new TCP connections per second.
- 100,000 active TCP connections (sampled per minute).
- 1 GB per hour for EC2 instances, containers and IP addresses as targets.

Q: What is the LCU metrics for UDP traffic on Network Load Balancer?

A: The LCU metrics for the UDP traffic is as follows:

- 400 new flows per second.
- 50,000 active UDP flows (sampled per minute).
- 1 GB per hour for EC2 instances, containers and IP addresses as targets.

Q: What is the LCU metrics for TLS traffic on Network Load Balancer?

A: The LCU metrics for the TLS traffic is as follows:

- 50 new TLS connections per second.
- 3,000 active TLS connections (sampled per minute).
- 1 GB per hour for EC2 instances, containers and IP addresses as targets.

Q: Will I be billed on all the dimensions (Processed Bytes, New Flows and Active Flows)?

A: No, for each protocol you are charged only on one of the three dimensions (the highest for the hour).

Q: Is new connections/flows per sec same as requests/sec?

A: No. Multiple requests can be sent in a single connection.

Q: Will I be billed on Classic Load Balancers by LCU?

A: No. Classic Load Balancers will continue to be billed for bandwidth and hourly charge.

Q: How do I know the number of LCUs a Network Load Balancer is using?

A: We will expose the usage of all three dimensions that constitutes a LCU via Amazon CloudWatch.

Q: Will I be billed on all the dimensions in an LCU?

A: No. The number of LCUs per hour will be determined based on maximum resource consumed amongst the three dimensions that constitutes a LCU.

Q: Will I be billed on partial LCUs?

A: Yes.

Q: Is a free tier offered on a Network Load Balancer for new AWS accounts?

A: Yes. For new AWS accounts, a free tier for a Network Load Balancer offers 750 hours and 15 LCUs. This free tier offer is only available to new AWS customers, and is available for 12 months following your AWS sign-up date.

Q: Can I use a combination of Network Load Balancer, Application Load Balancer and Classic Load Balancer as part of my free tier?

A: Yes. You can use Application and Network each for 15 LCUs and Classic for 15 GB respectively. The 750 load balancer hours are shared between Application, Network and

Classic Load Balancers.

Classic Load Balancer

Q: Which operating systems does the Classic Load Balancer support?

A: The Classic Load Balancer supports Amazon EC2 instances with any operating system currently supported by the Amazon EC2 service.

Q: Which protocols does the Classic Load Balancer support?

A: The Classic Load Balancer supports load balancing of applications using HTTP, HTTPS (Secure HTTP), SSL (Secure TCP) and TCP protocols.

Q: What TCP ports can I load balance?

A: You can perform load balancing for the following TCP ports:

- [EC2-VPC] 1-65535
- [EC2-Classic] 25, 80, 443, 465, 587, 1024-65535

Q: Does the Classic Load Balancer support IPv6 traffic?

A: Yes. Each Classic Load Balancer has an associated IPv4, IPv6, and dualstack (both IPv4 and IPv6) DNS name. IPv6 is not supported in VPC. You can use an Application Load Balancer for native IPv6 support in VPC.

Q: Can I configure my Amazon EC2 instances to only accept traffic from Classic Load Balancers?

A: Yes.

Q: Can I configure a security group for the front-end of Classic Load Balancers?

A: If you are using Amazon Virtual Private Cloud, you can configure security groups for the front-end of your Classic Load Balancers.

Q: Can I use a single Classic Load Balancer for handling HTTP and HTTPS requests?

A: Yes, you can map HTTP port 80 and HTTPS port 443 to a single Classic Load Balancer.

Q: How many connections will my load balanced Amazon EC2 instances need to accept from each Classic Load Balancer?

A: Classic Load Balancers do not cap the number of connections that they can attempt to establish with your load balanced Amazon EC2 instances. You can expect this number to scale with the number of concurrent HTTP, HTTPS, or SSL requests or the number of concurrent TCP connections that the Classic load balancers receive.

Q: Can I load balance Amazon EC2 instances launched using a Paid AMI?

A: You can load balance Amazon EC2 instances launched using a paid AMI from AWS Marketplace. However, Classic Load Balancers do not support instances launched using a paid AMI from Amazon DevPay site.

Q: Can I use Classic Load Balancers in Amazon Virtual Private Cloud?

A: Yes. See the Elastic Load Balancing web page.

Q: Can I get a history of Classic Load Balancer API calls made on my account for security analysis and operational troubleshooting purposes?

A: Yes. To receive a history of Classic Load Balancer API calls made on your account, simply turn on CloudTrail in the AWS Management Console.

Q: Do Classic Load Balancers support SSL termination?

A: Yes you can terminate SSL on Classic Load Balancers. You must install an SSL certificate on each load balancer. The load balancers use this certificate to terminate the connection and then decrypt requests from clients before sending them to the back-end instances.

Q: What are the steps to get a SSL certificate?

A: You can either use AWS Certificate Manager to provision a SSL/TLS certificate or you can obtain the certificate from other sources by creating the certificate request, getting the certificate request signed by a CA, and then uploading the certificate using the AWS Identity and Access Management (IAM) service.

Q: How do Classic Load Balancers integrate with AWS Certificate Manager (ACM)?

A: Classic Load Balancers are now integrated with AWS Certificate Management (ACM). Integration with ACM makes it very simple to bind a certificate to each load balancer thereby making the entire SSL offload process very easy. Typically purchasing, uploading, and renewing SSL/TLS certificates is a time-consuming manual and complex process. With

ACM integrated with Classic Load Balancers, this whole process has been shortened to simply requesting a trusted SSL/TLS certificate and selecting the ACM certificate to provision it with each load balancer.

Q: How do I enable cross-zone load balancing in Classic Load Balancer?

A: You can enable cross-zone load balancing using the console, the AWS CLI, or an AWS SDK. See Cross-Zone Load Balancing documentation for more details.

Q: Am I charged for regional AWS data-transfer when I enable cross-zone load balancing in Classic Load Balancer?

A: No, you are not charged for regional data transfer between Availability Zones when you enable cross-zone load balancing for your Classic Load Balancer.

Amazon Braket FAQs

General

Q: What is Amazon Braket?

Amazon Braket is a fully managed service that helps you get started with quantum computing by providing a development environment to explore and design quantum algorithms, test them on simulated quantum computers, and run them on your choice of different quantum hardware technologies.

Q: Why should our company be thinking about quantum computing today?

Quantum computing is still an early stage technology, and designing useful quantum applications requires new skills and potentially radically different approaches to problem solving. Building this expertise will take time and requires access to quantum technologies and programming tools. Amazon Braket and the Amazon Quantum Solutions Lab help organizations assess the state of current technologies, identify how they might impact their business, and prepare for the future.

Q: What can I do with Amazon Braket?

Amazon Braket lets you design your own quantum algorithms from scratch or choose from a set of pre-built algorithms. Once you define your algorithm, Amazon Braket provides a fully managed simulation service to help troubleshoot and verify your implementation. When you are ready, you can run your algorithm on your choice of different quantum computers, including gate based and quantum annealing superconductors, and ion trap hardware. To make it easier to develop hybrid algorithms that combine classical and quantum tasks, Amazon Braket helps manage classical compute resources and establish low-latency connections to the quantum hardware. Once your tests are complete, you will be automatically notified and your results will be stored in

Amazon S3. Amazon Braket publishes event logs and performance metrics such as completion status and execution time to Amazon CloudWatch.

Q: What is a hybrid quantum/classical algorithm?

Hybrid quantum/classical algorithms combine quantum operations with optimization processes and other tasks running on classical compute instances. This allows you to create iterative systems that help mitigate the effect of errors inherent in todays’ quantum computing systems. Amazon Braket supports the execution of hybrid algorithms as fully managed jobs, orchestrating the necessary resources to maximize efficiency and reduce cost.

Q: Why do I need to simulate my algorithm?

Software-based simulators running on classical hardware can accelerate quantum and hybrid algorithm development, saving cost and making it easy to troubleshoot code and optimize designs prior to running them on hardware-based quantum computers.

Q: Can quantum computers be used to crack encryption?

No, the performance of current quantum computers is significantly below the level required to run algorithms that could factor encryption keys and enable an attacker to decrypt encrypted data. However, it is important to start planning for the future and to familiarize yourself with the various efforts to develop and standardize quantum-safe cryptographic algorithms and key exchange protocols. AWS is an active participant in this field and has already implemented candidate cipher suites and made them available to customers as hybrid key exchange schemes within our own implementation of TLS. To learn more read our blog here.

Build

Q: What is the Amazon Braket development environment?

The Amazon Braket console provides a development environment that includes learning materials, fully managed notebooks to build and edit quantum algorithms, a technology agnostic developer framework for defining hybrid algorithms, and a choice of classical simulators for testing your designs.

Q: What is the Amazon Braket framework?

Amazon Braket provides a technology agnostic developer framework so you don’t have to code against different quantum programming environments. As different quantum technologies emerge you can be confident that your development experience will be consistent and that your work is compatible across technologies.

Q: How am I charged for using the Amazon Braket development environment?

With Amazon Braket, you pay only for what you use. For designing and building algorithms using managed notebooks, testing using classical simulators, running jobs on quantum and classical hardware, and training and deploying quantum algorithms you are charged only for the time that you are using those resources. Public pricing will be available when the service is Generally Available.

Test

Q: What is the Amazon Braket simulator?

Amazon Braket offers a choice of different simulators that includes both Schrödinger and tensor network based classical simulators for testing quantum and hybrid algorithms. With a simple selection you can choose the approach that is best suits your requirements. For quick validation of circuit designs you can run basic simulations directly in your notebook. For larger designs you can perform on-demand simulations using the Amazon Braket service. For longer running, more complex algorithms Amazon Braket orchestrates Amazon EC2 clusters and GPU resources to run fully managed, high-performance simulation tasks.

Q: Why are there different types of simulators?

The efficient simulation of quantum operations using classical computers is an active area of research with different simulation methods suiting different circuit designs and different quantum technologies and processes. With new advances in the field, we plan to expand the range of available simulators to help ensure that you can choose between these leading-edge tools, even as we enter the regime where quantum hardware can no longer be fully simulated.

Q: What can I do with Amazon Braket simulation?

The Amazon Braket simulators enable you to test your quantum algorithms at a fraction of the cost of using quantum hardware and, in some cases, without having to wait to access specific quantum machines. Simulation is a convenient way to quickly debug quantum circuits and to troubleshoot and optimize hybrid algorithms before progressing to run them on your choice of quantum hardware. There is a choice of simulators to suit the work you are doing.

Q: What simulations does it support?

Amazon Braket offers a choice of different simulators that includes both Schrödinger and tensor network based classical simulators. The simulators can be used for testing individual circuit designs or as part of hybrid quantum/classical algorithms.

Q: What instance types are supported?

Amazon Braket runs simulations as fully managed jobs, automatically determining the optimum compute instance type to run your simulation and managing those resources on your behalf. You don't have to understand the various AWS instance types or how well they support your simulation job, Amazon Braket handles that for you.

Q: What is a simulation job?

A simulation job is no different than a quantum job. Both are Amazon Braket API requests to execute a quantum operation. All requests specify which backend system should run the operation. The choice of backends includes the various simulators and quantum computers that are available through the service. By changing an API parameter you select whether your jobs run on a simulator or your choice of quantum hardware.

Q: How do I see the results?

After completion, you will be automatically notified and your results will be stored in Amazon S3. In addition to providing the results of the execution, Amazon Braket also publishes event logs and performance metrics such as completion status and execution time to Amazon CloudWatch.

Q: How am I charged for running the simulator?

With Amazon Braket, you pay only for the simulator resources that you use. Public pricing will be available when the service is Generally Available.

Run

Q: How do I run my application on a quantum machine?

Quantum jobs are defined as either a quantum algorithm, typically a circuit design, or other types of problem definition statements depending on the nature of the application. These can be defined and run from within a managed Jupyter notebook, or initiated via the Amazon Braket console. Alternatively, you can design algorithms using your preferred developer environment and submit jobs for execution directly using the Amazon Braket API.

Q: How do I choose which machine to run on?

Some types of quantum computers are particularly well suited to solving specific sets of problems, for example, quantum annealers are typically used to solve combinatorial optimization problems, whereas other types are designed to be universal quantum computers. There are many factors that determine which type of machine will meet your needs, such as qubit count, qubit fidelity (error rate), qubit connectivity, coherence time, and cost. Full specifications are provided in the Amazon Braket console, and for further guidance you can engage with the Amazon Quantum Solutions Lab.

Q: What quantum hardware providers do you support?

We currently offer access to gate based superconductor computers from Rigetti, quantum annealing superconductor computers from D-Wave, and ion trap computers from IonQ. Over time, we plan to add more choices to give you the opportunity to test a wider variety of technology types and other providers within those various categories.

Q: Do my quantum jobs run immediately or do I have to wait?

Quantum computing is still an early stage technology and capacity on current machines is limited. Different types of quantum computers have different operational characteristics. If your chosen quantum hardware is available, your job will run immediately, if not, your job will be queued until it becomes available and you will be notified when the job is complete.

Q: How do I run hybrid algorithms?

When you run hybrid algorithms you can chose to manage the classical components of the algorithm yourself and use Amazon Braket for just the quantum operations or you can run the entire hybrid algorithm as a fully managed job, in which case, we orchestrate the classical compute resources on your behalf, returning the result for the overall operation when it completes.

Q: How am I charged for running my quantum jobs?

With Amazon Braket, you pay only for the time that you use the quantum hardware you requested. Pricing varies according to the hardware provider that you chose to use. Public pricing will be available at when the service is Generally Available.

Q: How do I see the results?

After completion of your quantum job or hybrid algorithm execution, you will be automatically notified and your results will be stored in Amazon S3. In addition to providing the results of the execution, Amazon Braket also publishes event logs and performance metrics such as completion status and execution time to Amazon CloudWatch.

AWS RoboMaker FAQs

Cloud extensions for ROS

Q: What are RoboMaker cloud extensions for ROS?

A: RoboMaker provides cloud extensions for ROS so that you can offload to the cloud the more resource-intensive computing processes that are typically required for intelligent robotics applications and free up local compute resources. RoboMaker provides each of these cloud service extensions as open source ROS packages, so you can build functions on your robot by taking advantage of cloud APIs, all in a familiar software framework.

Q: What are the supported cloud extensions?

A: Currently supported cloud extensions are Amazon CloudWatch Logs, Amazon CloudWatch Metrics, Amazon Polly, Amazon Lex, Amazon Kinesis Videos Streams, and Amazon Rekognition.

Q: Am I charged for using RoboMaker cloud extensions?

A: RoboMaker cloud extensions are open source under Apache 2.0 license and free of charge. You will, however, be charged separately when you invoke corresponding services such as Amazon Rekognition or Amazon Kinesis Video Streams from these extensions. See the Pricing Page for details.

Development environment

Q: What is RoboMaker development environment?

A: RoboMaker development environment is a customized environment in AWS Cloud9 for robotics development. This environment comes with ROS pre-installed and includes sample applications. This environment is also integrated with other RoboMaker capabilities such as simulation so that you can use these capabilities from the interface of the development environment.

Q: What instance types does RoboMaker development environment support?

A: RoboMaker development environment supports all AWS Cloud9 instance types except for T2 instance, which does not have sufficient resource to run ROS.

Q: Am I charged for using RoboMaker development environment?

A: You are billed for the underlying EC2 instance and EBS storage of the RoboMaker development environment and there is no additional charge on top. See the Pricing Page for details.

Simulation

Q: What is RoboMaker simulation?

A: RoboMaker simulation is a fully managed service that enables you to run simulation jobs without provisioning or managing any infrastructure.

Q: What can I do with RoboMaker simulation?

A: You can use RoboMaker simulation for various use cases. For example, you can use RoboMaker simulation to generate synthetic image or lidar data for algorithm development or testing, to train machine learning models, for regression testing robot applications, and as a testing tool during application development.

Q: What engines does RoboMaker simulation support?

A: The default simulation engine in RoboMaker is the open source Gazebo engine. The default physics engine is ODE (Open Dynamics Engine). The default rendering engine is OGRE (Object-Oriented Graphics Rendering Engine).

Q: What tools does RoboMaker simulation support?

A: RoboMaker simulation supports Gazebo client for interacting with a running simulation job, rviz for visualizing sensor data, rqt for running various GUI tools, and command line for interacting with the running robot application.

Q: What logging and monitoring features does RoboMaker simulation have?

A: RoboMaker simulation is integrated with Amazon CloudWatch Metrics so you can monitor service metrics such as real-time factors from the CloudWatch console. In addition, you can leverage the Amazon CloudWatch Metrics package provided from RoboMaker cloud extensions in your robot application to emit real-time metrics such as robot battery level, velocity, and collision during a simulation job run.

RoboMaker simulation is also integrated with Amazon CloudWatch Logs. You can have stdout statements in your robot application and RoboMaker simulation will log these messages to your CloudWatch Log group during a simulation run. This feature helps you debug your robot application during application development.

You can enable RoboMaker simulation to log ros bag during a simulation job run. The log will be delivered to your S3 bucket. You can use ros bag to analyze and debug message passing of the robot application.

Q: What is a simulation job?

A: Simulation job is the main entity in RoboMaker simulation. You run your simulation workloads on RoboMaker by creating a simulation job.

Q: What is a robot application?

A: A robot application refers to the ROS based application that you will run on a physical robot. You need to build your robot application source code to X86 architecture in order to work with RoboMaker simulation.

Q: What is a simulation application?

A: A simulation application includes a 3D simulation world and Gazebo plugins that control the movement of a robot in a simulation world. The default format of Gazebo simulation world is .sdf. Similar to robot application, you need to build your simulation application source code to X86 architecture in order to work with RoboMaker simulation.

Q: What is application versioning?

A: RoboMaker supports versioning of robot applications and simulation applications so you can control which implementation your robots and simulations use. A version is a numbered snapshot of your robot application or simulation application you can create for use in different parts of your development workflow, such as development, beta deployment, and production.

Q: Can I delete a particular version of a robot application or simulation application?

A: Yes, you can delete a particular version of an application or all the versions of an application.

Q: Does RoboMaker store my robot application and simulation application?

A: No, RoboMaker does not store your robot application or simulation application. You will upload your applications to your S3 bucket and refer to the S3 object path during robot application and simulation application creation.

Q: What is the failure behavior of a simulation job?

A: You can configure the failure behavior of your simulation job to be fail or continue. In fail mode, your simulation job will fail and the underlying infrastructure will be released if there is any failure during the simulation job run. In continue mode, your simulation job will fail but the underlying infrastructure will be kept if there is any failure during the simulation job run. The continue mode allows you to further troubleshoot a simulation job failure by using the command line tool. Note that RoboMaker simulation charges will continue to incur in continue mode until you terminate the job.

Q: When do I restart a simulation job?

A: Restarting a simulation job can be used for quick iterative development. Restarting a simulation job is faster than creating a new simulation job because it doesn't require resource provisioning and configuration behind the scenes. During robotics application development, you can use the restart simulation job feature to quickly test your code as you iterate through code changes.

Q: How is restarting a simulation job different from cloning a simulation job?

A: Restarting a simulation job is used for quick iteration test and the updated simulation job runs on the same infrastructure resources provisioned and managed by RoboMaker. Cloning a simulation job uses the same configurations of the existing job to create a new job. It triggers infrastructure resource provisioning and configuration behind the scenes, which might take a few minutes.

Q: Why do I need to provide an IAM role in a simulation job?

A: The IAM role provides RoboMaker simulation access to the resources you specified in your simulation job. For example, your S3 bucket for RoboMaker to write simulation logs into.

Q: What is simulation duration?

A: Simulation duration defines how long the simulation job should run until it gets terminated. The duration is based on wall clock time versus simulation time.

Q: Does my simulation job run in real-time?

A: RoboMaker simulation attempts to run your simulation job as fast as possible. You can also use the "real time update rate" and "max step size" settings in your Gazebo world configuration to control how fast the simulation job should run.

Q: Does RoboMaker simulation charge based on wall clock time or simulation time?

A: Your simulation job can run faster, slower than real-time, or at real-time. RoboMaker simulation charges based on wall clock time so the speed of the simulation time doesn't impact how you are charged. See the Pricing Page for details.

Fleet management

Q: What is RoboMaker fleet management?

A: RoboMaker fleet management allows you to create robots, register robots to a fleet, and then deploy a robot application into a fleet.

Q: What minimum hardware specifications are required for using RoboMaker fleet management?

A: RoboMaker fleet management is built on top of AWS Greengrass and has the same minimum hardware specifications requirement as AWS Greengrass. See the corresponding AWS Greengrass FAQ for details.

Q: How do I associate a physical robot with RoboMaker fleet management?

A: You will first create a robot from the RoboMaker console and then download corresponding certificates for that robot and AWS Greengrass-based agent. You will then follow steps to install the agent and certificate on your physical robot. Once installed and connected, your physical robot will be associated with the robot you created in RoboMaker console.

Q: What is a fleet?

A: A fleet is a group of robots. You can freely register or deregister existing robots to a fleet. A deployment job in RoboMaker Fleet Management is defined for a particular fleet and each robot can only belong to one fleet.

Q: How am I charged by using RoboMaker fleet management?

A: RoboMaker fleet management is integrated with AWS Greengrass and you are charged standard AWS Greengrass pricing. There is no additional charges for using RoboMaker fleet management. See the Pricing Page for details.

AWS Ground Station FAQs

General

Q. What is AWS Ground Station?
AWS Ground Station is a fully managed service that will enable customers to easily command, control, and downlink data from satellites. You can schedule access to AWS Ground Station antennas on a per-minute basis and pay only for your time used. When using AWS Ground Station, you can ingest data from the satellite, monitor the satellite health and status, and transmit commands to change the satellite's operations. Incoming data is streamed to an Amazon EC2 instance where it can then be stored or processed using other AWS services.

Q. What are the key benefits of AWS Ground Station?
AWS Ground Station is the industry's first satellite Ground Station as a Service located within the AWS Global Infrastructure footprint to offer on-demand, elastic access to ground station satellite antennas without long-term contracts. Traditionally, you have needed to make significant capital investments to build ground stations for these communications, including the costs for servers and storage to process incoming data. This limits your ability to respond quickly to new business opportunities or significant events (such as major weather events) and requires you to operate and maintain a global network of ground antennas. With AWS Ground Station, AWS manages the ground station infrastructure, so you can focus on innovating and rapidly experimenting with new applications that ingest satellite data and dynamically scale their server and storage use. And, you can easily integrate this data with other AWS services, either in the same region or in another AWS Region using Amazon's international, high-capacity backbone network.

Q. How do I get started using AWS Ground Station?
To get started, navigate to the AWS Ground Station console in the AWS Management Console. Here, you identify the satellites you need to communicate with and schedule "Contacts" with the satellite. Each Contact consists of a selected satellite, start and end time, and the ground location. You can review confirmed Contact times in your console and cancel or reschedule up to 15 minutes prior to the scheduled contact time. After scheduling a Contact, use the AWS Ground Station EC2 AMI to launch EC2 instances for the Contact: a Command instance to receive operational telemetry from the satellite and transmit changes to the satellite's planned future activities, and a Downlink instance to receive bulk mission data from the satellite. These instances will communicate with the AWS Ground

Station antenna gateway using an ENI connection that exists between the EC2 instances and the satellite antenna for the duration of the contact.

Q. Can I schedule time on an antenna in one location (e.g. Oregon) while I am working from another location (e.g. Sydney)?
Yes, AWS Ground Station is a global network of antenna systems that are available to users around the world. In fact, it will be common for many users to schedule time on antennas at every location in the global network.

Q: Where are Ground Station antennas located?

A: AWS Ground Station service is available immediately in US East (Ohio), US West (Oregon), Middle East (Bahrain) and Europe (Stockholm) and will expand to additional regions and locations in the coming year.

Features

Q. What is a Contact?
A Contact is a reservation to communicate with a specific satellite from an identified ground location between certain times.

Q. What happens when my scheduled Contact time arrives?
Prior to the scheduled contact time, use the Amazon EC2 AMI to launch the required instances to communicate with the satellite. Just prior to the scheduled contact, these instances will be able to establish connections to the AWS Ground Station antenna gateway over an ENI connection. Once the contact begins, the instances will begin sending and receiving data from the satellite.

Q. What happens if my requested Contact cannot be reserved?
In the event your preferred Contact cannot be granted (for example, due to existing reservations for antenna time at the chosen location), the AWS Ground Station Management Console will provide available alternatives for your review.

Q. How does AWS Ground Station make sure nobody else commands my satellite?
AWS Ground Station uses multiple measures to protect satellites from unauthorized contact. Prior to allowing contact with a satellite for the first time, AWS onboards and identifies the satellite owner and associates the satellite and satellite owner with a designated customer account. The customer is then able to use that account with AWS Ground Station to schedule satellite contacts or authorize other AWS accounts to schedule contacts. Prior to each contact, AWS validates that the contact will cause no radio interference. Access to the AWS Ground Station antenna gateway is limited to the

authorized EC2 instances for the period of the contact. Customers are in complete control of the encryption keys used to authorize and encrypt data to the satellite.

Q. What types of satellites can AWS Ground Station communicate with?
Low Earth Orbit (LEO), Non-Geostationary Earth Orbit (NGSO), and Medium Earth Orbit (MEO).

Q. What radio frequencies/bands can AWS Ground Station use to communicate with satellites?
AWS Ground Station antennas are capable of using common frequency bands to communicate with satellites, including: S- and X-.

Q. How fast is AWS Ground Station?
AWS Ground Station supports narrowband uplink speeds up to 54 MHz and downlink speeds up to 500MHz.

Q. How far in advance can I schedule Contacts?
Reserved Minute Contacts may be reserved up to 21 days in advance and rescheduled up to 1 day prior to a scheduled contact. On demand contacts can be scheduled as much as 7 days and as little as 15 minutes in advance and cannot be rescheduled.

Q. How does AWS Ground Station relate to/work with other AWS products?
It's easy to use AWS Ground Station with other AWS services to process and store satellite-originated data, either in the region local to each AWS Ground Station antenna or in another AWS region (using Amazon's international backbone network.) Data can be stored locally on EC2 instances using Amazon Elastic Block Store (EBS), in a shared filesystem using Amazon Elastic File System (EFS), or in Amazon S3. In S3, you can configure lifecycle policies to automatically migrate older, less-frequently-accessed data to less expensive storage classes, including S3 Infrequent Access and Amazon Glacier.
You can use Amazon Kinesis Data Streams to fully manage data ingestion and provide consistent APIs for integrating data analysis into your applications. It provides seamless integration with Amazon Rekognition, enabling automatic recognition of objects (such as cars or airplanes). You can also use Amazon SageMaker to build custom machine learning applications that apply to your data.

Q: What transmit and receive operational frequencies does Ground Station support?

A: Existing Ground Station antenna systems are capable of supporting the following frequencies:

- S-Band transmit from 2025 to 2120 MHz
- S-Band receive from 2200 to 2300 MHz

- X-Band receive from 7750 to 8400 MHz

Pricing

Q: What if I schedule a contact but then need to cancel the contact?
Reserved Minutes customers may cancel contacts up to 24 hours before contact start with no fee or penalty. If Reserved Minutes customers need to cancel contacts less than 24 hours before contact start time, there is a cancellation fee equal to the price of the contact. On Demand customers may cancel contacts up to 15 minutes before contact start with a cancellation fee equal to the price of that contact.

Q: When will I be charged for my Ground Station usage?
Customers will be charged after each contact completes. Customers will see the charge on their monthly AWS bill each month for their Reserved Minutes commitment plus any On Demand minutes used that month.

AWS IAM FAQs

General

Q: What is AWS Identity and Access Management (IAM)?
You can use AWS IAM to securely control individual and group access to your AWS resources. You can create and manage user identities ("IAM users") and grant permissions for those IAM users to access your resources. You can also grant permissions for users outside of AWS (federated users).

Q: How do I get started with IAM?
To start using IAM, you must subscribe to at least one of the AWS services that is integrated with IAM. You then can create and manage users, groups, and permissions via IAM APIs, the AWS CLI, or the IAM console, which gives you a point-and-click, web-based interface. You can also use the visual editor to create policies.

Q: What problems does IAM solve?
IAM makes it easy to provide multiple users secure access to your AWS resources. IAM enables you to:

- Manage IAM users and their access: You can create users in AWS's identity management system, assign users individual security credentials (such as access keys, passwords, multi-factor authentication devices), or request temporary security credentials to provide users access to AWS services and resources. You can specify permissions to control which operations a user can perform.
- Manage access for federated users: You can request security credentials with configurable expirations for users who you manage in your corporate directory, allowing you to provide your employees and applications secure access to resources in your AWS account without creating an IAM user account for them. You specify the permissions for these security credentials to control which operations a user can perform.

Q: Who can use IAM?
Any AWS customer can use IAM. The service is offered at no additional charge. You will be charged only for the use of other AWS services by your users.

Q: What is a user?
A user is a unique identity recognized by AWS services and applications. Similar to a login user in an operating system like Windows or UNIX, a user has a unique name and can

identify itself using familiar security credentials such as a password or access key. A user can be an individual, system, or application requiring access to AWS services. IAM supports users (referred to as "IAM users") managed in AWS's identity management system, and it also enables you to grant access to AWS resources for users managed outside of AWS in your corporate directory (referred to as "federated users").

Q: What can a user do?
A user can place requests to web services such as Amazon S3 and Amazon EC2. A user's ability to access web service APIs is under the control and responsibility of the AWS account under which it is defined. You can permit a user to access any or all of the AWS services that have been integrated with IAM and to which the AWS account has subscribed. If permitted, a user has access to all of the resources under the AWS account. In addition, if the AWS account has access to resources from a different AWS account, its users may be able to access data under those AWS accounts. Any AWS resources created by a user are under control of and paid for by its AWS account. A user cannot independently subscribe to AWS services or control resources.

Q: How do users call AWS services?
Users can make requests to AWS services using security credentials. Explicit permissions govern a user's ability to call AWS services. By default, users have no ability to call service APIs on behalf of the account.

IAM user management

Q: How are IAM users managed?
IAM supports multiple methods to:

- Create and manage IAM users.
- Create and manage IAM groups.
- Manage users' security credentials.
- Create and manage policies to grant access to AWS services and resources.

You can create and manage users, groups, and policies by using IAM APIs, the AWS CLI, or the IAM console. You also can use the visual editor and the IAM policy simulator to create and test policies.

Q: What is a group?
A group is a collection of IAM users. Manage group membership as a simple list:

- Add users to or remove them from a group.
- A user can belong to multiple groups.

- Groups cannot belong to other groups.
- Groups can be granted permissions using access control policies. This makes it easier to manage permissions for a collection of users, rather than having to manage permissions for each individual user.
- Groups do not have security credentials, and cannot access web services directly; they exist solely to make it easier to manage user permissions. For details, see Working with Groups and Users.

Q: What kinds of security credentials can IAM users have?
IAM users can have any combination of credentials that AWS supports, such as an AWS access key, X.509 certificate, SSH key, password for web app logins, or an MFA device. This allows users to interact with AWS in any manner that makes sense for them. An employee might have both an AWS access key and a password; a software system might have only an AWS access key to make programmatic calls; IAM users might have a private SSH key to access AWS CodeCommit repositories; and an outside contractor might have only an X.509 certificate to use the EC2 command-line interface. For details, see Temporary Security Credentials in the IAM documentation.

Q: Which AWS services support IAM users?
You can find the complete list of AWS services that support IAM users in the AWS Services That Work with IAM section of the IAM documentation. AWS plans to add support for other services over time.

Q: Can I enable and disable user access?
Yes. You can enable and disable an IAM user's access keys via the IAM APIs, AWS CLI, or IAM console. If you disable the access keys, the user cannot programmatically access AWS services.

Q: Who is able to manage users for an AWS account?
The AWS account holder can manage users, groups, security credentials, and permissions. In addition, you may grant permissions to individual users to place calls to IAM APIs in order to manage other users. For example, an administrator user may be created to manage users for a corporation—a recommended practice. When you grant a user permission to manage other users, they can do this via the IAM APIs, AWS CLI, or IAM console.

Q: Can I structure a collection of users in a hierarchical way, such as in LDAP?
Yes. You can organize users and groups under paths, similar to object paths in Amazon S3—for example /mycompany/division/project/joe.

Q: Can I define users regionally?
Not initially. Users are global entities, like an AWS account is today. No region is required to

be specified when you define user permissions. Users can use AWS services in any geographic region.

Q: How are MFA devices configured for IAM users?
You (the AWS account holder) can order multiple MFA devices. You can then assign these devices to individual IAM users via the IAM APIs, AWS CLI, or IAM console.

Q: What kind of key rotation is supported for IAM users?
User access keys and X.509 certificates can be rotated just as they are for an AWS account's root access identifiers. You can manage and rotate programmatically a user's access keys and X.509 certificates via the IAM APIs, AWS CLI, or IAM console.

Q: Can IAM users have individual EC2 SSH keys?
Not in the initial release. IAM does not affect EC2 SSH keys or Windows RDP certificates. This means that although each user has separate credentials for accessing web service APIs, they must share SSH keys that are common across the AWS account under which users have been defined.

Q: Where can I use my SSH keys?
Currently, IAM users can use their SSH keys only with AWS CodeCommit to access their repositories.

Q: Do IAM user names have to be email addresses?
No, but they can be. User names are just ASCII strings that are unique within a given AWS account. You can assign names using any naming convention you choose, including email addresses.

Q: Which character sets can I use for IAM user names?
You can only use ASCII characters for IAM entities.

Q: Are user attributes other than user name supported?
Not at this time.

Q: How are user passwords set?
You can set an initial password for an IAM user via the IAM console, AWS CLI, or IAM APIs. User passwords never appear in clear text after the initial provisioning, and are never displayed or returned via an API call. IAM users can manage their passwords via the **My Password** page in the IAM console. Users access this page by selecting the **Security Credentials** option from the drop-down list in the upper right corner of the AWS Management Console.

Q: Can I define a password policy for my user's passwords?
Yes, you can enforce strong passwords by requiring minimum length or at least one

number. You can also enforce automatic password expiration, prevent re-use of old passwords, and require a password reset upon the next AWS sign-in. For details, see Setting an Account Policy Password for IAM Users.

Q: Can I set usage quotas on IAM users?
No. All limits are on the AWS account as a whole. For example, if your AWS account has a limit of 20 Amazon EC2 instances, IAM users with EC2 permissions can start instances up to the limit. You cannot limit what an individual user can do.

IAM role management

Q: What is an IAM role?
An IAM role is an IAM entity that defines a set of permissions for making AWS service requests. IAM roles are not associated with a specific user or group. Instead, trusted entities assume roles, such as IAM users, applications, or AWS services such as EC2.

Q: What problems do IAM roles solve?
IAM roles allow you to delegate access with defined permissions to trusted entities without having to share long-term access keys. You can use IAM roles to delegate access to IAM users managed within your account, to IAM users under a different AWS account, or to an AWS service such as EC2.

Q: How do I get started with IAM roles?
You create a role in a way similar to how you create a user—name the role and attach a policy to it. For details, see Creating IAM Roles.

Q: How do I assume an IAM role?
You assume an IAM role by calling the AWS Security Token Service (STS) AssumeRole APIs (in other words, AssumeRole, AssumeRoleWithWebIdentity, and AssumeRoleWithSAML). These APIs return a set of temporary security credentials that applications can then use to sign requests to AWS service APIs.

Q: How many IAM roles can I assume?
There is no limit to the number of IAM roles you can assume, but you can only act as one IAM role when making requests to AWS services.

Q: Who can use IAM roles?
Any AWS customer can use IAM roles.

Q: How much do IAM roles cost?
IAM roles are free of charge. You will continue to pay for any resources a role in your AWS account consumes.

Q: How are IAM roles managed?
You can create and manage IAM roles via the IAM APIs, AWS CLI, or IAM console, which gives you a point-and-click, web-based interface.

Q: What is the difference between an IAM role and an IAM user?
An IAM user has permanent long-term credentials and is used to directly interact with AWS services. An IAM role does not have any credentials and cannot make direct requests to AWS services. IAM roles are meant to be assumed by authorized entities, such as IAM users, applications, or an AWS service such as EC2.

Q: When should I use an IAM user, IAM group, or IAM role?
An IAM user has permanent long-term credentials and is used to directly interact with AWS services. An IAM group is primarily a management convenience to manage the same set of permissions for a set of IAM users. An IAM role is an AWS Identity and Access Management (IAM) entity with permissions to make AWS service requests. IAM roles cannot make direct requests to AWS services; they are meant to be assumed by authorized entities, such as IAM users, applications, or AWS services such as EC2. Use IAM roles to delegate access within or between AWS accounts.

Q: Can I add an IAM role to an IAM group?
Not at this time.

Q: How many policies can I attach to an IAM role?
For inline policies: You can add as many inline policies as you want to a user, role, or group, but the total aggregate policy size (the sum size of all inline policies) per entity cannot exceed the following limits:

- User policy size cannot exceed 2,048 characters.
- Role policy size cannot exceed 10,240 characters.
- Group policy size cannot exceed 5,120 characters.

For managed policies: You can add up to 10 managed policies to a user, role, or group. The size of each managed policy cannot exceed 6,144 characters.

Q: How many IAM roles can I create?
You are limited to 1,000 IAM roles under your AWS account. If you need more roles, submit the IAM limit increase request form with your use case, and we will consider your request.

Q: To which services can my application make requests?
Your application can make requests to all AWS services that support role sessions.

Q: What is IAM roles for EC2 instances?
IAM roles for EC2 instances enables your applications running on EC2 to make requests to

AWS services such as Amazon S3, Amazon SQS, and Amazon SNS without you having to copy AWS access keys to every instance. For details, see IAM Roles for Amazon EC2.

Q: What are the features of IAM roles for EC2 instances?
IAM roles for EC2 instances provides the following features:

- AWS temporary security credentials to use when making requests from running EC2 instances to AWS services.
- Automatic rotation of the AWS temporary security credentials.
- Granular AWS service permissions for applications running on EC2 instances.

Q: What problem does IAM roles for EC2 instances solve?
IAM roles for EC2 instances simplifies management and deployment of AWS access keys to EC2 instances. Using this feature, you associate an IAM role with an instance. Then your EC2 instance provides the temporary security credentials to applications running on the instance, and the applications can use these credentials to make requests securely to the AWS service resources defined in the role.

Q: How do I get started with IAM roles for EC2 instances?
To understand how roles work with EC2 instances, you need to use the IAM console to create a role, launch an EC2 instance that uses that role, and then examine the running instance. You can examine the instance metadata to see how the role credentials are made available to an instance. You can also see how an application that runs on an instance can use the role. For more details, see How Do I Get Started?

Q: Can I use the same IAM role on multiple EC2 instances?
Yes.

Q: Can I change the IAM role on a running EC2 instance?
Yes. Although a role is usually assigned to an EC2 instance when you launch it, a role can also be assigned to an EC2 instance that is already running. To learn how to assign a role to a running instance, see IAM Roles for Amazon EC2. You can also change the permissions on the IAM role associated with a running instance, and the updated permissions take effect almost immediately.

Q: Can I associate an IAM role with an already running EC2 instance?
Yes. You can assign a role to an EC2 instance that is already running. To learn how to assign a role to an already running instance, see IAM Roles for Amazon EC2.

Q: Can I associate an IAM role with an Auto Scaling group?
Yes. You can add an IAM role as an additional parameter in an Auto Scaling launch configuration and create an Auto Scaling group with that launch configuration. All EC2

instances launched in an Auto Scaling group that is associated with an IAM role are launched with the role as an input parameter. For more details, see What Is Auto Scaling? in the Auto Scaling Developer Guide.

Q: Can I associate more than one IAM role with an EC2 instance?
No. You can only associate one IAM role with an EC2 instance at this time. This limit of one role per instance cannot be increased.

Q: What happens if I delete an IAM role that is associated with a running EC2 instance?
Any application running on the instance that is using the role will be denied access immediately.

Q: Can I control which IAM roles an IAM user can associate with an EC2 instance?
Yes. For details, see Permissions Required for Using Roles with Amazon EC2.

Q: Which permissions are required to launch EC2 instances with an IAM role?
You must grant an IAM user two distinct permissions to successfully launch EC2 instances with roles:

- Permission to launch EC2 instances.
- Permission to associate an IAM role with EC2 instances.

For details, see Permissions Required for Using Roles with Amazon EC2.

Q: Who can access the access keys on an EC2 instance?
Any local user on the instance can access the access keys associated with the IAM role.

Q: How do I use the IAM role with my application on the EC2 instance?
If you develop your application with the AWS SDK, the AWS SDK automatically uses the AWS access keys that have been made available on the EC2 instance. If you are not using the AWS SDK, you can retrieve the access keys from the EC2 instance metadata service. For details, see Using an IAM Role to Grant Permissions to Applications Running on Amazon EC2 Instances.

Q: How do I rotate the temporary security credentials on the EC2 instance?
The AWS temporary security credentials associated with an IAM role are automatically rotated multiple times a day. New temporary security credentials are made available no later than five minutes before the existing temporary security credentials expire.

Q: Can I use IAM roles for EC2 instances with any instance type or Amazon Machine Image?
Yes. IAM roles for EC2 instances also work in Amazon Virtual Private Cloud (VPC), with spot and reserved instances.

Q: What is a service-linked role?

A service-linked role is a type of role that links to an AWS service (also known as a linked service) such that only the linked service can assume the role. Using these roles, you can delegate permissions to AWS services to create and manage AWS resources on your behalf.

Q: Can I assume a service-linked role?

No. A service-linked role can be assumed only by the linked service. This is the reason why the trust policy of a service-linked role cannot be modified.

Q: Can I delete a service-linked role?

Yes. If you no longer want an AWS service to perform actions on your behalf, you can delete its service-linked role. Before you delete the role, you must delete all AWS resources that depend on the role. This step ensures that you do not inadvertently delete a role required for your AWS resources to function properly.

Q: How do I delete a service-linked role?

You can delete a service-linked role from the IAM console. Choose **Roles** in the navigation pane, choose the service-linked role that you want to delete, and choose **Delete role**. (**Note**: For Amazon Lex, you must use the Amazon Lex console to delete the service-linked role.)

Permissions

Q: How do permissions work?

Access control policies are attached to users, groups, and roles to assign permissions to AWS resources. By default, IAM users, groups, and roles have no permissions; users with sufficient permissions must use a policy to grant the desired permissions.

Q: How do I assign permissions using a policy?

To set permissions, you can create and attach policies using the AWS Management Console, the IAM API, or the AWS CLI. Users who have been granted the necessary permissions can create policies and assign them to IAM users, groups, and roles.

Q: What are managed policies?

Managed policies are IAM resources that express permissions using the IAM policy language. You can create, edit, and manage separately from the IAM users, groups, and roles to which they are attached. After you attach a managed policy to multiple IAM users, groups, or roles, you can update that policy in one place and the permissions automatically extend to all attached entities. Managed policies are managed either by you (these are called customer managed policies) or by AWS (these are called AWS managed policies). For more information about managed policies, see Managed Policies and Inline Policies.

Q: How do I create a customer managed policy?
You can use the visual editor or the JSON editor in the IAM console. The visual editor is a point-and-click editor that guides you through the process of granting permissions in a policy without requiring you to write the policy in JSON. You can create policies in JSON by using the CLI and SDK.

Q: How do I assign commonly used permissions?
AWS provides a set of commonly used permissions that you can attach to IAM users, groups, and roles in your account. These are called AWS managed policies. One example is read-only access for Amazon S3. When AWS updates these policies, the permissions are applied automatically to the users, groups, and roles to which the policy is attached. AWS managed policies automatically appear in the **Policies** section of the IAM console. When you assign permissions, you can use an AWS managed policy or you can create your own customer managed policy. Create a new policy based on an existing AWS managed policy, or define your own.

Q: How do group-based permissions work?
Use IAM groups to assign the same set of permissions to multiple IAM users. A user can also have individual permissions assigned to them. The two ways to attach permissions to users work together to set overall permissions.

Q: What is the difference between assigning permissions using IAM groups and assigning permissions using managed policies?
Use IAM groups to collect IAM users and define common permissions for those users. Use managed policies to share permissions across IAM users, groups, and roles. For example, if you want a group of users to be able to launch an Amazon EC2 instance, and you also want the role on that instance to have the same permissions as the users in the group, you can create a managed policy and assign it to the group of users and the role on the Amazon EC2 instance.

Q: How are IAM policies evaluated in conjunction with Amazon S3, Amazon SQS, Amazon SNS, and AWS KMS resource-based policies?
IAM policies are evaluated together with the service's resource-based policies. When a policy of any type grants access (without explicitly denying it), the action is allowed. For more information about the policy evaluation logic, see IAM Policy Evaluation Logic.

Q: Can I use a managed policy as a resource-based policy?
Managed policies can only be attached to IAM users, groups, or roles. You cannot use them as resource-based policies.

Q: How do I set granular permissions using policies?

Using policies, you can specify several layers of permission granularity. First, you can define specific AWS service actions you wish to allow or explicitly deny access to. Second, depending on the action, you can define specific AWS resources the actions can be performed on. Third, you can define conditions to specify when the policy is in effect (for example, if MFA is enabled or not).

Q: How can I easily remove unnecessary permissions?
To help you determine which permissions are needed, the IAM console now displays service last accessed data that shows the hour when an IAM entity (a user, group, or role) last accessed an AWS service. Knowing if and when an IAM entity last exercised a permission can help you remove unnecessary permissions and tighten your IAM policies with less effort.

Q: Can I grant permissions to access or change account-level information (for example, payment instrument, contact email address, and billing history)?
Yes, you can delegate the ability for an IAM user or a federated user to view AWS billing data and modify AWS account information. For more information about controlling access to your billing information, see Controlling Access.

Q: Who can create and manage access keys in an AWS account?
Only the AWS account owner can manage the access keys for the root account. The account owner and IAM users or roles that have been granted the necessary permissions can manage access keys for IAM users.

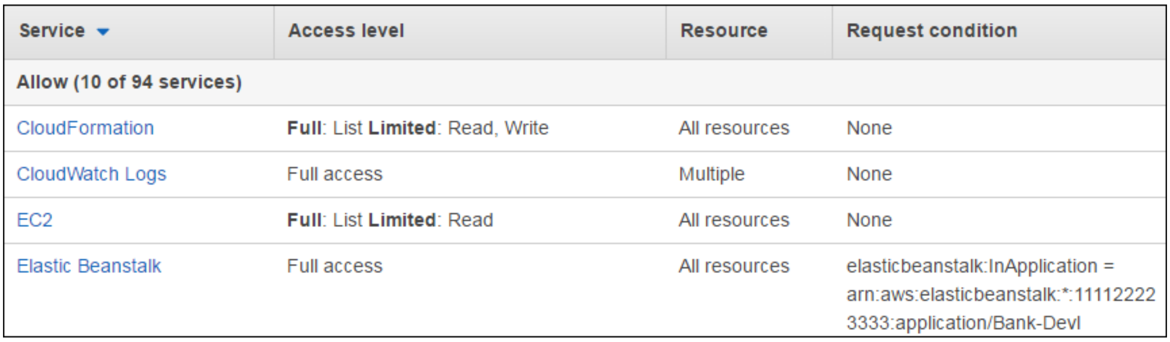
Q: Can I grant permissions to access AWS resources owned by another AWS account?
Yes. Using IAM roles, IAM users and federated users can access resources in another AWS account via the AWS Management Console, the AWS CLI, or the APIs. See Manage IAM Roles for more information.

Q: What does a policy look like?
The following policy grants access to add, update, and delete objects from a specific folder, example_folder, in a specific bucket, example_bucket.

```
{
 "Version":"2012-10-17",
 "Statement":[
  {
   "Effect":"Allow",
   "Action":[
    "s3:PutObject",
    "s3:GetObject",
    "s3:GetObjectVersion",
    "s3:DeleteObject",
    "s3:DeleteObjectVersion"
   ],

   "Resource":"arn:aws:s3:::example_bucket/example_folder/*"
  }
 ]
}
```

Q: What is a policy summary?
If you are using the IAM console and choose a policy, you will see a policy summary. A policy summary lists the access level, resources, and conditions for each service defined in a policy (see the following screenshot for an example). The access level (View, Read, Write, or Permissions management) is defined by actions granted for each service in the policy. You can view the policy in **JSON** by choosing the JSON button.



Service ▾	Access level	Resource	Request condition
Allow (10 of 94 services)			
CloudFormation	**Full**: List **Limited**: Read, Write	All resources	None
CloudWatch Logs	Full access	Multiple	None
EC2	**Full**: List **Limited**: Read	All resources	None
Elastic Beanstalk	Full access	All resources	elasticbeanstalk:InApplication = arn:aws:elasticbeanstalk:*:111122223333:application/Bank-Devl

Policy simulator

Q: What is the IAM policy simulator?
The IAM policy simulator is a tool to help you understand, test, and validate the effects of your access control policies.

Q: What can the policy simulator be used for?
You can use the policy simulator in several ways. You can test policy changes to ensure they have the desired effect before committing them to production. You can validate existing policies attached to users, groups, and roles to verify and troubleshoot permissions. You can also use the policy simulator to understand how IAM policies and resource-based policies work together to grant or deny access to AWS resources.

Q: Who can use the policy simulator?
The policy simulator is available to all AWS customers.

Q: How much does the policy simulator cost?
The policy simulator is available at no extra cost.

Q: How do I get started?
Go to https://policysim.aws.amazon.com, or click the link on the IAM console under "Additional Information." Specify a new policy or choose an existing set of policies from a user, group, or role that you'd like to evaluate. Then select a set of actions from the list of AWS services, provide any required information to simulate the access request, and run the simulation to determine whether the policy allows or denies permissions to the selected actions and resources. To learn more about the IAM policy simulator, watch our Getting Started video or see the documentation.

Q: What kinds of policies does the IAM policy simulator support?
The policy simulator supports testing of newly entered policies and existing policies attached to users, groups, or roles. In addition, you can simulate whether resource-level policies grant access to a particular resource for Amazon S3 buckets, Amazon Glacier vaults, Amazon SNS topics, and Amazon SQS queues. These are included in the simulation when an Amazon Resource Name (ARN) is specified in the **Resource** field in **Simulation Settings** for a service that supports resource policies.

Q: If I change a policy in the policy simulator, do those changes persist in production?
No. To apply changes to production, copy the policy that you've modified in the policy simulator and attach it to the desired IAM user, group, or role.

Q: Can I use the policy simulator programmatically?
Yes. You can use the policy simulator using the AWS SDKs or AWS CLI in addition to the policy simulator console. Use the iam:SimulatePrincipalPolicy API to programmatically test your existing IAM policies. To test the effects of new or updated policies that are not yet attached to a user, group, or role, call the iam:SimulateCustomPolicy API.

Signing in

Q: How does an IAM user sign in?
To sign in to the AWS Management Console as an IAM user, you must provide your account ID or account alias in addition to your user name and password. When your administrator created your IAM user in the console, they should have provided you with your user name and the URL to your account sign-in page. That URL includes your account ID or account alias.

https://My_AWS_Account_ID.signin.aws.amazon.com/console/

You can also sign in at the following general sign-in endpoint and type your account ID or account alias manually:

https://console.aws.amazon.com/

For convenience, the AWS sign-in page uses a browser cookie to remember the IAM user name and account information. The next time the user goes to any page in the AWS Management Console, the console uses the cookie to redirect the user to the account sign-in page.

Note: IAM users can still use the URL link provided to them by their administrator to sign in to the AWS Management Console.

Q: What is an AWS account alias?
The account alias is a name you define to make it more convenient to identify your account. You can create an alias using the IAM APIs, AWS Command Line Tools, or the IAM console. You can have one alias per AWS account.

Q: Which AWS sites can IAM users access?
IAM users can sign in to the following AWS sites:

- AWS Management Console
- AWS Forums
- AWS Support Center
- AWS Marketplace

Q: Can IAM users sign in to other Amazon.com properties with their credentials?
No. Users created with IAM are recognized only by AWS services and applications.

Q: Is there an authentication API to verify IAM user sign-ins?
No. There is no programmatic way to verify user sign-ins.

Q: Can users SSH to EC2 instances using their AWS user name and password?
No. User security credentials created with IAM are not supported for direct authentication to customer EC2 instances. Managing EC2 SSH credentials is the customer's responsibility within the EC2 console.

Temporary security credentials

Q: What are temporary security credentials?
Temporary security credentials consist of the AWS access key ID, secret access key, and security token. Temporary security credentials are valid for a specified duration and for a specific set of permissions. Temporary security credentials are sometimes simply referred to as tokens. Tokens can be requested for IAM users or for federated users you manage in your own corporate directory. For more information, see Common Scenarios for Temporary Credentials.

Q: What are the benefits of temporary security credentials?

Temporary security credentials allow you to:

- Extend your internal user directories to enable federation to AWS, enabling your employees and applications to securely access AWS service APIs without needing to create an AWS identity for them.
- Request temporary security credentials for an unlimited number of federated users.
- Configure the time period after which temporary security credentials expire, offering improved security when accessing AWS service APIs through mobile devices where there is a risk of losing the device.

Q: How can I request temporary security credentials for federated users?

You can call the GetFederationToken, AssumeRole, AssumeRoleWithSAML, or AssumeRoleWithWebIdentity STS APIs.

Q: How can IAM users request temporary security credentials for their own use?

IAM users can request temporary security credentials for their own use by calling the AWS STS GetSessionToken API. The default expiration for these temporary credentials is 12 hours; the minimum is 15 minutes, and the maximum is 36 hours.

You can also use temporary credentials with Multi-Factor Authentication (MFA)-Protected API Access.

Q: How can I use temporary security credentials to call AWS service APIs?

If you're making direct HTTPS API requests to AWS, you can sign those requests with the temporary security credentials that you get from AWS Security Token Service (AWS STS). To do this, do the following:

- Use the access key ID and secret access key that are provided with the temporary security credentials the same way you would use long-term credentials to sign a request. For more information about signing HTTPS API requests, see Signing AWS API Requests in the AWS General Reference.
- Use the session token that is provided with the temporary security credentials. Include the session token in the "x-amz-security-token" header. See the following example request.
  - For Amazon S3, via the "x-amz- security-token" HTTP header.
  - For other AWS services, via the SecurityToken parameter.

Q: Which AWS services accept temporary security credentials?

For a list of supported services, see AWS Services That Work with IAM.

Q: What is the maximum size of the access policy that I can specify when requesting temporary security credentials (either GetFederationToken or AssumeRole)?
The policy plaintext must be 2048 bytes or shorter. However, an internal conversion compresses it into a packed binary format with a separate limit.

Q: Can a temporary security credential be revoked prior to its expiration?
No. When requesting temporary credentials, we recommend the following:

- When creating temporary security credentials, set the expiration to a value that is appropriate for your application.
- Because root account permissions cannot be restricted, use an IAM user and not the root account for creating temporary security credentials. You can revoke permissions of the IAM user that issued the original call to request it. This action almost immediately revokes privileges for all temporary security credentials issued by that IAM user

Q: Can I reactivate or extend the expiration of temporary security credentials?
No. It is a good practice to actively check the expiration and request a new temporary security credential before the old one expires. This rotation process is automatically managed for you when temporary security credentials are used in roles for EC2 instances.

Q: Are temporary security credentials supported in all regions?
Customers can request tokens from AWS STS endpoints in all regions, including AWS GovCloud (US) and China (Beijing) regions. Temporary credentials from AWS GovCloud (US) and China (Beijing) can be used only in the region from which they originated. Temporary credentials requested from any other region such as US East (N. Virginia) or EU (Ireland) can be used in all regions except AWS GovCloud (US) and China (Beijing).

Q: Can I restrict the use of temporary security credentials to a region or a subset of regions?
No. You cannot restrict the temporary security credentials to a particular region or subset of regions, except the temporary security credentials from AWS GovCloud (US) and China (Beijing), which can be used only in the respective regions from which they originated.

Q: What do I need to do before I can start using an AWS STS endpoint?
AWS STS endpoints are active by default in all regions and you can start using them without any further actions.

Q: What happens if I try to use a regional AWS STS endpoint that has been deactivated for my AWS account?
If you attempt to use a regional AWS STS endpoint that has been deactivated for your AWS account, you will see an **AccessDenied** exception from AWS STS with the following

message: “AWS STS is not activated in this region for account: *AccountID*. Your account administrator can activate AWS STS in this region using the IAM console.”

Q: What permissions are required to activate or deactivate AWS STS regions from the Account Settings page?
Only users with at least iam:* permissions can activate or deactivate AWS STS regions from the **Account Settings** page in the IAM console. Note that the AWS STS endpoints in US East (N. Virginia), AWS GovCloud (US), and China (Beijing) regions are always active and cannot be deactivated.

Q: Can I use the API or CLI to activate or deactivate AWS STS regions?
No. There is no API or CLI support at this time to activate or deactivate AWS STS regions. We plan to provide API and CLI support in a future release.

Identity federation

Q: What is identity federation?
AWS Identity and Access Management (IAM) supports identity federation for delegated access to the AWS Management Console or AWS APIs. With identity federation, external identities are granted secure access to resources in your AWS account without having to create IAM users. These external identities can come from your corporate identity provider (such as Microsoft Active Directory or from the AWS Directory Service) or from a web identity provider (such as Amazon Cognito, Login with Amazon, Facebook, Google, or any OpenID Connect-compatible provider).

Q: What are federated users?
Federated users (external identities) are users you manage outside of AWS in your corporate directory, but to whom you grant access to your AWS account using temporary security credentials. They differ from IAM users, which are created and maintained in your AWS account.

Q: Do you support SAML?
Yes, AWS supports the Security Assertion Markup Language (SAML) 2.0.

Q: What SAML profiles does AWS support?
The AWS single sign-on (SSO) endpoint supports the IdP-initiated HTTP-POST binding WebSSO SAML Profile. This enables a federated user to sign in to the AWS Management Console using a SAML assertion. A SAML assertion can also be used to request temporary security credentials using the AssumeRoleWithSAML API. For more information, see About SAML 2.0-Based Federation.

Q: Can federated users access AWS APIs?

Yes. You can programmatically request temporary security credentials for your federated users to provide them secure and direct access to AWS APIs. We have provided a sample application that demonstrates how you can enable identity federation, providing users maintained by Microsoft Active Directory access to AWS service APIs. For more information, see Using Temporary Security Credentials to Request Access to AWS Resources.

Q: Can federated users access the AWS Management Console?

Yes. There are a couple ways to achieve this. One way is by programmatically requesting temporary security credentials (such as GetFederationToken or AssumeRole) for your federated users and including those credentials as part of the sign-in request to the AWS Management Console. After you have authenticated a user and granted them temporary security credentials, you generate a sign-in token that is used by the AWS single sign-on (SSO) endpoint. The user's actions in the console are limited to the access control policy associated with the temporary security credentials. For more details, see Creating a URL that Enables Federated Users to Access the AWS Management Console (Custom Federation Broker).

Alternatively, you can post a SAML assertion directly to AWS sign-in (https://signin.aws.amazon.com/saml). The user's actions in the console are limited to the access control policy associated with the IAM role that is assumed using the SAML assertion. For more details, see Enabling SAML 2.0 Federated Users to Access the AWS Management Console.

Using either approach allows a federated user to access the console without having to sign in with a user name and password. We have provided a sample application that demonstrates how you can enable identity federation, providing users maintained by Microsoft Active Directory access to the AWS Management Console.

Q: How do I control what a federated user is allowed to do when signed in to the console?

When you request temporary security credentials for your federated user using an AssumeRole API, you can optionally include an access policy with the request. The federated user's privileges are the intersection of permissions granted by the access policy passed with the request and the access policy attached to the IAM role that was assumed. The access policy passed with the request cannot elevate the privileges associated with the IAM role being assumed. When you request temporary security credentials for your federated user using the GetFederationToken API, you must provide an access control policy with the request. The federated user's privileges are the intersection of the permissions granted by the access policy passed with the request and the access policy attached to the IAM user that was used to make the request. The access policy passed with the request cannot elevate the privileges associated with the IAM user used to make the request. These federated user permissions apply to both API access and actions taken within the AWS Management Console.

Q: What permissions does a federated user need to use the console?
A user requires permissions to the AWS service APIs called by the AWS Management Console. Common permissions required to access AWS services are documented in Using Temporary Security Credentials to Request Access to AWS Resources.

Q: How do I control how long a federated user has access to the AWS Management Console?
Depending on the API used to create the temporary security credentials, you can specify a session limit between 15 minutes and 36 hours (for GetFederationToken and GetSessionToken) and between 15 minutes and 12 hours (for AssumeRole* APIs), during which time the federated user can access the console. When the session expires, the federated user must request a new session by returning to your identity provider, where you can grant them access again. Learn more about setting session duration.

Q: What happens when the identity federation console session times out?
The user is presented with a message stating that the console session has timed out and that they need to request a new session. You can specify a URL to direct users to your local intranet web page where they can request a new session. You add this URL when you specify an Issuer parameter as part of your sign-in request. For more information, see Enabling SAML 2.0 Federated Users to Access the AWS Management Console.

Q: How many federated users can I give access to the AWS Management Console?
There is no limit to the number of federated users who can be given access to the console.

Q: What is web identity federation?
Web identity federation allows you to create AWS-powered mobile apps that use public identity providers (such as Amazon Cognito, Login with Amazon, Facebook, Google, or any OpenID Connect-compatible provider) for authentication. With web identity federation, you have an easy way to integrate sign-in from public identity providers (IdPs) into your apps without having to write any server-side code and without distributing long-term AWS security credentials with the app.
For more information about web identity federation and to get started, see About Web Identity Federation.

Q: How do I enable web identity federation with accounts from public IdPs?
For best results, use Amazon Cognito as your identity broker for almost all web identity federation scenarios. Amazon Cognito is easy to use and provides additional capabilities such as anonymous (unauthenticated) access, and synchronizing user data across devices and providers. However, if you have already created an app that uses web identity federation by manually calling the AssumeRoleWithWebIdentity API, you can continue to use it and your apps will still work.

Here are the basic steps to enable identify federation using one of the supported web IdPs:
Sign up as a developer with the IdP and configure your app with the IdP, who gives you a unique ID for your app.
If you use an IdP that is compatible with OIDC, create an identity provider entity for it in IAM.
In AWS, create one or more IAM roles.
In your application, authenticate your users with the public IdP.
In your app, make an unsigned call to the AssumeRoleWithWebidentity API to request temporary security credentials.
Using the temporary security credentials you get in the AssumeRoleWithWebidentity response, your app makes signed requests to AWS APIs.
Your app caches the temporary security credentials so that you do not have to get new ones each time the app needs to make a request to AWS.
For more detailed steps, see Using Web Identity Federation APIs for Mobile Apps.

Q: How does identity federation using AWS Directory Service differ from using a third-party identity management solution?
If you want your federated users to be able to access only the AWS Management Console, using AWS Directory Service provides similar capabilities compared to using a third-party identity management solution. End users are able to sign in using their existing corporate credentials and access the AWS Management Console. Because AWS Directory Service is a managed service, customers do not need to set up or manage federation infrastructure, but rather need to create an AD Connector directory to integrate with their on-premises directory. If you are interested in providing your federated users access to AWS APIs, use a third-party offering, or deploy your own proxy server.

Billing

Q: Does AWS Billing provide aggregated usage and cost breakdowns by user?
No, this is not currently supported.

Q: Does the IAM service cost anything?
No, this is a feature of your AWS account provided at no additional charge.

Q: Who pays for usage incurred by users under an AWS Account?
The AWS account owner controls and is responsible for all usage, data, and resources under the account.

Q: Is billable user activity logged in AWS usage data?
Not currently. This is planned for a future release.

Q: How does IAM compare with Consolidated Billing?
IAM and Consolidated Billing are complementary features. Consolidated Billing enables you to consolidate payment for multiple AWS accounts within your company by designating a single paying account. The scope of IAM is not related to Consolidated Billing. A user exists within the confines of an AWS account and does not have permissions across linked accounts. For more details, see Paying Bills for Multiple Accounts Using Consolidated Billing.

Q: Can a user access the AWS accounts billing information?
Yes, but only if you let them. In order for IAM users to access billing information, you must first grant access to the Account Activity or Usage Reports. See Controlling Access.

Additional questions

Q: What happens if a user tries to access a service that has not yet been integrated with IAM?
The service returns an “Access denied” error.

Q: Are IAM actions logged for auditing purposes?
Yes. You can log IAM actions, STS actions, and AWS Management Console sign-ins by activating AWS CloudTrail. To learn more about AWS logging, see AWS CloudTrail.

Q: Is there any distinction between people and software agents as AWS entities?
No, both of these entities are treated like users with security credentials and permissions. However, people are the only ones to use a password in the AWS Management Console.

Q: Do users work with AWS Support Center and Trusted Advisor?
Yes, IAM users have the ability to create and modify support cases as well as use Trusted Advisor.

Q: Are there any default quota limits associated with IAM?
Yes, by default your AWS account has initial quotas set for all IAM-related entities. For details see Limitations on IAM Entities and Objects.

These quotas are subject to change. If you require an increase, you can access the Service Limit Increase form via the Contact Us page, and choose IAM Groups and Users from the Limit Type drop-down list.

Multi-factor authentication

Q. What is AWS MFA?
AWS multi-factor authentication (AWS MFA) provides an extra level of security that you can apply to your AWS environment. You can enable AWS MFA for your AWS account and for individual AWS Identity and Access Management (IAM) users you create under your account.

Q. How does AWS MFA work?
There are two primary ways to authenticate using an AWS MFA device:

- **AWS Management Console users**: When a user with MFA enabled signs in to an AWS website, they are prompted for their user name and password (the first factor–what they know), and an authentication response from their AWS MFA device (the second factor–what they have). All AWS websites that require sign-in, such as the AWS Management Console, fully support AWS MFA. You can also use AWS MFA together with Amazon S3 secure delete for additional protection of your S3 stored versions.
- **AWS API users**: You can enforce MFA authentication by adding MFA restrictions to your IAM policies. To access APIs and resources protected in this way, developers can request temporary security credentials and pass optional MFA parameters in their AWS Security Token Service (STS) API requests (the service that issues temporary security credentials). MFA-validated temporary security credentials can be used to call MFA-protected APIs and resources. Note: AWS STS and MFA-protected APIs do not currently support U2F security key as MFA.

Q. How do I help protect my AWS resources with MFA?
Follow two easy steps:

1. Get a MFA device. You have three options:

- Purchase a hardware YubiKey security key from Yubico, a third-party provider.
- Purchase a hardware device from Gemalto, a third-party provider.
- Install a virtual MFA–compatible application on a device such as your smartphone.

> Visit the AWS MFA page for details about how to acquire a hardware or virtual MFA device.

2. After you have a MFA device, you must activate it in the IAM console. You can also use the AWS CLI to activate virtual MFA and hardware MFA (Gemalto device) for an IAM user. **Note**: AWS CLI does not currently support activation of U2F security keys.

Q. Is there a fee associated with using AWS MFA?
AWS does not charge any additional fees for using AWS MFA with your AWS account. However, if you want to use a physical MFA device then you will need to purchase the MFA device that is compatible with AWS MFA either from Gemalto or Yubico, third party providers. For more details, please visit Yubico or Gemalto's website.

Q. Can I have multiple MFA devices active for my AWS account?
Yes. Each IAM user can have its own MFA device. However, each identity (IAM user or root account) can be associated with only one MFA device.

Q. Can I use my U2F security key with multiple AWS accounts?

Yes. AWS allows you to use the same U2F security key with several root and IAM users across multiple accounts.

Q. Can I use virtual, hardware, or SMS MFA with multiple AWS accounts?
No. The MFA device or mobile phone number associated to virtual, hardware, and SMS MFA is bound to an individual AWS identity (IAM user or root account). If you have a TOTP-compatible application installed on your smartphone, you can create multiple virtual MFA devices on the same smartphone. Each one of the virtual MFA devices is bound to a single identity, just like hardware MFA (Gemalto) device. If you dissociate (deactivate) the MFA device, you can then reuse it with a different AWS identity. The MFA device associated to hardware MFA cannot currently be used by more than one identity simultaneously.

Q. I already have a hardware MFA device (Gemalto) from my place of work or from another service I use, can I re-use this device with AWS MFA?
No. AWS MFA relies on knowing a unique secret associated with your hardware MFA (Gemalto) device in order to support its use. Because of security constraints that mandate such secrets never be shared between multiple parties, AWS MFA cannot support the use of your existing Gemalto device. Only a compatible hardware MFA device purchased from Gemalto can be used with AWS MFA. You can re-use an existing U2F security key with AWS MFA, as U2F security keys do not share any secrets between multiple parties.

Purchasing an MFA Device

Q. I'm having a problem with an order for a MFA device using the third-party provider's website. Where can I get help?
Yubico or Gemalto's customer service can assist you.

Q. I received a defective or damaged MFA device from the third party provider. Where can I get help?
Yubico or Gemalto's customer service can assist you.

Q. I just received a MFA device from the third party provider. What should I do?
You simply need to activate the MFA device to enable AWS MFA for your AWS account. See the IAM console to perform this task.

Provisioning a Virtual MFA Device

Q. What is a virtual MFA device?
A virtual MFA device is an entry created in a TOTP compatible software application that can generate six-digit authentication codes. The software application can run on any compatible computing device, such as a smartphone.

Q. What are the differences between a virtual MFA device and physical MFA devices?
Virtual MFA devices use the same protocols as the physical MFA devices. Virtual MFA devices are software based and can run on your existing devices such as smartphones. Most virtual MFA applications also allow you to enable more than one virtual MFA device, which makes them more convenient than physical MFA devices.

Q. Which virtual MFA applications can I use with AWS MFA?
You can use applications that generate TOTP-compliant authentication codes, such as the Google Authenticator application, with AWS MFA. You can provision virtual MFA devices either automatically by scanning a QR code with the device's camera or by manual seed entry in the virtual MFA application.

Visit the MFA page for a list of supported virtual MFA applications.

Q. What is a QR code?
A QR code is a two-dimensional barcode that is readable by dedicated QR barcode readers and most smartphones. The code consists of black squares arranged in larger square patterns on a white background. The QR code contains the required security configuration information to provision a virtual MFA device in your virtual MFA application.

Q. How do I provision a new virtual MFA device?
You can configure a new virtual MFA device in the IAM console for your IAM users as well as for your AWS root account. You can also use the aws iam create-virtual-mfa-device command in the AWS CLI or the CreateVirtualMFADevice API to provision new virtual MFA devices under your account. The aws iam create-virtual-mfa-device and the CreateVirtualMFADevice API return the required configuration information, called a seed, to configure the virtual MFA device in your AWS MFA compatible application. You can either grant your IAM users the permissions to call this API directly or perform the initial provisioning for them.

Q. How should I handle and distribute the seed material for virtual MFA devices?

You should treat seed material like any other secret (for example the AWS secret keys and passwords).

Q. How can I enable an IAM user to manage virtual MFA devices under my account?
Grant the IAM user the permission to call the CreateVirtualMFADevice API. You can use this

API to provision new virtual MFA devices.

SMS MFA

Q. Can I still request preview access to the SMS MFA?

We are no longer accepting new participants for the SMS MFA preview. We encourage you to use MFA on your AWS account by using a U2F security key, hardware device, or virtual (software-based) MFA device.

Q. When will the preview for SMS MFA end?

On February 1, 2019, AWS will no longer require IAM users to enter an MFA six-digit code if the IAM user is setup with “An SMS MFA device”. These users will also no longer be provided an SMS code when they sign in. We encourage you to use MFA through a U2F security key, hardware device, or virtual (software-based) MFA device. You can continue using this feature until January 31, 2019.

Enabling AWS MFA Devices

Q. Where do I enable AWS MFA?
You can enable AWS MFA for an AWS account and your IAM users in the IAM console, the AWS CLI, or by calling the AWS API. **Note**: AWS CLI and AWS API do not currently support enabling U2F security key.

Q. What information do I need to activate a hardware or virtual MFA device?
If you are activating the MFA device with the IAM console then you only need the device. If you are using the AWS CLI or the IAM API then you need the following:

1. The serial number of the MFA device. The format of the serial number depends on whether you are using a hardware device or a virtual device:

- Hardware MFA device: The serial number is on the bar-coded label on the back of the device.
- Virtual MFA device: The serial number is the Amazon Resource Name (ARN) value returned when you run the iam-virtualmfadevicecreate command in the AWS CLI or call the CreateVirtualMFADevice API.

2. Two consecutive MFA codes displayed by the MFA device.

Q. My MFA device seems to be working normally, but I am not able to activate it. What should I do?
Please contact us for help.

Using AWS MFA

Q. If I enable AWS MFA for my AWS root account or my IAM users, do they always have to use MFA to sign in to the AWS Management Console?
Yes. The AWS root credential user and IAM users must have their MFA device with them any time they need to sign in to any AWS website.

If your MFA device is lost, damaged, stolen, or not working, you can sign in using alternative factors of authentication, deactivate the MFA device, and activate a new device. As a security best practice, we recommend that you change your root account's password.

If your IAM users lose or damage their MFA device, or if it is stolen or stops working, you can disable AWS MFA yourself by using the IAM console or the AWS CLI.

Q. If I enable AWS MFA for my AWS root account or IAM users, do they always need to complete the MFA challenge to directly call AWS APIs?
No, it’s optional. However, you must complete the MFA challenge if you plan to call APIs that are secured by MFA-protected API access.

If you are calling AWS APIs using access keys for your AWS root account or IAM user, you do not need to enter an MFA code. For security reasons, we recommend that you remove all access keys from your AWS root account and instead call AWS APIs with the access keys for an IAM user that has the required permissions.

Note: U2F security keys currently do not work with MFA-protected APIs and currently cannot be used as MFA for AWS APIs.

Q. How do I sign in to the AWS Portal and AWS Management Console using my MFA device?
Follow these two steps:

1. If you are signing in as an AWS root account, sign in as usual with your user name and password when prompted. To sign in as an IAM user, use the account-specific URL and provide your user name and password when prompted.
2. If you have enabled virtual, hardware, or SMS MFA, enter the six-digit MFA code that appears on your MFA device. If you have enabled U2F security key, insert the key into the USB port of your computer, wait for the key to blink, and then touch the button or gold disk on your key.

Q. Does AWS MFA affect how I access AWS Service APIs?
AWS MFA changes the way IAM users access AWS Service APIs only if the account administrator(s) choose to enable MFA-protected API access. Administrators may enable

this feature to add an extra layer of security over access to sensitive APIs by requiring that callers authenticate with an AWS MFA device. For more information, see the MFA-protected API access documentation in more detail.

Other exceptions include S3 PUT bucket versioning, GET bucket versioning, and DELETE object APIs, which allow you to require MFA authentication to delete or change the versioning state of your bucket. For more information see the S3 documentation discussing Configuring a Bucket with MFA Delete in more detail.

For all other cases, AWS MFA does not currently change the way you access AWS service APIs.

Note: U2F security keys currently do not work with MFA-protected APIs and currently cannot be used as MFA for AWS APIs.

Q. For virtual and hardware MFA, can I use a given MFA code more than once?
No. For security reasons, you can use each MFA code provided by your virtual and hardware MFA device only once.

Q. I was recently asked to resync my MFA device because my MFA codes were being rejected. Should I be concerned?
No, this can happen occasionally. Virtual and hardware MFA relies on the clock in your MFA device being in sync with the clock on our servers. Sometimes, these clocks can drift apart. If this happens, when you use the MFA device to sign in to access secure pages on the AWS website or the AWS Management Console, AWS automatically attempts to resync the MFA device by requesting that you provide two consecutive MFA codes (just as you did during activation).

U2F security keys do not go out of sync and do not need a resync.

Q. My MFA device seems to be working normally, but I am not able to use it to sign in to the AWS Management Console. What should I do?
If you are using virtual or hardware MFA, we suggest you resynchronize MFA devices for your IAM user's credentials. If you already tried to resync and are still having trouble signing in, you can sign in using alternate factors of authentication and reset your MFA device.

If you are using U2F security keys, you can sign in using alternate factors of authentication and reset your MFA device.

If you are still encountering issues, contact us for help.

Q. My MFA device is lost, damaged, stolen, or not working, and now I can’t sign in to the AWS Management Console. What should I do?

If your MFA device is associated with an AWS root account:

You can reset your MFA device on the AWS Management Console by first signing in with your password and then verifying the email address and phone number associated with your root account.

If your MFA device is lost, damaged, stolen or not working, you can sign in using alternative factors of authentication, deactivate the MFA device, and activate a new MFA device. As a security best practice, we recommend that you change your root account's password.

If you need a new MFA device, you can purchase a new MFA device from a third-party provider, Yubico or Gemalto, or provision a new virtual MFA device under your account by using the IAM console.

If you have tried the preceding approaches and are still having trouble signing in, contact AWS Support.

Q. How do I disable AWS MFA?

To disable AWS MFA for your AWS account, you can deactivate your MFA device using the Security Credentials page. To disable AWS MFA for your IAM users, you need to use the IAM console or the AWS CLI.

Q. Can I use AWS MFA in GovCloud?
Yes, you can use AWS virtual MFA and hardware MFA devices in GovCloud.

MFA-protected API access

Q. What is MFA-protected API access?
MFA-protected API access is optional functionality that lets account administrators enforce additional authentication for customer-specified APIs by requiring that users provide a second authentication factor in addition to a password. Specifically, it enables administrators to include conditions in their IAM policies that check for and require MFA authentication for access to selected APIs. Users making calls to those APIs must first get temporary credentials that indicate the user entered a valid MFA code.

Q. Can I use my U2F security key with MFA-protected APIs?

No. MFA-protected APIs currently do not support U2F security keys.

Q. What problem does MFA-protected API access solve?
Previously, customers could require MFA for access to the AWS Management Console, but could not enforce MFA requirements on developers and applications interacting directly

with AWS service APIs. MFA-protected API access ensures that IAM policies are universally enforced regardless of access path. As a result, you can now develop your own application that uses AWS and prompts the user for MFA authentication before calling powerful APIs or accessing sensitive resources.

Q. How do I get started with MFA-protected API access?
You can get started in two simple steps:

1. Assign an MFA device to your IAM users. You can purchase a hardware key fob, or download a free TOTP-compatible application for your smartphone, tablet, or computer. See the MFA detail page for more information on AWS MFA devices.
2. Enable MFA-protected API access by creating permission policies for the IAM users and/or IAM groups from which you want to require MFA authentication. To learn more about access policy language syntax, see the access policy language documentation.

Q. How do developers and users access APIs and resources secured with MFA-protected API access?
Developers and users interact with MFA-protected API access both in the AWS Management Console and at the APIs.

In the AWS Management Console, any MFA-enabled IAM user must authenticate with their device to sign in. Users that do not have MFA do not receive access to MFA-protected APIs and resources.

At the API level, developers can integrate AWS MFA into their applications to prompt users to authenticate using their assigned MFA devices before calling powerful APIs or accessing sensitive resources. Developers enable this functionality by adding optional MFA parameters (serial number and MFA code) to requests to obtain temporary security credentials (such requests are also referred to as “session requests”). If the parameters are valid, temporary security credentials that indicate MFA status are returned. See the temporary security credentials documentation for more information.

Q. Who can use MFA-protected API access?
MFA-protected API access is available for free to all AWS customers.

Q. Which services does MFA-protected API access work with?
MFA-protected API access is supported by all AWS services that support temporary security credentials. For a list of supported services, see AWS Services that Work with IAM and review the column labeled Supports temporary security credentials.

Q. What happens if a user provides incorrect MFA device information when requesting temporary security credentials?
The request to issue temporary security credentials fails. Temporary security credential

requests that specify MFA parameters must provide the correct serial number of the device linked to the IAM user as well as a valid MFA code.

Q. Does MFA-protected API access control API access for AWS root accounts?
No, MFA-protected API access only controls access for IAM users. Root accounts are not bound by IAM policies, which is why we recommend that you create IAM users to interact with AWS service APIs rather than use AWS root account credentials.

Q. Do users have to have an MFA device assigned to them in order to use MFA-protected API access?
Yes, a user must first be assigned a unique hardware or virtual MFA device.

Q. Is MFA-protected API access compatible with S3 objects, SQS queues, and SNS topics?
Yes.

Q. How does MFA-protected API access interact with existing MFA use cases such as S3 MFA Delete?
MFA-protected API access and S3 MFA Delete do not interact with each other. S3 MFA Delete currently does not support temporary security credentials. Instead, calls to the S3 MFA Delete API must be made using long-term access keys.

Q. Does MFA-protected API access work in the GovCloud (US) region?
Yes.

Q. Does MFA-protected API access work for federated users?
Customers cannot use MFA-protected API access to control access for federated users. The GetFederatedSession API does not accept MFA parameters. Since federated users can’t authenticate with AWS MFA devices, they are unable to access resources designated using MFA-protected API access.

Pricing

Q. What will I be charged for using AWS IAM?

IAM is a feature of your AWS account offered at no additional charge. You will be charged only for the use of other AWS services by your users.

Amazon Cognito FAQs

- General
- Add User Signup & Sign-in to your mobile and web apps
- Federate identities and provide secure access to AWS resources
- Store and Sync Data Across Devices
- Pricing

Get Started with Amazon Cognito

General

Q: What is Amazon Cognito?
Amazon Cognito lets you easily add user sign-up and authentication to your mobile and web apps. Amazon Cognito also enables you to authenticate users through an external identity provider and provides temporary security credentials to access your app's backend resources in AWS or any service behind Amazon API Gateway. Amazon Cognito works with external identity providers that support SAML or OpenID Connect, social identity providers (such as Facebook, Twitter, Amazon) and you can also integrate your own identity provider.

In addition, Amazon Cognito enables you to synchronize data across a user's devices so that their app experience remains consistent when they switch between devices or upgrade to a new device. Your app can save data locally on users' devices allowing your applications to work even when the devices are offline and then automatically synchronize the data when the device is back online.

With Amazon Cognito, you can focus on creating great app experiences instead of worrying about building, securing, and scaling a solution to handle user management, authentication, and sync across platforms and devices.

Q: Who should use Amazon Cognito?
Amazon Cognito is designed for developers who want to add user management and sync functionality to their mobile and web apps. Developers can use Cognito Identity to add sign-up and sign-in to their apps and to enable their users to securely access their app's resources. Cognito also enables developers to sync data across devices, platforms, and applications.

Q: How do I start using Amazon Cognito?
You can easily get started by visiting the AWS Console. If you do not have an Amazon Web

Services account, you can create an account when you sign in to the console. Once you have created a user pool for user management or an identity pool for federated identities or sync operations, you can download and integrate the AWS Mobile SDK with your app. Alternatively you can call the Cognito server-side APIs directly, instead of using the SDK. See our developer guide for more information.

Q: Does Amazon Cognito expose server-side APIs?
Yes. Cognito exposes server-side APIs. You can create your own custom interface to Cognito by calling these APIs directly. The server-side APIs are described in the Developer Guide.

Q: Which platforms does Amazon Cognito support?
Support for Cognito is included in the optional AWS Mobile SDK, which is available for iOS, Android, Unity, and Kindle Fire. Cognito is also available in the AWS SDK for JavaScript. Cognito Your User Pools is currently supported in the AWS Mobile SDKs for iOS and Android and in the JavaScript AWS SDK for Cognito. Visit our resource page to download the SDKs.

Q: Do I have to use the AWS Mobile SDK?
No. Cognito exposes its control and data APIs as web services. You can implement your own client library calling the server-side APIs directly.

Add User Sign-up & Sign-in to your mobile and web apps

Q: Can I have my own identity provider to support user sign-up and sign-in?

Yes, you can easily and securely add sign-up and sign-in functionality to your apps with Cognito Identity. Your users can sign-up and sign-in using email, phone number, or user name. You can also implement enhanced security features, such as email verification, phone number verification, and multi-factor authentication. Cognito Identity also enables you to customize workflows by, for example, adding app-specific logic to user registration for fraud detection and user validation through AWS Lambda. To learn more, visit our docs.

Q: What is a User Pool?

A User Pool is your user directory that you can configure for your web and mobile apps. A User Pool securely stores your users' profile attributes. You can create and manage a User Pool using the AWS console, AWS CLI, or AWS SDK.

Q: What user profile information is supported by Cognito Identity?

Developers can use either standard OpenID Connect-based user profile attributes (such as user name, phone number, address, time zone, etc.) or customize to add app-specific user attributes.

Q: Can I enable my application’s users to sign up or sign in with an email address or phone number?

Yes, you can use the aliasing feature to enable your users to sign up or sign in with an email address and a password or a phone number and a password. To learn more, visit our docs.

Q: Can I set up password policies?

Yes, you can set up password policies, such as strength of password and character type requirements, when setting up or configuring your user pool.

Q: Can I verify the email addresses and phone numbers of my application’s users?

Yes, with Cognito Identity you can require your users’ email addresses and phone numbers to be verified prior to providing them access to your application. During sign-up, a verification code will be sent to the user’s phone number or email address, and the user must input the verification code to complete sign-up and become confirmed.

Q: Does Cognito Identity support SMS-based multi-factor authentication (MFA)?

Yes, you can enable the end users of your application to sign in with SMS-based MFA. With SMS-based MFA enabled, your users will be prompted for their password (the first factor—what they know), and for a security code that can only be received on their mobile phone via SMS (the second factor—what they have).

Q: Is it possible to customize user sign-up and sign-in workflows?

Yes, you can customize sign-up and sign-in by adding app-specific logic to the user sign-up and sign-in flows using AWS Lambda. For example, you can create AWS Lambda functions to identify fraud or perform additional validations on user data. You are able to trigger developer-provided Lambda functions at pre-registration, at post-confirmation, at pre-authentication, during authentication to customize the challenges, and at post-authentication. You can also use Lambda functions to customize messages sent as part of email or phone number verification and multi-factor authentication.

Q: Can I remember the devices associated with my application's users in a Cognitio user pool?

Yes, you can opt to remember devices used to access your application, and you associate these remembered devices with your application's users in a Cognito user pool. You can also opt to use remembered devices to supress second factor challenges for your users when you have set up multi-factor authentication.

Q: How can I migrate my existing users into an Amazon Cognito user pool?

You can use our import tool to migrate your existing users into an Amazon Cognito user pool. User attribute values are imported from a .csv file, which can be uploaded through the console, our APIs, or CLI. When imported users first sign in, they confirm their account and create a new password with a code sent to their email address or phone. There is no additional cost for using the import tool. To learn more, see the import tool documentation.

The import tool does not migrate passwords. If you want to retain your users' current passwords, you might consider an alternative approach to migrate users one at a time as they sign-in to your app during a transition period. With this approach, your app first tries to sign-in the user with your Cognito user pool. If that user doesn't exist in the user pool, your app will sign the user in with your existing identity system and temporarily retain the username and password used to do so. After a user successfully signs in with your existing identity system, your app would then use the same username and password to create the user in your Cognito user pool. This approach requires maintaining your existing identity system during the transition period, but after the transition period ends, you can use our import tool to import the remaining users (without their passwords).

Federate identities and provide secure access to AWS resources

Q: Can I use Cognito Identity to federate identities and secure access to AWS resources?
Yes, Cognito Identity enables you to authenticate users through an external identity provider and provides temporary security credentials to access your app's backend resources in AWS or any service behind Amazon API Gateway. Amazon Cognito works with external identity providers that support SAML or OpenID Connect, social identity providers (such as Facebook, Twitter, Amazon) and you can also integrate your own identity provider.

Q: Which public identity providers can I use with Amazon Cognito Identity?
You can use Amazon, Facebook, Twitter, Digits, Google and any other OpenID Connect compatible identity provider.

Q: What is an Identity Pool?
Identity pools are the containers that Cognito Identity uses to keep your apps' federated identities organized. Identity Pool associates federated identities from social identity providers with a unique user specific identifier. Identity Pools do not store any user profiles. An identity pool can be associated with one or many apps. If you use two different identity pools for two apps then the same end user will have a different unique identifier in each Identity Pool.

Q: How does the login flow work with public identity providers?
Your mobile app authenticates with an Identity Provider (IdP) using the provider's SDK. Once the end user is authenticated with the IdP, the OAuth or OpenID Connect token or the SAML assertion returned from the IdP is passed by your app to Cognito Identity, which returns a new Cognito ID for the user and a set of temporary, limited-privilege AWS credentials.

Q: Can I register and authenticate my own users?
Cognito Identity can integrate with your existing authentication system. With a simple API call you can retrieve a Cognito ID for your end users based on your own unique identifier for your users. Once you have retrieved the Cognito ID and OpenID Token Cognito Identity provides, you can use the Cognito Identity client SDK to access AWS resources and synchronize user data. Cognito Identity is a fully managed identity provider to make it easier for you to implement user sign-up and sign-in for your mobile and web apps.

Q: How does Cognito Identity help me control permissions and access AWS services securely?
Cognito Identity assigns your users a set of temporary, limited privilege credentials to access your AWS resources so you do not have to use your AWS account credentials. The permissions for each user are controlled through AWS IAM roles that you create. You can define rules to choose the IAM role for each user, or if you are using groups in a Cognito user pool, you can assign IAM roles based on groups. Cognito Identity also allows you to define a separate IAM role with limited permissions for guest users who are not authenticated. In addition, you can use the unique identifier that Cognito generates for your users to control access to specific resources. For example you can create a policy for an S3 bucket that only allows each user access to their own folder within the bucket.

Q: When using public identity providers, does Amazon Cognito Identity store users' credentials?

No, your app communicates directly with the supported public identity provider (Amazon, Facebook, Twitter, Digits, Google, or an Open ID Connect-compliant provider) to authenticate users. Cognito Identity does not receive or store user credentials. Cognito Identity uses the token from the identity provider to obtain a unique identifier for the user and then hashes it using a one-way hash so that the same user can be recognized again in the future without storing the actual user identifier.

Q: Does Cognito Identity receive or store confidential information about my users from the identity providers?
No. Cognito Identity does not receive any confidential information (such as email address, friends list, etc.) from the identity providers.

Q: Do I still need my own backend authentication systems with Cognito Identity?
No. Cognito Identity supports login through Amazon, Facebook, Twitter, Digits, and Google, as well as providing support for unauthenticated users. With Cognito Identity you can support federated authentication, profile data sync store and AWS access token distribution without writing any backend code.

Q: What if I don't want to force my users to log in?
Cognito Identity supports the creation and token vending process for unauthenticated users as well as authenticated users. This removes the friction of an additional login screen

in your app, but still enables you to use temporary, limited privilege credentials to access AWS resources.

Q: What are unauthenticated users?
Unauthenticated users are users who do not authenticate with any identity provider, but instead access your app as a guest. You can define a separate IAM role for these users to provide limited permissions to access your backend resources.

Q: Does Cognito Identity support separate identities for different users on the same device?
Yes. Cognito Identity supports separate identities on a single device, such as a family iPad. Each identity is treated separately and you have complete control over how your app logs users in and out and how local and remote app data is stored.

Q: How do I store data associated with Cognito Identity?
You can programmatically create a data set associated with Cognito Identity and start saving data in the form of key/value pairs. The data is stored both locally on the device and in the Cognito sync store. Cognito can also sync this data across all of the end user’s devices.

Q: Does the number of identities in the Cognito Identity console tell me how many users are using my app?
The number of identities in the Cognito Identity console shows you how many identities were created via the Cognito Identity APIs. For Authenticated Identities (those logging in with a login provider such as Facebook or an OpenID Connect provider), each call to Cognito Identity’s GetId API will only ever create a single identity for each user. However, for Unauthenticated identities, each time the client in an app calls the GetId API will generate a new identity. Therefore, if your app calls GetId for unauthenticated identities multiple times for a single user it will appear that a single user has multiple identities. So it is important that you cache the response from GetId when using unauthenticated identities and not call it multiple times per user.

The Mobile SDK provides the logic to cache the Cognito Identity automatically so you don't have to worry about this. If you're looking for a complete analytics solution for your app, including the ability to track unique users, please look at Amazon Mobile Analytics.

Store and sync data across devices

Q: What is the Amazon Cognito sync store?
The Amazon Cognito Sync store is a key/value pair store linked to an Amazon Cognito identity. There is no limit to the number of identities you can create in your identity pools and sync store. Each Amazon Cognito identity within the sync store has its own user information store.

Q: Is data saved directly to the Amazon Cognito sync store?
No. The optional AWS Mobile SDK saves your data to an SQLite database on the local device, this way the data is always accessible to your app. The data is pushed to the Amazon Cognito sync store by calling the synchronize() method and, if push synchronization is enabled, all other devices linked to an identity are notified of the data change in the sync store via Amazon SNS.

Q: How is data stored in the Amazon Cognito sync store?
Data associated with an Amazon Cognito identity are organized as key/value pairs. A key is a label e.g. “MusicVolume”, and a value e.g. “11”. Key/value pairs are grouped and categorized using data sets. Data sets are a logical partition of key/value pairs and the most granular entity used by Amazon Cognito to perform sync operations.

Q: What is the maximum size of a user information store within the Amazon Cognito sync store?
Each user information store can have a maximum size of 20MB. Each data set within the user information store can contain up to 1MB of data. Within a data set you can have up to 1024 keys.

Q: What kind of data can I store in a data set?
Both keys and values within a data set are alphanumeric strings. There is no limit to the length of the strings other than the total amount of values in a dataset cannot exceed 1MB. Binary data can be stored as a base64 encoded string as a value provided it does not exceed the 1MB limit.

Q: Why are data sets limited to 1MB?
Limiting the data set size to 1MB increases the chances of a synchronization task completing successfully even when bandwidth is limited without lots of retries that consume battery life and data plans.

Q: Are user identities and user information stores shared across developers?
No, a user identity and information store is tied to a specific AWS account. If there are multiple apps from different publishers on a particular device that use Amazon Cognito, each app will use the information store created by each publisher.

Q: How can I analyze and query the data stored in the Cognito Sync store?
With Cognito Streams, you can push sync store data to a Kinesis stream in your AWS account. You can then consume this stream and store the data in a way that makes it easy for you to analyze such as a Amazon Redshift database, an RDS instance you own or even an S3 file. We have published sample Kinesis consumer application to show how to store the updates data in Amazon Redshift.

Q: Why should I use Kinesis stream instead of a database export?
By streaming the data to Kinesis you can receive all of the history of changes to your

datasets in real-time. This means you receive all the changes an end user makes to a dataset and gives you the flexibility to store this data in a tool of your choice.

Q: What if I already have data stored in Cognito?
When you enable the Kinesis stream feature you will be able to start a bulk publish. This process asynchronously sends all of the data currently stored in your Cognito sync store to the Kinesis stream you selected.

Q: What is the price of this feature?
Cognito pushes the data to a Kinesis stream you own. There is no difference in Cognito’s per-synchronization price if this feature is enabled. You will be charged Kinesis’ standard rates for your shards.

Q: Can I validate data before it is saved?
Amazon Cognito Events allows developers to run an AWS Lambda function in response to important events in Cognito. The Sync Trigger event is an event that occurs when any dataset is synchronized. Developers can write an AWS Lambda function to intercept the synchronization event. The function can evaluate the changes to the underlying Dataset and manipulate the data before it is stored in the cloud and synchronized back to the user's other devices. Alternatively, the AWS Lambda function could fail the sync operation so that the data is not synchronized to the user's other devices.

Q: How is data synchronized with Amazon Cognito?
You can programmatically trigger the sync of data sets between client devices and the Amazon Cognito sync store by using the synchronize() method in the AWS Mobile SDK. The synchronize() method reads the latest version of the data available in the Amazon Cognito sync store and compares it to the local, cached copy. After comparison, the synchronize() method writes the latest updates as necessary to the local data store and the Amazon Cognito sync store. By default Amazon Cognito maintains the last-written version of the data. You can override this behavior and resolve data conflicts programmatically. In addition, push synchronization allows you to use Amazon Cognito to send a silent push notification to all devices associated with an identity to notify them that new data is available.

Q: What is a silent push notification?
Amazon Cognito uses the Amazon Simple Notification Service (SNS) to send silent push notifications to devices. A silent push notification is a push message that is received by your application on a user's device that will not be seen by the user.

Q: How do I use push synchronization?
To enable push synchronization you need to declare a platform application using the Amazon SNS page in the AWS Management Console. Then, from the identity pool page in the Amazon Cognito page of the AWS Management Console, you can link the SNS platform application to your Cognito identity pool. Amazon Cognito automatically utilizes the SNS platform application to notify devices of changes.

Q: How are conflicts in the synchronization process handled?
By default Amazon Cognito maintains the last-written version of the data. You can override this behavior by choosing to respond to a callback from the AWS Mobile SDK which will contain both versions of the data. Your app can then decide which version of the data (the local one or the one in the Amazon Cognito sync store) to keep and save to the Amazon Cognito sync store.

Pricing

Q: How much does Cognito Identity cost?
With Amazon Cognito, you pay only for what you use. There are no minimum fees and no upfront commitments.

If you are using the Cognito Identity to create a User Pool, you pay based on your monthly active users (MAUs) only. A user is counted as a MAU if within a calendar month there is an identity operation related to that user, such as sign-up, sign-in, token refresh, and password change. You are not charged for subsequent sessions or for inactive users with in that calendar month. Separate charges apply for optional use of SMS messaging as described below.

The Your User Pool feature has a free tier of 50,000 MAUs each month. The Cognito Identity free tier does not expire at the end of your 12 month AWS Free Tier term, and it is available to both existing and new AWS customers indefinitely

Federated Identities and secure access control for AWS resources are always free with Cognito Identity.

Q: How much does Cognito Sync cost?
Sync charges are based on the total amount of data saved in the Amazon Cognito sync store and the number of sync operations performed. A sync operation compares the local data store on a device to the Amazon Cognito sync store in the cloud and synchronizes the two data stores.

As part of the AWS Free Tier, eligible AWS customers receive 10 GB of cloud sync store and 1,000,000 sync operations per month for the first 12 months. Outside the Free Tier, Amazon Cognito costs $0.15 for each 10,000 sync operations and $0.15 per GB of sync store per month.

Q: What is a sync operation?
When you call the synchronize() method using the AWS Mobile SDK, this counts as a sync operation. If you are calling the server APIs directly, a sync operation is initiated when a new sync session token is emitted and is completed with a successful write or a timeout of the session token. Whether you use the SDK synchronize() method or call the server API’s directly, sync operations are charged at the same rate.

Q. What are Monthly Active Users (MAUs)?
A user is considered active and counted as a MAU when there is an operation (e.g., sign-in, token refresh, sign-up, or password change) associated with the user during the billing month. Therefore, you are not charged for subsequent operations during the billing month or for inactive users. Typically, your total number of users as well as your number of operations will be significantly larger than your total number of MAUs.

Q. What does it cost to use SMS messages with Cognito?
Use of SMS messaging to verify phone numbers, to send codes for forgotten or reset passwords, or for multi-factor authentication is charged separately. See the Worldwide SMS Pricing page for more information.

Q: Is Amazon Cognito part of the AWS Free Tier?
Yes. As part of the AWS Free Tier, Cognito offers 10GB of sync store and 1,000,000 sync operations in a month for up to the first 12 months of usage. Your user pool for Cognito Identity is free for the first 50,000 MAUs, and we offer volume-based tiers thereafter. The Federated Identities feature for authenticating users and generating unique identifiers is always free with Cognito Identity.

Q: Does every write or read from the app count as a sync operation?
No. You decide when to call the synchronize() method. Every write or read from the device is to the local SQlite store. This way you are in complete control of your costs.

Q: What does push synchronization cost
Cognito utilizes Amazon SNS to send silent push notifications. There is no additional charge for using Cognito for push synchronization, but normal Amazon SNS rates will apply for notifications sent to devices.

Q: What is the cost of using Lambda with Amazon Cognito Events?
There is no additional charge for using Cognito Events to trigger Lambda functions, but normal rates for your use of AWS Lambda and other AWS services will apply while your Lambda functions are executing. Please see the AWS Lambda pricing page for details.

Amazon Detective FAQs

General

Q: What is Amazon Detective?

Amazon Detective makes it easy to analyze, investigate, and quickly identify the root cause of potential security issues or suspicious activities. Amazon Detective automatically collects log data from your AWS resources and uses machine learning, statistical analysis, and graph theory to build a linked set of data that enables you to easily conduct faster and more efficient security investigations.

Q: What are the key benefits of Amazon Detective?

Amazon Detective simplifies the investigative process and helps security teams conduct faster and more effective investigations. Amazon Detective's prebuilt data aggregations, summaries, and context help you to quickly analyze and determine the nature and extent of possible security issues. Amazon Detective maintains up to a year of aggregated data and makes it easily available through a set of visualizations that shows changes in the type and volume of activity over a selected time window, and links those changes to security findings. There are no upfront costs and you pay only for the events analyzed, with no additional software to deploy or log feeds to enable.

Q: How much does Amazon Detective cost?

Amazon Detective is currently in preview. During the preview, Amazon Detective is available at no cost for those approved for access. Amazon Detective pricing is applicable at General Availability. It is based on the volume of data ingested from AWS CloudTrail logs, Amazon VPC Flow Logs, and Amazon GuardDuty findings. You are charged per Gigabyte (GB) ingested per account/region/month. Amazon Detective maintains up to a year of aggregated data for its analysis. Please see the Amazon Detective pricing page for the latest pricing information.

Q: Is Amazon Detective a regional or global service?

Amazon Detective needs to be enabled on a region by region basis and enables you to quickly analyze activity across all your accounts within each region. This ensures all data analyzed is regionally based and doesn't cross AWS regional boundaries.

Q: What regions does Amazon Detective support?

Amazon Detective is available during Preview in the following regions: US-East (Northern Virginia), US-East (Ohio), US-West (Oregon), EU (Ireland), and Asia Pacific (Tokyo).

Getting started with Amazon Detective

Q: How can I get started with Amazon Detective?

Amazon Detective is in preview. During the preview, Amazon Detective is available at no cost for those approved for access. Preview access can be requested here.

Q: How do I enable Amazon Detective?

You can enable Amazon Detective from within the AWS Management Console or by using the Amazon Detective API. If you are already using the Amazon GuardDuty or AWS Security Hub Consoles, you should enable Amazon Detective with the same account that is the Master account in Amazon GuardDuty or AWS Security Hub to enable the best cross-service experience.

Q: Can I manage multiple accounts with Amazon Detective?

Yes, Amazon Detective is a multi-account service that aggregates data from monitored member accounts under a single master account within the same region. You can configure multi-account monitoring deployments in same way that you configure master and member accounts in Amazon GuardDuty and AWS Security Hub.

Q: What data sources does Amazon Detective analyze?

Amazon Detective enables customers to view summaries and analytical data associated with AWS CloudTrail events as well as VPC Flow Logs. For customers that have Amazon GuardDuty enabled, Detective also processes Amazon GuardDuty findings.

Q: Can I use Amazon Detective if I do not have Amazon GuardDuty enabled?

Yes, you can monitor your accounts' AWS CloudTrail events and VPC flow activity even if you do not have GuardDuty enabled. Amazon Detective provides detailed summaries, analysis and visualizations for AWS accounts, EC2 instances, AWS users, roles, and IP Addresses. These can be very useful in developing an understanding of how an AWS environment and infrastructure is utilized from a management event and network flow perspective.

Q: How quickly does Amazon Detective start working?

Amazon Detective starts collecting log data as soon as it is enabled and provides visual summaries and analytics on the ingested data. Amazon Detective also provides comparisons of recent activity against historical baselines which are established after two weeks of account monitoring. If you are an Amazon GuardDuty customer, Amazon Detective will automatically ingest and process two weeks of historical log data upon activation. This enables you to start leveraging baseline comparisons and analytic insights immediately after enabling the service.

Q: Does Amazon Detective ingest historical data?

If you are an Amazon GuardDuty customer, Amazon Detective ingests and processes two weeks of historical log data upon activation to ensure baselines are established, so that you can get immediate value from Amazon Detective's analytics and visualizations.

Q: Can I export my raw log data from Amazon Detective?

Amazon Detective analyzes your AWS CloudTrail logs and VPC Flow Logs but does not make the raw logs available for export. AWS enables you to export these logs through other services.

Q: What data does Amazon Detective store, is it encrypted, and can I control what data sources are enabled?

Amazon Detective conforms to the AWS shared responsibility model, which includes regulations and guidelines for data protection. Once enabled, Amazon Detective will process data from AWS CloudTrail logs, VPC Flow Logs, and Amazon GuardDuty findings for any accounts where it has been turned on.

Q: Is there a performance or availability risk to my existing AWS workloads by enabling Amazon Detective?

Amazon Detective has no impact on the performance or availability of your AWS infrastructure since Amazon Detective retrieves the log data and findings directly from the AWS services.

Q: How does Amazon Detective differ from Amazon GuardDuty and AWS Security Hub?

Amazon GuardDuty is a threat detection service that continuously monitors for malicious activity and unauthorized behavior to protect your AWS accounts and workloads. With Security Hub, you have a single place that aggregates, organizes, and prioritizes your security alerts, or findings, from multiple AWS services, such as Amazon GuardDuty, Amazon Inspector, and Amazon Macie, as well as from AWS Partner solutions. Amazon Detective simplifies the process of investigating security findings and identifying the root cause. Amazon Detective analyzes trillions of events from multiple data sources such as VPC Flow Logs, AWS CloudTrail logs, and Amazon GuardDuty findings and automatically creates a graph model that provides you with a unified, interactive view of your resources, users, and the interactions between them over time.

Q: How can I stop Amazon Detective from looking at my logs and data sources?

Amazon Detective enables you to analyze and visualize security data from your AWS CloudTrail logs, VPC Flow logs, and Amazon GuardDuty findings. To stop Amazon Detective from analyzing these logs and findings for your accounts please disable the service by using the API or from the settings section in the AWS Console for Amazon Detective.

Working in the Amazon Detective console

Q: What guidance does Amazon Detective provide on how to investigate a security issue?

Amazon Detective provides a variety of visualizations that present context and insights about AWS resources such as AWS accounts, EC2 instances, users, roles, IP addresses, and Amazon GuardDuty findings. Each visualization is designed to answer specific questions that may come up as you analyze findings and the related activity. Each visualization provides textual guidance that clearly explains how to interpret the panel and use its information to answer your investigative questions.

Q: How is Amazon Detective integrated with other AWS security services like Amazon GuardDuty and AWS Security Hub?

Amazon Detective supports cross-service user workflows by supporting console integrations with Amazon GuardDuty and AWS Security Hub. These services provide links from within their consoles that redirect you from a selected finding directly to an Amazon Detective page containing a curated set of visualizations for investigating the selected finding. The findings detail page in Amazon Detective is already aligned to the timeframe of the finding and shows relevant data associated with the finding.

Q: How do I integrate Amazon Detective investigation results with remediation and response tools?

Various partner security solution providers have integrated with Amazon Detective to enable investigation steps within their automated playbooks and orchestrations. These products present links from within the response workflows that redirect users to Amazon Detective pages containing visualizations curated for investigating findings and resources identified within the workflow.

Amazon GuardDuty FAQs

Service Overview

Q: What is Amazon GuardDuty?

Q: What are the key benefits of Amazon GuardDuty?

Q: How much does Amazon GuardDuty cost?

Q. Does the estimated cost in the Amazon GuardDuty payer account show the total aggregated costs for linked accounts, or just that individual payer account?

Q: Is there a free trial?

Q: What is the difference between Amazon GuardDuty and Amazon Macie?

Q: Is Amazon GuardDuty a regional or global service?

Q: What regions does Amazon GuardDuty support?

Q: What partners work with Amazon GuardDuty?

Enabling GuardDuty

Q: How do I enable Amazon GuardDuty?

Q: Can I manage multiple accounts with Amazon GuardDuty?

Q: What data sources does Amazon GuardDuty analyze?

Q: How quickly does GuardDuty start working?

Q: Do I have to enable AWS CloudTrail, VPC Flow Logs, and DNS logs for Amazon GuardDuty to work?

Q: Is there any performance or availability impact to enabling Amazon GuardDuty on my account?

Q: Does Amazon GuardDuty manage or keep my logs?

Q: How can I stop Amazon GuardDuty from looking at my logs and data sources?

GuardDuty Findings

Q: What can Amazon GuardDuty detect?

Q: What is Amazon GuardDuty threat intelligence?

Q: Can I supply my own threat intelligence?

Q: How do machine learning and behavioral anomaly detections work?

Q: How are security findings delivered?

Q: What is the format of Amazon GuardDuty findings?

Q: How long are security findings made available in Amazon GuardDuty?

Q: Can I take automated preventative actions using Amazon GuardDuty?

Q: How are Amazon GuardDuty detections developed and managed?

Q: Can I write custom detections in Amazon GuardDuty?

Learn more about product pricing

See pricing examples and free trial details

Learn more »

Sign up for a free trial

Get access to the Amazon GuardDuty free trial.

Start free trial »

Start building in the console

Get started with Amazon GuardDuty in the AWS Console.

Sign in »

What is Amazon Inspector?

Amazon Inspector is an automated security assessment service that helps you test the network accessibility of your Amazon EC2 instances and the security state of your applications running on the instances.

What can I do with Amazon Inspector?

Amazon Inspector allows you to automate security vulnerability assessments throughout your development and deployment pipeline or against static production systems. This allows you to make security testing a more regular occurrence as part of development and IT operations. Amazon Inspector is an API-driven service that uses an optional agent, making it easy to deploy, manage, and automate. Amazon Inspector assessments are offered to you as pre-defined rules packages mapped to common security best practices and vulnerability definitions.

What makes up the Amazon Inspector service?

Amazon Inspector consists of a technology that analyzes your network configurations in AWS for reachability, an Amazon-developed agent that is installed in the operating system of your Amazon EC2 instances, and a security assessment service that uses telemetry from the agent and AWS configuration to assess instances for security exposures and vulnerabilities.

What is an assessment template?

An assessment template is a configuration that you create in Amazon Inspector to define your assessment run. This assessment template includes a rules package against which you want Amazon Inspector to evaluate your assessment target, the duration of the assessment run, Amazon Simple Notification Service (SNS) topics to which you want Amazon Inspector to send notifications about assessment run states and findings, and Amazon Inspector-specific attributes (key/value pairs) that you can assign to findings generated by the assessment run.

What is an assessment run?

An assessment run is the process of discovering potential security issues through the analysis of your assessment target's configuration, installed software, and behavior against specified rule packages. If the network reachability rules package is included, Inspector analyzes your network configurations in AWS to find accessibility of your EC2 instances over the network. If the Inspector agent is installed on the instance, the agent collects and sends on-host software and configuration data. Next, the Inspector service analyzes the data and compares it against the rule packages specified. A completed assessment run produces a list of findings for potential security issues.

Is there any performance impact during an Amazon Inspector assessment run?

There is no performance impact to your application when running an agentless assessment with the network reachability rules package. There is a minimal performance impact during the data collection phase of the assessment run when using the Amazon Inspector Agent.

What is an assessment target?

An assessment target represents a collection of Amazon EC2 instances that you want assessed, typically a set of instances that work together as a unit to help you accomplish your business goal(s). Amazon Inspector evaluates the security state of these EC2 instances. You can include all of your instances in an assessment target or specify a subset of instances by using Amazon EC2 tags.

What is a finding?

A finding is a potential security issue discovered during the Amazon Inspector assessment run of the specified target. Findings are displayed in the Amazon Inspector console or retrieved through the API, and contain both a detailed description of the security issue and a recommendation on how to fix it.

What is a rules package?

A rules package is a collection of security checks that can be configured as part of an assessment template and assessment run. Amazon Inspector has two types of rules packages, the network reachability rules package that checks for network accessibility of your Amazon EC2 instances, and host assessment rules packages that check for vulnerabilities and insecure configurations on the Amazon EC2 instance. Host assessment rules packages include Common Vulnerabilities and Exposures (CVE), Center for Internet Security (CIS) Operating System configuration benchmarks, and security best practices. See the Amazon Inspector documentation for a full list of rules packages available.

Can I define my own rules for assessment templates?

No. Only the pre-defined rules are currently allowed for assessment runs.

Which on-host software packages can Inspector analyze for vulnerabilities?

Amazon Inspector finds applications by querying the package manager or software installation system on the operating system where the agent is installed. This means that software that was installed through the package manager is assessed for vulnerabilities. The version and patch level of software that is not installed through these methods is not recognized by Inspector. For example, software installed via apt, yum, or Microsoft Installer will be assessed by Inspector. Software installed through make config / make install, or binary files copied directly to the system using automation software such as Puppet or Ansible will not be assessed by Inspector.

What is an assessment report, and what does it include?

An Amazon Inspector assessment report can be generated for an assessment run once it has been successfully completed. An assessment report is a document that details what is tested in the assessment run, and the results of the assessment. The results of your assessment are formatted into a standard report, which can be generated to share results within your team for remediation actions, to enrich compliance audit data, or to store for future reference.

You can select from two types of report for your assessment, a findings report or a full report. The findings report contains an executive summary of the assessment, the instances targeted, the rules packages tested, the rules that generated findings, and detailed information about each of these rules along with the list of instances that failed the check. The full report contains all the information in the findings report, and additionally provides the list of rules that were checked and passed on all instances in the assessment target.

What happens if some of my agents are unavailable when I run an assessment?

Amazon Inspector assessments with the network reachability rules package can be run without an agent for any Amazon EC2 instances. The agent is required for host assessment rules packages. Amazon Inspector will gather vulnerability data from all available agents and return any appropriate security findings for them. Inspector generates Exclusions to notify you of any EC2 instances without the agent installed or having an unhealthy agent.

How do agents become unavailable?

Amazon Inspector Agents could be unavailable for a number of reasons, such as: the EC2 instance is down or unresponsive; the targeted instance does not have the agent installed; the installed agent is unavailable or cannot return vulnerability data.

What is the pricing for Amazon Inspector?

Amazon Inspector pricing is based on number of Amazon EC2 instances included in each assessment, and depends on the rules packages you select for assessments. Inspector assessments can have any combination of host assessment rules packages and the network reachability rules package. Host assessment rules packages include Common Vulnerabilities and Exposures (CVE), Center for Internet Security benchmarks (CIS), and Security Best Practices. If your assessments include both host rules packages and the network reachability rules package, you will be billed for both separately. The on-demand billing period is one calendar month. See the Amazon Inspector pricing page for full pricing details.

Pricing example:

Consider a scenario where you run the following assessment runs in a month. In this example, all of your assessment runs include both host assessment rules packages and the network reachability rules package. And all of your EC2 instances have the Inspector Agent on them.

1 assessment run against 1 instance
1 assessment run against 10 instances
10 assessment runs against 2 instances each
30 assessment runs against 10 instances each

If the above represented the Amazon Inspector assessment runs activity in your account for a given billing period, you would be charged for 331 host agent-assessments and 331 network reachability instance-assessments.

The price of each individual host agent-assessment and network reachability instance-assessment is based on a tiered pricing model. For example, as you move up the volume of agent-assessments in a given billing period, you pay a lower price per agent-assessment.

The Amazon Inspector charges for your account for this billing period would be:

For host assessment rules packages -
First 250 agent-assessments @ $0.30 per agent-assessment
Next 81 agent-assessments @ $0.25 per agent-assessment

For network reachability rules package
First 250 instance-assessments @ $0.15 per instance-assessment
Next 81 instance-assessments @ $0.13 per instance-assessment

Adding up all the above, the Amazon Inspector bill would be $95.25 for host agent-assessments and $48.03 for network reachability instance-assessments, for a total of $143.28.

Is there a free trial for Amazon Inspector?

Yes. Accounts that have never run an Amazon Inspector assessment, you’re eligible for 250 agent-assessments with host rules packages and 250 instance-assessments with the network reachability rules package at no cost during your first 90 days.

What Operating Systems does Amazon Inspector support?

Please see the Amazon Inspector documentation for a current list of supported operating systems for the Inspector Agent. Note that the network reachability rules package can be run without an agent for any Amazon EC2 instances regardless of the operating system. If the Inspector Agent is installed, network reachability generates enhanced findings with information that identifies the software processes reachable on your EC2 instances.

In what regions is Amazon Inspector available?

Please see the Amazon Inspector documentation for a current list of supported regions.

Amazon Inspector sounds great, how do I get started?

Sign up for Amazon Inspector from the AWS Management Console. On the welcome page, you can enable scheduled network reachability assessments for your whole account with just one click. You can install the optional Inspector Agent on your EC2 instances to enable host assessment rules packages. You can also customize the EC2 instances to assess, rules package selection, and notifications of findings using the advanced setup option. Once an assessment run completes, Inspector will generate findings for security issues identified in your environment.

Does the Amazon Inspector Agent have to be installed on all of the EC2 instances I wish to assess?

No, Amazon Inspector assessments with the network reachability rules package can be run without an agent for any Amazon EC2 instances. The agent is required for host assessment rules packages.

How can I install the Amazon Inspector Agent?

There are several ways to install the agent. For simple installations, you can install it manually on each instance or do a one-time load using the AWS Systems Manager Run Command document (AmazonInspector-ManageAWSAgent). For larger deployments, you can automate agent installations using the EC2 User Data Function when configuring your instances or you can create automated installs of the agent using AWS Lambda. You can also launch an EC2 instance using the Amazon Linux AMI with the pre-installed Amazon Inspector Agent from the EC2 Console or the AWS Marketplace.

How do I check whether the Amazon Inspector Agent is installed and healthy on my EC2 instances?

You can view the status of the Amazon Inspector Agent for all the EC2 instances in your assessment target by using the ‘Preview Targets’ functionality available in the Inspector console and through the PreviewAgents API query. Agent status includes whether the agent is installed on the EC2 instance and the health of the agent. Along with the Inspector Agent status on the targeted EC2 instance, the instance ID, public hostname, and public IP address (if defined) are also displayed, along with links into the EC2 console for each instance.

Does Amazon Inspector access other AWS services in my account?

Amazon Inspector needs to enumerate your EC2 instances and tags to identify the instances specified in the assessment target and to read your AWS network configurations. Amazon Inspector gets access to these through a service-linked role that is created on your behalf when you get started with Inspector as a new customer or in a new region. The Inspector service-linked role is managed by Amazon Inspector, so you don’t have to worry about inadvertently revoking permissions required by Amazon Inspector. For some existing customers, an IAM role that was registered while getting started with Inspector might be used for accessing other AWS services until the Inspector service-linked role is created. You can create the Inspector service-linked role through the Inspector console’s dashboard page.

I use a Network Address Translation (NAT) for my instances. Will Amazon Inspector work with these instances?

Yes. Instances that use a NAT are supported by Amazon Inspector with no action required from you.

I use a Proxy for my instances. Will Amazon Inspector work with these instances?

Yes. The Amazon Inspector Agent supports proxy environments. For Linux instances, we support HTTPS Proxy, and for Windows instances, we support WinHTTP proxy. See the Amazon Inspector User Guide for instructions to configure Proxy support for the Amazon Inspector Agent.

I would like to automate the assessment of my infrastructure on a regular basis. Do you provide an automated way to set up assessments?

Yes. Amazon Inspector provides a full API allowing automatic creation of application environments, creation of assessments, evaluation of policies, creation of policy exceptions, and filters as well as retrieval of the results. Amazon Inspector assessments can also be configured and triggered through AWS CloudFormation templates.

Can I schedule security assessments to run at certain dates and times?

Yes, you can set up a simple recurring schedule for assessments in your assessment template. And Inspector assessments can be triggered by any Amazon CloudWatch Event. You can set up custom schedules with either a fixed recurring rate or a more detailed Cron expression through CloudWatch Events.

Can I trigger security assessments to run based on an event?

Yes. You can use Amazon CloudWatch Events to create event patterns which monitor other AWS services for actions to trigger an assessment. For example, you can create an event which monitors AWS Auto Scaling for new Amazon EC2 Instances being launched, or monitors AWS CodeDeploy notifications for when a code deployment has been successfully completed. Once CloudWatch Events have been configured against Amazon Inspector templates, these assessment events will be displayed in the Inspector console as part of your assessment templates so you can see all of the automated triggers for that assessment.

Can I set up Amazon Inspector assessments through AWS CloudFormation?

Yes, you can create Amazon Inspector resource groups, assessment targets, and assessment templates using AWS CloudFormation templates. This allows you to automatically set up security assessments for your EC2 instances as they are deployed. In your CloudFormation template, you can also bootstrap installation of the Inspector Agent on EC2 instances by using agent installation commands in either AWS::CloudFormation::Init or EC2 user data. Alternatively, you can create EC2 instances in your CloudFormation template using an AMI with the Inspector Agent pre-installed.

Where can I find metrics information on my Amazon Inspector assessments?

Amazon Inspector automatically publishes metrics data on your assessments to Amazon CloudWatch. If you are a CloudWatch user, your Inspector assessment statistics will automatically be populated to CloudWatch. The Inspector metrics that are currently available are: number of assessment runs, agents targeted, and findings generated. For more details, see the Amazon Inspector documentation for details on the assessment metrics published to CloudWatch.

Can Amazon Inspector be integrated with other AWS services for logging and notifications?

Amazon Inspector integrates with Amazon SNS to provide notification for various events such as monitoring milestones, failures, or expiration of exceptions and integrates with AWS CloudTrail for logging of calls to Amazon Inspector.

What is the network reachability rules package?

The network reachability rules package that identifies ports and services on your Amazon EC2 instances that are reachable from outside your VPC. When you run an assessment with this rules package, Inspector queries AWS APIs to read network configurations in your account such as Amazon Virtual Private Clouds (VPCs), security groups, network access control lists (ACLs), and route tables. then analyzes these network configurations to prove accessibility of ports. Findings show you the network configurations that allow access to a reachable port to help you easily restrict access as needed. The Amazon Inspector agent is not needed for assessments with the network reachability rules package. For instances with the Inspector agent installed, network reachability findings are enhanced with information that identifies which processes are listening on accessible ports.

What is the advantage of using the Inspector Agent for network reachability rules package?

The Amazon Inspector agent is not needed for assessments with the network reachability rules package. For instances with the Inspector agent installed network reachability findings are enhanced with information that identifies which processes are listening on accessible ports.

What is the “CIS Operating System Security Configuration Benchmarks” rules package?

CIS Security Benchmarks are provided by the Center for Internet Security and are the only consensus-based, best-practice security configuration guides both developed and accepted by government, business, industry, and academia. Amazon Web Services is a CIS Security Benchmarks Member company and the list of Amazon Inspector certifications can be viewed here. CIS benchmark rules are designed to be pass/fail security checks. For every CIS check that fails, Inspector generates a finding with High severity. Additionally, an Informational finding is generated for each instance that lists all the CIS rules that are checked, and the pass/fail result for each rule.

What is the “Common Vulnerabilities and Exposures” rules package?

The Common Vulnerabilities and Exposures or CVE rules check for exposure to publicly known information security vulnerabilities and exposures. CVE rule details are available publicly at the National Vulnerability Database (NVD). We use the NVD's Common Vulnerability Scoring System (CVSS) as the primary source of severity information. In case a CVE is not scored by NVD but is present in Amazon Linux AMI Security Advisory (ALAS), we use the severity from Amazon Linux advisory. In case neither of these scores is available for a CVE, we do not report that CVE as a finding. We check daily for latest information from NVD and ALAS and update our rules packages accordingly.

What is the severity of a finding?

Each Amazon Inspector rule has an assigned severity level, which Amazon has classified as High, Medium, Low, or Informational. Severity is intended to help you prioritize your responses to findings.

How is the severity determined?

Severity of a rule is based on potential impact of the security issue found. Although some rules packages have Severity levels provided as part of the rules they provide, these can often differ by rules set. Amazon Inspector has normalized the severity for findings across all available rules packages by mapping the individual severities to common High, Medium, Low, and Informational classifications. For “High”, “Medium”, and “Low” severity findings, the higher the severity of the finding, the more security impact the underlying issue has. Findings that are classified as “Informational” are provided to advise you of security issues which might not have an immediate security impact.

For AWS supported rules packages, the severity is determined by the AWS security team.

The CIS Benchmarks rules package findings always have severity set to “High”.

For the Common Vulnerabilities & Exploits (CVE) rules package, Amazon Inspector has mapped the provided CVSS Base Scoring and ALAS Severity levels provided:

Amazon Inspector Severity	CVSS Base Score	ALAS Severity (if CVSS not scored)
High	>= 5	Critical or Important
Medium	< 5 and >= 2.1	Medium
Low	< 2.1 and >= 0.8	Low
Informational	< 0.8	N/A

When I describe findings via the API (DescribeFindings), each finding has a “numericSeverity” attribute. What does this attribute signify?

The “numericSeverity” attribute is the numeric representation of the severity of a finding. The numeric severity values map to Severity as follows:

> Informational = 0.0
> Low = 3.0

> Medium = 6.0
> High = 9.0

Does Amazon Inspector work with AWS partner solutions?

Yes, Amazon Inspector has public facing APIs that are available for customers and AWS partners to utilize. Several partners have integrated with Amazon Inspector incorporating findings into email, ticketing systems, pager platforms, or broader security dashboards. For detail on supporting partners, please visit the Amazon Inspector Partners page.

Is Amazon Inspector a HIPAA eligible service?

Yes, Amazon Inspector is a HIPAA eligible service and has been added to the AWS Business Associate Addendum (BAA). If you have an executed BAA with AWS, you can run Inspector on your EC2 instances that contain protected health information (PHI).

What compliance and assurance programs does Amazon Inspector support?

Inspector supports SOC 1, SOC 2, SOC 3, ISO 9001, ISO 27001, ISO 27017, ISO 27018, and HIPAA. Inspector meets the controls for FedRAMP and we're waiting for the completion of the audit report. If you want to learn more about the AWS services in scope by compliance program, please visit the AWS Services in Scope Page.

Amazon Macie FAQ

General

Q: What is Amazon Macie?
Amazon Macie is an ML-powered security service that helps you prevent data loss by automatically discovering, classifying, and protecting sensitive data stored in Amazon S3. Amazon Macie uses machine learning to recognize sensitive data such as personally identifiable information (PII) or intellectual property, assigns a business value, and provides visibility into where this data is stored and how it is being used in your organization. Amazon Macie continuously monitors data access activity for anomalies, and delivers alerts when it detects risk of unauthorized access or inadvertent data leaks.

Q: What can I do with Amazon Macie?
You can use Amazon Macie to protect against security threats by continuously monitoring your data and account credentials. Amazon Macie gives you an automated and low touch way to discover and classify your business data and detect sensitive information such as personally identifiable information (PII) and credential data. When alerts are generated, you can use Amazon Macie for incident response, using Amazon CloudWatch Events to swiftly take action to protect your data.

Data Analysis

Q: What data sources does Amazon Macie support?
The Amazon Macie service supports Amazon S3 and AWS CloudTrail management API and S3 object-level events for the buckets and prefixes enrolled with Amazon Macie.

Q: How does Amazon Macie work?
Amazon Macie is a security service that provides customers both visibility and security for the content that they store in Amazon S3. Amazon Macie helps customers understand their data by automatically and continuously discovering, classifying, and intelligently and accurately assigning a business value to customer’s data. Through understanding the asset value of content and how it is being accessed, Amazon Macie is able to create contextual and narrative security alerts on challenges that our customers face, only alerting when high value content is being accessed in a way that creates risk for their business. Examples include Amazon Macie’s ability to detect global access permissions inadvertently being set on sensitive data, detect uploading of API keys inside source code, and verify sensitive customer data is being stored and accessed in a manner that meets their compliance standards.

Customers can enable Amazon Macie quickly and easily without the need to manually define and periodically update complicated data classifications and inflexible user roles. Amazon Macie combines machine learning with user behavior analytics to detect activity that signals potential risk to business-critical data or assets. For example, Amazon Macie can alert on the download of large quantities of source code by a user account that typically does not access that data, or sudden changes in permissions of Amazon S3 buckets that house data. Once enabled, customers can start receiving security and compliance alerts immediately and create automated policies to protect your data when suspicious activity is detected. Amazon Macie's rich user interface provides accurate alerts with detailed evidence and actionable recommendations that allow you to spend time responding to the most relevant risks. Amazon Macie features a rich user interface that allows for security and compliance use cases; offering a set of APIs that will allow partners and customers to incorporate Amazon Macie's data classification and security anomaly detection capabilities directly into their own applications.

Security and Access

Q: What are some examples of suspicious activity that Amazon Macie can detect?
Amazon Macie analyzes activity of user, application, and service accounts associated with sensitive data that suggests risk to the business, such as inadvertent exposure of data, insider threats, or targeted attacks. Amazon Macie can alert on suspicious activity such as compromised user accounts enumerating and downloading large amounts of sensitive content from unusual IP addresses, or the download of large quantities of source code by a user account that typically does not access this type of sensitive content. A compliance-focused example of Amazon Macie includes detection of large quantities of high-risk documents shared publically or to the entire company, such as files containing personally identifiable information (PII), protected health information (PHI), intellectual properties (IP), legal or financial data. Additionally, customers also have the ability to use Amazon Macie's dashboard to define their own alerts and policy definitions based on their security needs.

Q: How does Amazon Macie secure your data?
As part of the data classification process, Amazon Macie identifies customers' objects in their S3 buckets, and streams the object contents into memory for analysis. When deeper analysis is required for complex file formats, Amazon Macie will download a full copy of the object, only keeping it for the short time it takes to fully analyze the object. Immediately after Amazon Macie has analyzed the file content for data classification, it deletes the stored content and only retains the metadata required for future analysis. At any time, customers can revoke Amazon Macie access to data in the Amazon S3 bucket.

Q: How does Amazon Macie automate security policies to protect data and enforce compliance workloads?
The first step towards building compliance policies including Payment Card Industry (PCI), Health Insurance Portability and Accountability Act (HIPAA), or General Data Protection Regulation (GDPR) is around identifying where sensitive data exists across an organization. Amazon Macie automates this discovery phase, with highly accurate, machine learning based detection of over 70 data types related to Personally Identifiable Information (PII), Personal Health Information (PHI), regulatory documents, API keys and secret key material. Customers can get started quickly by enabling Amazon Macie's compliance policies to alert on the existence of credential information embedded within source code and backups, or to automate policies about how PII and PHI can be safely stored and accessed. In addition to supporting data compliance use cases, Amazon Macie identifies changes to policies and access control lists that could indicate inadvertent overexposure of information, or suspicious access to content that could indicate a potential data breach. Amazon Macie allows customers to automate response and remediation through Amazon CloudWatch Events and AWS Lambda functions that can be built to meet the specific needs of your organization.

Integration

Q: Can partner and third-party solutions integrate with Amazon Macie?

Yes, Amazon Macie supports control plane API endpoints through the AWS SDK, allowing for integration by partners and third party solutions. Additionally, Amazon Macie sends all findings to CloudWatch Events, allowing for follow on integration by partners and third party solutions through any available triggers. This includes external case management and ticketing systems such as Atlassian JIRA, Splunk, HP ArcSight, and IBM Resilient Systems.

Languages

Q: Does Amazon Macie support multiple languages?
Natural Language Processing is a key feature of Amazon Macie, which needs to process and understand content to provide its full value. Amazon Macie’s NLP supports discovery and classification of content in multiple languages. Although some features of Amazon Macie will work on non-English content, such as user behavior analytics, Amazon Macie is optimized for English only at this time.

Getting Started

Q: How do I get started with Amazon Macie?
To get started with Amazon Macie, simply log in to the Amazon Macie console, run the provided CloudFormation templates to configure the necessary IAM roles and policies in your account, and select which S3 buckets to protect. Learn more about Amazon Macie and supported use cases by reading our **Blog** and **Documentation**.

AWS Artifact FAQs

General

1. WHAT IS AWS ARTIFACT?

AWS Artifact, available in the console, is a self-service audit artifact retrieval portal that provides our customers with on-demand access to AWS’ compliance documentation and AWS agreements.

You can use AWS Artifact Reports to download AWS security and compliance documents, such as AWS ISO certifications, Payment Card Industry (PCI), and System and Organization Control (SOC) reports.

You can use AWS Artifact Agreements to review, accept, and track the status of AWS agreements such as the Business Associate Addendum (BAA).

2. WHO HAS ACCESS TO AWS ARTIFACT?

All AWS Accounts have access to AWS Artifact. Root users and IAM users with admin permissions can download all audit artifacts available to their account by agreeing to the associated terms and conditions.

You will need to grant IAM users with non-admin permissions access to AWS Artifact using IAM permissions. This allows you to grant a user access to AWS Artifact, while restricting access to other services and resources within your AWS Account. For information on how to grant access using IAM, refer to this help topic in the **AWS Artifact documentation**.

3. HOW DO I GIVE OTHER USERS ACCESS TO AWS ARTIFACT AGREEMENTS?

Your administrative account has all of the permissions needed to use AWS Artifact, but different documents and agreements might require you to delegate permissions differently for various users. You can delegate permissions by using IAM policies. Refer to the following tables in the AWS Artifact User Guide to view the permissions that you can assign to IAM users based on the level of access that they need.

- Report Permissions
- Agreement Permissions
- Common AWS Artifact IAM Policies

4. WHAT IS AN AUDIT ARTIFACT?

An audit artifact is a piece of evidence that demonstrates that an organization is following a documented process or meeting a specific requirement. Audit artifacts are gathered and archived throughout the system development life cycle and are to be used as evidence in internal and/or external audits and assessments.

AWS Artifact currently provides customers with reports and agreements that may be used as audit artifacts.

5. HOW DO I SHARE AUDIT ARTIFACTS WITH MY AUDITORS?

You will often need to provide your auditors with access to AWS compliance reports. You can easily accomplish this by creating IAM user credentials specific to each auditor and configuring the credentials so that the auditor can only access the reports that are relevant to the audit that they are conducting. For more information, see this help topic in the **AWS Artifact documentation**.

6. HOW CAN I USE THESE ARTIFACTS TO MEET MY AUDIT REQUIREMENTS?

You can provide the AWS audit artifacts to your auditors or regulators as evidence of AWS security controls.

You can also use the responsibility guidance provided by some of the AWS audit artifacts to design your cloud architecture. This guidance helps determine the additional security controls you should put in place in order to support the specific use cases of your system.

7. IS THERE A LIMIT TO THE NUMBER OF ARTIFACTS I CAN DOWNLOAD?

No. You can access and download all available artifacts at any time, as many times as you need.

Compliance Reports

1. WHO SHOULD USE AWS ARTIFACT REPORTS?

AWS Artifact Reports can be used by all AWS customers to assess and validate the security and compliance of the AWS infrastructure and services that they use.

You should use AWS Artifact Reports if you are:

- Obligated to demonstrate the compliance of your cloud architectures during system design, development and audit life cycles. In order to demonstrate the historical and current compliance of your AWS infrastructure (specific to the services that you use), auditors and regulators require you to provide evidence in the form of audit artifacts.
- Required to or are interested in using audit artifacts to validate that your AWS implemented controls are operating effectively.
- Interested in continuously monitoring or auditing your suppliers.
- A member of a development team that is building secure cloud architectures and are in need of guidance in understanding your responsibility for complying with ISO, PCI, SOC, and other regulatory standards. Often, the work of your team will either enable your enterprise to use AWS or ensure that your enterprise can continue to use AWS.

You can provide the AWS audit artifacts to your auditors or regulators as evidence of AWS security controls.

You can also use the responsibility guidance provided by some of the AWS audit artifacts to design your cloud architecture. This guidance helps determine the additional security controls you should put in place in order to support the specific use cases of your system.

2. CAN I SHARE AWS COMPLIANCE REPORTS WITH MY CUSTOMERS?

Your customers can access AWS compliance reports using their own AWS Account. If they do not already have an account, you should direct them to create one. There is no charge associated with creating an account.

After logging into their account, your customers can access available reports in the AWS Console by navigating to **Artifact** under **Security, Identity & Compliance**. If your customer would like to access a report that requires an NDA, they can receive access by signing a click-through NDA inside of the Artifact Console.

For more information refer to **Getting Started with AWS Artifact**.

3. WHERE CAN I ACCESS THE AWS FEDRAMP SECURITY PACKAGES?

To learn more about the Federal Risk and Authorization Management Program (FedRAMP) and how to download AWS FedRAMP Security packages using AWS Artifact, visit the **FedRAMP compliance** webpage.

Agreements

1. WHAT IS AWS ARTIFACT AGREEMENTS, AND WHY SHOULD I USE IT?

AWS Artifact Agreements, a feature of the AWS Artifact service (our audit and compliance portal), enables you to review, accept, and manage agreements with AWS for your individual account, and also for all accounts that are part of your organization in **AWS Organizations**. You can also use AWS Artifact to terminate agreements you have previously accepted if they are no longer required.

2. WHAT AGREEMENTS ARE AVAILABLE IN AWS ARTIFACT AGREEMENTS?

Different types of agreements are available in AWS Artifact Agreements to address the needs of customers subject to specific regulations.For example, the Business Associate Addendum (BAA) is available for customers that need to comply with the Health Insurance Portability and Accountability Act (HIPAA). For a complete list of agreements available to your account, login to **AWS Artifact**.

Before you enter into an agreement on AWS Artifact Agreements, you must download and agree to the terms of the AWS Artifact nondisclosure agreement (NDA). Each agreement is confidential and cannot be shared with others outside of your company.

3. IF I ALREADY HAVE A SIGNED NDA WITH AWS OUTSIDE OF ARTIFACT, DO I NEED TO ACCEPT A NEW NDA IN AWS ARTIFACT AGREEMENTS?

Yes, you will need to accept the AWS Artifact NDA to access and download confidential documents in Artifact. That said, if you have an existing NDA with Amazon, and if your existing NDA covers the same confidential information as the information provided in Artifact, then your existing NDA will apply instead of the Artifact NDA.

4. WHO HAS ACCESS TO USE AWS ARTIFACT AGREEMENTS?

If you’re an administrator of an AWS account, you automatically have permissions to download, accept, and terminate agreements for that account. If you are the administrator of the master account of an organization in AWS Organizations, you can accept and terminate agreements on behalf of the master account and all member accounts in your organization. You should always review any agreement terms with your legal, privacy and/or compliance teams before accepting. You can use IAM to grant access to your agreement stakeholders (such as members of your legal, privacy and/or compliance teams), so that those users can download, review, and accept agreements.

If you’re not an administrator, you will need to be granted additional permissions to download, accept, and terminate agreements (usually, by your administrator). Administrators have the flexibility to grant varying levels of permissions to IAM users based on the business needs of the users.

For a complete list of AWS Artifact permissions, refer to **Controlling Access** and **Common Policies** in the *AWS Artifact User Guide*.

5. WHAT IS THE DIFFERENCE BETWEEN AWS ARTIFACT ACCOUNT AGREEMENTS AND AWS ARTIFACT ORGANIZATION AGREEMENTS?

When accepted, AWS Artifact Account Agreements (located under the **Account agreements** tab) apply only to the individual account you used to sign into AWS.

When accepted, AWS Artifact Organization Agreements (located under the **Organization agreements** tab) apply to all accounts in an organization created through **AWS Organizations**, including the organization's master account and all member accounts. Only the master account in an organization can accept agreements in AWS Artifact Organization Agreements.

6. WHAT IS THE BENEFIT OF USING AWS ARTIFACT ORGANIZATION AGREEMENTS?

AWS Artifact Organization Agreements simplifies agreement management for multiple AWS accounts by allowing you to accept a single agreement on behalf of all accounts within your organization. When an authorized user of a master account accepts an organization agreement, all existing and future member accounts will be covered under the terms of the agreement automatically.

7. WHAT DO I NEED TO DO IN ORDER TO USE AWS ARTIFACT ORGANIZATION AGREEMENTS?

If you are a user of the master account of an organization in AWS Organizations, you can accept an agreement on behalf of all current and future member accounts in your organization. The organization that you belong to must be enabled for **all features**. If your organization is configured for **consolidated billing features** only, see **Enabling All Features in Your Organization**.

To get started, you must be signed in to the master account with the following IAM permissions:

artifact:DownloadAgreement
artifact:AcceptAgreement
artifact:TerminateAgreement
organizations:DescribeOrganization
organizations:EnableAWSServiceAccess
organizations:ListAWSServiceAccessForOrganization
iam:ListRoles
iam:CreateRole
iam:AttachRolePolicy

For a complete list of AWS Artifact permissions, refer to **Controlling Access** and **Common Policies** in the *AWS Artifact User Guide*.

8. WHY DO I HAVE TO GRANT AWS PERMISSION TO CREATE A ROLE IN MY ACCOUNT BEFORE USING AWS ARTIFACT ORGANIZATION AGREEMENTS?

AWS needs permission to create an IAM role in your account so that the AWS Artifact service can **ListAccounts** to identify the complete list of member accounts in your organization when an agreement is accepted. When a member account joins or leaves your organization, AWS will be notified, and the list of accounts covered by your accepted agreement(s) will be updated.

9. HOW DO I KNOW IF MY ORGANIZATION IS USING AWS ARTIFACT ORGANIZATION AGREEMENTS?

Visit the **AWS Artifact Agreements console** and click on the **Organization agreements** tab. If the master account in your organization has accepted one or more organization agreements, they will be listed as active. You can do this either when logged in as the master account or as a member account in the organization.

Important: The IAM user signed into the AWS console must have permission to organizations:DescribeOrganization in order for AWS Artifact to retrieve information about your account’s organization agreements. For a complete list of AWS Artifact permissions, refer to **Controlling Access** and **Common Policies** in the *AWS Artifact User Guide*.

10. WHAT IS AN ORGANIZATION?

An organization is a collection of one or more member accounts that you can manage centrally with a single master account using **AWS Organizations**. Refer to the **AWS Organizations website** to learn more.

11. WHAT IS A MASTER ACCOUNT?

A master account is the AWS account you use to create your organization in **AWS Organizations**. When logged into the master account, you can use AWS Organizations to create member accounts in your organization, invite existing accounts to join your organization, and remove accounts from your organization.

Only master accounts can use AWS Artifact Organization Agreements to accept or terminate agreements on behalf of all accounts in an organization.

12. WHAT IS A MEMBER ACCOUNT?

A member account is an AWS account, other than the master account, that is part of an organization in **AWS Organizations**. If you are an administrator of the master account in an organization, you can create member accounts in the organization and invite existing accounts to join the organization. A member account can belong to only one organization at a time.

Member accounts can use AWS Artifact Account Agreements to accept or terminate agreements on behalf of that individual member account only. Member accounts can use AWS Artifact Organization Agreements to view the agreements accepted on the member account's behalf by the organization's master account.

13. IF MY ACCOUNT IS NOT PART OF AN ORGANIZATION, CAN I STILL USE AWS ARTIFACT ORGANIZATION AGREEMENTS?

No, AWS Artifact Organization Agreements is only available for accounts using AWS Organizations. If you would like to create or join an organization, follow the instructions in **Creating and Managing an AWS Organizations**.

14. HOW DOES AWS ARTIFACT AGREEMENTS WORK FOR RESELLER ACCOUNTS?

AWS Artifacts Agreements works the same for reseller accounts. Resellers can use IAM to control who has permissions to download, accept, and terminate agreements. By default, only users with administrative privileges can grant access.

15. HOW DO I ACCEPT AN AGREEMENT FOR ACCOUNTS IN SEPARATE AWS ORGANIZATIONS?

If you have accounts in separate organizations that you want covered by an agreement, you must log in to each organization's master account and accept the relevant agreements through AWS Artifact Organization Agreements.

If you would like to consolidate accounts into a single organization, you can invite AWS accounts to join your organization by following the instructions in **Inviting an Account to**

Your Organization.

16. CAN I USE AWS ARTIFACT ORGANIZATION AGREEMENTS TO ACCEPT AN AGREEMENT FOR ONLY SOME MEMBER ACCOUNTS WITHIN MY ORGANIZATION?

No. In AWS Artifact Organization Agreements (the **Organization agreements** tab) you can only accept agreements on behalf of all accounts within the organization.

If you would like to accept an agreement for only some member accounts, you must sign in to each account individually and accept the relevant agreement(s) through AWS Artifact Account Agreements (the **Account agreements** tab).

17. CAN I ACCEPT AN AGREEMENT IN THE ORGANIZATION AGREEMENT TAB IF MY ACCOUNT ALREADY HAS AN AGREEMENT OF THE SAME TYPE ACCEPTED IN THE ACCOUNT AGREEMENTS TAB?

Yes, master accounts and member accounts can have AWS Artifact Account Agreements (i.e. agreements under the **Account agreements** tab) and AWS Artifact Organization Agreements (i.e. agreements under the **Organization agreements** tab) of the same type in place at the same time.

If your account has an account agreement and an organization agreement of the same type in place at the same time, the organization agreement will apply instead of the account agreement. If, with respect to an individual account, an organization agreement is terminated (e.g. by removal of a member account from the organization), the account agreement in place for that individual account (viewable under the Account agreements tab) will remain active and will continue to apply.

18. IF MY ACCOUNT HAS THE SAME AGREEMENT ACCEPTED IN THE ACCOUNT AGREEMENTS TAB AND THE ORGANIZATION AGREEMENTS TAB, WHICH ONE APPLIES?

The organization agreement will apply because according to its terms, it applies instead of the account agreement when both are active. If the organization agreement is terminated, and if you have an account agreement of the same type in place (under the **Account agreements** tab), the account agreement will apply to that account. **Note**: Terminating the organization agreement does not terminate the account agreement.

19. IF A MEMBER ACCOUNT IS REMOVED FROM MY ORGANIZATION, WHAT HAPPENS TO THE ORGANIZATION AGREEMENTS THAT HAVE BEEN ACCEPTED ON ITS BEHALF?

When a member account is removed from an organization (e.g. by leaving the organization, or by being removed from the organization by the master account), any organization agreements accepted on its behalf will no longer apply to that member account.

Master account administrators should alert member accounts prior to removing those accounts from the organization so that member accounts can put new account agreements in place, if necessary. Before member account owners leave an organization, they should determine (with the assistance of legal, privacy, or compliance teams, if appropriate) whether it is necessary to put new agreements in place.

20. IF A MEMBER ACCOUNT IS REMOVED FROM MY ORGANIZATION, WILL THEY BE NOTIFIED?

Currently, member accounts are not notified when they are removed from an organization. We are developing functionality that will alert member accounts when they have been removed from an organization and are no longer covered by an organization agreement.

Master account administrators should alert member accounts prior to removing those accounts from the organization so that member accounts can put new account agreements in place, if necessary. Before member account owners leave an organization, they should determine (with the assistance of legal, privacy, or compliance teams, if appropriate) whether it is necessary to put new agreements in place.

Business Associate Addendum (BAA)

1. HOW DO I ACCEPT AN AWS BAA USING AWS ARTIFACT AGREEMENTS?

AWS Artifact Agreements enables you to review and accept the AWS BAA from the AWS Management Console for your account or your organization in **AWS Organizations**. You can accept the AWS BAA for your individual account under the **Account agreements** tab, or if you are a master account in an organization, you can accept the AWS BAA on behalf of all accounts in your organization under the **Organization agreements** tab. Upon accepting the AWS BAA in AWS Artifact Agreements, you will instantly designate your AWS account(s) for use in connection with protected health information (PHI). Additionally, you can use the AWS Artifact Agreements console to see which agreements are in place for your AWS account or organization and review the terms of those agreements.

2. HOW DO I DESIGNATE MY ACCOUNT AS A HIPAA ACCOUNT UNDER A BAA USING AWS ARTIFACT AGREEMENTS?

When you accept an online BAA within the **Account agreements** tab in AWS Artifact, the account you used to sign in to AWS is automatically designated as a HIPAA Account under that online account BAA. If you are a master account in **AWS Organizations** and accept an online BAA under the **Organization agreements** tab in AWS Artifact, all accounts within your organization are automatically designated as HIPAA Accounts. Member accounts that are later added to that organization will be automatically designated as HIPAA Accounts as well.

3. CAN I DESIGNATE MORE THAN ONE ACCOUNT AS A HIPAA ACCOUNT UNDER A BAA USING AWS ARTIFACT AGREEMENTS?

Yes, if you use **AWS Organizations**, the master account in your organization can use the **Organization agreements** tab in AWS Artifact Agreements to accept an organization BAA on behalf of all existing and future member accounts in your organization.

If you do not use AWS Organizations, or would only like to designate certain of your member accounts, you must sign in to each account separately and accept a BAA on behalf of that account.

4. WHAT IS THE DIFFERENCE BETWEEN THE AWS BAA THAT CAN BE ACCEPTED AS AN ACCOUNT AGREEMENT AND THE AWS BAA THAT CAN ACCEPTED AS AN ORGANIZATION AGREEMENT?

The difference is that the BAA in the **Organization agreements** tab, when accepted, applies to all accounts linked to your master account through AWS Organizations. In comparison, the BAA in the **Account agreements** tab only applies to the individual account you used to accept the account BAA, and no other accounts. If you have accepted both the account BAA and the organization BAA, the organization BAA will apply instead of the account BAA.

5. IF MY ACCOUNT HAS ALREADY ACCEPTED AN ACCOUNT BAA, CAN I ACCEPT THE ORGANIZATION BAA SO THAT ALL OF MY ACCOUNTS ARE COVERED?

Yes, using the master account of your organization you can use the **Organization agreements** tab in AWS Artifact Agreements to accept an organization BAA on behalf of all existing and future member accounts in your organization. When both the account and organization BAA are accepted, the organization BAA will apply instead of the account BAA.

6. HOW DO I TERMINATE A BAA USING AWS ARTIFACT AGREEMENTS?

If you no longer need to use your AWS account or organization accounts in connection with PHI, and if you accepted the BAA using AWS Artifact Agreements, you can use AWS Artifact Agreements to terminate that BAA.

If you accepted the BAA offline, refer to the 'Offline BAA' FAQs below.

7. WHAT HAPPENS WHEN I TERMINATE AN ONLINE BAA IN AWS ARTIFACT AGREEMENTS?

If you terminate an online BAA under the **Account agreements** tab in AWS Artifact, the account you used to sign into AWS will immediately cease to be a HIPAA Account and, unless it is also covered by an organization BAA (within the **Organization agreements** tab), it will no longer be covered by a BAA with AWS. You should only terminate a BAA if you are sure that you have removed all protected health information (PHI) from the account and will no longer use the account in connection with PHI.

If you are a user of a master account and terminate an online BAA within the **Organization agreements** tab in AWS Artifact, all accounts within your organization will immediately be removed as HIPAA Accounts and, unless they are covered by individual account BAAs (within the **Account agreements** tab), they will no longer be covered by a BAA with AWS. You should only terminate a BAA for an organization if you are sure that you have removed all protected health information (PHI) from ALL accounts within such organization and will no longer use any of the accounts in connection with PHI.

8. WHICH BAA APPLIES IF MY AWS ACCOUNT HAS AN ACCEPTED ACCOUNT BAA AND ORGANIZATION BAA?

If you have both an account BAA and an organization BAA in place at the same time, the terms of the organization BAA will apply instead of the terms of the account BAA. Terminating the organization BAA does not terminate the account BAA, so if you terminate the organization BAA, the account BAA will continue to apply to that account.

9. IF A MEMBER ACCOUNT LEAVES AN ORGANIZATION, DOES THE ORGANIZATION AGREEMENT STILL APPLY TO THE ACCOUNT?

No. When a member account leaves an organization, any accepted organization agreement(s) no longer apply to that account. If the member account wants one or more of the agreements to continue to apply after leaving the organization, the member account should accept the relevant account agreement(s) under the **Account agreements** tab in AWS Artifact prior to leaving the organization.

10. IF I HAVE A BAA WITH AWS, WHAT AWS SERVICES CAN I USE IN MY HIPAA ACCOUNT?

You may use any AWS service in an account designated as a HIPAA Account, but you may only include PHI in HIPAA Eligible Services. Our **HIPAA Eligible Services Reference** page contains the latest list of HIPAA Eligible Services.

11. CAN I ENTER INTO A BAA AGREEMENT WITHOUT USING AWS ARTIFACT?

Yes. If you prefer to enter into an offline BAA with AWS, please contact your AWS Account Manager or **contact us** to submit your request. However, we encourage you to take advantage of the speed, efficiency and visibility provided by AWS Artifact Agreements.

12. IF I PREVIOUSLY SIGNED AN OFFLINE BAA WITH AWS, HOW WILL THAT BE AFFECTED BY THE ONLINE BAA AVAILABLE IN AWS ARTIFACT

AGREEMENTS?

If you previously signed an offline BAA, the terms of that BAA will continue to apply to the accounts you designated as HIPAA Accounts under that offline BAA.

For any accounts that you have not already designated as a HIPAA Account under your offline BAA, you can use AWS Artifact Agreements to accept an online BAA for those accounts.

13. IF I PREVIOUSLY SIGNED AN OFFLINE BAA WITH AWS, CAN I ACCEPT AN ONLINE BAA IN AWS ARTIFACT AGREEMENTS?

Yes. The master account in your organization can use the **Organization agreements** tab in AWS Artifact Agreements to accept an organization BAA on behalf of all existing and future member accounts in your organization.

14. IF I HAVE A PREVIOUSLY SIGNED OFFLINE BAA WITH AWS, CAN I VIEW OR DOWNLOAD THAT OFFLINE BAA IN AWS ARTIFACT AGREEMENTS?

No. In order to protect the confidentiality of your offline BAA, you will not be able to download a copy of it in AWS Artifact Agreements. If you would like to view a copy of your previously signed offline BAA, you can reach out to your AWS Account Manager to request it.

15. IF I PREVIOUSLY SIGNED AN OFFLINE BAA WITH AWS, CAN I USE AWS ARTIFACT AGREEMENTS TO DESIGNATE ADDITIONAL ACCOUNTS AS HIPAA ACCOUNTS UNDER THAT OFFLINE BAA?

No. You can use AWS Artifact Agreements to accept an online BAA for a single account or for all accounts within your organization in AWS Organizations. These will be subject to the terms of the applicable online BAA, however, **not** your offline BAA.

If you want to designate additional HIPAA Accounts under your offline BAA, you can do so by following the process described in your offline BAA (e.g., sending an email to aws-hipaa@amazon.com). Once confirmed by AWS, the Artifact Agreements interface will change for the newly designated account to reflect that it has been designated as a HIPAA Account under your offline BAA.

16. IF I HAVE AN OFFLINE BAA WITH AWS, CAN I TERMINATE MY OFFLINE BAA IN THE AWS ARTIFACT AGREEMENTS INTERFACE?

No. You can use AWS Artifact Agreements to remove an account as a HIPAA Account under your offline BAA, but it will not terminate the offline BAA itself. To terminate an offline BAA, you need to provide written notice to AWS according to the terms of your offline BAA.

17. IF I DESIGNATED AN ACCOUNT AS A HIPAA ACCOUNT UNDER A PREVIOUSLY SIGNED OFFLINE BAA, CAN I USE AWS ARTIFACT AGREEMENTS TO REMOVE THAT ACCOUNT AS A HIPAA ACCOUNT UNDER MY OFFLINE BAA?

Yes. You can follow the steps prompted within AWS Artifact to remove your account as a HIPAA Account under your offline BAA. You should only remove an account as a HIPAA Account if you are sure that you have removed all protected health information (PHI) from the account and will no longer use the account in connection with PHI.

18. I WANT TO ACCEPT AN ORGANIZATIONS BAA BUT ONLY SOME OF MY MEMBER ACCOUNTS ARE PROCESSING PHI. DO THE OBLIGATIONS OF THE BAA ONLY APPLY TO ACCOUNTS PROCESSING PHI?

By its terms, the AWS BAA only applies to “HIPAA Accounts,” which are defined as AWS accounts that store or transmit PHI, that only use HIPAA Eligible Services to store or transmit that PHI, and to which you have applied the required security configurations specified in the AWS BAA, such as encryption of PHI at rest and in transit (refer to the AWS BAA for a full list of the required security configurations). Accounts that do not meet the definition of a HIPAA Account are not subject to the AWS BAA.

AWS Australian Notifiable Data Breach Addendum (ANDB Addendum)

1. HOW DO I ACCEPT AN ANDB ADDENDUM USING AWS ARTIFACT AGREEMENTS?

AWS Artifact Agreements enables you to review and accept an ANDB Addendum from the AWS Management Console for either your AWS account or, if you are a master account in an AWS organization, your AWS organization. You can accept the ANDB Addendum for your individual AWS account under the **Account agreements** tab, or if you are a master account in an organization, you can accept the ANDB Addendum on behalf of all existing and future AWS accounts in your AWS organization under the **Organization agreements** tab. Additionally, you can use the AWS Artifact Agreements console to see which agreements are in place for your AWS account or AWS organization and review the terms of those agreements.

2. WHAT IS THE DIFFERENCE BETWEEN THE ANDB ADDENDUM THAT CAN BE ACCEPTED AS AN ACCOUNT AGREEMENT AND THE ANDB ADDENDUM THAT CAN BE ACCEPTED AS AN ORGANIZATION AGREEMENT?

The difference is that the ANDB Addendum in the **Organization agreements** tab, when accepted, applies to all existing and future AWS accounts linked to your master account through **AWS Organizations**. In comparison, the ANDB Addendum in the **Account agreements** tab only applies to the individual AWS account you used to accept the ANDB Addendum, and no other AWS accounts. If you have accepted both the account ANDB Addendum and the organizations ANDB Addendum, the organizations ANDB Addendum will apply instead of the account ANDB Addendum.

3. IF MY AWS ACCOUNT HAS ALREADY ACCEPTED AN ACCOUNT ANDB ADDENDUM, CAN I ACCEPT THE ORGANIZATIONS ANDB ADDENDUM SO THAT ALL OF MY AWS ACCOUNTS ARE COVERED?

Yes, using the master account of your organization you can use the **Organization agreements** tab in AWS Artifact Agreements to accept an ANDB Addendum on behalf of all existing and future member accounts in your organization. When both the account ANDB Addendum and organizations ANDB Addendum are accepted, the organizations ANDB Addendum will apply instead of the account ANDB Addendum.

4. HOW DO I TERMINATE AN ANDB ADDENDUM USING AWS ARTIFACT AGREEMENTS?

You can use AWS Artifact Agreements to terminate an ANDB Addendum at any time.

To terminate an account ANDB Addendum, you can use the **Account agreements** tab in AWS Artifact and click on the “Terminate the AWS Australian Notifiable Data Breach Addendum for this Account” button.

To terminate an organizations ANDB Addendum, you can use the **Organization agreements** tab in AWS Artifact and click on the “Terminate AWS Australian Notifiable Data Breach Addendum for this Organization” button.

5. WHAT HAPPENS WHEN I TERMINATE AN ANDB ADDENDUM IN AWS ARTIFACT AGREEMENTS?

If you terminate an account ANDB Addendum under the **Account agreements** tab in AWS Artifact, the AWS account you used to sign into AWS Artifact will not be covered by an ANDB Addendum with AWS, unless it is also covered by an organizations ANDB Addendum (within the **Organization agreements** tab). You should only terminate an account ANDB Addendum either when (a) you are sure that you have removed all personal information from the AWS account and you will no longer use the AWS account in connection with personal information or (b) you join that AWS account as a member account in an AWS organization that has an organizations ANDB Addendum.

If you are a user of a master account and terminate an organizations ANDB Addendum within the **Organization agreements** tab in AWS Artifact, the AWS accounts in that AWS organization will not be covered by an ANDB Addendum with AWS, unless they are covered by an account ANDB Addendum (within the **Account agreements** tab). You should only terminate an organizations ANDB Addendum either when (a) you are sure that all personal information has been removed from the AWS accounts in that AWS organization and those AWS accounts will no longer be used in connection with personal information or (b) you have agreed account ANDB Addendums for those AWS accounts that are used in connection with personal information.

6. WHICH ANDB ADDENDUM APPLIES IF MY AWS ACCOUNT HAS AN ACCEPTED ACCOUNT ANDB ADDENDUM AND ORGANIZATIONS ANDB ADDENDUM?

If you have both an account ANDB Addendum and an organizations ANDB Addendum in place at the same time, the terms of the organizations ANDB Addendum will apply instead of the terms of the account ANDB Addendum. Terminating the organizations ANDB Addendum does not terminate the account ANDB Addendum, so if you terminate the organizations ANDB Addendum, the account ANDB Addendum will then apply to that AWS account.

7. IF A MEMBER ACCOUNT LEAVES AN AWS ORGANIZATION, DOES THE ORGANIZATIONS ANDB ADDENDUM STILL APPLY TO THAT AWS ACCOUNT?

No. When a member account leaves an AWS organization, any accepted organization agreement(s), such as the organizations ANDB Addendum no longer apply to that AWS account. If the member account wants one or more of the agreements to continue to apply after leaving the organization, the member account should accept the relevant account agreement(s) under the **Account agreements** tab in AWS Artifact prior to leaving the AWS organization.

8. IF I HAVE AN ANDB ADDENDUM WITH AWS, WHAT AWS SERVICES CAN I USE?

You may use any AWS service in an AWS account covered by an ANDB Addendum with AWS.

9. I WANT TO ACCEPT AN ORGANIZATIONS ANDB ADDENDUM BUT ONLY SOME OF MY MEMBER ACCOUNTS ARE PROCESSING PERSONAL INFORMATION. DO THE OBLIGATIONS OF THE ORGANIZATIONS ANDB ADDENDUM ONLY APPLY TO AWS ACCOUNTS PROCESSING PERSONAL INFORMATION?

By its terms, the organizations ANDB Addendum only applies to “ANDB Accounts,” which are defined as AWS accounts where the entity responsible for that account is subject to the Australian Privacy Act, and that AWS account includes “personal information” (as defined in the Australian Privacy Act) in AWS’ possession or control. Accounts that do not meet the definition of an ANDB Account are not subject to the organizations ANDB Addendum.

Troubleshooting

1. I AM ATTEMPTING TO DOWNLOAD AN AGREEMENT, BUT I DON’T SEE THE DOWNLOAD APPEAR. WHAT CAN I DO NEXT?

1. Make certain that you are using the most current version of your web browser and have Adobe Reader as well.
2. Enable pop-ups for your browser so the attachment can download.
3. Check your recent downloads folder.
4. Review the document and share within your organization, as needed.

2. I AM RECEIVING AN ERROR MESSAGE, WHAT DOES IT MEAN?

Error messages are usually the result of your IAM user not having sufficient permissions to perform the desired action in AWS Artifact. Refer to the table below for a complete list of error messages and how to resolve them:

Error message in AWS Artifact console	
You don’t have the permissions to accept the agreement	You need permissions to accep Artifact. Contact your account attach the following permissio artifact:AcceptAgreement For an example IAM policy com Agreement Permissions.

You don’t have the permissions to terminate the agreement	You need permissions to termi AWS Artifact. Contact your acc attach the following permissio artifact:TerminateAgreement For an example IAM policy con Agreement Permissions.
You don’t have the permissions to download the agreement	You need permissions to down AWS Artifact. Contact your acc attach the following permissio artifact:DownloadAgreement For an example IAM policy con Agreement Permissions.
You don't have the permissions to download this report	You need permissions to down Artifact. Contact your account attach the following permissio artifact:get. For an example IAI refer to Report Permissions.
You need additional approval from AWS to access this report	The report you are trying to dc additional permission from AW Please open a request for acces Account ID you are using to do Your access request will be resp business day. If approved, you download the report using the submitted. If AWS has addition your request, you will receive a business day.
Your organization must be enabled for all features	Your organization is configurec consolidated billing. To use org agreements in AWS Artifact, yo be enabled for all features. Lea
Before you can manage agreements for your organization, you need the following permissions: organizations:EnableAWSServiceAccess and organizations:ListAWSServiceAccessForOrganization.	Contact your account administ following permission to your IA iam:CreateRole

These permissions enable AWS Artifact to access organization information in AWS Organizations.	iam:AttachRolePolicy iam:ListRoles For an example IAM policy com Agreement Permissions.
Before you can manage agreements for your organization, you need the following permissions to list, create, and attach IAM roles: iam:ListRoles, iam:CreateRole, and iam:AttachRolePolicy.	Contact your account administ following permission to your IA organizations:EnableAWSServi organizations:ListAWSServiceA For an example IAM policy com Agreement Permissions.
You don't have the permissions to retrieve information about your AWS account's organization	Contact your account administ following permission to your IA organizations:DescribeOrganiz For an example IAM policy com Agreement Permissions.
Your account isn't in an organization	You can create or join an organ the instructions in Creating and Organizations.

AWS Certificate Manager FAQs

General

Q: What is AWS Certificate Manager (ACM)?

AWS Certificate Manager is a service that lets you easily provision, manage, and deploy public and private Secure Sockets Layer/Transport Layer Security (SSL/TLS) certificates for use with AWS services and your internal connected resources. SSL/TLS certificates are used to secure network communications and establish the identity of websites over the Internet as well as resources on private networks. AWS Certificate Manager removes the time-consuming manual process of purchasing, uploading, and renewing SSL/TLS certificates. With AWS Certificate Manager, you can quickly request a certificate, deploy it on AWS resources such as Elastic Load Balancers, Amazon CloudFront distributions, and APIs on API Gateway, and let AWS Certificate Manager handle certificate renewals. It also enables you to create private certificates for your internal resources and manage the certificate lifecycle centrally. Public and private SSL/TLS certificates provisioned through AWS Certificate Manager and used exclusively with ACM-integrated services, such as Elastic Load Balancing, Amazon CloudFront, and Amazon API Gateway, are free. You pay for the AWS resources you create to run your application. You pay a monthly fee for the operation of each private CA until you delete it, and for the private certificates you issue that are not used exclusively with ACM-integrated services.

Q: What is an SSL/TLS certificate?

SSL/TLS certificates allow web browsers to identify and establish encrypted network connections to web sites using the Secure Sockets Layer/Transport Layer Security (SSL/TLS) protocol. Certificates are used within a cryptographic system known as a public key infrastructure (PKI). PKI provides a way for one party to establish the identity of another party using certificates if they both trust a third-party - known as a certificate authority. The Concepts topic in the ACM User Guide provides additional background information and definitions.

Q: What are private certificates?

Private certificates identify resources within an organization, such as applications, services, devices, and users. In establishing a secure encrypted communications channel, each endpoint uses a certificate and cryptographic techniques to prove its identity to the other endpoint. Internal API endpoints, web servers, VPN users, IoT devices, and many other applications use private certificates to establish encrypted communication channels that are necessary for their secure operation.

Q: What is the difference between public and private certificates?

Both public and private certificates help customers identify resources on networks and secure communication between these resources. Public certificates identify resources on the public Internet, whereas private certificates do the same for private networks. One key difference is that applications and browsers trust public certificates automatically by default, whereas an administrator must explicitly configure applications to trust private certificates. Public CAs, the entities that issue public certificates, must follow strict rules, provide operational visibility, and meet security standards imposed by the browser and operating system vendors that decide which CAs their browsers and operating systems trust automatically. Private CAs are managed by private organizations, and private CA administrators can make their own rules for issuing private certificates, including practices for issuing certificates and what information a certificate can include. Refer to ACM Private Certificate Authority below to learn more about private certificates and private CAs.

Q: What are the benefits of using AWS Certificate Manager (ACM) and ACM Private Certificate Authority (CA)?

ACM makes it easier to enable SSL/TLS for a website or application on the AWS platform. ACM eliminates many of the manual processes previously associated with using and managing SSL/TLS certificates. ACM can also help you avoid downtime due to misconfigured, revoked, or expired certificates by managing renewals. You get SSL/TLS protection and easy certificate management. Enabling SSL/TLS for Internet-facing sites can help improve the search rankings for your site and help you meet regulatory compliance requirements for encrypting data in transit.

When you use ACM to manage certificates, certificate private keys are securely protected and stored using strong encryption and key management best practices. ACM lets you use the AWS Management Console, AWS CLI, or AWS Certificate Manager APIs to centrally manage all of the SSL/TLS ACM certificates in an AWS Region. ACM is integrated with other AWS services, so you can request an SSL/TLS certificate and provision it with your Elastic Load Balancing load balancer or Amazon CloudFront distribution from the AWS Management Console, through AWS CLI commands, or with API calls.

ACM Private CA is a managed private CA service that helps you easily and securely manage the lifecycle of your private certificates. ACM Private CA provides you a highly-available private CA service without the upfront investment and ongoing maintenance costs of operating your own private CA. ACM Private CA extends ACM's certificate management capabilities to private certificates, enabling you to manage public and private certificates centrally. ACM Private CA allows developers to be more agile by providing them APIs to create and deploy private certificates programmatically. You also have the flexibility to create private certificates for applications that require custom certificate lifetimes or resource names. With ACM Private CA, you can create, manage, and track private certificates for your connected resources in one place with a secure, pay as you go, managed private CA service.

CA administrators can use ACM Private CA to create a complete CA hierarchy, including online root and subordinate CAs, with no need for external CAs. A CA hierarchy provides strong security and restrictive access controls for the most-trusted root CA at the top of the trust chain, while allowing more permissive access and bulk certificate issuance for subordinate CAs lower in the chain. Customers can create secure and highly available CAs without building and maintaining their own on-premises CA infrastructure.

Q: What types of certificates can I create and manage with ACM?

ACM enables you to manage the lifecycle of your public and private certificates. ACM's capabilities depend on whether the certificate is public or private, how you obtain the certificate, and where you deploy it. See ACM Public Certificates to learn more about public certificates and refer to the ACM Private CA section below to learn more about private certificates and private CAs.

Public certificates - ACM manages the renewal and deployment of public certificates used with ACM-integrated services, including Amazon CloudFront, Elastic Load Balancing, and Amazon API Gateway.

Private certificates – ACM Private CA provides three ways to create and manage private certificates. 1) You can choose to delegate private certificate management to ACM. When used in this way, ACM can automatically renew and deploy private certificates used with ACM-integrated services, including Amazon CloudFront, Elastic Load Balancing, and Amazon API Gateway. You can easily deploy these private certificates using the AWS Management console, APIs, and command-line interface (CLI). 2) You can export private certificates from ACM and use them with EC2 instances, containers, on-premises servers, and IoT devices. ACM Private CA automatically renews these certificates and sends an Amazon CloudWatch notification when the renewal is completed. You can write client-side code to download renewed certificates and private keys and deploy them with your application. 3) ACM Private CA gives you the flexibility to create your own private keys, generate a certificate signing request (CSR), issue private certificates from your ACM Private

CA, and manage the keys and certificates yourself. You are responsible for renewing and deploying these private certificates.

Imported certificates – If you want to use a third-party certificate with Amazon CloudFront, Elastic Load Balancing, or Amazon API Gateway, you may import it into ACM using the AWS Management Console, AWS CLI, or ACM APIs. ACM does not manage the renewal process for imported certificates. You are responsible for monitoring the expiration date of your imported certificates and for renewing them before they expire. You can use the AWS Management Console to monitor the expiration dates of an imported certificates and import a new third-party certificate to replace an expiring one.

CA certificates – ACM private CA can issue certificates to identify private certificate authorities. &nbsp;These certificates allow CA administrators to create a private CA hierarchy, which provides strong security and restrictive access controls for the most-trusted root CA at the top of the trust chain, while allowing more permissive access and bulk certificate issuance for subordinate CAs lower in the chain.

Q: How can I get started with ACM?

To get started with AWS Certificate Manager, navigate to Certificate Manager in the AWS Management Console and use the wizard to request an SSL/TLS certificate. If you have already created an ACM Private CA, you can choose whether you want a public or private certificate, and then enter the name of your site. See ACM Private CA and ACM Public Certificates below to determine which kind of certificate you need and to learn more about ACM Private CA. You can also request a certificate using the AWS CLI or API. After the certificate is issued, you can use it with other AWS services that are integrated with ACM. For each integrated service, you simply select the SSL/TLS certificate you want from a drop-down list in the AWS Management Console. Alternatively, you can execute an AWS CLI command or call an AWS API to associate the certificate with your resource. The integrated service then deploys the certificate to the resource you selected. For more information about requesting and using certificates provided by AWS Certificate Manager, visit Getting Started in the AWS Certificate Manager User Guide. In addition to using private certificates with ACM-integrated services, you can also use private certificates on EC2 instances, on ECS containers, or anywhere. See Private Certificates for more details.

Q: With which AWS services can I use ACM certificates?

You can use public and private ACM certificates with the following AWS services:
- Elastic Load Balancing – Refer to the Elastic Load Balancing documentation
- Amazon CloudFront – Refer to the CloudFront documentation

- Amazon API Gateway – Refer to the API Gateway documentation
- AWS Elastic Beanstalk – Refer to the AWS Elastic Beanstalk documentation
- AWS CloudFormation – Support is currently limited to public certificates that use email validation. Refer to the AWS CloudFormation documentation

In addition, you can use private certificates issued with ACM Private CA with EC2 instances, containers, IoT devices, and on your own servers.

Q: In what Regions is ACM available?

Please visit the AWS Global Infrastructure pages to see the current Region availability for AWS services. To use an ACM certificate with Amazon CloudFront, you must request or import the certificate in the US East (N. Virginia) region. ACM certificates in this region that are associated with a CloudFront distribution are distributed to all the geographic locations configured for that distribution.

ACM Private Certificate Authority (CA)

Q: What is ACM Private CA?

Private certificates are used for identifying and securing communication between connected resources on private networks such as servers, mobile and IoT devices, and applications. ACM Private CA is a managed private CA service that helps you easily and securely manage the lifecycle of your private certificates. ACM Private CA provides you a highly-available private CA service without the upfront investment and ongoing maintenance costs of operating your own private CA. ACM Private CA extends ACM's certificate management capabilities to private certificates, enabling you to create and manage public and private certificates centrally. You can easily create and deploy private certificates for your AWS resources using the AWS Management Console or the ACM API. For EC2 instances, containers, IoT devices, and on-premises resources, you can easily create and track private certificates and use your own client-side automation code to deploy them. You also have the flexibility to create private certificates and manage them yourself for applications that require custom certificate lifetimes, key algorithms, or resource names. Learn more about ACM Private CA.

Q: What are private certificates?

Private certificates identify resources within an organization, such as applications, services, devices, and users. In establishing a secure encrypted communications channel, each endpoint uses a certificate and cryptographic techniques to prove its identity to the other endpoint. Internal API endpoints, web servers, VPN users, IoT devices, and many other applications use private certificates to establish encrypted communication channels that are necessary for their secure operation.

Q: What is a private certificate authority (CA)?

A private CA handles the issuance, validation, and revocation of private certificates within a private network (i.e. not the public internet). It is comprised of two major components: The first is the CA certificate, a cryptographic building block upon which certificates can be issued. The second is a set of run-time services for maintaining revocation information through the Certificate Revocation List (CRL). When resources attempt to connect with one another, they check the CRL for the status of the certificates that each entity presents. If the certificates are valid, a handshake is accomplished between the resources which cryptographically proves the identity of each entity to the other, and creates an encrypted communication channel (TLS/SSL) between them.

Q: How are private certificates and private CAs different from public certificates and public CAs?

The components of a private CA are the same as a public CA. However, public CAs must issue and validate certificates for resources on the public Internet, whereas private CAs do the same for private networks. One key difference is that applications and browsers trust public certificates automatically by default, whereas an administrator must explicitly configure applications to trust certificates issued by private CAs. Public CAs must follow strict rules, provide operational visibility, and meet security standards imposed by the browser and operating system vendors that decide which CAs their browsers and operating systems trust automatically. Private CA administrators can make their own rules for issuing private certificates, including practices for issuing certificates and what information a certificate can include.

Q: Why do organizations use private certificates instead of public certificates?

Private certificates provide the flexibility to identify nearly anything in an organization, without disclosing the name publicly. Wiki.internal, IP address 192.168.1.1, fire-sensor-123, and user123 are examples of names that might be used in private certificates. In contrast, public certificates are strictly limited to identifying resources with public DNS names, such as www.example.com. Private certificates can include information prohibited in public certificates. Some enterprise applications have leveraged the ability to add extra information into private certificates, and could not function with public certificates.

Q: What are self-signed certificates and why should organizations use certificates from a private CA instead?

Self-signed certificates are those which are issued without a CA. Unlike certificates issued from a secure root maintained by a CA, self-signed certificates act as their own root, and as a result they have significant limitations: they can be used to provide on the wire encryption but not to verify identity, and they cannot be revoked. They are unacceptable from a security perspective, but organizations use them nonetheless because they are easy to generate, require no expertise or infrastructure, and many applications accept them. There are no controls in place for issuing self-signed certificates. Organizations that use them incur greater risk of outages caused by certificate expirations because they have no way to track expiration dates. ACM Private CA solves these problems.

Q: How can I get started with ACM Private CA?

To get started with ACM Private CA, navigate to Certificate Manager in the AWS Management Console and select Private CAs on the left side of the screen. Choose Get started to start creating a private certificate authority. Visit Getting Started in the ACM Private CA User Guide to learn more.

Q: Where can I learn more about ACM Private CA?

Refer to the ACM Private CA Detail page, ACM Private CA User Guide, ACM Private CA API Reference and ACM in the AWS CLI Reference to learn more.

Root CA Hierarchies

Q. What is a root CA?

A root CA is a cryptographic building block and root of trust upon which certificates can be issued. It is comprised of a private key for signing (issuing) certificates and a root certificate that identifies the root CA and binds the private key to the name of the CA. The root certificate is distributed to the trust stores of each entity in an environment. Administrators construct trust stores to include only the CAs they trust, and they update or build the trust stores into the operating systems, instances, and host machine images of entities in their environment. When resources attempt to connect with one another, they check the certificates that each entity presents. If the certificates are valid and a chain can be constructed from the certificate to a root certificate installed in the trust store, a “handshake” is accomplished between the resources which cryptographically proves the identity of each entity to the other, and creates an encrypted communication channel (TLS/SSL) between them.

Q: What is a CA hierarchy?

A CA hierarchy is structure for organizing certificate authorities. A CA hierarchy provides strong security and restrictive access controls for the most-trusted root CA at the top of the trust chain, while allowing more permissive access and bulk certificate issuance for subordinate CAs lower in the chain.

Q: How can I use a CA hierarchy to establish trust in private certificates?

Beneath a root CA in a CA hierarchy are subordinate CAs. A subordinate CA can either directly issue certificates, act as an intermediate CA which signs other subordinate CAs to create organizational structure, act as an issuing CA which issues end-entity certificates, or act as both an intermediate and an issuing CA. Once the root is distributed to trust stores within an organization (see “Q. What is a root CA?”), certificates for which a chain can be established to a root certificate in the trust store will also be trusted. This process is called certificate path validation. A certificate meeting this description is said to “chain up” to a trusted root.

Q: How are CAs in a hierarchy managed?

Root CAs and other CAs near the top of a CA hierarchy typically have restrictive policies controlling certificate issuance and administrative access. These CAs are used infrequently and are tightly controlled and audited, resulting in a lower risk of compromise. Therefore, they are more trusted. Root CAs typically have longer lifetimes than CAs lower in the hierarchy, commensurate with the isolation and control policies governing their use.

Q: Where does ACM Private CA fit within a CA hierarchy?

ACM Private CA allows you to create a CA hierarchy that is five levels deep, including a root CA, three levels of subordinate (intermediate) CAs, and one issuing CA. You can also include ACM Private CAs in your CA hierarchy with on-premises CAs.

Q: Do I need to have a CA hierarchy to use ACM Private CA?

No. You can issue end-entity certificates from a root CA; however, in most cases security best practices call for a CA hierarchy with at least two tiers, including a root CA as the root of trust and a subordinate CA for issuing end-entity certificates. Refer to the ACM Private CA [Creating a CA] guide for more details.

Q: Where can I learn more about ACM Private CA?

Refer to the ACM Private CA Detail page, ACM Private CA User Guide, ACM Private CA API Reference and ACM in the AWS CLI Reference to learn more.

ACM Certificates

Q: What types of certificates does ACM manage?

ACM manages public, private, and imported certificates. Refer to [Q: How can I manage certificates with ACM?] for details on ACM's management capabilities for each type of certificate.

Q: Can ACM provide certificates with multiple domain names?

Yes. Each certificate must include at least one domain name, and you can add additional names to the certificate if you want to. For example, you can add the name "www.example.net" to a certificate for "www.example.com" if users can reach your site by either name. You must own or control all of the names included in your certificate request.

Q: What is a wildcard domain name?

A wildcard domain name matches any first level subdomain or hostname in a domain. A first-level subdomain is a single domain name label that does not contain a period (dot). For example, you can use the name *.example.com to protect www.example.com, images.example.com, and any other host name or first-level subdomain that ends with .example.com. Refer to the ACM User Guide for more details.

Q: Can ACM provide certificates with wildcard domain names?

Yes.

Q: Does ACM provide certificates for anything other than SSL/TLS?

Certificates managed in ACM are intended to be used with SSL/TLS. If you issue private certificates directly from an ACM Private CA and manage the keys and certificates without using ACM for certificate management, you can configure the subject, validity period, key algorithm and signature algorithm of these private certificates and use them with SSL/TLS and other applications.

Q: Can I use ACM certificates for code signing or email encryption?

No.

Q: Does ACM provide certificates used to sign and encrypt email (S/MIME certificates)?

Not at this time.

Q: What is the validity period for ACM certificates?

Certificates issued through ACM are valid for 13 months. If you issue private certificates directly from an ACM Private CA and manage the keys and certificates without using ACM for certificate management, you can choose any validity period, including an absolute end date or a relative time that is days, months, or years from the present time.

Q: What algorithms do ACM certificates use?

Certificates managed in ACM use RSA keys with a 2048-bit modulus and SHA-256. If you issue private certificates directly from an ACM Private CA and manage the keys and certificates without using ACM for certificate management, you can also issue and use elliptic curve (ECDSA) certificates. ACM does not currently have the ability to manage these certificates.

Q: How do I revoke a certificate?

You can request ACM to revoke a public certificate by visiting the AWS Support Center and creating a case. To revoke a private certificate issued by your ACM Private CA, refer to the ACM Private CA User Guide.

Q: Can I copy a certificate between AWS Regions?

You cannot copy ACM-managed certificates between regions at this time. You can copy private certificates that you export from ACM and certificates you issue directly from your ACM Private CA without using ACM for certificate and private key management.

Q: Can I use the same ACM certificate in more than one AWS Region?

It depends on whether you're using Elastic Load Balancing or Amazon CloudFront. To use a certificate with Elastic Load Balancing for the same site (the same fully qualified domain name, or FQDN, or set of FQDNs) in a different Region, you must request a new certificate for each Region in which you plan to use it. To use an ACM certificate with Amazon CloudFront, you must request the certificate in the US East (N. Virginia) region. ACM certificates in this region that are associated with a CloudFront distribution are distributed to all the geographic locations configured for that distribution.

Q: Can I provision a certificate with ACM if I already have a certificate from another provider for the same domain name?

Yes.

Q: Can I use certificates on Amazon EC2 instances or on my own servers?

You can use private certificates issued with ACM Private CA with EC2 instances, containers, and on your own servers. At this time, public ACM certificates can be used only with specific AWS services. See With which AWS services can I use ACM certificates?

Q: Does ACM allow local language characters in domain names, otherwise known as Internationalized Domain Names (IDNs)?

ACM does not allow Unicode encoded local language characters; however, ACM allows ASCII-encoded local language characters for domain names.

Q: Which domain name label formats does ACM allow?

ACM allows only UTF-8 encoded ASCII, including labels containing “xn—”, commonly known as Punycode for domain names. ACM does not accept Unicode input (u-labels) for domain names.

ACM Public Certificates

Q: What are public certificates?

Both public and private certificates help customers identify resources on networks and secure communication between these resources. Public certificates identify resources on the Internet.

Q: What type of public certificates does ACM provide?

ACM provides Domain Validated (DV) public certificates for use with websites and applications that terminate SSL/TLS. For more details about ACM certificates, see Certificate Characteristics.

Q: Are ACM public certificates trusted by browsers, operating systems, and mobile devices?

ACM public certificates are trusted by most modern browsers, operating systems, and mobile devices. ACM-provided certificates have 99% browser and operating system ubiquity, including Windows XP SP3 and Java 6 and later.

Q: How can I confirm that my browser trusts ACM public certificates?

Browsers that trust ACM certificates display a lock icon and do not issue certificate warnings when connected to sites that use ACM certificates over SSL/TLS, for example using HTTPS.

Public ACM certificates are verified by Amazon's certificate authority (CA). Any browser, application, or OS that includes the Amazon Root CA 1, Starfield Services Root Certificate Authority - G2, or Starfield Class 2 Certification Authority trusts ACM certificates.

Q: Does ACM provide public Organizational Validation (OV) or Extended Validation (EV) certificates?

Not at this time.

Q: Where does Amazon describe its policies and practices for issuing public certificates?

They are described in the Amazon Trust Services Certificate Policies and Amazon Trust Services Certification Practices Statement documents. Refer to the Amazon Trust Services repository for the latest versions.

Q: How can I notify AWS if the information in a public certificate changes?

You notify AWS by sending email to validation-questions[at]amazon.com.

Provisioning ACM Public Certificates

Q: How can I provision a public certificate from ACM?

You can use the AWS Management Console, AWS CLI, or ACM APIs/SDKs. To use the AWS Management Console, navigate to the Certificate Manager, choose Request a certificate, select Request a public certificate, enter the domain name for your site, and follow the instructions on the screen to complete your request. You can add additional domain names to your request if users can reach your site by other names. Before ACM can issue a certificate, it validates that you own or control the domain names in your certificate request. You can choose DNS validation or email validation when requesting a certificate. With DNS validation, you write a record to the public DNS configuration for your domain to establish that you own or control the domain. After you use DNS validation once to establish control of your domain, you can obtain additional certificates and have ACM renew existing certificates for the domain as long as the record remains in place and the certificate remains in use. You do not have to validate control of the domain again. If you choose email validation instead of DNS validation, emails are sent to the domain owner requesting approval to issue the certificate. After validating that you own or control each domain name in your request, the certificate is issued and ready to be provisioned with other AWS services, such as Elastic Load Balancing or Amazon CloudFront. Refer to the ACM Documentation for details.

Q: Why does ACM validate domain ownership for public certificates?

Certificates are used to establish the identity of your site and secure connections between browsers and applications and your site. To issue a publicly trusted certificate, Amazon must validate that the certificate requestor has control over the domain name in the certificate request.

Q: How does ACM validate domain ownership before issuing a public certificate for a domain?

Prior to issuing a certificate, ACM validates that you own or control the domain names in your certificate request. You can choose DNS validation or email validation when requesting a certificate. With DNS validation, you can validate domain ownership by adding a CNAME record to your DNS configuration. Refer to DNS validation for further details. If you do not have the ability to write records to the public DNS configuration for your domain, you can use email validation instead of DNS validation. With email validation, ACM sends emails to the registered domain owner, and the owner or an authorized representative can approve issuance for each domain name in the certificate request. Refer to Email validation for further details.

Q. Which validation method should I use for my public certificate: DNS or email?

We recommend that you use DNS validation if you have the ability to change the DNS configuration for your domain. Customers who are unable to receive validation emails from ACM and those using a domain registrar that does not publish domain owner email contact information in WHOIS should use DNS validation. If you cannot modify your DNS configuration, you should use email validation.

Q. Can I convert an existing public certificate from email validation to DNS validation?

No, but you can request a new, free certificate from ACM and choose DNS validation for the new one.

Q: How long does it take for a public certificate to be issued?

The time to issue a certificate after all of the domain names in a certificate request have been validated may be several hours or longer.

Q: What happens when I request a public certificate?

ACM attempts to validate ownership or control of each domain name in your certificate request, according to the validation method you chose, DNS or email, when making the request. The status of the certificate request is Pending validation while ACM attempts to validate that you own or control the domain. Refer to the DNS validation and Email validation sections below for more information about the validation process. After all of the domain names in the certificate request are validated, the time to issue certificates may be several hours or longer. When the certificate is issued, the status of the certificate request changes to Issued and you can start using it with other AWS services that are integrated with ACM.

Q: Does ACM check DNS Certificate Authority Authorization (CAA) records before issuing public certificates?

Yes. DNS Certificate Authority Authorization (CAA) records allow domain owners to specify which certificate authorities are authorized to issue certificates for their domain. When you request an ACM Certificate, AWS Certificate Manager looks for a CAA record in the DNS zone configuration for your domain. If a CAA record is not present, then Amazon can issue a certificate for your domain. Most customers fall into this category.

If your DNS configuration contains a CAA record, that record must specify one of the following CAs before Amazon can issue a certificate for your domain: amazon.com, amazontrust.com, awstrust.com, or amazonaws.com. Refer to Configure a CAA Record or Troubleshooting CAA Problems in the AWS Certificate Manager User Guide for more information.

Q: Does ACM support any other methods for validating a domain?

Not at this time.

DNS Validation (Public Certificates)

Q. What is DNS validation?

With DNS validation, you can validate your ownership of a domain by adding a CNAME record to your DNS configuration. DNS Validation makes it easy for you to establish that you own a domain when requesting SSL/TLS certificates from ACM.

Q. What are the benefits of DNS validation?

DNS validation makes it easy to validate that you own or control a domain so that you can obtain an SSL/TLS certificate. With DNS validation, you simply write a CNAME record to your DNS configuration to establish control of your domain name. To simplify the DNS validation process, the ACM management console can configure DNS records for you if you manage your DNS records with Amazon Route 53. This makes it easy to establish control of your domain name with a few mouse clicks. Once the CNAME record is configured, ACM automatically renews certificates that are in use (associated with other AWS resources) as long as the DNS validation record remains in place. Renewals are fully automatic and touchless.

Q. Who should use DNS validation?

Anyone who requests a certificate through ACM and has the ability to change the DNS configuration for the domain they are requesting should consider using DNS validation.

Q. Does ACM still support email validation?

Yes. ACM continues to support email validation for customers who can't change their DNS configuration.

Q. What records do I need to add to my DNS configuration to validate a domain?

You must add a CNAME record for the domain you want to validate. For example, to validate the name www.example.com, you add a CNAME record to the zone for example.com. The record you add contains a random token that ACM generates specifically for your domain and your AWS account. You can obtain the two parts of the CNAME record (name and label) from ACM. For further instructions, refer to the ACM User Guide.

Q. How can I add or modify DNS records for my domain?

For more information about how to add or modify DNS records, check with your DNS provider. The Amazon Route 53 DNS documentation provides further information for customers who use Amazon Route 53 DNS.

Q. Can ACM simplify DNS validation for Amazon Route 53 DNS customers?

Yes. For customers who are using Amazon Route 53 DNS to manage DNS records, the ACM console can add records to your DNS configuration for you when you request a certificate. Your Route 53 DNS hosted zone for your domain must be configured in the same AWS account as the one you are making the request from, and you must have sufficient permissions to make a change to your Amazon Route 53 configuration. For further instructions, refer to the ACM User Guide.

Q. Does DNS Validation require me to use a specific DNS provider?

No. You can use DNS validation with any DNS provider as long as the provider allows you to add a CNAME record to your DNS configuration.

Q. How many DNS records do I need if I want more than one certificate for the same domain?

One. You can obtain multiple certificates for the same domain name in the same AWS account using one CNAME record. For example, if you make 2 certificate requests from the same AWS account for the same domain name, you need only 1 DNS CNAME record.

Q. Can I validate multiple domain names with the same CNAME record?

No. Each domain name must have a unique CNAME record.

Q. Can I validate a wildcard domain name using DNS validation?

Yes.

Q. How does ACM construct CNAME records?

DNS CNAME records have two components: a name and a label. The name component of an ACM-generated CNAME is constructed from an underscore character (_) followed by a token, which is a unique string that is tied to your AWS account and your domain name. ACM prepends the underscore and token to your domain name to construct the name component. ACM constructs the label from an underscore character prepended to a different token which is also tied to your AWS account and your domain name. ACM prepends the underscore and token to a DNS domain name used by AWS for validations: acm-validations.aws. The following examples show the formatting of CNAMEs for www.example.com, subdomain.example.com, and *.example.com.

> _TOKEN1.www.example.com   CNAME   _TOKEN2.acm-validations.aws
> _TOKEN3.subdomain.example.com CNAME   _TOKEN4.acm-validations.aws
> _TOKEN5.example.com   CNAME   _TOKEN6.acm-validations.aws

Notice that ACM removes the wildcard label (*) when generating CNAME records for wildcard names. As a result, the CNAME record generated by ACM for a wildcard name (such as *.example.com) is the same record returned for the domain name without the wildcard label (example.com).

Q. Can I validate all subdomains of a domain using one CNAME record?

No. Each domain name, including host names and subdomain names, must be validated separately, each with a unique CNAME record.

Q. Why does ACM use CNAME records for DNS validation instead of TXT records?

Using a CNAME record allows ACM to renew certificates for as long as the CNAME record exists. The CNAME record directs to a TXT record in an AWS domain (acm-validations.aws) that ACM can update as needed to validate or re-validate a domain name, without any action from you.

Q. Does DNS validation work across AWS Regions?

Yes. You can create one DNS CNAME record and use it to obtain certificates in the same AWS account in any AWS Region where ACM is offered. Configure the CNAME record once and you can get certificates issued and renewed from ACM for that name without creating another record.

Q. Can I choose different validation methods in the same certificate?

No. Each certificate can have only one validation method.

Q. How do I renew a certificate validated with DNS validation?

ACM automatically renews certificates that are in use (associated with other AWS resources) as long as the DNS validation record remains in place.

Q. Can I revoke permission to issue certificates for my domain?

Yes. Simply remove the CNAME record. ACM does not issue or renew certificates for your domain using DNS validation after you remove the CNAME record and the change is distributed through DNS. The propagation time to remove the record depends on your DNS provider.

Q. What happens if I remove the CNAME record?

ACM cannot issue or renew certificates for your domain using DNS validation if you remove the CNAME record.

Email Validation (Public Certificates)

Q: What is email validation?

With email validation, an approval request email is sent to the registered domain owner for each domain name in the certificate request. The domain owner or an authorized representative (approver) can approve the certificate request by following the instructions in the email. The instructions direct the approver to navigate to the approval website and click the link in the email or paste the link from the email into a browser to navigate to the approval web site. The approver confirms the information associated with the certificate request, such as the domain name, certificate ID (ARN), and the AWS account ID initiating the request, and approves the request if the information is accurate.

Q: When I request a certificate and choose email validation, to which email addresses is the certificate approval request sent?

When you request a certificate using email validation, a WHOIS lookup for each domain name in the certificate request is used to retrieve contact information for the domain. Email is sent to the domain registrant, administrative contact, and technical contact listed for the domain. Email is also sent to five special email addresses, which are formed by prepending admin@, administrator@, hostmaster@, webmaster@ and postmaster@ to the domain name you're requesting. For example, if you request a certificate for server.example.com, email is sent to the domain registrant, technical contact, and administrative contact using contact information returned by a WHOIS query for the example.com domain, plus admin@server.example.com, administrator@server.example.com, hostmaster@server.example.com, postmaster@server.example.com, and webmaster@server.example.com.

The five special email addresses are constructed differently for domain names that begin with "www" or wildcard names beginning with an asterisk (*). ACM removes the leading "www" or asterisk and email is sent to the administrative addresses formed by pre-pending admin@, administrator@, hostmaster@, postmaster@, and webmaster@ to the remaining portion of the domain name. For example, if you request a certificate for www.example.com, email is sent to the WHOIS contacts, as described previously, plus admin@example.com rather than admin@www.example.com. The remaining four special email addresses are similarly formed.

After you request a certificate, you can display the list of email addresses to which the email was sent for each domain using the ACM console, AWS CLI, or APIs.

Q: Can I configure the email addresses to which the certificate approval request is sent?

No, but you can configure the base domain name to which you want the validation email to be sent. The base domain name must be a superdomain of the domain name in the certificate request. For example, if you want to request a certificate for server.domain.example.com but want to direct the approval email to admin@domain.example.com, you can do so using the AWS CLI or API. See ACM CLI Reference and ACM API Reference for further details.

Q: Can I use domains that have proxy contact information (such as Privacy Guard or WhoisGuard)?

Yes; however, email delivery may be delayed as a result of the proxy. Email sent through a proxy may end up in your spam folder. Refer to the ACM User Guide for troubleshooting suggestions.

Q: Can ACM validate my identity using the technical contact for my AWS account?

No. Procedures and policies for validating the domain owner's identity are very strict, and determined by the CA/Browser Forum which sets policy standards for publicly trusted certificate authorities. To learn more, please refer to the latest Amazon Trust Services Certification Practices Statement in the Amazon Trust Services Repository.

Q: What should I do if I did not receive the approval email?

Refer to the ACM User Guide for troubleshooting suggestions.

Private Key Protection

Q: How are the private keys of ACM-provided certificates managed?

A key pair is created for each certificate provided by ACM. AWS Certificate Manager is designed to protect and manage the private keys used with SSL/TLS certificates. Strong encryption and key management best practices are used when protecting and storing private keys.

Q: Does ACM copy certificates across AWS Regions?

No. The private key of each ACM certificate is stored in the Region in which you request the certificate. For example, when you obtain a new certificate in the US East (N. Virginia) Region, ACM stores the private key in the N. Virginia Region. ACM certificates are only copied across Regions if the certificate is associated with a CloudFront distribution. In that case, CloudFront distributes the ACM certificate to the geographic locations configured for your distribution.

Q: Can I audit the use of certificate private keys?

Yes. Using AWS CloudTrail you can review logs that tell you when the private key for the certificate was used.

Billing

Q: How will I be charged and billed for my use of ACM certificates?

Public and private certificates provisioned through AWS Certificate Manager for use with ACM-integrated services, such as Elastic Load Balancing, Amazon CloudFront, and Amazon API Gateway services are free. You pay for the AWS resources you create to run your application. AWS Certificate Manager Private Certificate Authority has pay as you go pricing. You pay a monthly fee for the operation of each ACM Private CA until you delete it. You also pay for the private certificates you create and export from ACM, such as those used with EC2 or on-premises servers or the ones you issue directly from your Private CA by creating the private key yourself. Refer to Pricing page for more details and examples.

Details

Q: Can I use the same certificate with multiple Elastic Load Balancing load balancers and multiple CloudFront distributions?

Yes.

Q: Can I use public certificates for internal Elastic Load Balancing load balancers with no public internet access?

Yes, but you can also consider using ACM Private CA to issue private certificates that ACM can renew without validation. See Managed Renewal and Deployment for details about how ACM handles renewals for public certificates that are not reachable from the Internet and private certificates.

Q: Will a certificate for www.example.com also work for example.com?

No. If you want your site to be referenced by both domain names (www.example.com and example.com), you must request a certificate that includes both names.

Q: Can I import a third-party certificate and use it with AWS services?

Yes. If you want to use a third-party certificate with Amazon CloudFront, Elastic Load Balancing, or Amazon API Gateway, you may import it into ACM using the AWS Management Console, AWS CLI, or ACM APIs. ACM does not manage the renewal process for imported certificates. You can use the AWS Management Console to monitor the expiration dates of an imported certificates and import a new third-party certificate to replace an expiring one.

Q: How can ACM help my organization meet my compliance requirements?

Using ACM helps you comply with regulatory requirements by making it easy to facilitate secure connections, a common requirement across many compliance programs such as PCI, FedRAMP, and HIPAA. For specific information about compliance, please refer to http://aws.amazon.com/compliance.

Q: Does ACM have a service level agreement (SLA)?

ACM does not have an SLA. The ACM Private Certificate Authority managed private CA service has an SLA.

Q: Does ACM provide a secure site seal or trust logo that I can display on my web site?

No. If you would like to use a site seal, you can obtain one from a third-party vendor. We recommend choosing a vendor that evaluates and asserts the security of your site, or your business practices, or both.

Q: Does Amazon allow its trademarks or logo to be used as a certificate badge, site seal, or trust logo?

No. Seals and badges of this type can be copied to sites that do not use the ACM service, and used inappropriately to establish trust under false pretenses. To protect our customers and the reputation of Amazon, we do not allow our logo to be used in this manner.

Logging

Q: What logging information is available from AWS CloudTrail?

You can identify which users and accounts called AWS APIs for services that support AWS CloudTrail, the source IP address the calls were made from, and when the calls occurred. For example, you can identify which user made an API call to associate a certificate provided by ACM with an Elastic Load Balancer and when the Elastic Load Balancing service decrypted the key with a KMS API call.

Managed Renewal and Deployment

Q: What is ACM managed renewal and deployment?

ACM managed renewal and deployment manages the process of renewing SSL/TLS ACM certificates and deploying certificates after they are renewed.

Q: What are the benefits of using ACM managed renewal and deployment?

ACM can manage renewal and deployment of SSL/TLS certificates for you. ACM makes configuring and maintaining SSL/TLS for a secure web service or application more operationally sound than potentially error-prone manual processes. Managed renewal and deployment can help you avoid downtime due to expired certificates. ACM operates as a service that is integrated with other AWS services. This means you can centrally manage and deploy certificates on the AWS platform by using the AWS management console, AWS CLI, or APIs. With ACM Private CA, you can create private certificates and you can export them. ACM renews exported certificates, allowing your client side automation code to download and deploy them.

Q: Which ACM certificates can be renewed and deployed automatically?

Public Certificates

ACM can renew and deploy public ACM certificates without any additional validation from the domain owner. If a certificate cannot be renewed without additional validation, ACM manages the renewal process by validating domain ownership or control for each domain name in the certificate. After each domain name in the certificate has been validated, ACM renews the certificate and automatically deploys it with your AWS resources. If ACM cannot validate domain ownership, we will let you (the AWS account owner) know.

If you chose DNS validation in your certificate request, ACM can renew your certificate indefinitely without any further action from you, as long as the certificate is in use (associated with other AWS resources) and your CNAME record remains in place. If you selected email validation when requesting a certificate, you can improve ACM's ability to automatically renew and deploy ACM certificates, by ensuring that the certificate is in use, that all domain names included in the certificate can be resolved to your site, and that all domain names are reachable from the Internet.

Private Certificates

ACM provides three options for managing private certificates issued with ACM Private CAs. ACM provides different renewal and deployment capabilities depending on how you are managing your private certificates. You can choose the best management option for each private certificate you issue.

> 1) ACM can fully automate renewal and deployment of private certificates issued with your ACM Private CAs and used with ACM-integrated services, such as Elastic Load Balancing and API Gateway. ACM can renew and deploy private certificates that are created and managed in ACM as long as the Private CA that issued the certificate remains in the Active state.
> 2) For private certificates you export from ACM for use with on-premises resources, EC2 instances, and IoT devices, ACM Private CA renews your certificate automatically. You are responsible for retrieving the new certificate and private key and deploying them with your application.
> 3) If you issue certificates directly from ACM Private CA and manage the keys and certificates yourself without using ACM for certificate management, ACM does not renew your certificate. You are responsible for renewing and deploying these private certificates.

Q: When does ACM renew certificates?

ACM begins the renewal process up to 60 days prior to the certificate's expiration date. The validity period for ACM certificates is currently 13 months. Refer to the ACM User Guide for more information about managed renewal.

Q: Will I be notified before my certificate is renewed and the new certificate is deployed?

No. ACM may renew or rekey the certificate and replace the old one without prior notice.

Q: Can ACM renew public certificates containing bare domains, such as "example.com" (also known as zone apex or naked domains)?

If you chose DNS validation in your certificate request for a public certificate, then ACM can renew your certificate without any further action from you, as long as the certificate is in use (associated with other AWS resources) and your CNAME record remains in place.

If you selected email validation when requesting a public certificate with a bare domain, ensure that a DNS lookup of the bare domain resolves to the AWS resource that is associated with the certificate. Resolving the bare domain to an AWS resource may be challenging unless you use Route 53 or another DNS provider that supports alias resource records (or their equivalent) for mapping bare domains to AWS resources. For more information, refer to the Route 53 Developer Guide.

Q: Does my site drop existing connections when ACM deploys the renewed certificate?

No, connections established after the new certificate is deployed use the new certificate, and existing connections are not affected.

AWS CloudHSM FAQs

General

Q: What is AWS CloudHSM?

The AWS CloudHSM service helps you meet corporate, contractual, and regulatory compliance requirements for data security by using dedicated Hardware Security Module (HSM) instances within the AWS cloud. AWS and AWS Marketplace partners offer a variety of solutions for protecting sensitive data within the AWS platform, but for some applications and data subject to contractual or regulatory mandates for managing cryptographic keys, additional protection may be necessary. CloudHSM complements existing data protection solutions and allows you to protect your encryption keys within HSMs that are designed and validated to government standards for secure key management. CloudHSM allows you to securely generate, store, and manage cryptographic keys used for data encryption in a way that keys are accessible only by you.

Q: What is a Hardware Security Module (HSM)?

A Hardware Security Module (HSM) provides secure key storage and cryptographic operations within a tamper-resistant hardware device. HSMs are designed to securely store cryptographic key material and use the key material without exposing it outside the cryptographic boundary of the hardware.

Q: What can I do with CloudHSM?

You can use the CloudHSM service to support a variety of use cases and applications, such as database encryption, Digital Rights Management (DRM), Public Key Infrastructure (PKI), authentication and authorization, document signing, and transaction processing.

Q: How does CloudHSM work?

When you use the AWS CloudHSM service you create a CloudHSM Cluster. Clusters can contain multiple HSM instances, spread across multiple Availability Zones in a region. HSM instances in a cluster are automatically synchronized and load-balanced. You receive dedicated, single-tenant access to each HSM instance in your cluster. Each HSM instance appears as a network resource in your Amazon Virtual Private Cloud (VPC). Adding and removing HSMs from your Cluster is a single call to the AWS CloudHSM API (or on the command line using the AWS CLI). After creating and initializing a CloudHSM Cluster, you can configure a client on your EC2 instance that allows your applications to use the cluster over a secure, authenticated network connection.

Amazon administrators monitor the health of your HSMs, but do not have any access to configure, manage, or use them. Your applications use standard cryptographic APIs, in conjunction with HSM client software installed on the application instance, to send cryptographic requests to the HSM. The client software maintains a secure channel to all of the HSMs in your cluster and sends requests on this channel, and the HSM performs the operations and returns the results over the secure channel. The client then returns the result to the application through the cryptographic API.

Q: I don’t currently have a VPC. Can I still use AWS CloudHSM?

No. To protect and isolate your AWS CloudHSM from other Amazon customers, CloudHSM must be provisioned inside an Amazon VPC. Creating a VPC is easy. Please see the VPC Getting Started Guide for more information.

Q: Does my application need to reside in the same VPC as the CloudHSM Cluster?

No, but the server or instance on which your application and the HSM client are running must have network (IP) reachability to all HSMs in the cluster. You can establish network connectivity from your application to the HSM in many ways, including operating your application in the same VPC, with VPC peering, with a VPN connection, or with Direct Connect. Please see the VPC Peering Guide and VPC User Guide for more details.

Q: Does CloudHSM work with on-premises HSMs?

Yes. While CloudHSM does not interoperate directly with on-premises HSMs, you can securely transfer exportable keys between CloudHSM and most commercial HSMs using one of several supported RSA key wrap methods.

Q: How can my application use CloudHSM?

We have integrated and tested CloudHSM with a number of third-party software solutions such as Oracle Database 11g and 12c and Web servers including Apache and Nginx for SSL offload. Please see the CloudHSM User Guide for more information.

If you are developing your own custom application, your application can use the standard APIs supported by CloudHSM, including PKCS#11 and Java JCA/JCE (Java Cryptography Architecture/Java Cryptography Extensions). Support for Microsoft CAPI/CNG is coming soon. Please refer to the CloudHSM User Guide for code samples and help with getting started.

If you are moving an existing workload from CloudHSM Classic or on-premises HSMs to CloudHSM, our CloudHSM migration guide provides information on how to plan and execute your migration.

Q: Can I use CloudHSM to store keys or encrypt data used by other AWS services?

Yes. You can do all encryption in your CloudHSM-integrated application. In this case, AWS services such as Amazon S3 or Amazon Elastic Block Store (EBS) would only see your data encrypted.

Q: Can other AWS services use CloudHSM to store and manage keys?

AWS services integrate with AWS Key Management Service, which in turn is integrated with AWS CloudHSM through the KMS custom key store feature. If you want to use the server-side encryption offered by many AWS services (such as EBS, S3, or Amazon RDS), you can do so by configuring a custom key store in AWS KMS.

Q: Can CloudHSM be used to perform personal identification number (PIN) block translation or other cryptographic operations used with debit payment

transactions?

Currently, CloudHSM provides general-purpose HSMs. Over time we may provide payment functions. If this is of interest to you, please let us know.

Q: How does AWS Key Management Service (KMS) compare to AWS CloudHSM?

AWS Key Management Service (KMS) is a multi-tenant, managed service that allows you to use and manage encryption keys. Both services offer a high level of security for your cryptographic keys. AWS CloudHSM provides a dedicated, FIPS 140-2 Level 3 HSM under your exclusive control, directly in your Amazon Virtual Private Cloud (VPC).

Q: When should I use AWS CloudHSM instead of AWS KMS?

You should consider using AWS CloudHSM if you require:

- Keys stored in dedicated, third-party validated hardware security modules under your exclusive control.
- FIPS 140-2 compliance.
- Integration with applications using PKCS#11, Java JCE, or Microsoft CNG interfaces.
- High-performance in-VPC cryptographic acceleration (bulk crypto).

Q: Will my Safenet-based HSMs be retired?

Yes. Gemalto has announced end of life for the HSMs that CloudHSM Classic provides. You must upgrade to the new CloudHSM by April 2020. Please refer to the CloudHSM Classic FAQ for detailed information. We have developed several resources to assist you in migrating from CloudHSM Classic to the new CloudHSM. You can get started using the CloudHSM Upgrade Guide.

Q: How do I get started with CloudHSM?

You can provision a CloudHSM Cluster in the CloudHSM Console, or with a few API calls through the AWS SDK or API. To learn more, please see the CloudHSM User Guide for information about getting started, the CloudHSM

Documentation for information about the CloudHSM API, or the Tools for Amazon Web Services page for more information about the SDK.

Q: How do I terminate CloudHSM service?

You can use the CloudHSM API or SDK to delete your HSMs and stop using the service. Please refer to the CloudHSM User Guide for further instructions.

Billing

Q: How will I be charged and billed for my use of the AWS CloudHSM service?

You will be charged an hourly fee for each hour (or partial hour) that an HSM is provisioned to a CloudHSM Cluster. A cluster with no HSMs in it is not billed, nor are you billed for our automatic storage of encrypted backups. For more information, please visit the CloudHSM pricing page. Note that network data transfers to and from your CloudHSM instances are charged separately. For more information please review data transfer pricing for EC2.

Q: Is there a Free Tier for the CloudHSM service?

No, there is no free tier available for CloudHSM.

Q: Do charges vary depending on how many users or keys I create on my HSM?

No, the hourly fee, which varies by region, does not depend on how much you use your HSM.

Provisioning and operations

Q: Are there any prerequisites for signing up for CloudHSM?

Yes. In order to start using CloudHSM there are a few prerequisites, including a Virtual Private Cloud (VPC) in the region where you want CloudHSM service.

Refer to the CloudHSM User Guide for more details.

Q: Do I need to manage the firmware on my HSM?

No. AWS manages the firmware on the hardware. Firmware is maintained by a third-party, and every firmware must be evaluated by NIST for FIPS 140-2 Level 3 compliance. Only firmware that has been cryptographically signed by the FIPS key (which AWS does not have access to) can be installed.

Q: How many HSMs should I have in my CloudHSM Cluster?

AWS strongly recommends that you use at least two HSMs in two different Availability Zones for any production workload. For mission-critical workloads, we recommend at least three HSMs in at least two separate AZs. The CloudHSM client will automatically handle any HSM failures and load balance across two or more HSMs transparently to your application.

Q: Who is responsible for key durability?

AWS takes automatic encrypted backups of your CloudHSM Cluster on a daily basis, and additional backups when cluster lifecycle events occur (such as adding or removing an HSM).For the 24-hour period between backups, you are solely responsible for the durability of key material created or imported to your cluster. We strongly recommend ensuring that any keys created are synchronized to at least two HSMs in two different Availability Zones to ensure the durability of your keys. See the CloudHSM User Guide for more detail on verifying key synchronization.

Q: How do I set up a high availability (HA) configuration?

High availability is provided automatically when you have at least two HSMs in your CloudHSM Cluster. No additional configuration is required. In the event an HSM in your cluster fails, it will be replaced automatically, and all clients will be updated to reflect the new configuration without interrupting any processing. Additional HSMs can be added to the cluster via the AWS API or SDK, increasing availability without interrupting your application.

Q: How many HSM instances can be contained in a CloudHSM Cluster?

A single CloudHSM Cluster can contain up to 32 HSMs. Customers can create up to 28 instances, subject to account service limits. The remaining capacity is reserved for internal use, for example when replacing failed HSM instances.

Q: Can I back up the contents of a CloudHSM?

Your CloudHSM Cluster is backed up on a daily basis by AWS. Keys can also be exported (“wrapped”) out of your cluster and stored on-premises as long as they were not generated as “non-exportable”. No other backup options are available at this time, though we expect to provide a more comprehensive on-premises backup capability soon.

Q: Is there an SLA for CloudHSM?

Yes, you can find the service level agreement (SLA) for AWS CloudHSM here.

Security

Q: Do I share my CloudHSM with other AWS customers?

No. As part of the service you receive single-tenant access to the HSM. Underlying hardware may be shared with other customers, but the HSM is accessible only to you.

Q: How does AWS manage the HSM without having access to my encryption keys?

Separation of duties and role-based access control is inherent in the design of CloudHSM. AWS has a limited credential to the HSM that permits us to monitor and maintain the health and availability of the HSM, take encrypted backups, and to extract and publish audit logs to your CloudWatch Logs. AWS is unable to see, access or use your keys, or cause your HSM to perform any cryptographic operation using your keys.

Please see the CloudHSM User Guide for more information on the separation of duties, and the capabilities each class of user has on the HSM.

Q: Can I monitor my HSM?

Yes. CloudHSM publishes multiple CloudWatch metrics for CloudHSM Clusters and for individual HSM instances. You can use the AWS CloudWatch Console, API or SDK to obtain or alarm on these metrics.

Q: What is the 'entropy source' (source of randomness) for CloudHSM?

Each HSM has a FIPS-validated Deterministic Random Bit Generator (DRBG) that is seeded by a True Random Number Generator (TRNG) within the HSM hardware module that conforms to SP800-90B. This is a high-quality entropy source capable of producing 20Mb/sec of entropy per HSM.

Q: What happens if someone tampers with the HSM hardware?

CloudHSM has both physical and logical tamper detection and response mechanisms that trigger key deletion (zeroization) of the hardware. The hardware is designed to detect tampering if its physical barrier is breached. HSM instances are also protected against brute-force login attacks. After a fixed number of unsuccessful attempts to access an HSM with Crypto Officer (CO) credentials, the HSM instance will lock the CO out. Similarly, after a fixed number of unsuccessful attempts to access an HSM with Crypto User (CU) credentials, the user will be locked and must be unlocked by a CO.

Q: What happens in case of failure?

Amazon monitors and maintains the HSM and network for availability and error conditions. If an HSM fails or loses network connectivity, the HSM will be automatically replaced. You can check the health of an individual HSM using the CloudHSM API, SDK, or CLI Tools, and you can check the overall health of the service at any time using the AWS Service Health Dashboard.

Q: Could I lose my keys if a single HSM instance fails?

Yes. It is possible to lose keys that were created since the most recent daily backup if the CloudHSM cluster that you are using fails and you are not using two or more HSMs. Amazon strongly recommends that you use two or more

HSMs, in separate Availability Zones, in any production CloudHSM Cluster to avoid loss of cryptographic keys.

Q: Can Amazon recover my keys if I lose my credentials to my HSM?

No. Amazon does not have access to your keys or credentials and therefore has no way to recover your keys if you lose your credentials.

Q: How do I know that I can trust CloudHSM?

CloudHSM is built on hardware that is validated at Federal Information Processing Standard (FIPS) 140-2 Level 3. You can find information about the FIPS 140-2 Security Profile for the hardware used by CloudHSM, and the firmware it runs, at our compliance page.

Q: Does the CloudHSM service support FIPS 140-2 Level 3?

Yes, CloudHSM provides FIPS 140-2 Level 3 validated HSMs. You can follow the procedure in the CloudHSM User Guide under Verify the Authenticity of Your HSM to confirm that you have an authentic HSM on the same model hardware specified in the NIST Security Policy described in the previous question.

Q: How do I operate a CloudHSM in FIPS 140-2 mode?

CloudHSM is always in FIPS 140-2 mode. This can be verified by using the CLI tools as documented in the CloudHSM User Guide and running the getHsmInfo command, which will indicate the FIPS mode status.

Q: Can I get a history of all CloudHSM API calls made from my account?

Yes. AWS CloudTrail records AWS API calls for your account. The AWS API call history produced by CloudTrail lets you perform security analysis, resource change tracking, and compliance auditing. Learn more about CloudTrail at the CloudTrail home page, and turn it on via CloudTrail's AWS Management Console.

Q: Which events are not logged in CloudTrail?

CloudTrail does not include any of the HSM device or access logs. These are provided directly to your AWS account via CloudWatch Logs. See the CloudHSM User Guide for more details.

Compliance

Q: Which AWS compliance initiatives include CloudHSM?

Please refer to the AWS Compliance site for more information about which compliance programs cover CloudHSM. Unlike other AWS services, compliance requirements regarding CloudHSM are often met directly by the FIPS 140-2 Level 3 validation of the hardware itself, rather than as part of a separate audit program.

Q: Why is FIPS 140-2 Level 3 important?

FIPS 140-2 Level 3 is a requirement of certain use cases, including document signing, payments, or operating as a public Certificate Authority for SSL certificates.

Q: How can I request compliance reports that include CloudHSM in scope?

You can request compliance reports through your Business Development representative. If you don't have one, you can request one here.

Performance and capacity

Q: How many crypto operations per second can CloudHSM perform?

The performance of the individual HSMs varies based on the specific workload. The table below shows approximate single-HSM performance for several common cryptographic algorithms. You can create up to 28 HSM instances in each CloudHSM Cluster, so you can achieve up to ~28x the performance of the table listed below per cluster. Performance can vary based on exact configuration and data sizes, so we encourage load testing your application with CloudHSM to determine exact scaling needs.

RSA 2048-bit sign/verify	1,100/sec
EC P256	315 point mul/sec
AES 256	300Mb/sec full-duplex bulk encryption
2048-bit RSA Key Generation	~0.5/sec
Random Number Generation (CSPRNG)	20Mb/sec

Q: How many keys can be stored on a CloudHSM instance?

A CloudHSM cluster can store approximately 3,300 keys of any type or size.

AWS CloudHSM for Oracle TDE

Q: Does CloudHSM support Amazon RDS Oracle TDE?

No. Amazon RDS Oracle TDE is not supported; however, Oracle TDE is supported for Oracle Databases (11g and 12c) operating in EC2. See the CloudHSM User Guide for additional details. You can also use AWS Key Management Service (KMS) with Custom Key Store to secure Amazon RDS data using keys generated and stored in your AWS CloudHSM cluster.

AWS CloudHSM client, API and SDK

Q: What is the CloudHSM Client?

The CloudHSM Client is a software package supplied by AWS that allows you and your applications to interact with CloudHSM Clusters.

Q: Does the CloudHSM Client give AWS access to my CloudHSM Cluster?

No. All communication between the client and your HSM is encrypted end to end. AWS cannot see or intercept this communication, and has no visibility into your cluster access credentials.

Q: What are the CloudHSM Command Line Interface (CLI) Tools?

The CloudHSM Client comes with a set of CLI tools that allow you to administrate and use the HSM from the command line. Linux and Microsoft Windows are supported today. Support for Apple macOS is on our roadmap. These tools are available in the same package as the CloudHSM Client.

Q: How can I download and get started with the CloudHSM Command Line Interface Tools?

You'll find instructions in the CloudHSM User Guide.

Q: Do the CloudHSM CLI Tools provide AWS with access to the contents of the HSM?

No. The CloudHSM Tools communicate directly with your CloudHSM Cluster via the CloudHSM Client over a secured, mutually authenticated channel. AWS cannot observe any communication between the client, tools, and HSM, it is encrypted end-to-end.

Q: On what operating systems can I use the CloudHSM Client and CLI Tools?

Multiple Linux flavors (modern versions of Amazon Linux, Redhat, Centos, and Ubuntu) and Microsoft Windows are supported today. Support for Apple macOS is on our roadmap. Please let us know if there are other operating systems on which you would like to use the CloudHSM Client and CLI Tools.

Q: What are the network connectivity requirements for using the CloudHSM Command Line Interface Tools?

The host on which you are running the CloudHSM Client and/or using the CLI Tools must have network reachability to all of the HSMs in your CloudHSM Cluster.

Q: What can I do with the CloudHSM API & SDK?

You can create, modify, delete, and obtain the status of CloudHSM Clusters and HSMs. What you can do with the AWS CloudHSM API is limited to operations that AWS can perform with its restricted access. The API cannot access the contents of the HSM or modify any users, policies, or other settings. To learn more, please see the CloudHSM Documentation for information about the API, or the Tools for Amazon Web Services page for more information about the SDK.

Migrating to new CloudHSM

Q: How should I plan my migration to AWS CloudHSM?

Start by ensuring that the algorithms and modes you require are supported by CloudHSM. Your account manager can submit feature requests to us if needed. Next, determine your key rotation strategy. Suggestions for common use cases are in the next Q/A. We have also published an in-depth migration guide for CloudHSM. You're now ready to get started with the new CloudHSM.

Q: How can I rotate my keys?

Your rotation strategy will depend on the type of application. Common examples are below.

- Private keys for signing: Generally the private key on the HSM corresponds to an intermediate certificate, which is in turn signed by an offline enterprise root. You will rotate keys by issuing a new intermediate certificate. Create a new private key and generate the corresponding CSR using OpenSSL on CloudHSM. Next, sign the CSR with the same offline enterprise root. You may have to register this new certificate with any partners who do not automatically verify the entire certificate chain. Moving forward, you would sign all new requests (such as for documents, code or other certificates) with the new private key, corresponding to the new certificate. You can continue to verify signatures from the original private key using the corresponding public key. No revocation is necessary. This process is analogous to the process you would follow to retire or archive a signing key.

- Oracle Transparent Data Encryption: You can transfer your wallet by first switching from a hardware keystore (your original HSM) to a software keystore, and then back to a hardware keystore (the new CloudHSM).
- Symmetric key for envelope encryption: Envelope encryption refers to the key architecture where one key on the HSM encrypts/decrypts many data keys on the application host. You likely already have a key rotation process in place to go through and decrypt the data keys with the old wrapping key and re-encrypt them with the new wrapping key. The only difference during migration will be that the new wrapping key will be created and used on CloudHSM instead of your original HSM. If you do not already have a key rotation tool and process in place, you will need to create one.

Q: What if I can't rotate my keys?

Each application and use case is different. Solutions to common scenarios are discussed in the migration guide for CloudHSM. For additional questions, open a support case with details of your application, the type of HSM you are using today, the type of keys you are using, and whether these keys are exportable or not. We will help you determine an appropriate migration path.

Support and maintenance

Q: How is routine maintenance performed on HSM instances?

AWS' routine maintenance procedure for CloudHSM is designed to avoid simultaneous downtime in multiple AZs in the same region.

AWS monitors and maintains the HSM instances. We may need to remove an HSM instance from service for upgrade, replacement, or test purposes. Such operations are expected to take less than twenty minutes in the case of a replacement, and should not interfere with the performance of your CloudHSM Cluster under normal circumstances. An application that is actively using a specific HSM in the cluster when it is replaced may experience a momentary disruption while the CloudHSM Client retries the operation on a different HSM in the cluster.

AWS will not perform routine maintenance on HSMs in multiple AZs within the same region within the same 24-hour period.

In unforeseen circumstances, it is possible that AWS might perform emergency maintenance without prior notice. AWS will try to avoid this situation, as well as situations where emergency maintenance is performed within the same 24-hour period on HSMs in multiple AZs in the same region.

AWS strongly recommends that you use CloudHSM Clusters with two or more HSMs in separate Availability Zones to avoid any potential disruption.

Q: I am having a problem with CloudHSM. What do I do?

Contact AWS Support.

For questions about AWS CloudHSM Classic, please see AWS CloudHSM Classic FAQs.

Learn more about product pricing

See pricing examples and calculate your costs.

Learn more »

Sign up for a free account

Instantly get access to the AWS Free Tier.

Sign up »

Start building in the console

Get started building with AWS CloudHSM in the AWS Console.

Sign in »

AWS Directory Service FAQs

General

Q: What is AWS Directory Service?

AWS Directory Service is a managed service offering, providing directories that contain information about your organization, including users, groups, computers, and other resources. As a managed offering, AWS Directory Service is designed to reduce management tasks, thereby allowing you to focus more of your time and resources on your business. There is no need to build out your own complex, highly-available directory topology because each directory is deployed across multiple Availability Zones, and monitoring automatically detects and replaces domain controllers that fail. In addition, data replication and automated daily snapshots are configured for you. There is no software to install and AWS handles all of the patching and software updates.

Q: What can I do with AWS Directory Service?

AWS Directory Service makes it easy for you to setup and run directories in the AWS cloud, or connect your AWS resources with an existing on-premises Microsoft Active Directory. Once your directory is created, you can use it to manage users and groups, provide single sign-on to applications and services, create and apply group policy, join Amazon EC2 instances to a domain, as well as simplify the deployment and management of cloud-based Linux and Microsoft Windows workloads. AWS Directory Service enables your end users to use their existing corporate credentials when accessing AWS applications, such as Amazon WorkSpaces, Amazon WorkDocs and Amazon WorkMail, as well as directory-aware Microsoft workloads, including custom .NET and SQL Server-based applications. Finally, you can use your existing corporate credentials to administer AWS resources via AWS Identity and Access Management (IAM) role-based access to the AWS Management Console, so you do not need to build out more identity federation infrastructure.

Q: How do I create a directory?

You can use the AWS Management Console or the API to create a directory. All you need to provide is some basic information such as a fully qualified domain name (FQDN) for your directory, Administrator account name and password, and the VPC you want the directory to be attached to.

Q: Can I join an existing Amazon EC2 instance to an AWS Directory Service directory?

Yes, you can use the AWS Management Console or the API to add existing EC2 instances running Linux or Windows to a AWS Managed Microsoft AD.

Q: Are APIs supported for AWS Directory Service?

Public APIs are supported for creating and managing directories. You can now programmatically manage directories using public APIs. The APIs are available via the AWS CLI and SDK. Learn more about the APIs in the AWS Directory Service documentation.

Q: Does AWS Directory Service support CloudTrail logging?

Yes. Actions performed via the AWS Directory Service APIs or management console will be included in your CloudTrail audit logs.

Q: Can I receive notifications when the status of my directory changes?

Yes. You can configure Amazon Simple Notification Service (SNS) to receive email and text messages when the status of your AWS Directory Service changes. Amazon SNS uses topics to collect and distribute messages to subscribers. When AWS Directory Service detects a change in your directory's status, it will publish a message to the associated topic, which is then sent to topic subscribers. Visit the documentation to learn more.

Q: How much does AWS Directory Service cost?

See the pricing page for more information.

Q: Can I tag my directory?

Yes. AWS Directory Service supports cost allocation tagging. Tags make it easier for you to allocate costs and optimize spending by categorizing and grouping AWS resources. For example, you can use tags to group resources by administrator, application name, cost center, or a specific project.

Q: In which AWS regions is AWS Directory Service available?

Refer to Regional Products and Services for details of AWS Directory Service availability by region.

AWS Managed Microsoft AD

Q: How do I create a AWS Managed Microsoft AD directory?

You can launch the AWS Directory Service console from the AWS Management Console to create a AWS Managed Microsoft AD directory. Alternatively, you can use the AWS SDK or AWS CLI.

Q: How are AWS Managed Microsoft AD directories deployed?

AWS Managed Microsoft AD directories are deployed across two Availability Zones in a region by default and connected to your Amazon Virtual Private Cloud (VPC). Backups are automatically taken once per day, and the Amazon Elastic Block Store (EBS) volumes are encrypted to ensure that data is secured at rest. Domain controllers that fail are automatically replaced in the same Availability Zone using the same IP address, and a full disaster recovery can be performed using the latest backup.

Q: Can I configure the storage, CPU, or memory parameters of my AWS Managed Microsoft AD directory?

No. This functionality is not supported at this time.

Q: How do I manage users and groups for AWS Managed Microsoft AD?

You can use your existing Active Directory tools—running on Windows computers that are joined to the AWS Managed Microsoft AD domain—to manage users and groups in AWS Managed Microsoft AD directories. No special tools, policies, or behavior changes are required.

Q: How are my administrative permissions different between AWS Managed Microsoft AD and running Active Directory in my own Amazon EC2 Windows instances?

In order to deliver a managed-service experience, AWS Managed Microsoft AD must disallow operations by customers that would interfere with managing the service. Therefore, AWS does not provide Windows PowerShell access to directory instances, and it restricts access to directory objects, roles, and groups that require elevated privileges. AWS Managed Microsoft AD does not allow direct host access to domain controllers via Telnet, Secure Shell (SSH), or Windows Remote Desktop Connection. When you create an AWS Managed Microsoft AD directory, you are assigned an organizational unit (OU) and an administrative account with delegated administrative rights for the OU. You can create user accounts, groups, and policies within the OU by using standard Remote Server Administration Tools such as Active Directory Users and Groups.

Q: Can I use Microsoft Network Policy Server (NPS) with AWS Managed Microsoft AD?

Yes. The administrative account created for you when AWS Managed Microsoft AD is set up has delegated management rights over the Remote Access Service (RAS) and Internet Authentication Service (IAS) security group. This enables you to register NPS with AWS Managed Microsoft AD and manage network access policies for accounts in your domain.

Q: Does AWS Managed Microsoft AD support schema extensions?

Yes. AWS Managed Microsoft AD supports schema extensions that you submit to the service in the form of a LDAP Data Interchange Format (LDIF) file. You may extend but not modify the core Active Directory schema.

Q: Which applications are compatible with AWS Managed Microsoft AD?

Amazon Chime
Amazon Connect
Amazon EC2
Amazon RDS for SQL Server
Amazon QuickSight
Amazon WorkDocs
Amazon WorkMail
Amazon WorkSpaces
AWS Management Console
Active Directory Federation Services (AD FS)
Application Server (.NET)
Azure Active Directory (AD) Connect
Enterprise Certificate Authority
Remote Desktop Licensing Manager
SharePoint Server
SQL Server

Note that not all configurations of these applications may be supported.

Q: Can I migrate my existing, on-premises Microsoft Active Directory to AWS Managed Microsoft AD?

AWS does not provide any migration tools to migrate a self-managed Active Directory to AWS Managed Microsoft AD. You must establish a strategy for performing migration including password resets, and implement the plans using Remote Server Administration Tools.

Q: Can I configure conditional forwarders and trusts in the Directory Service console?

Yes. You can configure conditional forwarders and trusts for AWS Managed Microsoft AD using the Directory Service consoe as well as the API.

Q: Can I add additional domain controllers manually to my AWS Managed Microsoft AD?

Yes. You can add additional domain controllers to your managed domain using the AWS Directory Service console or API. Note that promoting Amazon EC2 instances to domain controllers manually is not supported.

Q: Can I use Microsoft Office 365 with user accounts managed in AWS Managed Microsoft AD?

Yes. You can synchronize identities from AWS Managed Microsoft AD to Azure AD using Azure AD Connect and use Microsoft Active Directory Federation Services (AD FS) for Windows 2016 with AWS Managed Microsoft AD to authenticate Office 365 users. For step-by-step instructions, see How to Enable Your Users to Access Office 365 with AWS Microsoft Active Directory Credentials.

Q: Can I use Security Assertion Markup Language (SAML) 2.0–based authentication with cloud applications using AWS Managed Microsoft AD?

Yes. You can use Microsoft Active Directory Federation Services (AD FS) for Windows 2016 with your AWS Managed Microsoft AD managed domain to authenticate users to cloud applications that support SAML.

Q: Can I encrypt communication between my applications and AWS Managed Microsoft AD using LDAPS?

Yes. AWS Managed Microsoft AD supports Lightweight Directory Access Protocol (LDAP) over Secure Socket Layer (SSL) / Transport Layer Security (TLS), also known as LDAPS, in both client and server roles. When acting as a server, AWS Managed Microsoft AD supports LDAPS over ports 636 (SSL) and 389 (TLS). You enable server-side LDAPS communication by installing a certificate on your AWS Managed Microsoft AD domain controllers from an AWS-based Active Directory Certificate Services certificate authority (CA). To learn more, see Enable Secure LDAP (LDAPS).

Q: Can I encrypt LDAP communications between AWS applications and my self-managed AD using AWS Managed Microsoft AD?

Yes. AWS Managed Microsoft AD supports Lightweight Directory Access Protocol (LDAP) over Secure Socket Layer (SSL) / Transport Layer Security (TLS), also known as LDAPS, in both client and server roles. When acting as a client, AWS Managed Microsoft AD supports LDAPS over ports 636 (SSL). You enable client-side LDAPS communication by registering certification authority (CA) certificates from your server certificate issuer into AWS. To learn more, see Enable Secure LDAP (LDAPS).

Q: How does AWS Managed Microsoft AD address Microsoft advisory ADV190023, which describes changes to default LDAP security settings on AD domain controllers?

AWS Managed Microsoft AD supports both LDAP signing and LDAP over SSL/TLS (LDAPS) when acting as LDAP clients communicating with self-managed Active Directory. Client-side LDAP signing requires no customer action to enable, and provides data integrity. Client-side LDAPS requires configuration, and provides data integrity and confidentiality. For more information, see this AWS Forums post.

Q: How many users, groups, computers, and total objects does AWS Managed Microsoft AD support?

AWS Managed Microsoft AD (Standard Edition) includes 1 GB of directory object storage. This capacity can support up to 5,000 users or 30,000 directory objects, including users, groups, and computers. AWS Managed Microsoft AD (Enterprise Edition) includes 17 GB of directory object storage, which can support up to 100,000 users or 500,000 objects.

Q: Can I use AWS Managed Microsoft AD as a primary directory?

Yes. You can use it as a primary directory to manage users, groups, computers, and Group Policy objects (GPOs) in the cloud. You can manage access and provide single sign-on (SSO) to AWS applications and services, and to third-party directory-aware applications running on Amazon EC2 instances in the AWS Cloud. In addition, you can use Azure AD Connect and AD FS to support SSO to cloud applications, including Office 365.

Q: Can I use AWS Managed Microsoft AD as a resource forest?

Yes. You can use AWS Managed Microsoft AD as a resource forest that contains primarily computers and groups with trust relationships to your on-premises directory. This enables your users to access AWS applications and resources with their on-premises AD credentials.

Seamless domain join

Q: What is seamless domain join?

Seamless domain join is a feature that allows you to join your Amazon EC2 for Windows Server instances seamlessly to a domain, at the time of launch and from the AWS Management Console. You can join instances to AWS Managed Microsoft AD that you launch in the AWS Cloud.

Q: How do I join an instance seamlessly to a domain?

When you create and launch an EC2 for Windows instance from the AWS Management Console, you have the option to select which domain your instance will join. To learn more, see the documentation.

Q: Can I join existing EC2 for Windows Server instances seamlessly to a domain?

You cannot use the seamless domain join feature from the AWS Management Console for existing EC2 for Windows Server instances, but you can join existing instances to a domain using the EC2 API or by using PowerShell on the instance. To learn more, see the documentation.

IAM integration

Q: How does AWS Directory Service enable single sign-on (SSO) to the AWS Management Console?

AWS Directory Service allows you to assign IAM roles to AWS Manage Microsoft AD or Simple AD users and groups in the AWS cloud, as well as an existing, on-premises Microsoft Active Directory users and groups using AD Connector. These roles will control users' access to AWS services based on IAM policies assigned to the roles. AWS Directory Service will provide a customer-specific URL for the AWS Management Console which users can use to sign in with their existing corporate credentials. See our documentation for more information on this feature.

Compliance

Q: Can I use AWS Managed Microsoft AD for AWS Cloud workloads that are subject to compliance standards?

Yes. AWS Managed Microsoft AD has implemented the controls necessary to enable you to meet the U.S. Health Insurance Portability and Accountability Act (HIPAA) requirements and is included as an in-scope service in the Payment Card Industry Data Security Standard (PCI DSS) Attestation of Compliance and Responsibility Summary.

Q: How can I access compliance and security reports?

To access a comprehensive list of documents relevant to compliance and security in the AWS Cloud, see AWS Artifact.

Q: What is the AWS Shared Responsibility Model?

Security, including HIPAA and PCI DSS compliance, is a shared responsibility between AWS and you. For example, it is your responsibility to configure your AWS Managed Microsoft AD password policies to meet PCI DSS requirements when using AWS Managed Microsoft AD. To learn more about the actions you may need to take to meet HIPAA and PCI DSS compliance requirements, see the compliance documentation for AWS Managed Microsoft AD, read the Architecting for HIPPA Security and Compliance on Amazon Web Services whitepaper, and see the AWS Cloud Compliance, HIPAA Compliance, and PCI DSS Compliance.

For questions about AD Connector or Simple AD, please see AWS Directory Service, Other Directory Options.

AWS Firewall Manager FAQs

General

What is AWS Firewall Manager?

What are the key benefits of AWS Firewall Manager?

What does AWS Firewall Manager configure?

Does AWS Firewall Manager configure VPC Security Groups or Network ACLs?

Which AWS resources can AWS Firewall Manager configure rules on?

How much does AWS Firewall Manager cost?

In which regions is AWS Firewall Manager available?

Enabling AWS Firewall Manager

What are the prerequisites for AWS Firewall Manager?

How do I use AWS Firewall Manager?

Can I create a Firewall Manager policy but not remediate automatically?

How many accounts can AWS Firewall Manager manage?

How many resources can AWS Firewall Manager manage?

Can I create protection policies across regions?

Can I exclude accounts or resources from the scope of the policy?

Dashboard and Visibility

How can I view the compliance status to a particular policy?

Does AWS Firewall Manager provide notifications when a resource is non-compliant?

How can I view all threats across my organization?

AWS Key Management Service FAQs

The following FAQ does not apply to AWS KMS in the AWS China (Beijing) Region, operated by Sinnet and the AWS China (Ningxia) Region, operated by NWCD. Please visit this FAQ link for content relevant to these two China regions.

General

Q: What is AWS Key Management Service (KMS)?
AWS KMS is a managed service that enables you to easily create and control the keys used for cryptographic operations. The service provides a highly available key generation, storage, management, and auditing solution for you to encrypt or digitally sign data within your own applications or control the encryption of data across AWS services.

Q: Why should I use AWS KMS?
If you are responsible for securing your data across AWS services, you should use it to centrally manage the encryption keys that control access to your data. If you are a developer who needs to encrypt data in your applications, you should use the AWS Encryption SDK with AWS KMS to easily generate, use and protect symmetric encryption keys in your code. If you are a developer who needs to digitally sign or verify data using asymmetric keys, you should use the service to create and manage the private keys you'll need. If you're looking for a scalable key management infrastructure to support your developers and their growing number of applications, you should use it to reduce your licensing costs and operational burden. If you're responsible for proving data security for regulatory or compliance purposes, you should use it because it facilitates proving your data is consistently protected. It's also in scope for a broad set of industry and regional compliance regimes.

Q: How do I get started with AWS KMS?
The easiest way is to get started using the service is to choose to encrypt your data within supported AWS services using AWS managed master keys that are automatically created in your account for each service. If you want full control over the management of your keys, including the ability to share access to keys across accounts or services, you can create your own customer master keys (CMKs) in AWS KMS. You can also use the CMKs that you create directly within your own applications. AWS KMS can be accessed from the KMS console that is grouped under Security, Identity and Compliance on the AWS Services home page of the AWS Console. AWS KMS APIs can also be accessed directly through the AWS KMS Command Line Interface or AWS SDK for programmatic access. AWS KMS APIs can also be used indirectly to encrypt data within your own applications by using the AWS Encryption SDK. Visit the Getting Started page to learn more.

Q: In what Regions is AWS KMS available?
Availability is listed on our global Products and Services by Region page.

Asymmetric keys are currently only available in Northern Virginia, Oregon, Sydney, Ireland, and Tokyo.

Q: What key management features are available in AWS KMS?
You can perform the following key management functions:

- Create symmetric and asymmetric keys where the key material is only ever used within the service
- Create symmetric keys where the key material is generated and used within a custom key store under your control*
- Import your own symmetric key material for use within the service
- Create both symmetric and asymmetric data key pairs for local use within your applications
- Define which IAM users and roles can manage keys
- Define which IAM users and roles can use keys to encrypt and decrypt data
- Choose to have keys that were generated by the service to be automatically rotated on an annual basis
- Temporarily disable keys so they cannot be used by anyone
- Re-enable disabled keys
- Schedule the deletion of keys that you no longer use

- Audit use of keys by inspecting logs in AWS CloudTrail

* The use of custom key stores requires CloudHSM resources to be available in your account.

Q: How does AWS KMS work?
You start using the service by requesting the creation of a CMK. You control the lifecycle of the CMK as well as who can use or manage it. The key material for a CMK is generated within hardware security modules (HSMs) managed by AWS KMS. Alternatively, you can import key material from your own key management infrastructure and associate it with a CMK. You can also have the key material generated and used in an AWS CloudHSM cluster as a part of the custom key store feature in AWS KMS.

Once you have created a CMK using any of the three supported options, you can submit data directly to the service AWS KMS to be signed, verified, encrypted, or decrypted using these CMK. You set usage policies on these keys that determine which users can perform which actions under which conditions.

AWS services and client-side toolkits that integrate with AWS KMS use a method known as envelope encryption to protect your data. Under this method, AWS KMS generates data keys which are used to encrypt data locally in the AWS service or your application. The data keys are themselves encrypted under a CMK you define. Data keys are not retained or managed by AWS KMS. AWS services encrypt your data and store an encrypted copy of the data key along with the encrypted data. When a service needs to decrypt your data, it requests AWS KMS to decrypt the data key using your CMK. If the user requesting data from the AWS service is authorized to decrypt under your CMK, the AWS service will receive the decrypted data key from AWS KMS. The AWS service then decrypts your data and returns it in plaintext. All requests to use your CMKs are logged in AWS CloudTrail so you can understand who used which key under what context and when they used it.

Q: Where is my data encrypted if I use AWS KMS?
There are typically three scenarios for how data is encrypted using AWS KMS. Firstly, you can use AWS KMS APIs directly to encrypt and decrypt data using your CMKs stored in the service. Secondly, you can choose to have AWS services encrypt your data using your CMKs stored in the service. In this case data is

encrypted using data keys that are protected by your CMKs. Thirdly, you can use the AWS Encryption SDK that is integrated with AWS KMS to perform encryption within your own applications, whether they operate in AWS or not.

Q: Which AWS cloud services are integrated with AWS KMS?
AWS KMS is seamlessly integrated with most other AWS services to make encrypting data in those services as easy as checking a box. In some cases data is encrypted by default using keys that are stored in AWS KMS but owned and managed by the AWS service in question. In many cases the CMKs are owned and managed by you within your account. Some services give you the choice of managing the keys yourself or allowing the service to manage the keys on your behalf. See the list of AWS services currently integrated with AWS KMS. See the AWS KMS Developer's Guide for more information on how integrated services use AWS KMS.

Q: Why use envelope encryption? Why not just send data to AWS KMS to encrypt directly?
While AWS KMS does support sending data up to 4 KB to be encrypted directly, envelope encryption can offer significant performance benefits. When you encrypt data directly with AWS KMS it must be transferred over the network. Envelope encryption reduces the network load since only the request and delivery of the much smaller data key go over the network. The data key is used locally in your application or encrypting AWS service, avoiding the need to send the entire block of data to AWS KMS and suffer network latency.

Q: What's the difference between a CMK I create and CMKs created automatically for me by other AWS services?
You have the option of selecting a specific CMK to use when you want an AWS service to encrypt data on your behalf. These are known as customer managed CMKs and you have full control over them. You define the access control and usage policy for each key and you can grant permissions to other accounts and services to use them. If you don't specify a CMK, the service in question will create an AWS managed CMK the first time you try to create an encrypted resource within that service. AWS will manage the policies associated with AWS managed CMKs on your behalf. You can track AWS managed keys in your account and all usage is logged in AWS CloudTrail, but you have no direct control over the keys themselves.

Q: Why should I create my own customer master keys?
Creating your own CMK gives you more control than you have with AWS managed CMKs. When you create a symmetric customer managed CMK, you can choose to use key material generated by AWS KMS, generated within an AWS CloudHSM cluster (custom key store), or import your own key material. You can define an alias and description for the key and opt-in to have the key automatically rotated once per year if it was generated by AWS KMS. You also define all the permissions on the key to control who can use or manage the key. With asymmetric customer managed CMKs, there are a couple of caveats to management: the key material can only be generated within AWS KMS HSMs and there is no option for automatic key rotation.

Q: Can I bring my own keys to AWS KMS?
Yes. You can import a copy of your key from your own key management infrastructure to AWS KMS and use it with any integrated AWS service or from within your own applications. You cannot import asymmetric CMKs into AWS KMS.

Q: When would I use an imported key?
You can use an imported key to get greater control over the creation, lifecycle management, and durability of your key in AWS KMS. Imported keys are designed to help you meet your compliance requirements which may include the ability to generate or maintain a secure copy of the key in your infrastructure, and the ability to immediately delete the imported copy of the key from AWS infrastructure.

Q: What type of keys can I import?
You can import 256-bit symmetric keys.

Q: How is the key that I import into AWS KMS protected in transit?
During the import process, your key must be wrapped by an AWS KMS-provided public key using one of two RSA PKCS#1 schemes. This ensures that your encrypted key can only be decrypted by AWS KMS.

Q: What's the difference between a key I import and a key I generate in AWS KMS?
There are two main differences:

1. You are responsible for maintaining a copy of your imported keys in your key management infrastructure so that you can re-import them at any time. AWS, however, ensures the availability, security, and durability of keys generated by AWS KMS on your behalf until you schedule the keys for deletion.
2. You may set an expiration period for an imported key. AWS KMS will automatically delete the key material after the expiration period. You may also delete imported key material on demand. In both cases the key material itself is deleted but the CMK reference in AWS KMS and associated metadata are retained so that the key material can be re-imported in the future. Keys generated by AWS KMS do not have an expiration time and cannot be deleted immediately; there is a mandatory 7 to 30 day wait period. All customer managed CMKs, irrespective of whether the key material was imported, can be manually disabled or scheduled for deletion. In this case the CMK itself is deleted, not just the underlying key material.

Q: Can I rotate my keys?
Yes. You can choose to have AWS KMS automatically rotate CMKs every year, provided that those keys were generated within AWS KMS HSMs. Automatic key rotation is not supported for imported keys, asymmetric keys, or keys generated in an AWS CloudHSM cluster using the AWS KMS custom key store feature. If you choose to import keys to AWS KMS or asymmetric keys or use a custom key store, you can manually rotate them by creating a new CMK and mapping an existing key alias from the old CMK to the new CMK.

Q: Do I have to re-encrypt my data after keys in AWS KMS are rotated?
If you choose to have AWS KMS automatically rotate keys, you don’t have to re-encrypt your data. AWS KMS automatically keeps previous versions of keys to use for decryption of data encrypted under an old version of a key. All new encryption requests against a key in AWS KMS are encrypted under the newest version of the key.

If you manually rotate your imported or custom key store keys, you may have to re-encrypt your data depending on whether you decide to keep old versions of keys available.

Q: Can I delete a key from AWS KMS?
Yes. You can schedule a customer master key and associated metadata that you created in AWS KMS for deletion, with a configurable waiting period from 7 to 30 days. This waiting period allows you to verify the impact of deleting a key on your applications and users that depend on it. The default waiting period is 30 days. You can cancel key deletion during the waiting period. The key cannot be used if it is scheduled for deletion until you cancel the deletion during the waiting period. The key gets deleted at the end of the configurable waiting period if you don’t cancel the deletion. Once a key is deleted, you can no longer use it. All data protected under a deleted master key is inaccessible.

For customer master keys with imported key material, you can delete the key material without deleting the customer master key id or metadata in two ways. First, you can delete your imported key material on demand without a waiting period. Second, at the time of importing the key material into the customer master key, you may define an expiration time for how long AWS can use your imported key material before it is deleted. You can re-import your key material into the customer master key if you need to use it again.

Q: What should I do if my imported key material has expired or I accidentally deleted it?
You can re-import your copy of the key material with a valid expiration period to AWS KMS under the original customer master key so it can be used.

Q: Can I be alerted that I need to re-import the key?
Yes. Once you import your key to a customer master key, you will receive an Amazon CloudWatch Metric every few minutes that counts down the time to expiration of the imported key. You will also receive an Amazon CloudWatch Event once the imported key under your customer master key expires. You can build logic that acts on these metrics or events and automatically re-imports the key with a new expiration period to avoid an availability risk.

Q: Can I use AWS KMS to help manage encryption of data outside of AWS cloud services?
Yes. AWS KMS is supported in AWS SDKs, AWS Encryption SDK, the Amazon DynamoDB Client-side Encryption, and the Amazon S3 Encryption Client to facilitate encryption of data within your own applications wherever they run.

Visit the AWS Crypto Tools and Developing on AWS website for more information.

Q: Is there a limit to the number of keys I can create in AWS KMS?
You can create up to 10000 CMKs per account per region. As both enabled and disabled CMKs count towards the limit, we recommend deleting disabled keys that you no longer use. AWS managed CMKs created on your behalf for use within supported AWS services do not count against this limit. There is no limit to the number of data keys that can be derived using a CMK and used in your application or by AWS services to encrypt data on your behalf. You may request a limit increase for CMKs by visiting the AWS Support Center.

Q: What types of symmetric key types and algorithms are supported?
AWS KMS supports 256-bit keys when creating a CMK. Generated data keys returned to the caller can be 256-bit, 128-bit, or an arbitrary value up to 1024-bits. When AWS KMS uses a 256-bit CMK on your behalf, the AES algorithm in Galois Counter Mode (AES-GCM) is used.

Q: What kind of asymmetric key types are supported?
AWS KMS supports the following asymmetric key types - RSA 2048, RSA 3072, RSA 4096, ECC NIST P-256, ECC NIST P-384, ECC NIST-521, and ECC SECG P-256k1.

Q: What kinds of asymmetric encryption algorithms are supported?
AWS KMS supports the RSAES_OAEP_SHA_1 and RSAES_OAEP_SHA_256 encryption algorithms with RSA 2048, RSA 3072, and RSA 4096 key types. Encryption algorithms cannot be used with the elliptic curve key types (ECC NIST P-256, ECC NIST P-384, ECC NIST-521, and ECC SECG P-256k1).

Q: What kinds of asymmetric signing algorithms are supported?
When using RSA key types, AWS KMS supports the RSASSA_PSS_SHA_256, RSASSA_PSS_SHA_384, RSASSA_PSS_SHA_512, RSASSA_PKCS1_V1_5_SHA_256, RSASSA_PKCS1_V1_5_SHA_384, and RSASSA_PKCS1_V1_5_SHA_512 signing algorithms.
When using elliptic curve key types, AWS KMS supports the ECDSA_SHA_256, ECDSA_SHA_384, and ECDSA_SHA_512 signing algorithms.

Q: Can symmetric CMKs be exported out of the service in plain text?
No. A symmetric CMK or the private portion of an asymmetric CMK cannot be exported in plain text from the HSMs. The public portion of an asymmetric CMK can be exported from the console or by calling the “GetPublicKey” API.

Q: Can data keys and data key pairs be exported out of the HSMs in plain text?
Yes. The symmetric data keys can be exported using either the “GenerateDataKey” API or the “GenerateDataKeyWithoutPlaintext” API. And the private and public portion of asymmetric data key pairs can both be exported out of AWS KMS using either the “GenerateDataKeyPair” API or the “GenerateDataKeypairWithoutPlaintext” API.

Q: How are data keys and data key pairs protected for storage outside the service?
The symmetric data key or the private portion of the asymmetric data key is encrypted under the symmetric CMK you define when you request AWS KMS to generate the data key.

Q: How do I use the public portion of an asymmetric CMK?
The public portion of the asymmetric key material is generated in AWS KMS and can be used for digital signature verification by calling the “Verify” API, or for public key encryption by calling the “Encrypt” API. The public key can also be used outside of AWS KMS for verification or encryption. You can call the GetPublicKey API to retrieve the public portion of the asymmetric CMK.

Q: What is the size limit for data sent to AWS KMS for asymmetric operations?
The size limit is 4KB. If you want to digitally sign data larger than 4KB, you have the option to create a message digest of the data and send it to AWS KMS. The digital signature is created over the digest of the data and returned. You specify whether you are sending the full message or a message digest as a parameter in the Sign API request. Any data submitted to the Encrypt, Decrypt, or Re-Encrypt APIs that require use of asymmetric operations must also be less than 4KB.

Q: How can I distinguish between asymmetric or symmetric CMKs I have created?
In the console, each key will have a new field called “Key Type”. It will have

either the value “Asymmetric Key” or “Symmetric Key” to indicate the type of key. The “DescribeKey” API will return a “KeyUsage” field that will specify if the key can be used to sign or encrypt.

Q: Is automatic rotation of asymmetric CMKs supported?
No. Automatic key rotation is not supported for asymmetric CMKs. You can manually rotate them by creating a new CMK and mapping an existing key alias from the old CMK to the new CMK.

Q: Can a single asymmetric CMK be used for both encryption and signing?
No. When creating a CMK, you must specify whether the key can be used for decrypt or sign operations. An RSA key type can be used for signing or encryption operations, but not both. Elliptic curve key types can only be used for signing operations.

Q: Are there service limits related to asymmetric keys?
Yes. The request per second rate limits are different for different key types and algorithms. Please refer to the AWS KMS limits page for details.

Q: Do asymmetric keys work with AWS KMS custom key stores or the Import Key feature?
No. You cannot use the custom key store functionality with asymmetric keys nor can you import asymmetric keys into AWS KMS.

Q: Can I use asymmetric CMKs for digital signing applications that require digital certificates?
Not directly. AWS KMS doesn’t store or associate digital certificates with asymmetric CMKs it creates. You could choose to have a certificate authority such as ACM PCA issue a certificate for the public portion of your asymmetric CMK. This will allow the entities that are consuming your public key to verify that the public key indeed belongs to you.

Q: For what use scenarios should I use ACM Private Certificate Authority vs. AWS KMS?
The primary reason to use the ACM Private Certificate Authority (CA) service is to provide a public key infrastructure (PKI) for the purpose of identifying entities and securing network connections. PKI provides processes and mechanisms, primarily using X.509 certificates, to put structure around public

key cryptographic operations. Certificates provide an association between an identity and a public key. The certification process in which a certificate authority issues a certificate allows the trusted certificate authority to assert the identity of another entity by signing a certificate. PKI provides identity, distributed trust, key lifecycle management, and certificate status vended through revocation. These functions add important processes and infrastructure to the underlying asymmetric cryptographic keys and algorithms provided by AWS KMS.

ACM Private CA allows you to issue certificates to identify web and application servers, service meshes, VPN users, internal API endpoints, and IoT devices. Certificates let you establish the identity of these resources and create encrypted TLS/SSL communications channels. If you are considering using asymmetric keys for TLS termination on web or application servers, Elastic Load Balancers, API Gateway endpoints, EC2 instances or containers, you should consider using ACM Private CA for issuing certificates and providing a PKI infrastructure.

In contrast, AWS KMS lets you generate, manage, and use asymmetric keys for digital signing and/or encryption operations that don't require certificates. While certificates can enable verification of sender and recipient identity between untrusted parties, the kind of raw asymmetric operations offered by AWS KMS are typically useful when you have other mechanisms to prove identity or don't need to prove it at all to get the security benefit you desire.

Q: Can I use my applications' cryptographic API providers such as OpenSSL, JCE, Bouncy Castle, or CNG with AWS KMS?
There is no native integration offered by AWS KMS for any other cryptographic API providers. You must use AWS KMS APIs directly or through the AWS SDK to integrate signing and encryption capabilities into your applications.

Q: Does AWS KMS offer a Service Level Agreement (SLA)?
Yes. The AWS KMS SLA provides for a service credit if a customer's monthly uptime percentage is below our service commitment in any billing cycle.

Custom Key Store

Q: What is a custom key store?
The AWS KMS custom key store feature combines the controls provided by AWS CloudHSM with the integration and ease of use of AWS KMS. You can configure your own CloudHSM cluster and authorize AWS KMS to use it as a dedicated key store for your keys rather than the default AWS KMS key store. When you create keys in AWS KMS you can chose to generate the key material in your CloudHSM cluster. CMKs that are generated in your custom key store never leave the HSMs in the CloudHSM cluster in plaintext and all AWS KMS operations that use those keys are only performed in your HSMs. In all other respects CMKs stored in your custom key store are consistent with other AWS KMS CMKs.

Additional guidance for deciding if using a custom key store it is right for you can be found in this blog.

Q: Why would I need to use a custom key store?
Since you control your AWS CloudHSM cluster, you have the option to manage the lifecycle of your CMKs independently of AWS KMS. There are four reasons why you might find a custom key store useful. Firstly, you might have keys that are explicitly required to be protected in a single tenant HSM or in an HSM over which you have direct control. Secondly, you might have keys that are required to be stored in an HSM that has been validated to FIPS 140-2 level 3 overall (the HSMs used in the standard AWS KMS key store are either validated or in the process of being validated to level 2 with level 3 in multiple categories). Thirdly, you might need the ability to immediately remove key material from AWS KMS and to prove you have done so by independent means. Finally, you might have a requirement to be able to audit all use of your keys independently of AWS KMS or AWS CloudTrail.

Q: Do custom key stores affect how keys are managed?
There are two differences when managing keys in a custom key store compared to the default AWS KMS key store. You cannot import key material into your custom key store and you cannot have AWS KMS automatically rotate keys. In all other respects, including the type of keys that can be generated, the way that keys use aliases and how policies are defined, keys that are stored in a custom key store are managed in the same way as any other AWS KMS customer managed CMK.

Q: Can I use a custom key store to store an AWS managed customer master key?
No, only customer managed CMKs can be stored and managed in an AWS KMS custom key store. AWS managed CMKs that are created on your behalf by other AWS services to encrypt your data are always generated and stored in the AWS KMS default key store.

Q: Do custom key stores affect how keys are used?
No, API requests to AWS KMS to use a CMK to encrypt and decrypt data are handled in the same way. Authentication and authorization processes operate independently of where the key is stored. All activity using a key in a custom key store is also logged to AWS CloudTrail in the same way. However, the actual cryptographic operations happen exclusively in either the custom key store or the default AWS KMS key store.

Q: How can I audit the use of keys in a custom key store?
In addition to the activity that is logged to AWS CloudTrail by AWS KMS the use of a custom key store provides three further auditing mechanisms. First, AWS CloudHSM also logs all API activity to CloudTrail, for example to create clusters and to add or remove HSMs. Second, each cluster also captures its own local logs to record user and key management activity. Third, each CloudHSM instance copies the local user and key management activity logs to AWS CloudWatch.

Q: What impact does using a custom key store have on availability of keys?
The use of an AWS KMS custom key store makes you responsible for ensuring that your keys are available for use by AWS KMS. Errors in configuration of CloudHSM and accidental deletion of key material within an AWS CloudHSM cluster could impact availability. The number of HSMs you use and your choice of availability zones (AZs) can also affect the resilience of your cluster. As in any key management system, it is important to understand how the availability of keys can impact the recovery of your encrypted data.

Q: What are the performance limitations associated with a custom key store?
The rate at which keys stored in an AWS KMS custom key store can be used via AWS KMS API calls are lower than for keys stored in the default AWS KMS key store. See the AWS KMS Developer Guide for the current performance limits.

Q: What are the costs associated with using a custom key store?
AWS KMS prices are unaffected by the use of a custom key store. However, each custom key store does require that your AWS CloudHSM cluster contains at least two HSMs. These HSMs are charged at the standard AWS CloudHSM prices. There are no additional charges for using a custom key store.

Q: What additional skills and resources are required to configure a custom key stores?
AWS KMS users that wish to use a custom key store will need to set up an AWS CloudHSM cluster, add HSMs, manage HSMs users and potentially restore HSMs from backup. These are security sensitive tasks and you should ensure that you have the appropriate resources and organizational controls in place.

Q: Can I import keys into a custom key store?
No, the ability to import your own key material into an AWS KMS custom key store is not supported. Keys that are stored in a custom key store can only be generated in the HSMs that form your AWS CloudHSM cluster.

Q: Can I migrate keys between the default AWS KMS keys store and a custom key store?
No, the ability to migrate keys between the different types of AWS KMS key store is not currently supported. All keys must be created in the key store in which they will be used, except in situations where you import you own key material into the default AWS KMS key store.

Q: Can I rotate keys stored in a custom key store?
The ability to automatically rotate key material in an AWS KMS custom key store is not supported. Key rotation must be performed manually by creating new keys and re-mapping AWS KMS key aliases used by your application code to use the new keys for future encryption operations.

Q: Can I use my AWS CloudHSM cluster for other applications?
Yes, AWS KMS does not require exclusive access to your AWS CloudHSM cluster. If you already have a cluster you can use it as a custom key store and continue to use it for your other applications. However, if your cluster is supporting high, non-AWS KMS, workloads you may experience reduced throughput for operations using CMKs in your custom key store. Similarly, a high AWS KMS request rate to your custom key store could impact your other applications.

Q: How can I learn more about AWS CloudHSM?
Visit the AWS CloudHSM web site for an overview of the service and for more details on configuring and using the service read the AWS CloudHSM User Guide.

Billing

Q: How will I be charged and billed for my use of AWS KMS?
With AWS KMS, you pay only for what you use, there is no minimum fee. There are no set-up fees or commitments to begin using the service. At the end of the month, your credit card will automatically be charged for that month’s usage.

You are charged for all CMKs you create and for API requests made to the service each month above a free tier.

For current pricing information, please visit the AWS KMS pricing page.

Q: Is there a free tier?
Yes. With the AWS Free Usage Tier you can get started with AWS KMS for free* in all regions. AWS managed CMKs that are created on your behalf by AWS services are free to store in your account. There is a free tier for usage that provides a free number of requests to the service each month. For current information on pricing, including the free tier, please visit the AWS KMS pricing page.

*API requests involving asymmetric CMKs and API requests to the GenerateDataKeyPair and GenerateDataKeyPairWithoutPlaintext APIs are excluded from the free tier.

Q: Do your prices include taxes?
Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. You can learn more here.

Security

Q: Who can use and manage my keys in AWS KMS?
AWS KMS enforces usage and management policies that you define. You choose to allow AWS Identity and Access Management (IAM) users and roles from your account or other accounts to use and manage your keys.

Q: How does AWS secure the CMKs that I create?
AWS KMS is designed so that no one, including AWS employees, can retrieve your plaintext CMKs from the service. The service uses hardware security modules (HSMs) that have been validated under FIPS 140-2, or are in the process of being validated, to protect the confidentiality and integrity of your keys regardless of whether you use AWS KMS or AWS CloudHSM to create your keys or you import them into the service yourself. Your plaintext CMKs never leave the HSMs, are never written to disk and are only ever used in the volatile memory of the HSMs for the time needed to perform your requested cryptographic operation. AWS KMS keys are never transmitted outside of the AWS regions in which they were created. Updates to software on the service hosts and to the AWS KMS HSM firmware is controlled by multi-party access control that is audited and reviewed by an independent group within Amazon as well as a NIST-certified lab in compliance with FIPS 140-2.

More details about these security controls can be found in the AWS KMS Cryptographic Details whitepaper. You can also review the FIPS 140-2 certificate for AWS KMS HSM along with the associated Security Policy to get more details about how AWS KMS HSM meets the security requirements of FIPS 140-2. In addition, you can download a copy of the Service Organization Controls (SOC) report from AWS Artifact to learn more about security controls used by the service to protect your CMKs.

Q: How do I migrate my existing CMKs to use FIPS 140-2 validated HSMs?
All CMKs regardless of their creation date or origin are automatically protected using HSMs that have been validated under FIPS 140-2, or are in the process of being validated. No action is required on your part to use the FIPS 140-2 validated HSMs.

Q: Which AWS regions have FIPS 140-2 validated HSMs?
FIPS 140-2 validated HSMs are available in all AWS regions where AWS KMS is

offered.

Q: What is the difference between the FIPS 140-2 validated endpoints and the FIPS 140-2 validated HSMs in AWS KMS?
AWS KMS is a two-tier service. The API endpoints receive client requests over an HTTPS connection using only TLS ciphersuites that support perfect forward secrecy. These API endpoints authenticate and authorize the request before passing the request for a cryptographic operation to the AWS KMS HSMs or your AWS CloudHSM cluster if you're using the KMS custom key store feature.

Q: How do I make API requests to AWS KMS using the FIPS 140-2 validated endpoints?
You configure your applications to connect to the unique regional FIPS 140-2 validated HTTPS endpoints. AWS KMS FIPS 140-2 validated HTTPS endpoints are powered by the OpenSSL FIPS Object Module. You can review the security policy of the OpenSSL module at https://www.openssl.org/docs/fips/SecurityPolicy-2.0.13.pdf. FIPS 140-2 validated API endpoints are available in all commercial regions where AWS KMS is available.

Q: Can I use AWS KMS to help me comply with the encryption and key management requirements in the Payment Card Industry Data Security Standard (PCI DSS 3.2.1)?
Yes. AWS KMS has been validated as having the functionality and security controls to help you meet the encryption and key management requirements (primarily referenced in sections 3.5 and 3.6) of the PCI DSS 3.2.1.

For more details on PCI DSS compliant services in AWS, you can read the PCI DSS FAQs.

Q: How does AWS KMS secure the data keys I export and use in my application?
You can request that AWS KMS generate data keys and return them for use in your own application. The data keys are encrypted under a master key you define in AWS KMS so that you can safely store the encrypted data key along with your encrypted data. Your encrypted data key (and therefore your source data) can only be decrypted by users with permissions to use the original master key to decrypt your encrypted data key.

Q: Can I export a CMK and use it in my own applications?
No. CMKs are created and used only within the service to help ensure their security, enable your policies to be consistently enforced, and provide a centralized log of their use.

Q: What geographic region are my keys stored in?
Keys generated by AWS KMS are only stored and used in the region in which they were created. They cannot be transferred to another region. For example; keys created in the EU-Central (Frankfurt) region are only stored and used within the EU-Central (Frankfurt) region.

Q: How can I tell who used or changed the configuration of my keys in AWS KMS?
Logs in AWS CloudTrail will show all AWS KMS API requests, including both management requests (e.g. create, rotate, disable, policy edits) and cryptographic requests (e.g. encrypt/decrypt). Turn on AWS CloudTrail in your account to view these logs.

Q: How does AWS KMS compare to AWS CloudHSM?
AWS CloudHSM provides you with a FIPS 140-2 Level 3 overall validated single-tenant HSM cluster in your Amazon Virtual Private Cloud (VPC) to store and use your keys. You have exclusive control over how your keys are used via an authentication mechanism independent from AWS. You interact with keys in your AWS CloudHSM cluster similar to the way you interact with your applications running in Amazon EC2. You can use AWS CloudHSM to support a variety of use cases, such as Digital Rights Management (DRM), Public Key Infrastructure (PKI), document signing, and cryptographic functions using PKCS#11, Java JCE, or Microsoft CNG interfaces.

AWS KMS allows you to create and control the encryption keys used by your applications and supported AWS services in multiple regions around the world from a single console. The service uses an FIPS HSM that has been validated under FIPS 140-2, or are in the process of being validated, to protect the security of your keys. Centralized management of all your keys in AWS KMS lets you enforce who can use your keys under which conditions, when they get rotated, and who can manage them. AWS KMS integration with AWS CloudTrail gives you the ability to audit the use of your keys to support your regulatory and compliance activities. You interact with AWS KMS from your applications using the AWS SDK if you want to call the service APIs directly, via other AWS services that are integrated with AWS KMS or by using the AWS Encryption SDK if you want to perform client-side encryption.

AWS Resource Access Manager FAQs

Q: What is AWS Resource Access Manager?

A: AWS Resource Access Manager (RAM) provides you the ability to securely share your resources across AWS accounts or within your Organization. You can now centrally procure resources and use RAM to share resources with other accounts, eliminating the need to provision and manage resources in every account. When you share a resource with another account, that account is granted access to the resource and any policies and permissions in that account will apply to the shared resource.

Q: What types of AWS resources can I share using RAM?

The list of resources you can share with RAM are listed in the shareable resources section of the RAM user guide.

Q: How can I get started with RAM?

A: You can get started with RAM by creating a Resource Share using the API/CLI or the AWS Management Console. You can easily share resources by adding resources and accounts to a Resource Share. A Resource Share grants accounts access to resources.

Q: Who can I share resources with?

A: You can share resources with any AWS account, and if you are part of AWS Organizations, then you can also share resources with Organizational Units (OUs) or your entire Organization. If you share resources with accounts that are outside of your Organization, then those accounts will receive an invitation to the Resource Share and can start using the shared resources upon accepting the invitation.

Q: How can I view resources that have been shared with my account?

A: You can view resources that have been shared with your account in the RAM console or by using the RAM APIs. The resources that have been shared with your account will also appear in the respective resource console pages and the respective List/Describe APIs for those resource types. For example, when an AWS Deliverator Rule is shared with an account then that rule will appear on the Deliverator page of the Amazon Route53 console along with the other rules owned by that account and the shared rule will also be returned in the response of the ListDelivertorRules API.

Q: Will I incur any charges for sharing my resources with other accounts?

No. You can share resources at no additional cost.

Q: Can I tag a Resource Share?

Yes, you can tag a Resource Share at the time of creation or any time after creation.

Q: How can I control access to resources shared with me?

You can specify IAM policies to control access to resources shared with you.

Q: Can I stop sharing a resource?

Yes, you can stop sharing a resource by removing it from the Resource Share or by deleting the Resource Share.

Q: How can I know about changes to Resource Shares?

All calls to RAM APIs are logged in AWS CloudTrail. In addition, CloudWatch Events are triggered whenever there are changes to Resource Shares. Please click here to learn more about AWS CloudTrail and here to learn about AWS CloudWatch.

AWS Secrets Manager FAQs

General

What is AWS Secrets Manager?

Why should I use AWS Secrets Manager?

What can I do with AWS Secrets Manager?

What secrets can I manage in AWS Secrets Manager?

What secrets can I rotate with AWS Secrets Manager?

How can my application use these secrets?

How do I get started with AWS Secrets Manager?

In what regions is AWS Secrets Manager available?

Rotation

How does AWS Secrets Manager implement database credential rotation without impacting applications?

Will rotating database credentials impact open connections?

How do I know when AWS Secrets Manager rotates a database credential?

Security

How does AWS Secrets Manager keep my secrets secure?

Who can use and manage secrets in AWS Secrets Manager?

How does AWS Secrets Manager encrypt my secrets?

Billing

How will I be charged and billed for my use of AWS Secrets Manager?

Is there a free trial?

Learn more about product pricing

See pricing examples and calculate your costs.

Learn more »

Sign up for a free account

Instantly get access to the AWS Free Tier.

Sign up »

Start building in the console

Get started building with AWS Secrets Manager in the AWS Console.

Sign in »

AWS Security Hub FAQs

General

Q: What is AWS Security Hub?
AWS Security Hub provides you with a comprehensive view of your security state within AWS and your compliance with security industry standards and best practices. Security Hub centralizes and prioritizes security and compliance findings from across AWS accounts, services, and supported third-party partners to help you analyze your security trends and identify the highest priority security issues.

Q: What are the key benefits of AWS Security Hub?
AWS Security Hub eliminates the complexity and reduces the effort of managing and improving the security and compliance of your AWS accounts and workloads. AWS Security Hub is enabled within a particular region in minutes and the service helps you answer fundamental security and compliance questions you may have on a daily basis. Key benefits include:

> **Save time with centralized and normalized findings** - Security Hub collects findings from the security services enabled across your AWS accounts, such as intrusion detection findings from Amazon GuardDuty, vulnerability scans from Amazon Inspector, and sensitive data identification findings from Amazon Macie. Security Hub also collects findings from partner security products using a standardized AWS Security Finding Format, eliminating the need for time-consuming data parsing and normalization efforts. Customers can designate a master account that can see all findings across their accounts.
> **Improve compliance with automated checks** - Security Hub generates its own findings by running continuous and automated account and resource-level configuration checks against the rules in the supported industry best practices and standards (for example, the CIS AWS Foundations Benchmark).
> **Quickly take actions on findings** - Security Hub aggregates findings into pre-built dashboards that provide bar graphs, line charts, and tables that show you the current security and compliance status of your environment as well as trends. Now you can easily identify potential issues, and take the necessary next steps. For example, you can send findings to ticketing, chat, email, or automated remediation systems using integration with Amazon CloudWatch Events.

Q: How much does AWS Security Hub cost?
There are two pricing dimensions for Security Hub: number of compliance checks per account/region/month and number of finding ingestion events per account/region/month. Pricing is $0.001 per compliance check per account/region/month for first 100,000 checks; $0.0008 per check for the next 400,000 checks; and $0.0005 per check for above 500,000 checks. There is a perpetual free tier of 10,000 finding ingestion events per account/region/month and the pricing is $0.00003 per finding ingestion event per account/region/month after the first 10,000. Customers are not charged for finding ingestion events generated by Security Hub's compliance checks. All accounts and regions will have a 30-day free trial. Please see the AWS Security Hub pricing page for latest pricing information.

Note that AWS Config is required to be enabled in the account(s) using Security Hub. AWS Security Hub compliance checks use the configuration items recorded by AWS Config. If you are not already using AWS Config, please see the Config pricing page for the latest information on the price per configuration item recorded. There is no additional charge for the AWS Config rules enabled by Security Hub compliance checks.

Q: Is AWS Security Hub a regional or global service?
AWS Security Hub is a regional service. This ensures all findings data analyzed is regionally based and doesn't cross AWS regional boundaries. Customer must enable Security Hub in each region to view findings in that region.

Q: What regions does AWS Security Hub support?
The regional availability of AWS Security Hub is listed here: AWS Region Table

Q: What partners work with AWS Security Hub?
There are many technology partners that support the standardized findings format and have integrated with AWS Security Hub. See AWS Security Hub partners.

Getting started with AWS Security Hub

Q: How do I enable AWS Security Hub?
When you open the Security Hub console for the first time, simply choose Get Started, and then choose Enable. AWS Security Hub uses a service-linked role that includes the permissions and trust policy that Security Hub requires to detect and aggregate findings, and to configure the requisite AWS Config infrastructure needed to run compliance checks. In order for Security Hub to run compliance checks in an account, you must have AWS Config enabled in that account.

Q: Does AWS Security Hub help manage security across multiple AWS accounts?
Yes, you can manage multiple accounts within a region by configuring the multi-account hierarchy within Security Hub or by importing an existing hierarchy from services like Amazon GuardDuty.

Q: What is a finding?
A finding is a potential security issue. Security Hub aggregates, normalizes, and prioritizes security alerts, or findings, from AWS and third-party services, as well as generating its own findings as the result of running continuous and automated configuration checks. A finding ingestion event is when a new finding is ingested into Security Hub or when a finding update is ingested into Security Hub.

Q: What is an insight?
An insight is a collection of related findings. Security Hub offers managed insights using filters that you can further tailor for your unique environment. For example, insights help to identify EC2 instances that are missing security patches for important vulnerabilities, or S3 buckets with public read or write permissions. Managed and custom Security Hub insights help you track security issues in your AWS environment.

Q: What is a compliance standard vs. a control vs. a compliance check?
A compliance standard is a collection of controls based on regulatory frameworks or industry best practices. Security Hub conducts automated compliance checks against controls. Each compliance check consists of an evaluation of a rule against a single resource. A single control may involve multiple resources (e.g., IAM users) and a compliance check is performed against each resource. For example, Security Hub supports the CIS AWS Foundations Benchmark standard, which consists of 43 controls. Once Security Hub is enabled, it immediately begins running continuous and automated compliance checks against each control and each relevant resource associated with the control.

Q: What findings sources does AWS Security Hub analyze?
AWS Security Hub analyzes your security alerts, or findings, from these AWS services: Amazon GuardDuty, Amazon Inspector, and Amazon Macie. In addition, see the list of AWS Security Hub Partner solutions that are integrated with Security Hub and support the standardized findings format.

Working in AWS Security Hub

Q: How can I see what are my most important security issues in AWS Security Hub?
There are multiple ways to see your most important security issues. The Security Hub dashboard provides views on which resources have the most findings, how your volume of

security findings are evolving over time, which insights are generating the most findings. You can go to the insights page and use the managed insights to identify high priority issues. You can also create your own custom insights.

Q: Can Security Hub tell me how I measure against security best practices or compliance standards?
Yes. Security Hub creates a score to show you how you're doing against compliance standards and displays it on the main Security Hub dashboard. When you click through to the compliance standard, you will see a summary of the controls that need attention. Security Hub shows how the control was evaluated and informational best practices on how to mitigate the issue.

Q: If I score 100% on a compliance standard, does that mean that I will pass an audit for that compliance standard?
No. Security Hub is focused on automated compliance checks. Most compliance standards have various controls that can’t be checked in an automated fashion, and those are out of scope for Security Hub. Security Hub compliance checks can help you prepare for an audit, but they do not imply that you would pass an audit associated with the compliance standard.

Q: How can Security Hub prioritize the security data that I need the most?
Security Hub uses two mechanisms to help prioritize findings: insights and compliance standards. Insights are grouped or correlated findings that help you identify higher priority findings faster. Examples of insights are “Show me all my EC2 instances potentially infected with malware” and “Show me any possible cases of data exfiltration on EC2 instances.” Compliance standards are sets of controls that are based on regulatory requirements or best practices. AWS has defined specific compliance checks (that align to the controls within standards. An example of a supported Security Hub standard is the CIS AWS Foundations Benchmark.

Q: How can Security Hub integrate with my existing security operations and remediation processes?
Security Hub supports workflow options by enabling the export of findings via CloudWatch events. You can use CloudWatch events to setup integrations with chat systems such as Slack, automated remediation pipelines via Lambda or partner security orchestration tools, SIEMs, and ticketing systems such as ServiceNow.

Q: Will Security Hub replace the consoles of our other security services, such as Amazon GuardDuty, Amazon Inspector, or Amazon Macie?
No. Security Hub is complementary and additive to the AWS security services. In fact, Security Hub will link back into the other consoles to help you gain additional context.

Security Hub does not replicate the setup, configuration, or specialized features available within each security service.

Q: I deployed the CIS AWS Foundations Benchmark QuickStart, but the Security Hub CIS Compliance Standard is showing that I am failing some checks, why is that?
The QuickStart solution is designed as a single account and single region template for some hardening controls that cover checks 1.1, 2.1 through 2.7, and 3.1 through 3.14. The QuickStart includes a pre-requisite template that deploys a trail in a single region only. Since the CIS checks 1.1, 2.1 through 2.5, 2.7, and 3.1 through 3.14 require a multi-region trail, these checks fail in Security Hub CIS Compliance Standard. [Note that the CIS QuickStart solution implements hardening controls for only the following checks: 1.1, 2.1 through 2.7, and 3.1 through 3.14. The remaining checks are not addressed by the CIS QuickStart.] In addition, the QuickStart “Monitoring” checks 3.2, 3.4, 3.5, and 3.8 through 3.14 are implemented using CloudWatch events instead of CloudWatch metric filters, which also causes failures of these checks in Security Hub CIS Compliance Standard.

Learn more about product pricing

See pricing examples and free trial details

Learn more »

Sign up for a free account

Instantly get access to the AWS Free Tier.

Sign up »

Start building in the console

Get started with AWS Security Hub in the AWS Console.

Sign in »

AWS Shield FAQs

General

Q. What is AWS Shield?

AWS Shield is a managed service that provides protection against Distributed Denial of Service (DDoS) attacks for applications running on AWS. AWS Shield Standard is automatically enabled to all AWS customers at no additional cost. AWS Shield Advanced is an optional paid service. AWS Shield Advanced provides additional protections against more sophisticated and larger attacks for your applications running on Amazon EC2, Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator, and Route 53.

Q. What is AWS Shield Standard?

AWS Shield Standard provides protection for all AWS customers against common and most frequently occurring infrastructure (layer 3 and 4) attacks like SYN/UDP floods, reflection attacks, and others to support high availability of your applications on AWS.

Q. What is AWS Shield Advanced?

AWS Shield Advanced provides enhanced protections for your applications running on protected Amazon EC2, Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator, and Route 53 resources against more sophisticated and larger attacks. AWS Shield Advanced protection provides always-on, flow-based monitoring of network traffic and active application monitoring to provide near real-time notifications of suspected DDoS incidents. AWS Shield Advanced also employs advanced attack mitigation and routing techniques for automatically mitigating attacks. Customers, with Business or Enterprise support, can also engage the DDoS Response Team (DRT) 24x7 to manage and mitigate their application layer DDoS attacks. The DDoS cost protection for scaling protects your AWS bill against higher fees due to usage spikes from protected Amazon EC2, Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator, and Amazon Route 53 during a DDoS attack.

Q. What is DDoS cost protection for scaling?

AWS Shield Advanced includes DDoS cost protection, a safeguard from scaling charges as a result of a DDoS attack that causes usage spikes on protected Amazon EC2, Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator, or Amazon Route 53. If any of the AWS Shield Advanced protected resources scale up in response to a DDoS attack, you can request credits via the regular AWS Support channel.

Q. Can I use AWS Shield to protect web sites not hosted in AWS?

Yes, AWS Shield is integrated with Amazon CloudFront, which supports custom origins outside of AWS.

Q. Can I use IPv6 with all AWS Shield features?

Yes. All of AWS Shield's detection and mitigations work with IPv6 and IPv4 without any discernable changes to performance, scalability, or availability of the service.

Q. How can I test AWS Shield?

AWS Acceptable Use Policy describes permitted and prohibited behavior on AWS, and it includes descriptions of prohibited security violations and network abuse. However, because penetration testing and other simulated events are frequently indistinguishable from these activities, we have established a policy for customers to request permission to conduct penetration tests and vulnerability scans to or originating from the AWS environment. Visit our Penetration testing page to request permissions.

Q. In which AWS regions is AWS Shield Standard available?

AWS Shield Standard is available on all AWS services in every AWS Region and AWS edge location worldwide.

Please refer to Regional Products and Services for details of AWS Shield Standard availability by region.

Q. In which AWS regions is AWS Shield Advanced available?

AWS Shield Advanced is available globally on all Amazon CloudFront, AWS Global Accelerator, and Amazon Route 53 edge locations worldwide. You can protect your web applications hosted anywhere in the world by deploying Amazon CloudFront in front of your application. Your origin servers can be Amazon S3, Amazon EC2, Elastic Load Balancing, or a custom server outside of AWS. You can also enable AWS Shield Advanced directly on Elastic Load Balancing or Amazon EC2 in the following AWS Regions - Northern Virginia, Ohio, Oregon, Northern California, Montreal, São Paulo, Ireland, Frankfurt, London, Paris, Stockholm, Singapore, Tokyo, Sydney, Seoul, and Mumbai.

Please refer to Regional Products and Services for up-to-date details of AWS Shield Advanced availability by region.

Q. Is AWS Shield HIPAA eligible?

Yes, AWS has expanded its HIPAA compliance program to include AWS Shield as a HIPAA eligible service. If you have an executed Business Associate Agreement (BAA) with AWS, you can use AWS Shield to safeguard your web applications running on AWS from Distributed Denial of Service (DDoS) attacks. For more information, see HIPAA Compliance.

Configuring protections

Q. What types of attacks can AWS Shield help me stop?

AWS Shield helps protects your website from all types of DDoS attacks including Infrastructure layer attacks (like UDP floods), State exhaustion attacks (like TCP SYN floods), and Application layer attacks (like HTTP GET or POST floods). See the AWS WAF and AWS Shield Advanced Developer Guide for examples.

Q. What types of attacks can AWS Shield Standard help protect me from?

AWS Shield Standard automatically provides protection for web applications running on AWS against the most common, frequently occurring Infrastructure layer attacks like UDP floods, and State exhaustion attacks like TCP SYN floods. Customers can also use AWS WAF to protect against Application layer attacks like HTTP POST or GET floods. Find more details on how to deploy application layer protections in the AWS WAF and AWS Shield Advanced Developer Guide.

Q. How many resources can I enable for AWS Shield Standard protection?

There is no limit on the number of resources subject to AWS Shield Standard protection. You can get the full benefits of AWS Shield Standard protections by following the best practices of DDoS resiliency on AWS.

Q. How many resources can I enable for AWS Shield Advanced protection?

You can enable up to 1000 AWS resources of each supported resource type (Classic / Application Load Balancers, Amazon CloudFront distributions, Amazon Route 53 hosting zones, Elastic IPs, AWS Global Accelerator accelerators) for AWS Shield Advanced protection. If you want to enable more than 1000, you can request for a limit increase by creating an AWS Support case.

Q. Can I activate AWS Shield Advanced protection via API?

Yes. AWS Shield Advanced can be activated via APIs. You can also add or remove AWS resources from AWS Shield Advanced protection via APIs.

Q. How quickly are attacks mitigated?

Typically, 99% of infrastructure layer attacks detected by AWS Shield are mitigated in less than 1 second for attacks on Amazon CloudFront and Amazon Route 53, and less than 5 minutes for attacks on Elastic Load Balancing. The remaining 1% of infrastructure attacks are typically mitigated in under 20 minutes. Application layer attacks are mitigated by writing rules on AWS WAF, which are inspected and mitigated inline with incoming traffic.

Q. Can I protect resources outside of AWS?

Yes a number of our customers choose to use AWS endpoints in front of their backend instances. Most commonly, these endpoints are our globally distributed services of CloudFront and Route 53. These services are also our best practice suggestions for DDoS resiliency. Customers can then protect these CloudFront distributions and Route 53 hosted zones with Shield Advanced. Please note that you need to lock down their backend resources to only accept traffic from these AWS endpoints.

Responding to attacks

Q. What tools does AWS Shield Standard provide me to mitigate DDoS attacks?

AWS Shield Standard automatically protects your web applications running on AWS against the most common, frequently occurring DDoS attacks. You can get the full benefits of AWS Shield Standard by following the best practices of DDoS resiliency on AWS.

Q. What tools does AWS Shield Advanced provide me to mitigate DDoS attacks?

AWS Shield Advanced manages mitigation of layer 3 and layer 4 DDoS attacks. This means that your designated applications are protected from attacks like UDP Floods, or TCP SYN floods. In addition, for application layer (layer 7) attacks, AWS Shield Advanced can detect attacks like HTTP floods and DNS floods. You can use AWS WAF to apply your own mitigations, or, if you have Business or Enterprise support, you can engage the 24X7 AWS DDoS Response Team (DRT), who can write rules on your behalf to mitigate Layer 7 DDoS attacks.

Q. Do I need a special support plan to contact the AWS DDoS Response Team?

Yes, you need Business or Enterprise support plan in order to escalate to or engage the AWS DDoS Response Team (DRT). See the AWS Support website for more details about AWS Support plans.

Q. How can I contact the AWS DDoS Response Team?

You can engage the AWS DDoS Response Team (DRT) via regular AWS support, or contact AWS Support.

Q. How quickly can I engage the AWS DDoS Response Team (DRT)?

Response times for DRT depends on the AWS Support plan you are subscribed to. We will make every reasonable effort to respond to your initial request within the corresponding timeframes. See the AWS Support website for more details about AWS Support plans.

Visibility and reporting

Q. Does AWS Shield notify me when attacks happen?

Yes. With AWS Shield Advanced you will get notification of DDoS attacks via CloudWatch metrics.

Q. How quickly will I get an attack notifications?

Typically, AWS Shield Advanced provides notification of an attack within a few minutes of attack detection.

Q. Can I get a history of all DDoS attacks on my AWS resources?

Yes. With AWS Shield Advanced you will be able to see the history of all incidents in the trailing 13 months.

Q. Can I see attacks across AWS?

Yes, AWS Shield Advanced customers get access to the Global threat environment dashboard, which gives a anonymized and sampled view of all DDoS attacks seen on AWS within the last 2 weeks.

Q. How can I see if my AWS WAF rules are working?

AWS WAF includes two different ways to see how your website is being protected: one-minute metrics are available in CloudWatch and Sampled Web Requests are available in the AWS WAF API or management console. Additionally you can enable comprehensive logs that are delivered through Amazon Kinesis Firehose to a destination of your choice. These allow you to see which requests were blocked, allowed, or counted and what rule was matched on a given request (i.e., this web request was blocked due to an IP address condition, etc.). For more information see the AWS WAF and AWS Shield Advanced Developer Guide.

Q. I need to do a pen-test to evaluate the service and my application. What is the approved procedure?

Please refer to Penetration testing on AWS. However, this does not include a “DDoS load test”, which is not authorized on AWS. If you'd like to do a live DDoS test, you can request approval for the same by raising a ticket through AWS Support. Approval for the same involves agreement on the conditions of the test between AWS, the customer and the DDoS test vendor. Please note that we only work with approved DDoS test vendors, and whole process takes 3-4 weeks.

Billing

Q. How am I charged for AWS Shield Standard?

AWS Shield Standard is built into the AWS services that you already use for your web applications. There are no additional costs for AWS Shield Standard.

Q. How am I charged for AWS Shield Advanced?

With AWS Shield Advanced, you pay a monthly fee of $3,000 per month per organization. In addition, you also pay for AWS Shield Advanced Data Transfer usage fees for AWS resources enabled for advanced protection. AWS Shield Advanced charges are in addition to standard fees on Amazon EC2, Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator, and Amazon Route 53. Please see the AWS Shield Pricing page for more details.

Q. Can I choose to only protect some of my resources with AWS Shield Advanced?

Yes, AWS Shield Advanced allows you the flexibility to choose the resources that you'd like to protect. You will only be charged for AWS Shield Advanced Data Transfer on these protected resources.

Q. How can I enable AWS Shield Advanced across multiple AWS Accounts?

If your organization has multiple AWS accounts, then you can subscribe multiple AWS Accounts to AWS Shield Advanced by individually enabling it on each account using the AWS Management Console or API. You will pay the monthly fee once as long as the AWS accounts are all under a single consolidated billing, and you own all the AWS accounts and resources in those accounts.

AWS Single Sign-On FAQs

General

What is AWS Single Sign-On (AWS SSO)?

AWS SSO is an AWS service that enables you to use your existing credentials from your Microsoft Active Directory to access your cloud-based applications, such as AWS accounts and business applications (Office 365, Salesforce, Box), by using single sign-on (SSO).

What are the benefits of AWS SSO?

You can use AWS SSO to quickly and easily assign and manage your employees' access to multiple AWS accounts, SAML-enabled cloud applications (such as Salesforce, Office 365, and Box), and custom-built in-house applications, all from a central place. Employees can be more productive by signing in with their existing corporate Active Directory credentials or credentials that you configure in AWS SSO to access their applications from their personalized user portal. Now, employees won't need to remember multiple sets of credentials and access URLs to cloud applications, and new employees can be productive starting on day one. After you've added users to the appropriate group in your directory, they will automatically gain access to accounts and applications that are enabled for members of that group. You'll get better visibility into cloud application use because you can monitor and audit sign-in activity centrally from AWS CloudTrail.

What problems does AWS SSO solve?

AWS SSO eliminates the administrative complexity of custom SSO solutions you use to provision and manage identities across AWS accounts and business applications. As you use multiple AWS accounts and add accounts regularly, setting up SSO with Active Directory Federation Services (AD FS) to access these accounts requires learning the custom AD FS claims programming language. You also need to prepare the AWS accounts with necessary permissions to access these accounts. AWS SSO is available at no additional cost, and it reduces the complexity of repetitive setup and disparate management by tightly integrating with AWS. If you use separate passwords to access different AWS accounts or cloud applications, AWS SSO simplifies the user experience and improves security by eliminating individual passwords needed for each AWS account or cloud business application. AWS SSO also solves the problem of limited visibility of the access to your cloud applications by integrating with AWS CloudTrail and providing a central place for you to audit SSO access to AWS accounts and SAML-enabled cloud applications, such as Office 365, Salesforce, and Box.

Why should I use AWS SSO?

You should use AWS SSO to help your employees become productive quickly by granting them access to AWS accounts and business cloud applications, without writing custom scripts or investing in general-purpose SSO solutions. You should also use AWS SSO to reduce the administrative complexity and cost of setting up and managing SSO access.

AWS SSO is the place where your employees can access your AWS accounts and the applications they need in the course of their work from the AWS SSO user portal, regardless of where these applications were built or are hosted.

What can I do with AWS SSO?

You can use AWS SSO to quickly and easily assign your employees access to AWS accounts managed with AWS Organizations, business cloud applications (such as Salesforce, Office 365, and Box), and custom applications that support Security Assertion Markup Language (SAML) 2.0. Employees can sign in with their existing corporate credentials or credentials they configure in AWS SSO to access their business applications from a single user portal. AWS SSO also allows you to audit users’ access to cloud services by using AWS CloudTrail.

Who should use AWS SSO?

AWS SSO is for administrators who manage multiple AWS accounts and business applications, want to centralize user access management to these cloud services, and want to provide employees a single location to access these accounts and applications without them having to remember yet another password.

How do I start using AWS SSO?

As a new AWS SSO customer, you:

1. Sign in to the AWS Management Console of the master account in your AWS account and navigate to the AWS SSO console.
2. Select the directory you use for storing the identities of your users and groups from the AWS SSO console. AWS SSO provides you a directory by default that you can use to manage users and groups in AWS SSO. You can also change directory to connect to a Microsoft AD directory by clicking through a list of Managed Microsoft AD and AD

Connector instances that AWS SSO discovers in your account automatically. If you want to connect to a Microsoft AD directory, see Getting Started with AWS Directory Service.

3. Grant users SSO access to AWS accounts in your organization by selecting the AWS accounts from a list populated by AWS SSO, and then selecting users or groups from your directory and the permissions you want to grant them.
4. Give users access to business cloud applications by:
   1. Selecting one of the applications from the list of pre-integrated applications supported in AWS SSO.
   2. Configuring the application by following the configuration instructions.
   3. Selecting the users or groups that should be able to access this application.
5. Give your users the AWS SSO sign-in web address that was generated when you configured the directory so that they can sign in to AWS SSO and access accounts and business applications.

How much does AWS SSO cost?

AWS SSO is offered at no extra charge.

In which AWS regions is AWS SSO is available?

See the AWS Region Table for AWS SSO availability by Region.

Directories and Applications Support

What directories can I use with AWS SSO?

You can use the directory that AWS SSO provides you by default to create and manage users in AWS SSO. Alternatively, you can connect AWS SSO to Microsoft Active Directory, running either on-premises or in the AWS Cloud. AWS SSO supports AWS Directory Service for Microsoft Active Directory, also known as AWS Managed Microsoft AD, and AD Connector. AWS SSO does not support Simple AD. See AWS Directory Service Getting Started to learn more.

Can I use my Amazon Cognito User Pools as the connected directory in AWS SSO?

Not at this time. Today, AWS SSO supports creating and managing users in AWS SSO or connecting to a Microsoft Active Directory. Other directory types may be added over time based on customer feedback and demand.

Which cloud-based applications can I connect to using AWS SSO?

You can connect the following applications to AWS SSO:

1. AWS Management Console: You can set up SSO access to the AWS Management Console.
2. Third-party SaaS applications: AWS SSO comes preintegrated with commonly used business applications. For a comprehensive list, see the AWS SSO console.
3. Custom SAML applications: AWS SSO supports applications that allow identity federation using SAML 2.0. For applications that are not preintegrated with AWS SSO, you can set up SSO by using the AWS SSO custom application wizard.

I manage users and groups in Active Directory on premises. How do I connect my directory to AWS SSO?

You have two options for connecting Active Directory–hosted on premises to AWS SSO: (1) Use a AWS Managed Microsoft AD trust relationship, or (2) use AD Connector.

AWS Managed Microsoft AD creates a fully managed Active Directory in the AWS Cloud and can be used to set up a forest trust relationship between your on-premises directory and AWS Managed Microsoft AD. To set up a trust relationship, see When to Create a Trust Relationship.

AD Connector is a directory gateway that can redirect directory requests to your on-premises Microsoft Active Directory without caching any information in the cloud. To connect an on-premises directory using AD Connector, see AD Connector.

I manage users and groups in AWS Identity and Access Management (IAM). Can I connect my directory to AWS SSO?

AWS SSO does not support AWS IAM users and groups at this time.

Can I connect more than one directory to AWS SSO?

No. At any given time, you can have only one directory connected to AWS SSO. But, you can change the directory that is connected to a different one.

SSO Access to AWS Accounts

Which AWS accounts can I connect to AWS SSO?

You can add any AWS account managed using AWS Organizations to AWS SSO. You need to enable all features in your organizations to manage your accounts SSO.

How do I set up SSO to AWS accounts in an organizational unit (OU) within my organization?

You can pick accounts within the organization or filter accounts by OU.

How do I control what permissions my users get when they use SSO to access their account ?

When granting SSO access to your users, you can limit the users' permissions by picking a permission set. Permission sets are a collection of permissions that you can create in AWS SSO, modelling them based on AWS managed policies for job functions or any AWS managed policies. AWS managed policies for job functions are designed to closely align to common job functions in the IT industry. If required, you can also fully customize the permission set to meet your security requirements. AWS SSO applies these permissions to the selected accounts automatically. As you change the permission sets, AWS SSO enables you to apply the changes to the relevant accounts easily. When your users access the accounts through the AWS SSO user portal, these permissions restrict what they can do within those accounts. You can also grant multiple permission sets to your users. When they access the account through the user portal, they can pick which permission set they want to assume for that session.

For which AWS accounts can I get AWS Command Line Interface (CLI) credentials?

You can get AWS CLI credentials for any AWS account and user permissions that your AWS SSO administrator has assigned to you. These CLI credentials can be used for programmatic access to the AWS account.

How long are the AWS Command Line Interface credentials from the AWS SSO user portal valid?

AWS CLI Credentials fetched through the AWS SSO user portal are valid for 60 minutes. You can get a fresh set of credentials as often as needed.

SSO Access to Business Applications

How do I set up SSO to business applications, such as Salesforce?

From the AWS SSO console, navigate to the applications pane, choose Configure new application, and choose an application from the list of cloud applications that are preintegrated with AWS SSO. Follow the on-screen instructions to configure the application. Your application is now configured and you may assign access to it. Choose the groups or users that you want to provide with access to the application and Choose Assign Access to complete the process.

My company uses business applications that are not in AWS SSO's preintegrated application list. Can I still use AWS SSO?

Yes. If your application supports SAML 2.0, you can configure your application as a custom SAML 2.0 application. From the AWS SSO console, navigate to the applications pane, choose Configure new application, and choose Custom SAML 2.0 application. Follow the instructions to configure the application. Your application is now configured and you may assign access to it. Choose the groups or users that you want to provide with access to the application, and choose Assign Access to complete the process.

My application supports OpenID Connect (OIDC) only. Can I set up SSO with AWS SSO?

No. AWS SSO supports only SAML 2.0–based applications.

Does AWS SSO support single sign-on to native mobile and desktop applications?

No. AWS SSO supports single sign-on to business applications through web browsers only.

Miscellaneous

What data will AWS SSO store on my behalf?

AWS SSO will store data about which AWS accounts and cloud applications are assigned to which users and groups, as well as what permissions have been granted for accessing AWS accounts. AWS SSO will also create and manage IAM roles in individual AWS accounts for each permission set you grant access for your users.

Does AWS SSO support multifactor authentication (MFA)?

Yes. You can enable or require users to set up a multi-factor application on their phones or you can require users to provide an additional factor for signing in to AWS SSO by operating a Remote Authentication Dial-In User Service (RADIUS) server and configuring the RADIUS server to work with Active Directory or AD Connector.

How do my employees get started using AWS SSO?

Employees can get started with AWS SSO by visiting the AWS SSO user portal that is generated when you configure your directory in AWS SSO. If you manage your users in AWS SSO, your employees can use their email address and password they configured with AWS SSO to sign into the user portal. If you connect to a Microsoft Active Directory, your employees can sign in to user portal with their Active Directory user name and password and then view the accounts and applications assigned to them. To access an account or application, employees choose the associated icon from the AWS SSO user portal.

Is there an API available for AWS SSO?

No. You can use the AWS SSO console to perform all necessary operations.

AWS WAF FAQs

General

1. What is AWS WAF?

AWS WAF is a web application firewall that helps protect web applications from attacks by allowing you to configure rules that allow, block, or monitor (count) web requests based on conditions that you define. These conditions include IP addresses, HTTP headers, HTTP body, URI strings, SQL injection and cross-site scripting.

2. How does AWS WAF block or allow traffic?

As the underlying service receives requests for your web sites, it forwards those requests to AWS WAF for inspection against your rules. Once a request meets a condition defined in your rules, AWS WAF instructs the underlying service to either block or allow the request based on the action you define.

3. How does AWS WAF protect my web site or application?

AWS WAF is tightly integrated with Amazon CloudFront and the Application Load Balancer (ALB), services that AWS customers commonly use to deliver content for their websites and applications. When you use AWS WAF on Amazon CloudFront, your rules run in all AWS Edge Locations, located around the world close to your end users. This means security doesn't come at the expense of performance. Blocked requests are stopped before they reach your web servers. When you use AWS WAF on Application Load Balancer, your rules run in region and can be used to protect internet-facing as well as internal load balancers.

4. Can I use AWS WAF to protect web sites not hosted in AWS?

Yes, AWS WAF is integrated with Amazon CloudFront, which supports custom origins outside of AWS.

5. What types of attacks can AWS WAF help me to stop?

AWS WAF helps protects your website from common attack techniques like SQL injection and Cross-Site Scripting (XSS). In addition, you can create rules that can block attacks from specific user-agents, bad bots, or content scrapers. See the AWS WAF Developer Guide for examples.

6. Can I get a history of all AWS WAF API calls made on my account for security, operational or compliance auditing?

Yes. To receive a history of all AWS WAF API calls made on your account, you simply turn on AWS CloudTrail in the CloudTrail's AWS Management Console. For more information, visit AWS CloudTrail home page or visit the AWS WAF Developer Guide.

7. Does AWS WAF support IPv6?

Yes, support for IPv6 allows the AWS WAF to inspect HTTP/S requests coming from both IPv6 and IPv4 addresses.

8. Does IPSet match condition for an AWS WAF Rule support IPv6?

Yes, you can setup new IPv6 match condition(s) for new and existing WebACLs, as per the documentation.

9. Can I expect to see IPv6 address appear in the AWS WAF sampled requests where applicable?

Yes. The sampled requests will show the IPv6 address where applicable.

10. Can I use IPv6 with all AWS WAF features?

Yes. You will be able to use all the existing features for traffic both over IPv6 and IPv4 without any discernable changes to performance, scalability or availability of the service.

11. What services does AWS WAF support?

AWS WAF can be deployed on Amazon CloudFront, the Application Load Balancer (ALB), and Amazon API Gateway. As part of Amazon CloudFront it can be part of your Content Distribution Network (CDN) protecting your resources and content at the Edge locations. As part of the Application Load Balancer it can protect your origin web servers running behind the ALBs. As part of Amazon API Gateway, it can help secure and protect your REST APIs.

12. In what Regions is AWS WAF on ALB available in?

AWS WAF on ALB is available in the following AWS Regions.

13. Is AWS WAF HIPAA eligible?

Yes, AWS has expanded its HIPAA compliance program to include AWS WAF as a HIPAA eligible service. If you have an executed Business Associate Agreement (BAA) with AWS, you can use AWS WAF to protect your web applications from common web exploits. For more information, see HIPAA Compliance.

14. How does AWS WAF pricing work? Are there any upfront costs?

AWS WAF charges based on the number of web access control lists (web ACLs) that you create, the number of rules that you add per web ACL, and the number of web requests that you receive. There are no upfront commitments. AWS WAF charges are in addition to Amazon CloudFront pricing, the Application Load Balancer (ALB) pricing, and/or Amazon API Gateway pricing.

15. What is Rate-based Rule in AWS WAF?

Rate-based Rules are a new type of Rule that can be configured in AWS WAF. This feature allows you to specify the number of web requests that are allowed by a client IP in a trailing, continuously updated, 5 minute period. If an IP address breaches the configured limit, new requests will be blocked until the request rate falls below the configured threshold.

16. How does a Rate-based rule compare to a regular AWS WAF Rule?

Rate-based Rules are similar to regular Rules, with one addition: the ability to configure a rate-based threshold. If, for example, the threshold for the Rate-based Rule is set to (say) 2,000, the rule will block all IPs that have more than

2,000 requests in the last 5 minute interval. A Rate-based Rule can also contain any other AWS WAF Condition that is available for a regular rule.

17. What does the Rate-based Rule cost?

A Rate-based Rule costs the same as a regular AWS WAF Rule which is $1 per rule per WebACL per month

18. What are the use cases for the Rate-based Rule?

Here are some popular use cases customers can address with Rate-based rules:

- I want to blacklist or count an IP address when that IP address exceeds the configured threshold rate (configurable in web requests per trailing 5 minute period)
- I want to know which IP address are currently being blacklisted because they exceeded the configured threshold rate
- I want IP addresses that have been added to the blacklist to be automatically removed when they are no longer violating the configured threshold rate
- I want to exempt certain high-traffic source IP ranges from being blacklisted by my Rate-based rules

19. Are the existing matching conditions compatible with the Rate-base Rule?

Yes. Rate-based rules are compatible with existing AWS WAF match conditions. This allows you to further refine your match criteria and limit rate-based mitigations to specific URLs of your website or traffic coming from specific referrers (or user agents) or add other custom match criteria.

20. Can I use Rate-based rule to mitigate Web layer DDoS attacks?

Yes. This new rules type is designed to protect you from use cases such web-layer DDoS attacks, brute force login attempts and bad bots.

21. What visibility features does Rate-based Rules offer?

Rate-based Rules support all the visibility features currently available on the regular AWS WAF Rules. Additionally, they will get visibility into the IP addresses blocked as a result of the Rate-based Rule.

22. Can I use Rate-based rule to limit access to a certain parts of my Webpage?

Yes. Here is an example. Suppose that you want to limit requests to the login page on your website. To do this, you could add the following string match condition to a rate-based rule:

- The Part of the request to filter on is “URI”.
- The Match Type is “Starts with”.
- A Value to match is “/login” (this need to be whatever identifies the login page in the URI portion of the web request)

Additionally, you would specify a Rate Limit of, say, 15,000 requests per 5 minutes. Adding this rate-based rule to a web ACL will limit requests to your login page per IP address without affecting the rest of your site.

23. Can I exempt certain high-traffic source IP ranges from being blacklisted by my Rate-based Rule(s)?

Yes. You can do this by having an IP Whitelist condition within the Rate-base Rule.

24. How accurate is your GeoIP database?

The accuracy of the IP Address to country lookup database varies by region. Based on recent tests, our overall accuracy for the IP address to country mapping is 99.8%.

Managed Rules for AWS WAF

1. What are AWS WAF Managed Rules?

AWS WAF Managed Rules are an easy way to deploy pre-configured rules to protect your applications common threats like application vulnerabilities like OWASP, bots, or Common Vulnerabilities and Exposures (CVE). All Managed Rules are automatically updated by AWS Marketplace security Sellers.

2. How can I subscribe to Managed Rules?

You can subscribe to a Managed Rule provided by a Marketplace security Seller from the AWS WAF console or from the AWS Marketplace. All subscribed Managed Rules will be available for you to add to an AWS WAF web ACL.

3. Can I use Managed Rules along with my existing AWS WAF rules?

Yes, you can use Managed Rules along with your custom AWS WAF rules. You can add Managed Rules to your existing AWS WAF web ACL to which you might have already added your own rules.

4. Does a Managed Rule have multiple AWS WAF rules?

Yes, each Managed Rule could have multiple AWS WAF rules. The number of rules depends on each security seller and their Marketplace product.

5. Will Managed Rules add to my existing AWS WAF limit on number of rules?

The number of rules inside a Managed Rule does not impact your AWS WAF limits. But each Managed Rule added to your web ACL will count as 1 rule.

6. How can I disable a Managed Rule?

You can add a Managed Rule to a web ACL or remove it from the web ACL anytime. The Managed Rules are disabled once you disassociate a Managed Rule from any web ACLs.

7. How can I test a Managed Rule?

AWS WAF allows you to configure a “count” action for a Managed Rule, which counts the number of web requests that are matched by the rules inside the Managed Rule. You can look at the number of counted web requests to estimate how many of your web requests would be blocked if you enable the Managed Rule.

AWS WAF configuration

1. Can I configure custom error pages?

Yes, you can configure CloudFront to present a custom error page when requests are blocked. Please see the CloudFront Developer Guide for more information

2. How long does it take AWS WAF to propagate my rules?

After an initial setup, adding or changing to rules typically takes around a minute to propagate worldwide.

3. How can I see if my rules are working?

AWS WAF includes two different ways to see how your website is being protected: one-minute metrics are available in CloudWatch and Sampled Web Requests are available in the AWS WAF API or management console. These allow you to see which requests were blocked, allowed, or counted and what rule was matched on a given request (i.e., this web request was blocked due to an IP address condition, etc.). For more information see the AWS WAF Developer Guide.

4. How can I test my rules?

AWS WAF allows you to configure a “count” action for rules, which counts the number of web requests that meet your rule conditions. You can look at the number of counted web requests to estimate how many of your web requests would be blocked or allowed if you enable the rule.

5. How long are Real-Time Metrics and Sampled Web Requests stored?

Real-Time Metrics are stored in Amazon CloudWatch. Using Amazon CloudWatch you can configure the time period in which you want to expire events. Sampled Web Requests are stored for up to 2 hours.

6. Can AWS WAF inspect HTTPS traffic?

Yes. AWS WAF helps protect applications and can inspect web requests transmitted over HTTP or HTTPS.

Amazon S3 FAQs

General S3 FAQs

Q: What is Amazon S3?

Amazon S3 is object storage built to store and retrieve any amount of data from anywhere on the Internet. It's a simple storage service that offers an extremely durable, highly available, and infinitely scalable data storage infrastructure at very low costs.

Q: What can I do with Amazon S3?

Amazon S3 provides a simple web service interface that you can use to store and retrieve any amount of data, at any time, from anywhere on the web. Using this web service, you can easily build applications that make use of Internet storage. Since Amazon S3 is highly scalable and you only pay for what you use, you can start small and grow your application as you wish, with no compromise on performance or reliability.

Amazon S3 is also designed to be highly flexible. Store any type and amount of data that you want; read the same piece of data a million times or only for emergency disaster recovery; build a simple FTP application, or a sophisticated web application such as the Amazon.com retail web site. Amazon S3 frees developers to focus on innovation instead of figuring out how to store their data.

Q: How can I get started using Amazon S3?

To sign up for Amazon S3, click this link. You must have an Amazon Web Services account to access this service; if you do not already have one, you will be prompted to create one when you begin the Amazon S3 sign-up process. After signing up, please refer to the Amazon S3 documentation and sample code in the Resource Center to begin using Amazon S3.

Q: What can developers do with Amazon S3 that they could not do with an on-premises solution?

Amazon S3 enables any developer to leverage Amazon's own benefits of massive scale with no up-front investment or performance compromises. Developers are now free to innovate knowing that no matter how successful their businesses become, it will be inexpensive and simple to ensure their data is quickly accessible, always available, and secure.

Q: What kind of data can I store in Amazon S3?

You can store virtually any kind of data in any format. Please refer to the Amazon Web Services Licensing Agreement for details.

Q: How much data can I store in Amazon S3?

The total volume of data and number of objects you can store are unlimited. Individual Amazon S3 objects can range in size from a minimum of 0 bytes to a maximum of 5 terabytes. The largest object that can be uploaded in a single PUT is 5 gigabytes. For objects larger than 100 megabytes, customers should consider using the Multipart Upload capability.

Q: What storage classes does Amazon S3 offer?

Amazon S3 offers a range of storage classes designed for different use cases. These include S3 Standard for general-purpose storage of frequently accessed data; S3 Intelligent-Tiering for data with unknown or changing access patterns; S3 Standard-Infrequent Access (S3 Standard-IA) and S3 One Zone-Infrequent Access (S3 One Zone-IA) for long-lived, but less frequently accessed data; and Amazon S3 Glacier (S3 Glacier) and Amazon S3 Glacier Deep Archive (S3 Glacier Deep Archive) for long-term archive and digital preservation. You can learn more about these storage classes on the Amazon S3 Storage Classes page.

Q: What does Amazon do with my data in Amazon S3?

Amazon will store your data and track its associated usage for billing purposes. Amazon will not otherwise access your data for any purpose outside of the

Amazon S3 offering, except when required to do so by law. Please refer to the Amazon Web Services Licensing Agreement for details.

Q: Does Amazon store its own data in Amazon S3?

Yes. Developers within Amazon use Amazon S3 for a wide variety of projects. Many of these projects use Amazon S3 as their authoritative data store and rely on it for business-critical operations.

Q: How is Amazon S3 data organized?

Amazon S3 is a simple key-based object store. When you store data, you assign a unique object key that can later be used to retrieve the data. Keys can be any string, and they can be constructed to mimic hierarchical attributes. Alternatively, you can use S3 Object Tagging to organize your data across all of your S3 buckets and/or prefixes.

Q: How do I interface with Amazon S3?

Amazon S3 provides a simple, standards-based REST web services interface that is designed to work with any Internet-development toolkit. The operations are intentionally made simple to make it easy to add new distribution protocols and functional layers.

Q: How reliable is Amazon S3?

Amazon S3 gives any developer access to the same highly scalable, highly available, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites. The S3 Standard storage class is designed for 99.99% availability, the S3 Standard-IA storage class is designed for 99.9% availability, the S3 One Zone-IA storage class is designed for 99.5% availability, and the S3 Glacier and S3 Glacier Deep Archive class are designed for 99.99% availability and SLA of 99.9%. All of these storage classes are backed by the Amazon S3 Service Level Agreement.

Q: How will Amazon S3 perform if traffic from my application suddenly spikes?

Amazon S3 was designed from the ground up to handle traffic for any Internet application. Pay-as-you-go pricing and unlimited capacity ensures that your incremental costs don’t change and that your service is not interrupted. Amazon S3’s massive scale enables us to spread load evenly, so that no individual application is affected by traffic spikes.

Q: Does Amazon S3 offer a Service Level Agreement (SLA)?

Yes. The Amazon S3 SLA provides for a service credit if a customer's monthly uptime percentage is below our service commitment in any billing cycle.

Q: What is a Provisioned Capacity Unit (PCU) and when should it use PCU?

Provisioned Capacity guarantees that your retrieval capacity for Expedited retrievals will be available when you need it. Each unit of capacity ensures that at least 3 expedited retrievals can be performed every 5 minutes and provides up to 150MB/s of retrieval throughput. Retrieval capacity can be provisioned if you have specific Expedited retrieval rate requirements that need to be met. Without provisioned capacity, Expedited retrieval requests will be accepted if capacity is available at the time the request is made. You can purchase provisioned capacity using the console, SDK, or the CLI. Each unit of provisioned capacity costs $100 per month from the date of purchase.

AWS Regions

Q: Where is my data stored?

You specify an AWS Region when you create your Amazon S3 bucket. For S3 Standard, S3 Standard-IA, and S3 Glacier storage classes, your objects are automatically stored across multiple devices spanning a minimum of three Availability Zones, each separated by miles across an AWS Region. Objects stored in the S3 One Zone-IA storage class are stored redundantly within a single Availability Zone in the AWS Region you select. Please refer to Regional Products and Services for details of Amazon S3 service availability by AWS Region.

Q: What is an AWS Region?

An AWS Region is a geographic location where AWS provides multiple, physically separated and isolated Availability Zones which are connected with low latency, high throughput, and highly redundant networking.

Q: What is an AWS Availability Zone (AZ)?

An AWS Availability Zone is a physically isolated location within an AWS Region. Within each AWS Region, S3 operates in a minimum of three AZs, each separated by miles to protect against local events like fires, floods, etc.

Amazon S3 Standard, S3 Standard-Infrequent Access, and S3 Glacier storage classes replicate data across a minimum of three AZs to protect against the loss of one entire AZ. This remains true in Regions where fewer than three AZs are publicly available. Objects stored in these storage classes are available for access from all of the AZs in an AWS Region.

The Amazon S3 One Zone-IA storage class replicates data within a single AZ. Data stored in this storage class is susceptible to loss in an AZ destruction event.

Q: How do I decide which AWS Region to store my data in?

There are several factors to consider based on your specific application. You may want to store your data in a Region that…

- ...is near to your customers, your data centers, or your other AWS resources in order to reduce data access latencies.
- ...is remote from your other operations for geographic redundancy and disaster recovery purposes.
- ...enables you to address specific legal and regulatory requirements.
- ...allows you to reduce storage costs. You can choose a lower priced region to save money. For S3 pricing information, please visit the S3 pricing page.

Q: In which parts of the world is Amazon S3 available?

Amazon S3 is available in AWS Regions worldwide, and you can use Amazon S3 regardless of your location. You just have to decide which AWS Region(s) you want to store your Amazon S3 data. See the AWS Regional Availability Table for a list of AWS Regions in which S3 is available today.

Billing

Q: How much does Amazon S3 cost?

With Amazon S3, you pay only for what you use. There is no minimum fee. You can estimate your monthly bill using the AWS Simple Monthly Calculator.

We charge less where our costs are less. Some prices vary across Amazon S3 Regions. Billing prices are based on the location of your bucket. There is no Data Transfer charge for data transferred within an Amazon S3 Region via a COPY request. Data transferred via a COPY request between AWS Regions is charged at rates specified in the pricing section of the Amazon S3 detail page. There is no Data Transfer charge for data transferred between Amazon EC2 and Amazon S3 within the same region, for example, data transferred within the US East (Northern Virginia) Region. However, data transferred between Amazon EC2 and Amazon S3 across all other regions is charged at rates specified on the Amazon S3 pricing page, for example, data transferred between Amazon EC2 US East (Northern Virginia) and Amazon S3 US West (Northern California).

Q: How will I be charged and billed for my use of Amazon S3?

There are no set-up fees or commitments to begin using the service. At the end of the month, your credit card will automatically be charged for that month’s usage. You can view your charges for the current billing period at any time on the Amazon Web Services web site, by logging into your Amazon Web Services account, and clicking “Account Activity” under “Your Web Services Account”.

With the AWS Free Usage Tier*, you can get started with Amazon S3 for free in all regions except the AWS GovCloud Region. Upon sign-up, new AWS customers receive 5 GB of Amazon S3 Standard storage, 20,000 Get Requests, 2,000 Put Requests, 15GB of data transfer in, and 15GB of data transfer out each month for one year.

Amazon S3 charges you for the following types of usage. Note that the calculations below assume there is no AWS Free Tier in place.

Storage Used:

Amazon S3 storage pricing is summarized on the Amazon S3 Pricing page.

The volume of storage billed in a month is based on the average storage used throughout the month. This includes all object data and metadata stored in buckets that you created under your AWS account. We measure your storage usage in "TimedStorage-ByteHrs," which are added up at the end of the month to generate your monthly charges.

Storage Example:

Assume you store 100GB (107,374,182,400 bytes) of data in Amazon S3 Standard in your bucket for 15 days in March, and 100TB (109,951,162,777,600 bytes) of data in Amazon S3 Standard for the final 16 days in March.

At the end of March, you would have the following usage in Byte-Hours: Total Byte-Hour usage = [107,374,182,400 bytes x 15 days x (24 hours / day)] + [109,951,162,777,600 bytes x 16 days x (24 hours / day)] = 42,259,901,212,262,400 Byte-Hours.

Let's convert this to GB-Months: 42,259,901,212,262,400 Byte-Hours / 1,073,741,824 bytes per GB / 744 hours per month = 52,900 GB-Months

This usage volume crosses two different volume tiers. The monthly storage price is calculated below assuming the data is stored in the US East (Northern Virginia) Region: 50 TB Tier: 51,200 GB x $0.023 = $1,177.60 50 TB to 450 TB Tier: 1,700 GB x $0.022 = $37.40

Total Storage Fee = $1,177.60 + $37.40 = $1,215.00

Network Data Transferred In:

Amazon S3 Data Transfer In pricing is summarized on the Amazon S3 Pricing page. This represents the amount of data sent to your Amazon S3 buckets.

Network Data Transferred Out:

Amazon S3 Data Transfer Out pricing is summarized on the Amazon S3 Pricing page. For Amazon S3, this charge applies whenever data is read from any of your buckets from a location outside of the given Amazon S3 Region.

Data Transfer Out pricing rate tiers take into account your aggregate Data Transfer Out from a given region to the Internet across Amazon EC2, Amazon S3, Amazon RDS, Amazon SimpleDB, Amazon SQS, Amazon SNS and Amazon VPC. These tiers do not apply to Data Transfer Out from Amazon S3 in one AWS Region to another AWS Region.

Data Transfer Out Example:
Assume you transfer 1TB of data out of Amazon S3 from the US East (Northern Virginia) Region to the Internet every day for a given 31-day month. Assume you also transfer 1TB of data out of an Amazon EC2 instance from the same region to the Internet over the same 31-day month.

Your aggregate Data Transfer would be 62 TB (31 TB from Amazon S3 and 31 TB from Amazon EC2). This equates to 63,488 GB (62 TB * 1024 GB/TB).

This usage volume crosses three different volume tiers. The monthly Data Transfer Out fee is calculated below assuming the Data Transfer occurs in the US East (Northern Virginia) Region:
10 TB Tier: 10,239 GB (10×1024 GB/TB – 1 (free)) x $0.09 = $921.51
10 TB to 50 TB Tier: 40,960 GB (40×1024) x $0.085 = $3,481.60
50 TB to 150 TB Tier: 12,288 GB (remainder) x $0.070 = $860.16

Total Data Transfer Out Fee = $921.51+ $3,481.60 + $860.16= $5,263.27

Data Requests:

Amazon S3 Request pricing is summarized on the Amazon S3 Pricing Chart.

Request Example:
Assume you transfer 10,000 files into Amazon S3 and transfer 20,000 files out of Amazon S3 each day during the month of March. Then, you delete 5,000 files on March 31st.

Total PUT requests = 10,000 requests x 31 days = 310,000 requests
Total GET requests = 20,000 requests x 31 days = 620,000 requests
Total DELETE requests = 5,000×1 day = 5,000 requests

Assuming your bucket is in the US East (Northern Virginia) Region, the Request fees are calculated below:
310,000 PUT Requests: 310,000 requests x $0.005/1,000 = $1.55
620,000 GET Requests: 620,000 requests x $0.004/10,000 = $0.25
5,000 DELETE requests = 5,000 requests x $0.00 (no charge) = $0.00

Data Retrieval:

Amazon S3 data retrieval pricing applies for the S3 Standard-Infrequent Access (S3 Standard-IA) and S3 One Zone-IA storage classes and is summarized on the Amazon S3 Pricing page.

Data Retrieval Example:
Assume in one month you retrieve 300GB of S3 Standard-IA, with 100GB going out to the Internet, 100GB going to EC2 in the same AWS region, and 100GB going to CloudFront in the same AWS Region.

Your data retrieval fees for the month would be calculated as 300GB x $0.01/GB = $3.00. Note that you would also pay network data transfer fees for the portion that went out to the Internet.

Please see here for details on billing of objects archived to Amazon S3 Glacier.

* * Your usage for the free tier is calculated each month across all regions except the AWS GovCloud Region and automatically applied to your bill – unused monthly usage will not roll over. Restrictions apply; See offer terms for more details.

Q: Why do prices vary depending on which Amazon S3 Region I choose?

We charge less where our costs are less. For example, our costs are lower in the US East (Northern Virginia) Region than in the US West (Northern California) Region.

Q: How am I charged for using Versioning?

Normal Amazon S3 rates apply for every version of an object stored or requested. For example, let's look at the following scenario to illustrate storage costs when utilizing Versioning (let's assume the current month is 31 days long):

1) Day 1 of the month: You perform a PUT of 4 GB (4,294,967,296 bytes) on your bucket.
2) Day 16 of the month: You perform a PUT of 5 GB (5,368,709,120 bytes) within the same bucket using the same key as the original PUT on Day 1.

When analyzing the storage costs of the above operations, please note that the 4 GB object from Day 1 is not deleted from the bucket when the 5 GB object is written on Day 15. Instead, the 4 GB object is preserved as an older version and the 5 GB object becomes the most recently written version of the object within your bucket. At the end of the month:

Total Byte-Hour usage
[4,294,967,296 bytes x 31 days x (24 hours / day)] + [5,368,709,120 bytes x 16 days x (24 hours / day)] = 5,257,039,970,304 Byte-Hours.

Conversion to Total GB-Months
5,257,039,970,304 Byte-Hours x (1 GB / 1,073,741,824 bytes) x (1 month / 744 hours) = 6.581 GB-Month

The fee is calculated based on the current rates for your region on the Amazon S3 Pricing page.

Q: How am I charged for accessing Amazon S3 through the AWS Management Console?

Normal Amazon S3 pricing applies when accessing the service through the AWS Management Console. To provide an optimized experience, the AWS Management Console may proactively execute requests. Also, some interactive operations result in more than one request to the service.

Q: How am I charged if my Amazon S3 buckets are accessed from another AWS account?

Normal Amazon S3 pricing applies when your storage is accessed by another AWS Account. Alternatively, you may choose to configure your bucket as a Requester Pays bucket, in which case the requester will pay the cost of requests and downloads of your Amazon S3 data.

You can find more information on Requester Pays bucket configurations in the Amazon S3 Documentation.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax.

Learn more about taxes on AWS services »

Security

Q: How secure is my data in Amazon S3?

Amazon S3 is secure by default. Upon creation, only the resource owners have access to Amazon S3 resources they create. Amazon S3 supports user authentication to control access to data. You can use access control mechanisms such as bucket policies and Access Control Lists (ACLs) to selectively grant permissions to users and groups of users. The Amazon S3 console highlights your publicly accessible buckets, indicates the source of public accessibility, and also warns you if changes to your bucket policies or bucket ACLs would make your bucket publicly accessible. You should enable Block Public Access for all accounts and buckets that you do not want publicly accessible.

You can securely upload/download your data to Amazon S3 via SSL endpoints using the HTTPS protocol. If you need extra security you can use the Server-Side Encryption (SSE) option to encrypt data stored at rest. You can configure your Amazon S3 buckets to automatically encrypt objects before storing them if the incoming storage requests do not have any encryption information.

Alternatively, you can use your own encryption libraries to encrypt data before storing it in Amazon S3.

Q: How can I control access to my data stored on Amazon S3?

Customers may use four mechanisms for controlling access to Amazon S3 resources: Identity and Access Management (IAM) policies, bucket policies, Access Control Lists (ACLs), and Query String Authentication. IAM enables organizations with multiple employees to create and manage multiple users under a single AWS account. With IAM policies, customers can grant IAM users fine-grained control to their Amazon S3 bucket or objects while also retaining full control over everything the users do. With bucket policies, customers can define rules which apply broadly across all requests to their Amazon S3 resources, such as granting write privileges to a subset of Amazon S3 resources. Customers can also restrict access based on an aspect of the request, such as HTTP referrer and IP address. With ACLs, customers can grant specific permissions (i.e. READ, WRITE, FULL_CONTROL) to specific users for an individual bucket or object. With Query String Authentication, customers can create a URL to an Amazon S3 object which is only valid for a limited time. For more information on the various access control policies available in Amazon S3, please refer to the Access Control topic in the Amazon S3 Developer Guide.

Q: Does Amazon S3 support data access auditing?

Yes, customers can optionally configure an Amazon S3 bucket to create access log records for all requests made against it. Alternatively, customers who need to capture IAM/user identity information in their logs can configure AWS CloudTrail Data Events.

These access log records can be used for audit purposes and contain details about the request, such as the request type, the resources specified in the request, and the time and date the request was processed.

Q: What options do I have for encrypting data stored on Amazon S3?

You can choose to encrypt data using SSE-S3, SSE-C, SSE-KMS, or a client library such as the Amazon S3 Encryption Client. All four enable you to store sensitive data encrypted at rest in Amazon S3.

SSE-S3 provides an integrated solution where Amazon handles key management and key protection using multiple layers of security. You should choose SSE-S3 if you prefer to have Amazon manage your keys.

SSE-C enables you to leverage Amazon S3 to perform the encryption and decryption of your objects while retaining control of the keys used to encrypt objects. With SSE-C, you don't need to implement or use a client-side library to perform the encryption and decryption of objects you store in Amazon S3, but you do need to manage the keys that you send to Amazon S3 to encrypt and decrypt objects. Use SSE-C if you want to maintain your own encryption keys, but don't want to implement or leverage a client-side encryption library.

SSE-KMS enables you to use AWS Key Management Service (AWS KMS) to manage your encryption keys. Using AWS KMS to manage your keys provides several additional benefits. With AWS KMS, there are separate permissions for the use of the master key, providing an additional layer of control as well as protection against unauthorized access to your objects stored in Amazon S3. AWS KMS provides an audit trail so you can see who used your key to access which object and when, as well as view failed attempts to access data from users without permission to decrypt the data. Also, AWS KMS provides additional security controls to support customer efforts to comply with PCI-DSS, HIPAA/HITECH, and FedRAMP industry requirements.

Using an encryption client library, such as the Amazon S3 Encryption Client, you retain control of the keys and complete the encryption and decryption of objects client-side using an encryption library of your choice. Some customers prefer full end-to-end control of the encryption and decryption of objects; that way, only encrypted objects are transmitted over the Internet to Amazon S3. Use a client-side library if you want to maintain control of your encryption keys, are able to implement or use a client-side encryption library, and need to have your objects encrypted before they are sent to Amazon S3 for storage.

For more information on using Amazon S3 SSE-S3, SSE-C, or SSE-KMS, please refer to the topic on Using Encryption in the Amazon S3 Developer Guide.

Q: Can I comply with EU data privacy regulations using Amazon S3?

Customers can choose to store all data in the EU by using the EU (Frankfurt), EU (Ireland), EU (London), or EU (Paris) region. It is your responsibility to ensure that you comply with EU privacy laws. Please see the AWS GDPR Center for more information.

Q: Where can I find more information about security on AWS?

For more information on security on AWS please refer to the AWS security page.

Q: What is an Amazon VPC Endpoint for Amazon S3?

An Amazon VPC Endpoint for Amazon S3 is a logical entity within a VPC that allows connectivity only to S3. The VPC Endpoint routes requests to S3 and routes responses back to the VPC. For more information about VPC Endpoints, read Using VPC Endpoints.

Q: Can I allow a specific Amazon VPC Endpoint access to my Amazon S3 bucket?

You can limit access to your bucket from a specific Amazon VPC Endpoint or a set of endpoints using Amazon S3 bucket policies. S3 bucket policies now support a condition, aws:sourceVpce, that you can use to restrict access. For more details and example policies, read Using VPC Endpoints.

Q: What is Amazon Macie?

Amazon Macie is an AI-powered security service that helps you prevent data loss by automatically discovering, classifying, and protecting sensitive data stored in Amazon S3. Amazon Macie uses machine learning to recognize sensitive data such as personally identifiable information (PII) or intellectual property, assigns a business value, and provides visibility into where this data is stored and how it is being used in your organization. Amazon Macie continuously monitors data access activity for anomalies, and delivers alerts when it detects risk of unauthorized access or inadvertent data leaks.

Q: What can I do with Amazon Macie?

You can use Amazon Macie to protect against security threats by continuously monitoring your data and account credentials. Amazon Macie gives you an automated and low touch way to discover and classify your business data. It provides controls via templated Lambda functions to revoke access or trigger password reset policies upon the discovery of suspicious behavior or unauthorized data access to entities or third-party applications. When alerts are generated, you can use Amazon Macie for incident response, using Amazon CloudWatch Events to swiftly take action to protect your data.

Q: How does Amazon Macie secure your data?

As part of the data classification process, Amazon Macie identifies customers' objects in their S3 buckets, and streams the object contents into memory for analysis. When deeper analysis is required for complex file formats, Amazon Macie will download a full copy of the object, only keeping it for the short time it takes to fully analyze the object. Immediately after Amazon Macie has analyzed the file content for data classification, it deletes the stored content and only retains the metadata required for future analysis. At any time, customers can revoke Amazon Macie access to data in the Amazon S3 bucket. For more information, go to the Amazon Macie User Guide.

Q: What is Access Analyzer for S3?

Access Analyzer for S3 is a feature that monitors your access policies, ensuring that the policies provide only the intended access to your S3 resources. Access Analyzer for S3 evaluates your bucket access policies and enables you to discover and swiftly remediate buckets with potentially unintended access.

Q. How does Access Analyzer for S3 work?

Access Analyzer for S3 alerts you when you have a bucket that is configured to allow access to anyone on the internet or that is shared with other AWS accounts. You receive insights or 'findings' into the source and level of public or shared access. For example, Access Analyzer for S3 will proactively inform you if read or write access were unintendedly provided through an access control list (ACL) or bucket policy. With these insights, you can immediately set or restore the intended access policy.

When reviewing results that show potentially shared access to a bucket, you can Block All Public Access to the bucket with a single click in the S3 Management console. You can also drill down into bucket level permission settings to configure granular levels of access.

For specific and verified use cases that require public access, such as static website hosting, you can acknowledge and archive the findings on a bucket to record that you intend for the bucket to remain public or shared. You can revisit and modify these bucket configurations at any time. For auditing purposes, Access Analyzer for S3 findings can be downloaded as a CSV report.

Q. How do I enable Access Analyzer for S3?

To get started with Access Analyzer for S3, visit the IAM console to enable the AWS Identity and Access Management (IAM) Access Analyzer. When you do this, Access Analyzer for S3 will automatically be visible in the S3 Management Console.

Access Analyzer for S3 is available at no additional cost in the S3 Management Console.

Durability & Data Protection

Q: How durable is Amazon S3?

Amazon S3 Standard, S3 Standard–IA, S3 One Zone-IA, and S3 Glacier are all designed to provide 99.999999999% durability of objects over a given year. This durability level corresponds to an average annual expected loss of 0.000000001% of objects. For example, if you store 10,000,000 objects with Amazon S3, you can on average expect to incur a loss of a single object once every 10,000 years. In addition, Amazon S3 Standard, S3 Standard-IA, and S3 Glacier are all designed to sustain data in the event of an entire S3 Availability Zone loss.

As with any environment, the best practice is to have a backup and to put in place safeguards against malicious or accidental deletion. For S3 data, that best practice includes secure access permissions, Cross-Region Replication, versioning, and a functioning, regularly tested backup.

Q: How are Amazon S3 and Amazon S3 Glacier designed to achieve 99.999999999% durability?

Amazon S3 Standard, S3 Standard-IA, and S3 Glacier storage classes redundantly store your objects on multiple devices across a minimum of three Availability Zones (AZs) in an Amazon S3 Region before returning SUCCESS. The S3 One Zone-IA storage class stores data redundantly across multiple devices within a single AZ. These services are designed to sustain concurrent device failures by quickly detecting and repairing any lost redundancy, and they also regularly verify the integrity of your data using checksums.

Q: What checksums does Amazon S3 employ to detect data corruption?

Amazon S3 uses a combination of Content-MD5 checksums and cyclic redundancy checks (CRCs) to detect data corruption. Amazon S3 performs these checksums on data at rest and repairs any corruption using redundant data. In addition, the service calculates checksums on all network traffic to detect corruption of data packets when storing or retrieving data.

Q: What is Versioning?

Versioning allows you to preserve, retrieve, and restore every version of every object stored in an Amazon S3 bucket. Once you enable Versioning for a bucket, Amazon S3 preserves existing objects anytime you perform a PUT, POST, COPY, or DELETE operation on them. By default, GET requests will retrieve the most recently written version. Older versions of an overwritten or deleted object can be retrieved by specifying a version in the request.

Q: Why should I use Versioning?

Amazon S3 provides customers with a highly durable storage infrastructure. Versioning offers an additional level of protection by providing a means of recovery when customers accidentally overwrite or delete objects. This allows you to easily recover from unintended user actions and application failures. You can also use Versioning for data retention and archiving.

Q: How do I start using Versioning?

You can start using Versioning by enabling a setting on your Amazon S3 bucket. For more information on how to enable Versioning, please refer to the Amazon S3 Technical Documentation.

Q: How does Versioning protect me from accidental deletion of my objects?

When a user performs a DELETE operation on an object, subsequent simple (un-versioned) requests will no longer retrieve the object. However, all versions of that object will continue to be preserved in your Amazon S3 bucket and can be retrieved or restored. Only the owner of an Amazon S3 bucket can permanently delete a version. You can set Lifecycle rules to manage the lifetime and the cost of storing multiple versions of your objects.

Q: Can I setup a trash, recycle bin, or rollback window on my Amazon S3 objects to recover from deletes and overwrites?

You can use Lifecycle rules along with Versioning to implement a rollback window for your Amazon S3 objects. For example, with your versioning-enabled bucket, you can set up a rule that archives all of your previous versions to the lower-cost Glacier storage class and deletes them after 100 days, giving you a 100-day window to roll back any changes on your data while lowering your storage costs.

Q: How can I ensure maximum protection of my preserved versions?

Versioning's Multi-Factor Authentication (MFA) Delete capability can be used to provide an additional layer of security. By default, all requests to your Amazon S3 bucket require your AWS account credentials. If you enable Versioning with MFA Delete on your Amazon S3 bucket, two forms of authentication are required to permanently delete a version of an object: your AWS account credentials and a valid six-digit code and serial number from an authentication device in your physical possession. To learn more about enabling Versioning with MFA Delete, including how to purchase and activate an authentication device, please refer to the Amazon S3 Technical Documentation.

Q: How am I charged for using Versioning?

Normal Amazon S3 rates apply for every version of an object stored or requested. For example, let's look at the following scenario to illustrate storage costs when utilizing Versioning (let's assume the current month is 31 days long):

1) Day 1 of the month: You perform a PUT of 4 GB (4,294,967,296 bytes) on your bucket.
2) Day 16 of the month: You perform a PUT of 5 GB (5,368,709,120 bytes) within the same bucket using the same key as the original PUT on Day 1.

When analyzing the storage costs of the above operations, please note that the 4 GB object from Day 1 is not deleted from the bucket when the 5 GB object is written on Day 15. Instead, the 4 GB object is preserved as an older version and the 5 GB object becomes the most recently written version of the object within your bucket. At the end of the month:

Total Byte-Hour usage
[4,294,967,296 bytes x 31 days x (24 hours / day)] + [5,368,709,120 bytes x 16 days x (24 hours / day)] = 5,257,039,970,304 Byte-Hours.

Conversion to Total GB-Months
5,257,039,970,304 Byte-Hours x (1 GB / 1,073,741,824 bytes) x (1 month / 744 hours) = 6.581 GB-Month

The fee is calculated based on the current rates for your region on the Amazon S3 Pricing Page.

S3 Access Points

Q: What is Amazon S3 Access Points?

Today, customers manage access to their S3 buckets using a single bucket policy that controls access for hundreds of applications with different permission levels.

Amazon S3 Access Points simplifies managing data access at scale for applications using shared data sets on S3. With S3 Access Points, you can now easily create hundreds of access points per bucket, representing a new way of provisioning access to shared data sets. Access Points provide a customized path into a bucket, with a unique hostname and access policy that enforces the specific permissions and network controls for any request made through the access point.

Q: How do S3 Access Points work?

Each S3 Access Point is configured with an access policy specific to a use case or application, and a bucket can have hundreds of access points. For example, you can create an access point for your S3 bucket that grants access for groups of users or applications for your data lake. An Access Point could support a single user or application, or groups of users or applications, allowing separate management of each access point. Each access point is associated with a single bucket and contains a network origin control, and a Block Public Access control. For example, you can create an access point with a network origin control that only permits storage access from your Virtual Private Cloud, a logically isolated section of the AWS Cloud. You can also create an access point with the access point policy configured to only allow access to objects with a defined prefix, such as "finance".

Because each access point contains a unique DNS name, you can now address existing and new buckets with any name of your choice that is unique within the AWS account and region. Using access points that are restricted to a VPC, you can now have an easy, auditable way to make sure S3 data stays within your VPC. Additionally, you can now use AWS Service Control Policies to require any new access point in their organization to be restricted to VPC only access.

Q: What is the difference between a bucket and an access point?

A bucket is the logical storage container for your objects while an access point provides access to the bucket and its contents. An access point is a separate Amazon resource created for a bucket with an Amazon Resource Name (ARN), hostname (in the format of https://[access_point_name]-[account ID].s3-accesspoint.[region].amazonaws.com), an access control policy, and a network origin control.

Q: Why should I use an access point?

S3 Access Points simplify how you manage data access for your application set to your shared data sets on S3. You no longer have to manage a single, complex bucket policy with hundreds of different permission rules that need to be written, read, tracked, and audited. With S3 Access Points, you can now create application-specific access points permitting access to shared data sets with policies tailored to the specific application.

Using Access Points, you can decompose one large bucket policy into separate, discrete access point policies for each application that needs to access the shared data set. This makes it simpler to focus on building the right access policy for an application, while not having to worry about disrupting what any other application is doing within the shared data set. You can also create a Service Control Policy (SCP) and require that all access points be restricted to a Virtual Private Cloud (VPC), firewalling your data to within your private networks. Using access points, you can easily test new access control policies before migrating applications to the access point, or copying the policy to an existing access point. With S3 Access Points you can specify VPC Endpoint policies that permit access only to access points (and thus buckets) owned by specific account IDs. This simplifies the creation of access policies that permit access to buckets within the same account, while rejecting any other S3 access via the VPC Endpoint. S3 Access points allow you to specify any name that is unique within the account and region. For example, you can now have a “test” access point in every account and region.

Q: How do I get started with S3 Access Points?

You can start creating Access Points on new buckets as well as your existing buckets through the AWS Management Console, the AWS Command Line Interface (CLI), the Application Programming Interface (API), and the AWS Software Development Kit (SDK) client. For example, if your bucket is in the Northern California region under AWS account ID 123456789012 and you want to give data access only to your applications running within VPC ‘vpc-1a2b3c4d,’ you can now set up a new access point “foo” with a “network origin control” value of vpc using the following command:

```
aws s3control create-access-point --bucket [bucket name] --name foo --account-id 123456789012 --vpc-configuration VpcId= vpc-1a2b3c4d
```

If your software uses a hostname to connect to your bucket, specify the new access point hostname (“foo-123456789012.s3-accesspoint.us-west-1.amazonaws.com”) and you will begin using the access point. If your software uses a bucket name, after updating to the latest AWS SDK release specify, the access point ARN (‘arn:aws:s3:us-west-1: 123456789012:accesspoint/foo’) as the bucket name to make requests to your data through this access point. Note that access points do not support the CopyObject API to create a copy of an object that is already stored in S3. We are currently working to support CopyObject with access points.

Q: How do I manage access points?

You can add, view, and delete access points as well as edit access point policies through the S3 console and the CLI. You will also be able to use CloudFormation templates to get started with access points. You can monitor and audit access point operations such as “create access point” and “delete access point” through AWS CloudTrail logs. You can control access point usage using AWS Organizations support for AWS SCPs.

Q: Does this change how I create buckets?

No. When you create a bucket, there will be no access points attached to the bucket.

Q: What happens to my existing S3 buckets that do not have any access points attached to them?

You can continue to access existing buckets directly using the bucket hostname. These buckets without access points will continue to function the same way as they always have. No changes are needed to manage them.

Q: When using an access point, how are requests authorized?

S3 access points have their own IAM access point policy. You write access point policies like you would a bucket policy, using the access point ARN as the resource. Access point policies can grant or restrict access to the S3 data requested through the access point. Amazon S3 evaluates all the relevant policies, including those on the user, bucket, access point, VPC Endpoint, and service control policies as well as Access Control Lists, to decide whether to authorize the request.

Q: How do I write access point policies?

You can write an access point policies just like a bucket policy, using IAM rules to govern permissions and the access point ARN in the policy document.

Q: How is restricting access to specific VPCs using network origin controls on access points different from restricting access to VPCs using the bucket policy?

You can continue to use bucket policies to limit bucket access to specified VPCs. Access points provide an easier, auditable way to lock down all or a subset of data in a shared data set to VPC-only traffic for all applications in your organization using API controls. You can use an AWS Organizations Service Control Policy (SCP) to mandate that any access point created in your organization set the “network origin control” API parameter value to “vpc”. Then, any new access point created automatically restricts data access to VPC-only traffic. No additional access policy is required to make sure that data requests are processed only from specified VPCs.

Q: How do I configure Block Public Access (BPA) settings on my access point?

You can configure the Block Public Access (BPA) settings uniquely on each access point at creation time. We are currently working to support changing BPA settings after creation time. Amazon S3 applies the most restrictive combination of the access point-level, bucket-level, and account-level settings.

Q: Can I enforce a “No Internet data access” policy for all access points in my organization?

Yes. To enforce a “No Internet data access” policy for access points in your organization, you would want to make sure all access points enforce VPC only access. To do so, you will write an AWS SCP that only supports the value “vpc” for the “network origin control” parameter in the create_access_point() API. If you had any Internet facing access points that you created previously, they can be removed. You will also need to modify the bucket policy in each of your buckets to further restrict Internet access directly to your bucket through the bucket hostname. Since other AWS services may be directly accessing your bucket, make sure you setup access to allow the AWS services you want by modifying the policy to permit these AWS services. Refer to the S3 documentation for examples of how to do this.

Q: Can I completely disable direct access to a bucket using the bucket hostname?

Not currently, but you can attach a bucket policy that rejects requests not made using an access point. Refer to the S3 Documentation for more details.

Q: Can I replace or remove an access point from a bucket?

Yes. When you remove an access point, any access to the associated bucket through other access points, and through the bucket hostname, will not be disrupted.

Q: How can I control access to access point management APIs (creating new access points, deleting access points)?

Similar to controlling access to bucket management APIs, you can control the use of access point management APIs through IAM user, group, and role policies permissions.

Q: Will I be able to view metrics on operations performed through an access point?

You can monitor and aggregate request metrics on operations performed through an access point using CloudTrail logs and S3 Server Access Logs, and bucket level CloudWatch metrics include requests made through access points.

Q: Is there a quota on how many access points I can create?

By default, each account can create 1,000 access points per region. Please visit AWS Service Quotas to request an increase in this quota.

Q: Can other AWS services and features use access points?

Yes, some AWS services support using access points, please refer to the S3 documentation for the current list. AWS services and features that currently do not support S3 Access Points can continue to use the bucket hostname to access your bucket. Note we are currently working to support Amazon EMR and the Apache Hadoop S3A client.

Q: What is the cost of Amazon S3 Access Points?

There is no additional charge for access points or buckets that use access points. Usual Amazon S3 request rates apply.

S3 Intelligent-Tiering

Q: What is S3 Intelligent-Tiering?

Amazon S3 Intelligent-Tiering (S3 Intelligent-Tiering) is an S3 storage class for data with unknown access patterns or changing access patterns that are difficult to learn. It is the first cloud storage class that delivers automatic cost savings by moving objects between two access tiers when access patterns change. One tier is optimized for frequent access and the other lower-cost tier is designed for infrequent access.

Objects uploaded or transitioned to S3 Intelligent-Tiering are automatically stored in the frequent access tier. S3 Intelligent-Tiering works by monitoring access patterns and then moving the objects that have not been accessed in 30 consecutive days to the infrequent access tier. If the objects are accessed later, S3 Intelligent-Tiering moves the object back to the frequent access tier. This means all objects stored in S3 Intelligent-Tiering are always available when needed. There are no retrieval fees, so you won’t see unexpected increases in storage bills when access patterns change.

Q: Why would I choose to use S3 Intelligent-Tiering?

S3 Intelligent-Tiering is for data with unknown access patterns or changing access patterns that are difficult to learn. It is ideal for data sets where you may not be able to anticipate access patterns. S3 Intelligent-Tiering can also be used to store new data sets where, shortly after upload, access is frequent, but decreases as the data set ages. Then you can move the data set to S3 One Zone-IA or archive it to S3 Glacier.

Q: What performance does S3 Intelligent-Tiering offer?

S3 Intelligent-Tiering provides the same performance as S3 Standard storage.

Q: How durable and available is S3 Intelligent-Tiering?

S3 Intelligent-Tiering is designed for the same 99.999999999% durability as S3 Standard. S3 Intelligent-Tiering is designed for 99.9% availability, and carries a service level agreement providing service credits if availability is less than our service commitment in any billing cycle.

Q: How do I get my data into S3 Intelligent-Tiering?

There are two ways to get data into S3 Intelligent-Tiering. You can directly PUT into S3 Intelligent-Tiering by specifying INTELLIGENT_TIERING in the x-amz-storage-class header or set lifecycle policies to transition objects from S3 Standard or S3 Standard-IA to S3 INTELLIGENT_TIERING.

Q: Are my S3 Intelligent-Tiering objects backed by the Amazon S3 Service Level Agreement?

Yes, S3 Intelligent-Tiering is backed with the Amazon S3 Service Level Agreement, and customers are eligible for service credits if availability is less than our service commitment in any billing cycle.

Q: How will my latency and throughput performance be impacted as a result of using S3 Intelligent-Tiering

You should expect the same latency and throughput performance as S3 Standard when using S3 Intelligent-Tiering.

Q: Is there a minimum duration for S3 Intelligent-Tiering?

S3 Intelligent-Tiering has a minimum storage duration of 30 days, which means that data that is deleted, overwritten, or transitioned to a different S3 Storage

Class before 30 days will incur the normal usage charge plus a pro-rated charge for the remainder of the 30-day minimum.

Q: Is there a minimum object size for S3 Intelligent-Tiering?

S3 Intelligent-Tiering has no minimum billable object size, but objects smaller than 128KB are not eligible for auto-tiering and will always be stored at the frequent access tier rate.

Q: Can I tier objects from S3 Intelligent-Tiering to the Amazon S3 Glacier storage class?

Yes. In addition to using lifecycle policies to migrate objects from S3 Intelligent-Tiering to S3 One Zone-IA, you can also set up lifecycle policies to archive objects to S3 Glacier.

Q: Can I have a bucket that has different objects in different storage classes?

Yes, you can have a bucket that has different objects stored in S3 Standard, S3 Intelligent-Tiering, S3 Standard-IA, and S3 One Zone-IA.

Q: Is S3 Intelligent-Tiering available in all AWS Regions in which Amazon S3 operates?

Yes

S3 Standard-Infrequent Access (S3 Standard-IA)

Q: What is S3 Standard-Infrequent Access?

Amazon S3 Standard-Infrequent Access (S3 Standard-IA) is an Amazon S3 storage class for data that is accessed less frequently but requires rapid access when needed. S3 Standard-IA offers the high durability, throughput, and low latency of the Amazon S3 Standard storage class, with a low per-GB storage price and per-GB retrieval fee. This combination of low cost and high performance make S3 Standard-IA ideal for long-term storage, backups, and as a data store for disaster recovery. The S3 Standard-IA storage class is set at the object level and can exist in the same bucket as the S3 Standard or S3 One Zone-IA storage classes, allowing you to use S3 Lifecycle policies to automatically transition objects between storage classes without any application changes.

Q: Why would I choose to use S3 Standard-IA?

S3 Standard-IA is ideal for data that is accessed less frequently, but requires rapid access when needed. S3 Standard-IA is ideally suited for long-term file storage, older sync and share storage, and other aging data.

Q: What performance does S3 Standard-IA offer?

S3 Standard-IA provides the same performance as the S3 Standard and S3 One Zone-IA storage classes.

Q: How durable and available is S3 Standard-IA?

S3 Standard-IA is designed for the same 99.999999999% durability as the S3 Standard and S3 Glacier storage classes. S3 Standard-IA is designed for 99.9% availability, and carries a service level agreement providing service credits if availability is less than our service commitment in any billing cycle.

Q: How do I get my data into S3 Standard-IA?

There are two ways to get data into S3 Standard-IA. You can directly PUT into S3 Standard-IA by specifying STANDARD_IA in the x-amz-storage-class header. You can also set Lifecycle policies to transition objects from the S3 Standard to the S3 Standard-IA storage class.

Q: Are my S3 Standard-IA objects backed by the Amazon S3 Service Level Agreement?

Yes, S3 Standard-IA is backed with the Amazon S3 Service Level Agreement, and customers are eligible for service credits if availability is less than our service commitment in any billing cycle.

Q: How will my latency and throughput performance be impacted as a result of using S3 Standard-IA?

You should expect the same latency and throughput performance as the S3 Standard storage class when using S3 Standard-IA.

Q: How am I charged for using S3 Standard-IA?

Please see the Amazon S3 pricing page for general information about S3 Standard-IA pricing.

Q: What charges will I incur if I change the storage class of an object from S3 Standard-IA to S3 Standard with a COPY request?

You will incur charges for an S3 Standard-IA COPY request and an S3 Standard-IA data retrieval.

Q: Is there a minimum storage duration charge for S3 Standard-IA?

S3 Standard-IA is designed for long-lived but infrequently accessed data that is retained for months or years. Data that is deleted from S3 Standard-IA within 30 days will be charged for a full 30 days. Please see the Amazon S3 pricing page for information about S3 Standard-IA pricing.

Q: Is there a minimum object storage charge for S3 Standard-IA?

S3 Standard-IA is designed for larger objects and has a minimum object storage charge of 128KB. Objects smaller than 128KB in size will incur storage charges as if the object were 128KB. For example, a 6KB object in S3 Standard-IA will incur S3 Standard-IA storage charges for 6KB and an additional minimum object size fee equivalent to 122KB at the S3 Standard-IA storage price. Please see the Amazon S3 pricing page for information about S3 Standard-IA pricing.

Q: Can I tier objects from S3 Standard-IA to S3 One Zone-IA or S3 Glacier?

Yes. In addition to using Lifecycle policies to migrate objects from S3 Standard to S3 Standard-IA, you can also set up Lifecycle policies to tier objects from S3 Standard-IA to S3 One Zone-IA or S3 Glacier.

S3 One Zone-Infrequent Access (S3 One Zone-IA)

Q: What is S3 One Zone-IA storage class?

S3 One Zone-IA storage class is an Amazon S3 storage class that customers can choose to store objects in a single availability zone. S3 One Zone-IA storage redundantly stores data within that single Availability Zone to deliver storage at 20% less cost than geographically redundant S3 Standard-IA storage, which stores data redundantly across multiple geographically separate Availability Zones.

S3 One Zone-IA offers a 99% available SLA and is also designed for eleven 9's of durability within the Availability Zone. But, unlike the S3 Standard and S3 Standard-IA storage classes, data stored in the S3 One Zone-IA storage class will be lost in the event of Availability Zone destruction.

S3 One Zone-IA storage offers the same Amazon S3 features as S3 Standard and S3 Standard-IA and is used through the Amazon S3 API, CLI and console. S3 One Zone-IA storage class is set at the object level and can exist in the same bucket as S3 Standard and S3 Standard-IA storage classes. You can use S3 Lifecycle policies to automatically transition objects between storage classes without any application changes.

Q: What use cases are best suited for S3 One Zone-IA storage class?

Customers can use S3 One Zone-IA for infrequently-accessed storage, like backup copies, disaster recovery copies, or other easily re-creatable data.

Q: What performance does S3 One Zone-IA storage offer?

S3 One Zone-IA storage class offers similar performance to S3 Standard and S3 Standard-Infrequent Access storage.

Q: How durable is the S3 One Zone-IA storage class?

S3 One Zone-IA storage class is designed for 99.999999999% of durability within an Availability Zone. However, S3 One Zone-IA storage is not designed to withstand the loss of availability or total destruction of an Availability Zone, in which case data stored in S3 One Zone-IA will be lost. In contrast, S3 Standard, S3 Standard-Infrequent Access, and S3 Glacier storage are designed to withstand loss of availability or the destruction of an Availability Zone. S3 One Zone-IA can deliver the same or better durability and availability than most modern, physical data centers, while providing the added benefit of elasticity of storage and the Amazon S3 feature set.

Q: What is the availability SLA for S3 One Zone-IA storage class?

S3 One Zone-IA offers a 99% availability SLA. For comparison, S3 Standard offers a 99.9% availability SLA and S3 Standard-Infrequent Access offers a 99% availability SLA. As with all S3 storage classes, S3 One Zone-IA storage class carries a service level agreement providing service credits if availability is less than our service commitment in any billing cycle. See the Amazon S3 Service Level Agreement.

Q: How will using S3 One Zone-IA storage affect my latency and throughput?

You should expect similar latency and throughput in S3 One Zone-IA storage class to Amazon S3 Standard and S3 Standard-IA storage classes.

Q: How am I charged for using S3 One Zone-IA storage class?

Like S3 Standard-IA, S3 One Zone-IA charges for the amount of storage per month, bandwidth, requests, early delete and small object fees, and a data retrieval fee. Amazon S3 One Zone-IA storage is 20% cheaper than Amazon S3 Standard-IA for storage by month, and shares the same pricing for bandwidth, requests, early delete and small object fees, and the data retrieval fee.

As with S3 Standard-Infrequent Access, if you delete a S3 One Zone-IA object within 30 days of creating it, you will incur an early delete charge. For example, if you PUT an object and then delete it 10 days later, you are still charged for 30 days of storage.

Like S3 Standard-IA, S3 One Zone-IA storage class has a minimum object size of 128KB. Objects smaller than 128KB in size will incur storage charges as if the object were 128KB. For example, a 6KB object in a S3 One Zone-IA storage class will incur storage charges for 6KB and an additional minimum object size fee equivalent to 122KB at the S3 One Zone-IA storage price. Please see the pricing page for information about S3 One Zone-IA pricing.

Q: Is an S3 One Zone-IA "Zone" the same thing as an AWS Availability Zone?

Yes. Each AWS Region is a separate geographic area. Each region has multiple, isolated locations known as Availability Zones. The Amazon S3 One Zone-IA storage class uses an individual AWS Availability Zone within the region.

Q: Are there differences between how Amazon EC2 and Amazon S3 work with Availability Zone-specific resources?

Yes. Amazon EC2 provides you the ability to pick the AZ to place resources, such as compute instances, within a region. When you use S3 One Zone-IA, S3 One Zone-IA assigns an AWS Availability Zone in the region according to available capacity.

Q: Can I have a bucket that has different objects in different storage classes and Availability Zones?

Yes, you can have a bucket that has different objects stored in S3 Standard, S3 Standard-IA and S3 One Zone-IA.

Q: Is S3 One Zone-IA available in all AWS Regions in which S3 operates?

Yes.

Q: How much disaster recovery protection do I forgo by using S3 One Zone-IA?

Each Availability Zone uses redundant power and networking. Within an AWS Region, Availability Zones are on different flood plains, earthquake fault zones, and geographically separated for fire protection. S3 Standard and S3 Standard-IA storage classes offer protection against these sorts of disasters by storing your data redundantly in multiple Availability Zones. S3 One Zone-IA offers protection against equipment failure within an Availability Zone, but it does not protect against the loss of the Availability Zone, in which case, data stored in S3 One Zone-IA would be lost. Using S3 One Zone-IA, S3 Standard, and S3

Standard-IA options, you can choose the storage class that best fits the durability and availability needs of your storage.

Amazon S3 Glacier

Q: Why is Amazon Glacier now called Amazon S3 Glacier?

Customers have long thought of Amazon Glacier, our backup and archival storage service, as a storage class of Amazon S3. In fact, a very high percentage of the data stored in Amazon Glacier today comes directly from customers using S3 Lifecycle policies to move cooler data into Amazon Glacier. Now, Amazon Glacier is officially part of S3 and will be known as Amazon S3 Glacier (S3 Glacier). All of the existing Glacier direct APIs continue to work just as they have, but we've now made it even easier to use the S3 APIs to store data in the S3 Glacier storage class.

Q: Does Amazon S3 provide capabilities for archiving objects to lower cost storage classes?

Yes, Amazon S3 enables you to utilize Amazon S3 Glacier's extremely low-cost storage service for data archival. Amazon S3 Glacier stores data for as little as $0.004 per gigabyte per month. To keep costs low yet suitable for varying retrieval needs, Amazon S3 Glacier provides three options for access to archives, ranging from a few minutes to several hours. Some examples of archive uses cases include digital media archives, financial and healthcare records, raw genomic sequence data, long-term database backups, and data that must be retained for regulatory compliance.

Q: How can I store my data using the Amazon S3 Glacier storage class?

If you have storage which should be immediately archived without delay, or if you make business decisions about when to transition objects to S3 Glacier that can't be expressed through an Amazon S3 Lifecycle policy, S3 PUT to Glacier allows you to use S3 APIs to upload to the S3 Glacier storage class on an object-by-object basis. There are no transition delays and you control the timing. This is also a good option if you want your application to make storage class decisions without having to set a bucket-level policy.

You can use Lifecycle rules to automatically archive sets of Amazon S3 objects to S3 Glacier based on object age. Use the Amazon S3 Management Console, the AWS SDKs, or the Amazon S3 APIs to define rules for archival. Rules specify a prefix and time period. The prefix (e.g. “logs/”) identifies the object(s) subject to the rule. The time period specifies either the number of days from object creation date (e.g. 180 days) or the specified date after which the object(s) should be archived. Any S3 Standard, S3 Standard-IA, or S3 One Zone-IA objects which have names beginning with the specified prefix and which have aged past the specified time period are archived to S3 Glacier. To retrieve Amazon S3 data stored in S3 Glacier, initiate a retrieval job via the Amazon S3 APIs or Management Console. Once the retrieval job is complete, you can access your data through an Amazon S3 GET object request.

For more information on using Lifecycle rules for archival to S3 Glacier, please refer to the Object Archival topic in the Amazon S3 Developer Guide.

Q: Can I use the Amazon S3 APIs or Management Console to list objects that I’ve archived to Amazon S3 Glacier?

Yes, like Amazon S3’s other storage classes (S3 Standard, S3 Standard-IA, and S3 One Zone-IA), S3 Glacier objects stored using Amazon S3’s APIs or Management Console have an associated user-defined name. You can get a real-time list of all of your Amazon S3 object names, including those stored using the S3 Glacier storage class, using the S3 LIST API or the S3 Inventory report.

Q: Can I use Amazon Glacier direct APIs to access objects that I’ve archived to Amazon S3 Glacier?

No. Because Amazon S3 maintains the mapping between your user-defined object name and Amazon S3 Glacier’s system-defined identifier, Amazon S3 objects that are stored using the S3 Glacier storage class are only accessible through the Amazon S3 APIs or the Amazon S3 Management Console.

Q: How can I retrieve my objects that are archived in Amazon S3 Glacier and will I be notified when the object is restored?

To retrieve Amazon S3 data stored in the S3 Glacier storage class, initiate a retrieval request using the Amazon S3 APIs or the Amazon S3 Management Console. The retrieval request creates a temporary copy of your data in the S3 RRS or S3 Standard-IA storage class while leaving the archived data intact in S3 Glacier. You can specify the amount of time in days for which the temporary copy is stored in S3. You can then access your temporary copy from S3 through an Amazon S3 GET request on the archived object.

With restore notifications, you can now be notified with an S3 Event Notification when an object has successfully restored from S3 Glacier and the temporary copy is made available to you. The bucket owner (or others, as permitted by an IAM policy) can arrange for notifications to be issued to Amazon Simple Queue Service (SQS) or Amazon Simple Notification Service (SNS). Notifications can also be delivered to AWS Lambda for processing by a Lambda function.

Q: How long will it take to restore my objects archived in S3 Glacier and can I upgrade an in-progress request to a faster restore speed?

When processing a retrieval job, Amazon S3 first retrieves the requested data from S3 Glacier, and then creates a temporary copy of the requested data in S3 (which typically takes a few minutes). The access time of your request depends on the retrieval option you choose: Expedited, Standard, or Bulk retrievals. For all but the largest objects (250MB+), data accessed using Expedited retrievals are typically made available within 1-5 minutes. Objects retrieved using Standard retrievals typically complete between 3-5 hours. Bulk retrievals typically complete within 5-12 hours. For more information about S3 Glacier retrieval options, please refer to the S3 Glacier FAQs.

S3 Restore Speed Upgrade is an override of an in-progress restore to a faster restore tier if access to the data becomes urgent. You can use S3 Restore Speed Upgrade by issuing another restore request to the same object with a new “tier” job parameter. When issuing an S3 Restore Speed Upgrade, you must choose a faster restore speed than the in-progress restore. Other parameters such as Object Expiry Time will not be changed. You can update the Object Expiry Time after the restore is complete. You pay for each restore request and the per-GB retrieval charge for the faster restore tier. For example, if you issued a Bulk tier restore and then issued an S3 Restore Speed Upgrade request at the Expedited tier to override the in-progress Bulk tier restore, you would be charged for two requests and the per-GB retrieval charge for the Expedited tier.

Q: What am I charged for archiving objects in Amazon S3 Glacier?

Amazon S3 Glacier storage class is priced based on monthly storage capacity and the number of Lifecycle transition requests into Amazon S3 Glacier. Objects that are archived to Amazon S3 Glacier have a minimum of 90 days of storage, and objects deleted before 90 days incur a pro-rated charge equal to the storage charge for the remaining days. See the Amazon S3 pricing page for current pricing.

Q: How is my storage charge calculated for Amazon S3 objects archived to Amazon S3 Glacier?

The volume of storage billed in a month is based on average storage used throughout the month, measured in gigabyte-months (GB-Months). Amazon S3 calculates the object size as the amount of data you stored plus an additional 32KB of Amazon S3 Glacier data plus an additional 8KB of S3 Standard storage class data. Amazon S3 Glacier requires an additional 32KB of data per object for Glacier’s index and metadata so you can identify and retrieve your data. Amazon S3 requires 8KB to store and maintain the user-defined name and metadata for objects archived to Amazon S3 Glacier. This enables you to get a real-time list of all of your Amazon S3 objects, including those stored using the Amazon S3 Glacier storage class, using the Amazon S3 LIST API or the S3 Inventory report. For example, if you have archived 100,000 objects that are 1GB each, your billable storage would be:

1.000032 gigabytes for each object x 100,000 objects = 100,003.2 gigabytes of Amazon S3 Glacier storage.
0.000008 gigabytes for each object x 100,000 objects = 0.8 gigabytes of Amazon S3 Standard storage.

The fee is calculated based on the current rates for your AWS Region on the Amazon S3 Pricing Page.

Q: How much data can I retrieve from Amazon S3 Glacier for free?

You can retrieve 10GB of your Amazon S3 Glacier data per month for free with the AWS free tier. The free tier allowance can be used at any time during the month and applies to Amazon S3 Glacier Standard retrievals.

Q: How am I charged for deleting objects from Amazon S3 Glacier that are less than 90 days old?

Amazon S3 Glacier is designed for use cases where data is retained for months, years, or decades. Deleting data that is archived to Amazon S3 Glacier is free if the objects being deleted have been archived in Amazon S3 Glacier for 90 days or longer. If an object archived in Amazon S3 Glacier is deleted or overwritten within 90 days of being archived, there will be an early deletion fee. This fee is prorated. If you delete 1GB of data 30 days after uploading it, you will be charged an early deletion fee for 60 days of Amazon S3 Glacier storage. If you delete 1 GB of data after 60 days, you will be charged for 30 days of Amazon S3 Glacier storage.

Q: How much does it cost to retrieve data from Amazon S3 Glacier?

There are three ways to restore data from Amazon S3 Glacier – Expedited, Standard, and Bulk Retrievals - and each has a different per-GB retrieval fee and per-archive request fee (i.e. requesting one archive counts as one request). For detailed S3 Glacier pricing by AWS Region, please visit the Amazon S3 Glacier pricing page.

Q: What is the backend infrastructure supporting the S3 Glacier storage class?

We prefer to focus on the customer outcomes of performance, durability, availability, and security. However, this question is often asked by our customers. We use a number of different technologies which allow us to offer the prices we do to our customers. Our services are built using common data storage technologies specifically assembled into purpose-built, cost-optimized systems using AWS-developed software. S3 Glacier benefits from our ability to optimize the sequence of inputs and outputs to maximize efficiency accessing the underlying storage.

Amazon S3 Glacier Deep Archive

Q: What is S3 Glacier Deep Archive?

S3 Glacier Deep Archive is a new Amazon S3 storage class that provides secure and durable object storage for long-term retention of data that is accessed once or twice in a year. From just $0.00099 per GB-month (less than one-tenth of one cent, or about $1 per TB-month), S3 Glacier Deep Archive offers the lowest cost storage in the cloud, at prices significantly lower than storing and maintaining data in on-premises magnetic tape libraries or archiving data off-site.

Q: What use cases are best suited for S3 Glacier Deep Archive?

S3 Glacier Deep Archive is an ideal storage class to provide offline protection of your company's most important data assets, or when long-term data retention is required for corporate policy, contractual, or regulatory compliance requirements. Customers find S3 Glacier Deep Archive to be a compelling choice to protect core intellectual property, financial and medical records, research results, legal documents, seismic exploration studies, and long-term backups, especially in highly regulated industries, such as Financial Services, Healthcare, Oil & Gas, and Public Sectors. In addition, there are organizations, such as media and entertainment companies, that want to keep a backup copy of core intellectual property. Frequently, customers using S3 Glacier Deep Archive are able to reduce or discontinue the use of on-premises magnetic tape libraries and off-premises tape archival services.

Q: How does S3 Glacier Deep Archive differ from S3 Glacier?

S3 Glacier Deep Archive expands our data archiving offerings, enabling you to select the optimal storage class based on storage and retrieval costs, and retrieval times. Choose S3 Glacier when you need to retrieve archived data typically in 1-5 minutes using Expedited retrievals. S3 Glacier Deep Archive, in contrast, is designed for colder data that is very unlikely to be accessed, but still requires long-term, durable storage. S3 Glacier Deep Archive is up to 75% less expensive than S3 Glacier and provides retrieval within 12 hours using the

Standard retrieval speed. You may also reduce retrieval costs by selecting Bulk retrieval, which will return data within 48 hours.

Q: How durable and available is S3 Glacier Deep Archive?

S3 Glacier Deep Archive is designed for the same 99.999999999% durability as the S3 Standard and S3 Glacier storage classes. S3 Glacier Deep Archive is designed for 99.9% availability, and carries a service level agreement providing service credits if availability is less than our service commitment in any billing cycle.

Q: Are my S3 Glacier Deep Archive objects backed by Amazon S3 Service Level Agreement?

Yes, S3 Glacier Deep Archive is backed with the Amazon S3 Service Level Agreement, and customers are eligible for service credits if availability is less than our service commitment in any billing cycle.

Q: How do I get started using S3 Glacier Deep Archive?

The easiest way to store data in S3 Glacier Deep Archive is to use the S3 API to upload data directly. Just specify “S3 Glacier Deep Archive” as the storage class. You can accomplish this using the AWS Management Console, S3 REST API, AWS SDKs, or AWS Command Line Interface.

You can also begin using S3 Glacier Deep Archive by creating policies to migrate data using S3 Lifecycle, which provides the ability to define the lifecycle of your object and reduce your cost of storage. These policies can be set to migrate objects to S3 Glacier Deep Archive based on the age of the object. You can specify the policy for an S3 bucket, or for specific prefixes. Lifecycle transitions are billed at the S3 Glacier Deep Archive Upload price.

Tape Gateway, a cloud-based virtual tape library feature of AWS Storage Gateway, now integrates with S3 Glacier Deep Archive, enabling you to store your virtual tape-based, long-term backups and archives in S3 Glacier Deep Archive, thereby providing the lowest cost storage for this data in the cloud. To get started, create a new virtual tape using AWS Storage Gateway Console or API, and set the archival storage target either to S3 Glacier or S3 Glacier Deep

Archive. When your backup application ejects the tape, the tape will be archived to your selected storage target.

Q: How do you recommend migrating data from my existing tape archives to S3 Glacier Deep Archive?

There are multiple ways to migrate data from existing tape archives to S3 Glacier Deep Archive. You can use the AWS Tape Gateway to integrate with existing backup applications using a virtual tape library (VTL) interface. This interface presents virtual tapes to the backup application. These can be immediately used to store data in Amazon S3, S3 Glacier, and S3 Glacier Deep Archive.

You can also use AWS Snowball or Snowmobile to migrate data. Snowball and Snowmobile accelerate moving terabytes to petabytes of data into and out of AWS using physical storage devices designed to be secure for transport. Using Snowball and Snowmobile helps to eliminate challenges that can be encountered with large-scale data transfers including high network costs, long transfer times, and security concerns.

Finally, you can use AWS Direct Connect to establish dedicated network connections from your premises to AWS. In many cases, Direct Connect can reduce your network costs, increase bandwidth throughput, and provide a more consistent network experience than Internet-based connections.

Q: How can I retrieve my objects stored in S3 Glacier Deep Archive?

To retrieve data stored in S3 Glacier Deep Archive, initiate a “Restore” request using the Amazon S3 APIs or the Amazon S3 Management Console. The Restore creates a temporary copy of your data in the S3 One Zone-IA storage class while leaving the archived data intact in S3 Glacier Deep Archive. You can specify the amount of time in days for which the temporary copy is stored in S3. You can then access your temporary copy from S3 through an Amazon S3 GET request on the archived object.

When restoring an archived object, you can specify one of the following options in the Tier element of the request body: Standard is the default tier and lets you access any of your archived objects within 12 hours, and Bulk lets you retrieve large amounts, even petabytes of data inexpensively and typically completes within 48 hours.

Q: How am I charged for using S3 Glacier Deep Archive?

S3 Glacier Deep Archive storage is priced based on the amount of data you store in GBs, the number of PUT/lifecycle transition requests, retrievals in GBs, and number of restore requests. This pricing model is similar to S3 Glacier. Please see the Amazon S3 pricing page for information about S3 Glacier Deep Archive pricing.

Q: How will S3 Glacier Deep Archive usage show up on my AWS bill and in the AWS Cost Management tool?

S3 Glacier Deep Archive usage and cost will show up as an independent service line item on your monthly AWS bill, separate from your Amazon S3 usage and costs. However, if you are using the AWS Cost Management tool, S3 Glacier Deep Archive usage and cost will be included under the Amazon S3 usage and cost in your detailed monthly spend reports, and not broken out as a separate service line item.

Q: Are there minimum storage duration and minimum object storage charges for S3 Glacier Deep Archive?

S3 Glacier Deep Archive is designed for long-lived but rarely accessed data that is retained for 7-10 years or more. Objects that are archived to S3 Glacier Deep Archive have a minimum of 180 days of storage, and objects deleted before 180 days incur a pro-rated charge equal to the storage charge for the remaining days. Please see the Amazon S3 pricing page for information about S3 Glacier Deep Archive pricing.

S3 Glacier Deep Archive has a minimum billable object storage size of 40KB. Objects smaller than 40KB in size may be stored but will be charged for 40KB of storage. Please see the Amazon S3 pricing page for information about S3 Glacier Deep Archive pricing.

Q: How does S3 Glacier Deep Archive integrate with other AWS Services?

Deep Archive is integrated with Amazon S3 features including S3 Storage Class Analysis, S3 Object Tagging, S3 Lifecycle policies, Composable objects, S3 Object Lock, and S3 Replication. With S3 storage management features, you can use a single Amazon S3 bucket to store a mixture of S3 Glacier Deep Archive, S3 Standard, S3 Standard-IA, S3 One Zone-IA, and S3 Glacier data. This allows storage administrators to make decisions based on the nature of the data and data access patterns. Customers can use Amazon S3 Lifecycle policies to automatically migrate data to lower-cost storage classes as the data ages, or S3 Cross-Region Replication or Same-Region Replication policies to replicate data to the same or a different region.

AWS Storage Gateway service integrates Tape Gateway with S3 Glacier Deep Archive storage class, allowing you to store virtual tapes in the lowest-cost Amazon S3 storage class, reducing the monthly cost to store your long-term data in the cloud by 75%. With this feature, Tape Gateway supports archiving your new virtual tapes directly to S3 Glacier and S3 Glacier Deep Archive, helping you meet your backup, archive, and recovery requirements. Tape Gateway helps you move tape-based backups to AWS without making any changes to your existing backup workflows. Tape Gateway supports most of the leading backup applications such as Veritas, Veeam, Commvault, Dell EMC NetWorker, IBM Spectrum Protect (on Windows OS), and Microsoft Data Protection Manager.

Q: What is the backend infrastructure supporting the S3 Glacier Deep Archive storage class?

In general, AWS does not disclose the backend infrastructure and architecture for our compute, networking, and storage services, as we are more focused on the customer outcomes of performance, durability, availability, and security. However, this question is often asked by our customers. We use a number of different technologies which allow us to offer the prices we do to our customers. Our services are built using common data storage technologies specifically assembled into purpose-built, cost-optimized systems using AWS-developed software. S3 Glacier Deep Archive benefits from our ability to optimize the sequence of inputs and outputs to maximize efficiency accessing the underlying storage.

Query in Place

Q: What is "Query in Place" functionality?

Amazon S3 allows customers to run sophisticated queries against data stored without the need to move data into a separate analytics platform. The ability to query this data in place on Amazon S3 can significantly increase performance and reduce cost for analytics solutions leveraging S3 as a data lake. S3 offers multiple query in place options, including S3 Select, Amazon Athena, and Amazon Redshift Spectrum, allowing you to choose one that best fits your use case. You can even use Amazon S3 Select with AWS Lambda to build serverless apps that can take advantage of the in-place processing capabilities provided by S3 Select.

Q: What is S3 Select?

S3 Select is an Amazon S3 feature that makes it easy to retrieve specific data from the contents of an object using simple SQL expressions without having to retrieve the entire object. You can use S3 Select to retrieve a subset of data using SQL clauses, like SELECT and WHERE, from objects stored in CSV, JSON, or Apache Parquet format. It also works with objects that are compressed with GZIP or BZIP2 (for CSV and JSON objects only), and server-side encrypted objects.

Q: What can I do with S3 Select?

You can use S3 Select to retrieve a smaller, targeted data set from an object using simple SQL statements. You can use S3 Select with AWS Lambda to build serverless applications that use S3 Select to efficiently and easily retrieve data from Amazon S3 instead of retrieving and processing entire object. You can also use S3 Select with Big Data frameworks, such as Presto, Apache Hive, and Apache Spark to scan and filter the data in Amazon S3.

Q: Why should I use S3 Select?

S3 Select provides a new way to retrieve specific data using SQL statements from the contents of an object stored in Amazon S3 without having to retrieve the entire object. S3 Select simplifies and improves the performance of scanning and filtering the contents of objects into a smaller, targeted dataset by up to 400%. With S3 Select, you can also perform operational investigations on log files in Amazon S3 without the need to operate or manage a compute cluster.

Q: What is Amazon Athena?

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL queries. Athena is serverless, so there is no infrastructure to setup or manage, and you can start analyzing data immediately. You don't even need to load your data into Athena, it works directly with data stored in any S3 storage class. To get started, just log into the Athena Management Console, define your schema, and start querying. Amazon Athena uses Presto with full standard SQL support and works with a variety of standard data formats, including CSV, JSON, ORC, Apache Parquet and Avro. While Athena is ideal for quick, ad-hoc querying and integrates with Amazon QuickSight for easy visualization, it can also handle complex analysis, including large joins, window functions, and arrays.

Q: What is Amazon Redshift Spectrum?

Amazon Redshift Spectrum is a feature of Amazon Redshift that enables you to run queries against exabytes of unstructured data in Amazon S3 with no loading or ETL required. When you issue a query, it goes to the Amazon Redshift SQL endpoint, which generates and optimizes a query plan. Amazon Redshift determines what data is local and what is in Amazon S3, generates a plan to minimize the amount of Amazon S3 data that needs to be read, requests Redshift Spectrum workers out of a shared resource pool to read and process data from Amazon S3.

Redshift Spectrum scales out to thousands of instances if needed, so queries run quickly regardless of data size. And, you can use the exact same SQL for Amazon S3 data as you do for your Amazon Redshift queries today and connect to the same Amazon Redshift endpoint using the same BI tools. Redshift Spectrum lets you separate storage and compute, allowing you to scale each independently. You can setup as many Amazon Redshift clusters as you need to query your Amazon S3 data lake, providing high availability and limitless concurrency. Redshift Spectrum gives you the freedom to store your data where you want, in the format you want, and have it available for processing when you need it.

Event Notification

Q: What are Amazon S3 Event Notifications?

Amazon S3 event notifications can be sent in response to actions in Amazon S3 like PUTs, POSTs, COPYs, or DELETEs. Notification messages can be sent through either Amazon SNS, Amazon SQS, or directly to AWS Lambda.

Q: What can I do with Amazon S3 event notifications?

Amazon S3 event notifications enable you to run workflows, send alerts, or perform other actions in response to changes in your objects stored in S3. You can use S3 event notifications to set up triggers to perform actions including transcoding media files when they are uploaded, processing data files when they become available, and synchronizing S3 objects with other data stores. You can also set up event notifications based on object name prefixes and suffixes. For example, you can choose to receive notifications on object names that start with “images/."

Q: What is included in an Amazon S3 event notification?

For a detailed description of the information included in Amazon S3 event notification messages, please refer to the Configuring Amazon S3 Event Notifications topic in the Amazon S3 Developer Guide.

Q: How do I set up Amazon S3 event notifications?

For a detailed description of how to configure event notifications, please refer to the Configuring Amazon S3 event notifications topic in the Amazon S3 Developer Guide. You can learn more about AWS messaging services in the Amazon SNS Documentation and the Amazon SQS Documentation.

Q: What does it cost to use Amazon S3 event notifications?

There are no additional charges for using Amazon S3 for event notifications. You pay only for use of Amazon SNS or Amazon SQS to deliver event notifications, or for the cost of running an AWS Lambda function. Visit the Amazon SNS, Amazon SQS, or AWS Lambda pricing pages to view the pricing details for these services.

Amazon S3 Transfer Acceleration

Q: What is S3 Transfer Acceleration?

Amazon S3 Transfer Acceleration enables fast, easy, and secure transfers of files over long distances between your client and your Amazon S3 bucket. S3 Transfer Acceleration leverages Amazon CloudFront's globally distributed AWS Edge Locations. As data arrives at an AWS Edge Location, data is routed to your Amazon S3 bucket over an optimized network path.

Q: How do I get started with S3 Transfer Acceleration?

To get started with S3 Transfer Acceleration enable S3 Transfer Acceleration on an S3 bucket using the Amazon S3 console, the Amazon S3 API, or the AWS CLI. After S3 Transfer Acceleration is enabled, you can point your Amazon S3 PUT and GET requests to the s3-accelerate endpoint domain name. Your data transfer application must use one of the following two types of endpoints to access the bucket for faster data transfer: .s3-accelerate.amazonaws.com or .s3-accelerate.dualstack.amazonaws.com for the "dual-stack" endpoint. If you want to use standard data transfer, you can continue to use the regular endpoints.

There are certain restrictions on which buckets will support S3 Transfer Acceleration. For details, please refer the Amazon S3 developer guide.

Q: How fast is S3 Transfer Acceleration?

S3 Transfer Acceleration helps you fully utilize your bandwidth, minimize the effect of distance on throughput, and is designed to ensure consistently fast data transfer to Amazon S3 regardless of your client's location. The amount of acceleration primarily depends on your available bandwidth, the distance between the source and destination, and packet loss rates on the network path. Generally, you will see more acceleration when the source is farther from the destination, when there is more available bandwidth, and/or when the object size is bigger.

One customer measured a 50% reduction in their average time to ingest 300 MB files from a global user base spread across the US, Europe, and parts of Asia to a bucket in the Asia Pacific (Sydney) region. Another customer observed cases where performance improved in excess of 500% for users in South East Asia and Australia uploading 250 MB files (in parts of 50MB) to an S3 bucket in the US East (N. Virginia) region.

Try the speed comparison tool to get a preview of the performance benefit from your location!

Q: Who should use S3 Transfer Acceleration?

S3 Transfer Acceleration is designed to optimize transfer speeds from across the world into S3 buckets. If you are uploading to a centralized bucket from geographically dispersed locations, or if you regularly transfer GBs or TBs of data across continents, you may save hours or days of data transfer time with S3 Transfer Acceleration.

Q: How secure is S3 Transfer Acceleration?

S3 Transfer Acceleration provides the same security as regular transfers to Amazon S3. All Amazon S3 security features, such as access restriction based on a client's IP address, are supported as well. S3 Transfer Acceleration communicates with clients over standard TCP and does not require firewall changes. No data is ever saved at AWS Edge Locations.

Q: What if S3 Transfer Acceleration is not faster than a regular Amazon S3 transfer?

Each time you use S3 Transfer Acceleration to upload an object, we will check whether S3 Transfer Acceleration is likely to be faster than a regular Amazon S3 transfer. If we determine that S3 Transfer Acceleration is not likely to be faster than a regular Amazon S3 transfer of the same object to the same destination AWS Region, we will not charge for the use of S3 Transfer Acceleration for that transfer, and we may bypass the S3 Transfer Acceleration system for that upload.

Q: Can I use S3 Transfer Acceleration with multipart uploads?

Yes, S3 Transfer Acceleration supports all bucket level features including multipart uploads.

Q: How should I choose between S3 Transfer Acceleration and Amazon CloudFront’s PUT/POST?

S3 Transfer Acceleration optimizes the TCP protocol and adds additional intelligence between the client and the S3 bucket, making S3 Transfer Acceleration a better choice if a higher throughput is desired. If you have objects that are smaller than 1GB or if the data set is less than 1GB in size, you should consider using Amazon CloudFront's PUT/POST commands for optimal performance.

Q: How should I choose between S3 Transfer Acceleration and AWS Snow Family (Snowball, Snowball Edge, and Snowmobile)?

The AWS Snow Family is ideal for customers moving large batches of data at once. The AWS Snowball has a typical 5-7 days turnaround time. As a rule of thumb, S3 Transfer Acceleration over a fully-utilized 1 Gbps line can transfer up to 75 TBs in the same time period. In general, if it will take more than a week to transfer over the Internet, or there are recurring transfer jobs and there is more than 25Mbps of available bandwidth, S3 Transfer Acceleration is a good option. Another option is to use both: perform initial heavy lift moves with an AWS Snowball (or series of AWS Snowballs) and then transfer incremental ongoing changes with S3 Transfer Acceleration.

Q: Can S3 Transfer Acceleration complement AWS Direct Connect?

AWS Direct Connect is a good choice for customers who have a private networking requirement or who have access to AWS Direct Connect exchanges. S3 Transfer Acceleration is best for submitting data from distributed client locations over the public Internet, or where variable network conditions make throughput poor. Some AWS Direct Connect customers use S3 Transfer Acceleration to help with remote office transfers, where they may suffer from poor Internet performance.

Q: Can S3 Transfer Acceleration complement the AWS Storage Gateway or a 3rd party gateway?

If you can configure the bucket destination in your 3rd party gateway to use an S3 Transfer Acceleration endpoint domain name you will see the benefit.

Visit this File section of the Storage Gateway FAQ to learn more about the AWS implementation.

Q: Can S3 Transfer Acceleration complement 3rd party integrated software?

Yes. Software packages that connect directly into Amazon S3 can take advantage of S3 Transfer Acceleration when they send their jobs to Amazon S3.

Learn more about Storage Partner Solutions »

Q: Is S3 Transfer Acceleration HIPAA eligible?

Yes, AWS has expanded its HIPAA compliance program to include Amazon S3 Transfer Acceleration as a HIPAA eligible service. If you have an executed Business Associate Agreement (BAA) with AWS, you can use Amazon S3 Transfer Acceleration to enable fast, easy, and secure transfers of files including protected health information (PHI) over long distances between your client and your Amazon S3 bucket.

Learn more about HIPAA Compliance »

Storage Management

S3 Object Tagging

Q: What are S3 object tags?

S3 object tags are key-value pairs applied to S3 objects which can be created, updated or deleted at any time during the lifetime of the object. With these, you'll have the ability to create Identity and Access Management (IAM) policies, setup S3 Lifecycle policies, and customize storage metrics. These object-level tags can then manage transitions between storage classes and expire objects in the background.

Q: How do I apply object tags to my objects?

You can add tags to new objects when you upload them or you can add them to existing objects. Up to ten tags can be added to each S3 object and you can use either the AWS Management Console, the REST API, the AWS CLI, or the AWS SDKs to add object tags.

Q: Why should I use object tags?

Object tags are a tool you can use to enable simple management of your S3 storage. With the ability to create, update, and delete tags at any time during the lifetime of your object, your storage can adapt to the needs of your business. These tags allow you to control access to objects tagged with specific key-value pairs, allowing you to further secure confidential data for only a select group or user. Object tags can also be used to label objects that belong to a specific project or business unit, which could be used in conjunction with S3 Lifecycle policies to manage transitions to other storage classes (S3 Standard-IA, S3 One Zone-IA, and S3 Glacier) or with S3 Replication to selectively replicate data between AWS Regions.

Q: How can I update the object tags on my objects?

Object tags can be changed at any time during the lifetime of your S3 object, you can use either the AWS Management Console, the REST API, the AWS CLI, or the AWS SDKs to change your object tags. Note that all changes to tags outside of the AWS Management Console are made to the full tag set. If you have five tags attached to a particular object and want to add a sixth, you need to include the original five tags in that request.

Q: Will my object tags be replicated if I use Cross-Region Replication?

Object tags can be replicated across AWS Regions using Cross-Region Replication. For customers with Cross-Region Replication already enabled, new permissions are required in order for tags to replicate. For more information about setting up Cross-Region Replication, please visit How to Set Up Cross-Region Replication in the Amazon S3 Developer Guide.

Q: How much do object tags cost?

Object tags are priced based on the quantity of tags and a request cost for adding tags. The requests associated with adding and updating Object Tags are priced the same as existing request prices. Please see the Amazon S3 pricing page for more information.

Storage Class Analysis

Q: What is Storage Class Analysis?

With Storage Class Analysis, you can analyze storage access patterns and transition the right data to the right storage class. This new S3 feature automatically identifies infrequent access patterns to help you transition storage to S3 Standard-IA. You can configure a Storage Class Analysis policy to monitor an entire bucket, prefix, or object tag. Once an infrequent access pattern is observed, you can easily create a new S3 Lifecycle age policy based on the results. Storage Class Analysis also provides daily visualizations of your storage usage on the AWS Management Console that you can export to an S3 bucket to analyze using business intelligence tools of your choice such as Amazon QuickSight.

Q: How do I get started with Storage Class Analysis?

You can use the AWS Management Console or the S3 PUT Bucket Analytics API to configure a Storage Class Analysis policy to identify infrequently accessed storage that can be transitioned to the S3 Standard-IA or S3 One Zone-IA storage class or archived to the S3 Glacier storage class. You can navigate to the “Management” tab in the S3 Console to manage Storage Class Analysis, S3 Inventory, and S3 CloudWatch metrics.

Q: How am I charged for using Storage Class Analysis?

Please see the Amazon S3 pricing page for general information about Storage Class Analysis pricing.

Q: How often is the Storage Class Analysis updated?

Storage Class Analysis is updated on a daily basis in the S3 Management Console. Additionally, you can configure Storage Class Analysis to export your report to an S3 bucket of your choice.

S3 Inventory

Q: What is S3 Inventory?

The S3 Inventory report provides a scheduled alternative to Amazon S3's synchronous List API. You can configure S3 Inventory to provide a CSV, ORC, or Parquet file output of your objects and their corresponding metadata on a daily or weekly basis for an S3 bucket or prefix. You can simplify and speed up business workflows and big data jobs with S3 Inventory. You can also use S3 inventory to verify encryption and replication status of your objects to meet business, compliance, and regulatory needs.

Q: How do I get started with S3 Inventory?

You can use the AWS Management Console or the PUT Bucket Inventory API to configure a daily or weekly inventory report for all the objects within your S3 bucket or a subset of the objects under a shared prefix. As part of the configuration, you can specify a destination S3 bucket for your S3 Inventory report, the output file format (CSV, ORC, or Parquet), and specific object metadata necessary for your business application, such as object name, size, last modified date, storage class, version ID, delete marker, noncurrent version flag, multipart upload flag, replication status, or encryption status.

Q: Can S3 Inventory report files be encrypted?

Yes, you can configure encryption of all files written by S3 inventory to be encrypted by SSE-S3 or SSE-KMS. For more information, refer to the user guide.

Q: How do I use S3 Inventory?

You can use S3 Inventory as a direct input into your application workflows or Big Data jobs. You can also query S3 Inventory using Standard SQL language with Amazon Athena, Amazon Redshift Spectrum, and other tools such as Presto, Hive, and Spark.

Learn more about querying S3 Inventory with Athena »

Q: How am I charged for using S3 Inventory?

Please see the Amazon S3 pricing page for S3 Inventory pricing. Once you configure encryption using SSE-KMS, you will incur KMS charges for encryption, refer to the KMS pricing page for detail.

S3 Batch Operations

Q: What is S3 Batch Operations?

S3 Batch Operations is a feature that you can use to automate the execution, management, and auditing of a specific S3 request or Lambda function across many objects stored in Amazon S3. You can use S3 Batch Operations to automate replacing tag sets on S3 objects, updating access control lists (ACL) for S3 objects, copying storage between buckets, initiating a restore from Glacier to S3, or performing custom operations with Lambda functions. S3 Batch Operations can be used from the S3 console, or through the AWS CLI and SDK.

Q: Why should I use S3 Batch Operations?

You should use S3 Batch Operations if you want to automate the execution of a single operation (like copying an object, or executing an AWS Lambda function) across many objects. With S3 Batch Operations, you can, with a few clicks in the S3 console or a single API request, make a change to billions of objects without having to write custom application code or run compute clusters for storage management applications. Not only does S3 Batch Operations administer your storage operation across many objects, S3 Batch Operations manages retries, displays progress, delivers notifications, provides a completion report, and sends events to AWS CloudTrail for all operations performed on your target objects. If you are interested in learning more about S3 Batch Operations, go to the Amazon S3 features page.

Q: How do I get started with S3 Batch Operations?

You can get started with S3 Batch Operations by going into the Amazon S3 console or using the AWS CLI or SDK to create your first S3 Batch Operations job. A S3 Batch Operations job consists of the list of objects to act upon and the type of operation to be performed. Start by selecting an S3 Inventory report or providing your own custom list of objects for S3 Batch Operations to act upon. An S3 Inventory report is a file listing all objects stored in an S3 bucket or prefix. Next, you choose from a set of S3 operations supported by S3 Batch Operations, such as replacing tag sets, changing ACLs, copying storage from one bucket to another, or initiating a restore from Glacier to S3. You can then customize your S3 Batch Operations jobs with specific parameters such as tag values, ACL grantees, and restoration duration. To further customize your storage actions, you can write your own Lambda function and invoke that code through S3 Batch Operations.

Once you create your S3 Batch Operations job, S3 Batch Operations will process your list of objects and send the job to the “awaiting confirmation” state if required. After you confirm the job details, S3 Batch Operations will begin executing the operation you specified. You can view your job’s progress programmatically or through the S3 console, receive notifications on completion, and review a completion report that itemizes the changes made to your storage.

If you are interested in learning more about S3 Batch Operations watch the tutorials videos and visit the documentation.

S3 Object Lock

Q: What is Amazon S3 Object Lock?

Amazon S3 Object Lock is a new Amazon S3 feature that blocks object version deletion during a customer-defined retention period so that you can enforce retention policies as an added layer of data protection or for regulatory compliance. You can migrate workloads from existing write-once-read-many

(WORM) systems into Amazon S3, and configure S3 Object Lock at the object- and bucket-levels to prevent object version deletions prior to pre-defined Retain Until Dates or Legal Hold Dates. S3 Object Lock protection is maintained regardless of which storage class the object resides in and throughout S3 Lifecycle transitions between storage classes.

Q: Why should you use Amazon S3 Object Lock?

You should use S3 Object Lock if you have regulatory requirements that specify that data must be WORM protected, or if you want to add an additional layer of protection to data in Amazon S3. S3 Object Lock can help you to meet regulatory requirements that specify that data should be stored in an immutable format, and also can protect against accidental or malicious deletion for data in Amazon S3.

Q: How does Amazon S3 Object Lock work?

Amazon S3 Object Lock blocks deletion of an object for the duration of a specified retention period. Coupled with S3 Versioning, which protects objects from being overwritten, you’re able to ensure that objects remain immutable for as long as WORM protection is applied. You can apply WORM protection by either assigning a Retain Until Date or a Legal Hold to an object using the AWS SDK, CLI, REST API, or the S3 Management Console. You can apply retention settings within a PUT request, or apply them to an existing object after it has been created.

The Retain Until Date defines the length of time for which an object will remain immutable. Once a Retain Until Date has been assigned to an object, that object cannot be modified or deleted until the Retain Until Date has passed. If a user attempts to delete an object before its Retain Until Date has passed, the operation will be denied.

S3 Object Lock can be configured in one of two Modes. When deployed in Governance Mode, AWS accounts with specific IAM permissions are able to remove WORM protection from an object. If you require stronger immutability in order to comply with regulations, you can use Compliance Mode. In Compliance Mode, WORM protection cannot be removed by any user, including the root account.

Alternatively, you can make an object immutable by applying a Legal Hold to that object. A Legal Hold places indefinite S3 Object Lock protection on an object, which will remain until it is explicitly removed. In order to place and remove Legal Holds, your AWS account must have write permission for the PutObjectLegalHold action. Legal Hold can be applied to any object in an S3 Object Lock enabled bucket, whether or not that object is currently WORM-protected by a retention period.

Q: What AWS electronic storage services have been assessed based on financial services regulations?

For customers in the financial services industry, S3 Object Lock provides added support for broker-dealers who must retain records in a non-erasable and non-rewritable format to satisfy regulatory requirements of SEC Rule 17a-4(f), FINRA Rule 4511, or CFTC Regulation 1.31. You can easily designate the records retention time frame to retain regulatory archives in the original form for the required duration, and also place legal holds to retain data indefinitely until the hold is removed.

Q: What AWS documentation supports the SEC 17a-4(f)(2)(i) and CFTC 1.31(c) requirement for notifying my regulator?

Provide notification to your regulator or “Designated Examining Authority (DEA)” of your choice to use Amazon S3 for electronic storage along with a copy of the Cohasset Assessment. For the purposes of these requirements, AWS is not a designated third party (D3P). Be sure to select a D3P and include this information in your notification to your DEA.

S3 CloudWatch Metrics

Q: How do I get started with S3 CloudWatch Metrics?

You can use the AWS Management Console to enable the generation of 1-minute CloudWatch request metrics for your S3 bucket or configure filters for the metrics using a prefix or object tag. Alternatively, you can call the S3 PUT Bucket Metrics API to enable and configure publication of S3 storage metrics. CloudWatch Request Metrics will be available in CloudWatch within 15 minutes after they are enabled. CloudWatch Storage Metrics are enabled by default for all buckets, and reported once per day.

Q: Can I align S3 CloudWatch request metrics to my applications or business organizations?

Yes, you can configure S3 CloudWatch request metrics to generate metrics for your S3 bucket or configure filters for the metrics using a prefix or object tag.

Q: What alarms can I set on my storage metrics?

You can use CloudWatch to set thresholds on any of the storage metrics counts, timers, or rates and trigger an action when the threshold is breached. For example, you can set a threshold on the percentage of 4xx Error Responses and when at least 3 data points are above the threshold trigger a CloudWatch alarm to alert a DevOps engineer.

Q: How am I charged for using S3 CloudWatch Metrics?

CloudWatch storage metrics are provided free. Cloudwatch request metrics are priced as custom metrics for Amazon CloudWatch. Please see the Amazon CloudWatch pricing page for general information about S3 CloudWatch metrics pricing.

S3 Lifecycle Management

Q: What is S3 Lifecycle management?

S3 Lifecycle management provides the ability to define the lifecycle of your object with a predefined policy and reduce your cost of storage. You can set a lifecycle transition policy to automatically migrate objects stored in the S3 Standard storage class to the S3 Standard-IA, S3 One Zone-IA, and/or S3 Glacier storage classes based on the age of the data. You can also set lifecycle expiration policies to automatically remove objects based on the age of the object. You can set a policy for multipart upload expiration, which expires incomplete multipart uploads based on the age of the upload.

Q: How do I set up an S3 Lifecycle management policy?

You can set up and manage Lifecycle policies in the AWS Management Console, S3 REST API, AWS SDKs, or AWS Command Line Interface (CLI). You can specify the policy at the prefix or at the bucket level.

Q: How much does it cost to use S3 Lifecycle management?

There is no additional cost to set up and apply Lifecycle policies. A transition request is charged per object when an object becomes eligible for transition according to the Lifecycle rule. Refer to the S3 Pricing page for pricing information.

Q: What can I do with Lifecycle management policies?

As data matures, it can become less critical, less valuable, and/or subject to compliance requirements. Amazon S3 includes an extensive library of policies that help you automate data migration processes between storage classes. For example, you can set infrequently accessed objects to move into lower cost storage classes (like S3 Standard-IA or S3 One Zone-IA) after a period of time. After another period, those objects can be moved into Amazon S3 Glacier for archive and compliance. If policy allows, you can also specify a lifecycle policy for object deletion. These rules can invisibly lower storage costs and simplify management efforts. These policies also include good stewardship practices to remove objects and attributes that are no longer needed to manage cost and optimize performance.

Q: How can I use Amazon S3 Lifecycle management to help lower my Amazon S3 storage costs?

With Amazon S3 Lifecycle policies, you can configure your objects to be migrated to from the S3 Standard storage class to S3 Standard-IA or S3 One Zone-IA and/or archived to S3 Glacier. You can also specify an S3 Lifecycle policy to delete objects after a specific period of time. You can use this policy-driven automation to quickly and easily reduce storage costs as well as save time. In each rule you can specify a prefix, a time period, a transition to S3 Standard-IA, S3 One Zone-IA, or S3 Glacier, and/or an expiration. For example, you could create a rule that archives into S3 Glacier all objects with the common prefix “logs/” 30 days from creation and expires these objects after 365 days from creation. You can also create a separate rule that only expires all objects with the prefix “backups/” 90 days from creation. S3 Lifecycle policies apply to both existing and new S3 objects, helping you optimize storage and maximize cost savings for all current data and any new data placed in S3 without time-consuming manual data review and migration. Within a lifecycle rule, the prefix field identifies the objects subject to the rule. To apply the rule to an individual object, specify the key name. To apply the rule to a set of objects, specify their common prefix (e.g. “logs/”). You can specify a transition action to have your objects archived and an expiration action to have your objects removed. For time period, provide the creation date (e.g. January 31, 2015) or the number of days from creation date (e.g. 30 days) after which you want your objects to be archived or removed. You may create multiple rules for different prefixes.

Q: How can I configure my objects to be deleted after a specific time period?

You can set an S3 Lifecycle expiration policy to remove objects from your buckets after a specified number of days. You can define the expiration rules for a set of objects in your bucket through the Lifecycle configuration policy that you apply to the bucket.

Learn more about S3 Lifecycle expiration policies »

Q: Why would I use an S3 Lifecycle policy to expire incomplete multipart uploads?

The S3 Lifecycle policy that expires incomplete multipart uploads allows you to save on costs by limiting the time non-completed multipart uploads are stored. For example, if your application uploads several multipart object parts, but never commits them, you will still be charged for that storage. This policy can lower your S3 storage bill by automatically removing incomplete multipart uploads and the associated storage after a predefined number of days.

Learn more about using S3 Lifecycle to expire incomplete mulitpart uploads »

Replication

Q: What is Amazon S3 Replication?

Amazon S3 Replication enables automatic, asynchronous copying of objects across Amazon S3 buckets. Buckets that are configured for object replication can be owned by the same AWS account or by different accounts. You can copy objects between different AWS Regions (S3 Cross-Region Replication), or within the same AWS Region (S3 Same-Region Replication).

Q: What is Amazon S3 Cross-Region Replication (CRR)?

CRR is an Amazon S3 feature that automatically replicates data between buckets across different AWS Regions. With CRR, you can set up replication at a bucket level, a shared prefix level, or an object level using S3 object tags. You can use CRR to provide lower-latency data access in different geographic regions. CRR can also help if you have a compliance requirement to store copies of data hundreds of miles apart. You can use CRR to change account ownership for the replicated objects to protect data from accidental deletion. To learn more about CRR, please visit the replication developer guide.

Q: What is Amazon S3 Same-Region Replication (SRR)?

SRR is an Amazon S3 feature that automatically replicates data between buckets within the same AWS Region. With SRR, you can set up replication at a bucket level, a shared prefix level, or an object level using S3 object tags. You can use SRR to make a second copy of your data in the same AWS Region. SRR helps you address data sovereignty and compliance requirements by keeping a copy of your data in a separate AWS account in the same region as the original. You can use SRR to change account ownership for the replicated objects to protect data from accidental deletion. You can also use SRR to easily aggregate logs from different S3 buckets for in-region processing, or to configure live replication between test and development environment. To learn more about SRR, please visit the replication developer guide.

Q: How do I enable Amazon S3 Replication (Cross-Region Replication and Same-Region Replication)?

Amazon S3 Replication (CRR and SRR) is configured at the S3 bucket level, a shared prefix level, or an object level using S3 object tags. You add a replication configuration on your source bucket by specifying a destination bucket in the same or different AWS region for replication.

You can use either the AWS Management Console, the REST API, the AWS CLI, or the AWS SDKs to enable replication. Versioning must be enabled for both the source and destination buckets to enable replication. To learn more, please visit overview of setting up Replication in the Amazon S3 Developer Guide.

Q: Can I use S3 Replication (CRR and SRR) with S3 Lifecycle rules?

With S3 Replication (CRR and SRR), you can establish replication rules to make copies of your objects into another storage class, in the same or a different region. Lifecycle actions are not replicated, and if you want the same lifecycle configuration applied to both source and destination buckets, enable the same lifecycle configuration on both.

For example, you can configure a lifecycle rule to migrate data from the S3 Standard storage class to the S3 Standard-IA or S3 One Zone-IA storage class or archive data to S3 Glacier on the destination bucket.

You can find more information about lifecycle configuration and replication on the S3 Replication developer guide.

Q: Can I use replication with objects encrypted by AWS Key Management Service (KMS)?

Yes, you can replicate KMS-encrypted objects by providing a destination KMS key in your replication configuration.

Learn more about replicating KMS-encrypted objects »

Q: Are objects securely transferred and encrypted throughout replication process?

Yes, objects remain encrypted throughout the replication process. The encrypted objects are transmitted securely via SSL from the source region to the destination region (CRR) or within the same region (SRR).

Q: Can I use replication across AWS accounts to protect against malicious or accidental deletion?

Yes, for CRR and SRR, you can set up replication across AWS accounts to store your replicated data in a different account in the target region. You can use Ownership Overwrite in your replication configuration to maintain a distinct ownership stack between source and destination, and grant destination account ownership to the replicated storage.

Q: What is Amazon S3 Replication Time Control?

Amazon S3 Replication Time Control is a feature of S3 Replication that helps you meet compliance or business requirements for predictable replication times. S3 Replication Time Control is designed to replicate most objects in seconds, 99% of objects within 5 minutes, and 99.99% of objects within 15 minutes. S3 Replication Time Control is backed by a Service Level Agreement (SLA) commitment that 99.9% of objects will be replicated in 15 minutes for each replication region pair during any billing month. Replication Time works with all S3 Replication features. To learn more, please visit the replication developer guide.

Q: How do I enable Amazon S3 Replication Time Control?

Amazon S3 Replication Time Control is enabled as an option in your S3 Replication configuration. You can create a new S3 Replication policy with S3 Replication Time Control, or enable the feature on an existing policy.

You can use either the S3 Management Console, the REST API, the AWS CLI, or the AWS SDKs to configure replication. To learn more, please visit overview of setting up Replication in the Amazon S3 Developer Guide.

Q: What are Amazon S3 Replication metrics and events?

Amazon S3 Replication metrics and events provides visibility into Amazon S3 Replication Time Control activity. With S3 Replication metrics, you can monitor the total number and size of objects that are pending replication, and the maximum replication time for each S3 Replication rule configured with S3 Replication Time Control. Replication metrics are available through the Amazon S3 Management Console and through Amazon CloudWatch. S3 Replication events will notify you in the rare instance when an object takes more than 15 minutes to replicate, and also when that object replicates successfully to their destination. Like other Amazon S3 events, S3 Replication events are available through Amazon Simple Queue Service (Amazon SQS), Amazon Simple Notification Service (Amazon SNS), or AWS Lambda.

Q: How do I enable Amazon S3 Replication Time Control metrics and events?

Amazon S3 Replication metrics and events are enabled automatically for each S3 Replication rule configured with S3 Replication Time Control. Once you enable Replication Time Control, you can access metrics through the Amazon S3 Management Console and Amazon CloudWatch. Like other Amazon S3 events, S3 Replication events are available through Amazon Simple Queue Service (Amazon SQS), Amazon Simple Notification Service (Amazon SNS), or AWS Lambda. To learn more, please visit enabling Replication metrics in the Amazon S3 Developer Guide.

Q: What is the Amazon S3 Replication Time Control Service Level Agreement (SLA)?

Amazon S3 Replication Time Control is designed to replicate 99.99% of your objects within 15 minutes, and is backed by a service level agreement. If fewer than 99.9% of your objects are replicated in 15 minutes for each replication region pair during a monthly billing cycle, the S3 RTC SLA provides for a service credit on any object that took longer than 15 minutes to replicate. The service credit covers a percentage of all replication-related charges associated with the objects that did not meet the SLA, including the RTC fee, replication bandwidth and request charges, and the cost associated with storing your replica in the destination region in the monthly billing cycle affected. To learn more, read the S3 Replication Time Control SLA.

Q: How do I know if I qualify for an Amazon S3 Replication Time Control SLA service credit?

You are eligible for an SLA credit for Amazon S3 Replication Time Control, if fewer than 99.9% of your objects are replicated in 15 minutes for each replication region pair during a monthly billing cycle. For full details on all of the terms and conditions of the SLA, as well as details on how to submit a claim, please see the S3 Replication Time Control SLA.

Q: What is the pricing for S3 Replication and S3 Replication Time Control?

For CRR and SRR, you pay the Amazon S3 charges for storage in the destination S3 storage class you select, in addition to the storage charges for the primary copy, replication PUT requests, and applicable infrequent access storage retrieval fees. For CRR, you also pay for inter-region Data Transfer OUT From Amazon S3 to your destination region. Pricing for the replicated copy of storage is based on the destination AWS Region, while pricing for requests and inter-region data transfers are based on the source AWS Region. For S3 Replication Time Control, you pay an additional Data Transfer fee and S3 Replication Metrics charges that are billed at the same rate as Amazon CloudWatch custom metrics. For more information, please visit the S3 pricing page.

If the source object is uploaded using the multipart upload feature, then it is replicated using the same number of parts and part size. For example, a 100 GB object uploaded using the multipart upload feature (800 parts of 128 MB each) will incur request cost associated with 802 requests (800 Upload Part requests + 1 Initiate Multipart Upload request + 1 Complete Multipart Upload request) when replicated. You will incur a request charge of $0.00401 (802 requests x $0.005 per 1,000 requests) and (if the replication was between different AWS regions) a charge of $2.00 ($0.020 per GB transferred x 100 GB) for inter-region data transfer. After replication, the 100 GB will incur storage charges based on the destination region.

Amazon S3 and IPv6

Q: What is IPv6?

Every server and device connected to the Internet must have a unique address. Internet Protocol Version 4 (IPv4) was the original 32-bit addressing scheme. However, the continued growth of the Internet means that all available IPv4 addresses will be utilized over time. Internet Protocol Version 6 (IPv6) is the new addressing mechanism designed to overcome the global address limitation on IPv4.

Q: What can I do with IPv6?

Using IPv6 support for Amazon S3, applications can connect to Amazon S3 without the need for any IPv6 to IPv4 translation software or systems. You can meet compliance requirements, more easily integrate with existing IPv6-based on-premises applications, and remove the need for expensive networking equipment to handle the address translation. You can also now utilize the existing source address filtering features in IAM policies and bucket policies with IPv6 addresses, expanding your options to secure applications interacting with Amazon S3.

Q: How do I get started with IPv6 on Amazon S3?

You can get started by pointing your application to Amazon S3's new "dual-stack" endpoint, which supports access over both IPv4 and IPv6. In most cases, no further configuration is required for access over IPv6, because most network clients prefer IPv6 addresses by default.

Q: Should I expect a change in Amazon S3 performance when using IPv6?

No, you will see the same performance when using either IPv4 or IPv6 with Amazon S3.

Q: What can I do if my clients are impacted by policy, network, or other restrictions in using IPv6 for Amazon S3?

Applications that are impacted by using IPv6 can switch back to the standard IPv4-only endpoints at any time.

Q: Can I use IPv6 with all Amazon S3 features?

No, IPv6 support is not currently available when using Website Hosting and access via BitTorrent. All other features should work as expected when accessing Amazon S3 using IPv6.

Q: Is IPv6 supported in all AWS Regions?

You can use IPv6 with Amazon S3 in all commercial AWS Regions except China (Beijing) and China (Ningxia). You can also use IPv6 in the AWS GovCloud (US) Region.

Ready to get started?

Check out the product features

Learn more about features for data management, security, access management, analytics, and more.

Learn more »

Sign up for a free account

Instantly get access to the AWS Free Tier and start experimenting with Amazon S3.

Sign up »

Start building in the console

Get started building with Amazon S3 in the AWS Console.

Get started »

Amazon EBS FAQs

General

Q: Are Amazon EBS volume and snapshot ID lengths changing in 2018?

Yes, please visit the EC2 FAQs page for more details.

Q: What happens to my data when an Amazon EC2 instance terminates?

Unlike the data stored on a local instance store (which persists only as long as that instance is alive), data stored on an Amazon EBS volume can persist independently of the life of the instance. Therefore, we recommend that you use the local instance store only for temporary data. For data requiring a higher level of durability, we recommend using Amazon EBS volumes or backing up the data to Amazon S3. If you are using an Amazon EBS volume as a root partition, set the Delete on termination flag to "No" if you want your Amazon EBS volume to persist outside the life of the instance.

Q: What kind of performance can I expect from Amazon EBS volumes?

Amazon EBS provides four current generation volume types: Provisioned IOPS SSD (io1), General Purpose SSD (gp2), Throughput Optimized HDD (st1) and Cold HDD (sc1). These volume types differ in performance characteristics and price, allowing you to tailor your storage performance and cost to the needs of your applications. The average latency between EC2 instances and EBS is single digit milliseconds. For more performance information see the EBS product details page.

For more information about Amazon EBS performance guidelines, see Increasing EBS Performance.

Q: Which volume should I choose?

Amazon EBS includes two major categories of storage: SSD-backed storage for transactional workloads (performance depends primarily on IOPS) and HDD-backed storage for throughput workloads (performance depends primarily on throughput, measured in MB/s). SSD-backed volumes are designed for transactional, IOPS-intensive database workloads, boot volumes, and workloads that require high IOPS. SSD-backed volumes include Provisioned IOPS SSD (io1) and General Purpose SSD (gp2). HDD-backed volumes are designed for throughput-intensive and big-data workloads, large I/O sizes, and sequential I/O patterns. HDD-backed volumes include Throughput Optimized HDD (st1) and Cold HDD (sc1).

Q: How do I modify the capacity, performance, or type of an existing EBS volume?

Changing a volume configuration is easy. The Elastic Volumes feature allows you to increase capacity, tune performance, or change your volume type with a single CLI call, API call or a few console clicks. For more information about Elastic Volumes, see the Elastic Volumes documentation.

Q: Are EBS Standard Volumes still available?

EBS Standard Volumes have been renamed to EBS Magnetic volumes. Any existing volumes will not have been changed as a result of this and there are no functional differences in the EBS Magnetic offering compared to EBS Standard. The name of this offering was changed to avoid confusion with our General Purpose SSD (gp2) volume type which is our recommended default volume type.

Q: Are Provisioned IOPS SSD (io1) volumes available for all Amazon EC2 instance types?

Yes, Provisioned IOPS SSD (io1) volumes are available for all Amazon EC2 Instance Types. To enable your EC2 instances to use the IOPS provisioned on an EBS volume consistently and predictably, you can launch selected EC2 instance types as EBS-optimized instances. EBS-optimized instances deliver dedicated throughput between Amazon EC2 and Amazon EBS, with options between 62.5 MB/s and 2,375 MB/s depending on the instance type used. To achieve optimum performance for workloads requiring more than 32,000 IOPS or 500 MB/s, you should use EC2 instances based on the Nitro system.

Performance

Q: What level of performance consistency can I expect to see from my Provisioned IOPS SSD (io1) volumes?

When attached to EBS-optimized instances, Provisioned IOPS SSD (io1) volumes are designed to deliver within 10% of the provisioned IOPS performance 99.9% of the time in a given year. Your exact performance depends on your application's I/O requirements.

Q: What level of performance latency can I expect to see from my Provisioned IOPS SSD (io1) volumes?

When attached to EBS-optimized instances, Provisioned IOPS volumes can achieve single digit millisecond latencies. Your exact performance depends on your application's I/O requirements.

Q: Does the I/O size of my application reads and writes affect the rate of IOPS I get from my Provisioned IOPS SSD (io1) volumes?

Yes. For a given allocation of resources, the IOPS rate you get depends on the I/O size of your application reads and writes.

Provisioned IOPS volumes have a base I/O size of 16KB. Volumes up to 32,000 IOPS have an enhanced throughput I/O size of 256KB, up to 500 MB/s throughput. Every increase in I/O size above the base size (16KB) or enhanced throughput size (256KB) linearly increase the resources you need to achieve the same IOPS rate.

For example, if you have provisioned a volume with 40,000 IOPS it will achieve up to 40,000 16KB writes per second, 20,000 32KB writes per second, or 10,000 64KB writes per second, and so on. If you have provisioned a volume with 500 IOPS it can achieve up to 500 256KB writes per second, 250 512KB writes per second, or 125 1024KB writes per second, and so on.

You can use Amazon CloudWatch to monitor your throughput and I/O sizes.

Q: What factors can affect the performance consistency I see with Provisioned IOPS SSD (io1) volumes?

Provisioned IOPS SSD (io1) volumes attached to EBS-optimized instances are designed to offer consistent performance, delivering within 10% of the provisioned IOPS performance 99.9% of the time over a given year. For maximum performance consistency with new volumes created from a snapshot, we recommend reading or writing to all of the blocks on your volume before placing it into service.

Another factor that can impact your performance is if your application isn’t sending enough I/O requests. This can be monitored by looking at your volume’s queue depth. The queue depth is the number of pending I/O requests from your application to your volume. For maximum consistency, a Provisioned IOPS volume must maintain an average queue depth (rounded to the nearest whole number) of one for every 1000 provisioned IOPS in a minute. For example, for a volume provisioned with 3000 IOPS, the queue depth average must be 3. For more information about ensuring consistent performance of your volumes, see Increasing EBS Performance.

Q: What level of performance consistency can I expect to see from my HDD-backed volumes?

When attached to EBS-optimized instances, Throughput Optimized HDD (st1) and Cold HDD (sc1) volumes are designed to deliver within 10% of the expected throughput performance 99% of the time in a given year. Your exact performance depends on your application’s I/O requirements and the performance of your EC2 instance.

Q: Does the I/O size of my application reads and writes affect the rate of throughput I get from my HDD-backed volumes?

Yes. The throughput rate you get depends on the I/O size of your application reads and writes. HDD-backed volumes process reads and writes in I/O sizes of 1MB. Sequential I/Os are merged and processed as 1 MB units while each non-sequential I/O is processed as 1MB even if the actual I/O size is smaller. Thus, while a transactional workload with small, random IOs, such as a database, won't perform well on HDD-backed volumes, sequential I/Os and large I/O sizes will achieve the advertised performance of st1 and sc1 for a longer period of time.

Q: What factors can affect the performance consistency of my HDD-backed volumes?

Throughput Optimized HDD (st1) and Cold HDD (sc1) volumes attached to EBS-optimized instances are designed to offer consistent performance, delivering within 10% of the expected throughput performance 99% of the time in a given year. There are several factors that could affect the level of consistency you see. For example, the relative balance between random and sequential I/O operations on the volume can impact your performance. Too many random small I/O operations will quickly deplete your I/O credits and lower your performance down to the baseline rate. Your throughput rate may also be lower depending on the instance selected. Although st1 can drive throughput up to 500 MB/s, performance will be limited by the separate instance-level limit for EBS traffic. Another factor is taking a snapshot which will decrease expected write performance down to the baseline rate, until the snapshot completes. This is specific to st1 and sc1.

Your performance can also be impacted if your application isn’t sending enough I/O requests. This can be monitored by looking at your volume’s queue depth and I/O size. The queue depth is the number of pending I/O requests from your application to your volume. For maximum consistency, HDD-backed volumes must maintain an average queue depth (rounded to the nearest whole number) of four or more for every 1 MB sequential I/O. For more information about ensuring consistent performance of your volumes, see Increasing EBS Performance.

Q: Can I stripe multiple volumes together to get better performance?

Yes. You can stripe multiple volumes together to achieve up to 80,000 IOPS or 2,375 MiB/s when attached to larger EC2 instances. However, performance for st1 and sc1 scales linearly with volume size so there may not be as much of a benefit to stripe these volumes together.

Q: How does Amazon EBS handle issues like storage contention?

EBS is a multi-tenant block storage service. We employ rate limiting as a mechanism to avoid resource contention. This starts with having defined performance criteria for the volumes – our volume types (gp2, PIOPS, st1, and sc1) all have defined performance characteristics in terms of IOPS and throughput. The next step is defining performance at the instance level. Each EBS Optimized instance has defined performance (both throughput and IOPS) for the set of EBS volumes attached to the instance. A customer can, therefore, size instances and volumes to get the desired level of performance. In addition, customers can use our reported metrics to observe instance level and volume level performance. They can set alarms to determine if what they are seeing does not match the expected performance – the metrics can also help determine if customers are configured at the right type of instance with the right amount of performance at the volume level or not. On the EBS end, we use the configured performance to inform how we allocate the appropriate instance and EBS infrastructure to support the volumes. By appropriately allocating infrastructure, we avoid resource contention. Additionally, we constantly monitor our infrastructure. This monitoring allows us to detect infrastructure failure (or imminent infrastructure failure) and therefore, move the volumes pro-actively to functioning hardware while the underlying infrastructure is either repaired or replaced (as appropriate).

Snapshots

Q: How can I use EBS direct APIs for Snapshots?

This feature can be used via the following APIs that can be called using AWS CLI or via AWS SDK.

- List Snapshot Blocks: The ListSnapshotBlocks API operation returns the block indexes and block tokens for blocks in the specified snapshot.
- List Changed Blocks: The ListChangedBlocks API operation returns the block indexes and block tokens for blocks that are different between two specified snapshots of the same volume/snapshot lineage.

- Get Snapshot Blocks: The GetSnapshotBlock API operation returns the data in a block for the specified snapshot ID, block index, and block token.

For more information, please refer to technical documentation.

Q: What block sizes are supported by getSnapshotBlock API?

GetSnapshotBlock will return data as 512KiB blocks.

Q: Will I be able to access my snapshots using the regular Amazon S3 API?

No, snapshots are only available through the Amazon EC2 API.

Q: Do volumes need to be un-mounted to take a snapshot?

No, snapshots can be done in real time while the volume is attached and in use. However, snapshots only capture data that has been written to your Amazon EBS volume, which might exclude any data that has been locally cached by your application or OS. To ensure consistent snapshots on volumes attached to an instance, we recommend detaching the volume cleanly, issuing the snapshot command, and then reattaching the volume. For Amazon EBS volumes that serve as root devices, we recommend shutting down the machine to take a clean snapshot.

Q: Does it take longer to snapshot an entire 16 TB volume as compared to an entire 1 TB volume?

By design, an EBS Snapshot of an entire 16 TB volume should take no longer than the time it takes to snapshot an entire 1 TB volume. However, the actual time taken to create a snapshot depends on several factors including the amount of data that has changed since the last snapshot of the EBS volume.

Q: Are snapshots versioned? Can I read an older snapshot to do a point-in-time recovery?

Each snapshot is given a unique identifier, and customers can create volumes based on any of their existing snapshots.

Q: How can I discover Amazon EBS snapshots that are shared with me?

You can find snapshots that are shared with you by selecting Private Snapshots from the list in the Snapshots section of the AWS Management Console. This section lists both snapshots that you own and snapshots that are shared with you.

Q: How can I find which Amazon EBS snapshots are shared globally?

You can find snapshots that are shared globally by selecting Public Snapshots from the list in the Snapshots section of the AWS Management Console.

Q: How can I find a list of Amazon public datasets stored in Amazon EBS Snapshots?

You can use the AWS Management Console to find public datasets stored as Amazon Snapshots. Log into the console, select the Amazon EC2 Service, select Snapshots and then filter on Public Snapshots. All information on public datasets is available in our AWS Public Datasets resource center.

Q: When would I use Fast Snapshot Restore (FSR)?

You should enable FSR on snapshots if you are concerned about latency of data access when you restore data from a snapshot to a volume and want to avoid the initial performance hit during initialization. FSR is intended to help with use cases such as virtual desktop infrastructure (VDI), backup & restore, test/dev volume copies, and booting from custom AMIs. By enabling FSR on your snapshot, you will see improved and predictable performance whenever you need to restore data from that snapshot.

Q: Does enabling FSR for my snapshot speed up snapshot creation?

No. FSR-enabled snapshots improve restoring backup data from your snapshot to your volumes. FSR-enabled snapshots do not speed up snapshot creation time.

Q: How do I enable Fast Snapshot Restore (FSR)?

To use the feature, invoke the new enable-fast-snapshot-restores API on a snapshot within the availability zone (AZ) where initialized volumes are to be restored.

The FSR-enabled snapshot may be in any one of the following states: enabling, optimizing, enabled, disabling, disabled. State transitions are published as CloudWatch events and the FSR state can be checked via the describe-fast-snapshot-restores API.

Enabling FSR on a snapshot does not change any existing snapshot API interactions, and existing workflows will not need to change. FSR can be enabled or disabled on account-owned snapshots only. FSR cannot be applied to shared snapshots. You can view the list of your FSR-enabled snapshots via API or the console.

Q: How do I use Fast Snapshot Restore (FSR)?

Volumes created from an FSR-enabled snapshot are fully initialized. However, there are limits on the number of volumes that can be created with immediate full performance. These limits are expressed in the form of a credit bucket that is associated with an FSR-enabled snapshot in a given AZ. The important things to know regarding credits:

1. A single volume create operation consumes a single credit
2. The number of credits is a function of the FSR-enabled snapshot size
3. Credits refill over time
4. Maximum credit bucket size is 10

To estimate your credit bucket size and fill rate, divide 1,024 by your snapshot size. For example, a 100 GiB FSR-enabled snapshot will have the maximum balance of 10 credits with a fill rate of 10 credits every hour. A 4 TiB snapshot will have a maximum balance of 1 with a fill rate of 1 credit every 4 hours.

It's important to note that the credit bucket size is a function of the FSR-enabled snapshot size, not the size of the volumes that are created. For example, it is possible to create up to ten 1TiB volumes from a 100GiB snapshot at once.

Lastly, each AZ in which the snapshot is FSR-enabled gets its own credit bucket independent of other AZs.

Q: How many concurrent volumes can I create and what happens when I surpass this limit?

The size of the create credit bucket represents the maximum number and the balance of the credit bucket represents the number of creates available. When filled, up to 10 initialized volumes can be created from an FSR-enabled snapshot at once. Both the maximum size of the credit bucket and the credit bucket balance are published as CloudWatch metrics. Volume creations beyond the limit will proceed as if FSR is not enabled on the snapshot.

Q: How do I know when a volume was created from an FSR-enabled snapshot?

When using FSR, a new EBS-specific attribute (fastRestored) is added in the DescribeVolumes API to denote the status at create time. When a volume is created from an FSR-enabled snapshot without sufficient volume-create credits, the create will succeed but the volume will not be initialized.

Encryption

Q: What is Amazon EBS encryption?

Amazon EBS encryption offers seamless encryption of EBS data volumes, boot volumes and snapshots, eliminating the need to build and maintain a secure key management infrastructure. EBS encryption enables data at rest security by encrypting your data using Amazon-managed keys, or keys you create and manage using the AWS Key Management Service (KMS). The encryption occurs on the servers that host EC2 instances, providing encryption of data as it moves between EC2 instances and EBS storage. For more details, see Amazon EBS encryption in the Amazon EC2 User Guide.

Q: What is the AWS Key Management Service (KMS)?

AWS KMS is a managed service that makes it easy for you to create and control the encryption keys used to encrypt your data. AWS Key Management Service is integrated with other AWS services including Amazon EBS, Amazon S3, and

Amazon Redshift, to make it simple to encrypt your data with encryption keys that you manage. AWS Key Management Service is also integrated with AWS CloudTrail to provide you with logs of all key usage to help meet your regulatory and compliance needs. To learn more about KMS, visit the AWS Key Management Service product page.

Q: Why should I use EBS encryption?

You can use Amazon EBS encryption to meet security and encryption compliance requirements for data at rest encryption in the cloud. Pairing encryption with existing IAM access control policies improves your company's defense-in-depth strategy.

Q: How are my Amazon EBS encryption keys managed?

Amazon EBS encryption handles key management for you. Each newly created volume gets a unique 256-bit AES key; Volumes created from the encrypted snapshots share the key. These keys are protected by our own key management infrastructure, which implements strong logical and physical security controls to prevent unauthorized access. Your data and associated keys are encrypted using the industry-standard AES-256 algorithm.

Q: Does EBS encryption support boot volumes?

Yes.

Q: Can I create an encrypted data volume at the time of instance launch?

Yes, using customer master keys (CMKs) that are either AWS-managed or customer-managed. You can specify the volume details and encryption through a RunInstances API call with the BlockDeviceMapping parameter or through the Launch Wizard in the EC2 Console.

Q: Can I create additional encrypted data volumes at the time of instance launch that are not part of the AMI?

Yes, you can create encrypted data volume with either default or custom CMK encryption at the time of instances launch. You can specify the volume details and encryption through BlockDeviceMapping object in RunInstances API call or through Launch Wizard in EC2 Console.

Q: Can I launch an encrypted EBS instance from an unencrypted AMI?

Yes. See technical documentation for details.

Q: Can I share encrypted snapshots and AMIs with other accounts?

Yes. You can share encrypted snapshots and AMIs using a customer-managed customer master key (CMK) with other AWS accounts. See technical documentation for details.

Q: Can I ensure that all new volumes created are always encrypted?

Yes, you can enable EBS encryption by default with a single setting per region. This ensures that all new volumes are always encrypted. Refer to technical documentation for more details.

Billing and metering

Q: Will I be billed for the IOPS provisioned on a Provisioned IOPS volume when it is disconnected from an instance?

Yes, you will be billed for the IOPS provisioned when it is disconnected from an instance. When a volume is detached, we recommend you consider creating a snapshot and deleting the volume to reduce costs. For more information, see the "Underutilized Amazon EBS Volumes" cost optimization check in Trusted Advisor. This item checks your Amazon Elastic Block Store (Amazon EBS) volume configurations and warns when volumes appear to be underused.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Multi-Attach

Q: Is there an additional fee to enable Multi-Attach?

No. Multi-Attach can be enabled on an EBS Provisioned IOPS io1 volume and there will be charges for the storage (GB-Mo) and IOPS (IOPS-Mo) provisioned.

Q: Can I boot an EC2 instance using a Multi-Attach enabled volume?

No.

Q: What happens if all of my attached instances do not have the ‘deleteOnTermination’ flag set?

The volume's deleteOnTermination behavior is determined by the configuration of the last attached instance that is terminated. To ensure predictable delete on termination behavior, enable or disable 'deleteOnTermination' for all of the instances to which the volume is attached.

If you want the volume to be deleted when the attached instances are terminated, enable ‘deleteOnTermination’ for all of instances to which the volume is attached. If you want to retain the volume after the attached instances have been terminated, disable ‘deleteOnTermination’ for all attached instances. For more information, see Multi-Attach technical documentation.

Q: Can my application use Multi-Attach?

If your application does not require storage layer coordination of write operations, such as a read-only application or it enforces application level IO fencing, then your application can use Multi-Attach.

Amazon EFS FAQs

General

Q. What is Amazon Elastic File System?

Amazon EFS is a fully-managed service that makes it easy to set up, scale, and cost-optimize file storage in the Amazon Cloud. With a few clicks in the AWS Management Console, you can create file systems that are accessible to Amazon EC2 instances via a file system interface (using standard operating system file I/O APIs) and support full file system access semantics (such as strong consistency and file locking).

Amazon EFS file systems can automatically scale from gigabytes to petabytes of data without needing to provision storage. Tens, hundreds, or even thousands of Amazon EC2 instances can access an Amazon EFS file system at the same time, and Amazon EFS provides consistent performance to each Amazon EC2 instance. Amazon EFS is designed to be highly durable and highly available. With Amazon EFS, there is no minimum fee or setup costs, and you pay only for what you use.

Q. What use cases does Amazon EFS support?

Amazon EFS is designed to provide performance for a broad spectrum of workloads and applications, including Big Data and analytics, media processing workflows, content management, web serving, and home directories.

Q. When should I use Amazon EFS vs. Amazon S3 vs. Amazon Elastic Block Store (EBS)?

Amazon Web Services (AWS) offers cloud storage services to support a wide range of storage workloads.

Amazon EFS is a file storage service for use with Amazon EC2. Amazon EFS provides a file system interface, file system access semantics (such as strong consistency and file locking), and concurrently-accessible storage for up to thousands of Amazon EC2 instances.

Amazon EBS is a block level storage service for use with Amazon EC2. Amazon EBS can deliver performance for workloads that require the lowest-latency access to data from a single EC2 instance.

Amazon S3 is an object storage service. Amazon S3 makes data available through an Internet API that can be accessed anywhere.

Learn more about what to evaluate when considering Amazon EFS.

Q. What regions is Amazon EFS currently available in?

Please refer to Regional Products and Services for details of Amazon EFS service availability by region.

Q. How do I get started using Amazon EFS?

To use Amazon EFS, you must have an AWS account. If you do not already have an AWS account, you can sign up for an AWS account and instantly get access to the AWS Free Tier.

Once you have created an AWS account, please refer to the Amazon EFS Getting Started guide to begin using Amazon EFS. You can create a file system via the AWS Management Console, the AWS Command Line Interface (AWS CLI), and Amazon EFS API (and various language-specific SDKs).

Q. How do I access a file system from an Amazon EC2 instance?

To access your file system, you mount the file system on an Amazon EC2 Linux-based instance using the standard Linux mount command and the file system's DNS name. To simplify accessing your EFS file systems, we recommend using the EFS mount helper utility. Once mounted, you can work with the files and directories in your file system just like you would with a local file system.

Amazon EFS uses the Network File System version 4 (NFS v4) protocol. For a step-by-step example of how to access a file system from an Amazon EC2 instance, please see the guide here.

Q. What Amazon EC2 instance types and AMIs work with Amazon EFS?

Amazon EFS is compatible with all Linux-based AMIs for Amazon EC2. You can mix and match the instance types connected to a single file system. For a step-by-step example of how to access a file system from an Amazon EC2 instance, please see the instance type guide here.

Q. How do I manage a file system?

Amazon EFS is a fully-managed service, so all of the file storage infrastructure is managed for you. When you use Amazon EFS, you avoid the complexity of deploying and maintaining complex file system infrastructure. An Amazon EFS file system grows and shrinks automatically as you add and remove files, so you do not need to manage storage procurement or provisioning.

You can administer a file system via the AWS Management Console, the AWS command-line interface (CLI), or the Amazon EFS API (and various language-specific SDKs). The Console, API, and SDK provide the ability to create and delete file systems, configure how file systems are accessed, create and edit file system tags, enable features like Provisioned Throughput and Lifecycle Management, and display detailed information about file systems.

Q. How do I load data into a file system?

AWS DataSync provides a fast and simple way to securely sync existing file systems with Amazon EFS. DataSync works over any network connection, including with AWS Direct Connect or AWS VPN. AWS Direct Connect provides a high bandwidth and lower latency dedicated network connection, over which you can mount your EFS file systems. You can use AWS DataSync to copy files between two EFS file systems, including those in different AWS regions and those belonging to different AWS accounts. You can also use standard Linux copy tools to move data files to Amazon EFS.

For more information about accessing a file system from an on-premises server, please see the On-premises Access section of this FAQ.

For more information about moving data to the Amazon cloud, please see the Cloud Data Migration page.

Storage classes and lifecycle management

Q. What storage classes does Amazon EFS offer?

Amazon EFS offers a Standard and an Infrequent Access storage class. The Standard storage class is designed for active file system workloads and you pay only for the file system storage you use per month. EFS Infrequent Access (EFS IA) is a lower cost storage class that's cost-optimized for files not accessed every day. Data stored on the EFS IA storage class costs 85% less than Standard and you pay a fee each time you read from or write to a file. EFS file systems transparently serve data from both storage classes.

Q. How do I move files to EFS IA?

Moving files to EFS IA starts by enabling EFS Lifecycle Management and choosing an age-off policy. Lifecycle Management automatically moves your data to the EFS IA storage class according to the lifecycle policy you choose. For example, you can automatically move files into EFS IA fourteen days of not being accessed.

Q. When should I enable Lifecycle Management?

Enable Lifecycle Management when your file system contains files that are not accessed every day to reduce your storage costs. EFS IA is ideal for EFS customers who need their full data set to be readily accessible and want to automatically save on storage costs as their files become less frequently accessed. Examples include satisfying audits, performing historical analysis, or backup and recovery.

Q. What happens when I disable EFS Lifecycle Management?

When you disable Lifecycle Management, files will no longer be moved to the Infrequent Access storage class, and any files that have already moved to EFS IA will remain there.

Q. What Amazon EFS features are supported when using EFS IA storage class?

All Amazon EFS features are supported when using the EFS IA storage class. Files smaller than 128 KiB are not eligible for Lifecycle Management and will always be stored on EFS Standard.

Q. Is there a latency difference between the EFS Standard and EFS Infrequent Access storage classes?

When reading from or writing to EFS IA, your first-byte latency is higher than that of EFS Standard. EFS Standard is designed to provide single-digit latencies on average, and EFS IA is designed to provide double-digit latencies on average.

Q. What throughput can I drive against files stored in the EFS Infrequent Access storage class?

The throughput you can drive against an EFS file system scales linearly with the amount of data stored on the EFS Standard storage class. All EFS file systems, regardless of size, can burst to 100 MiB/s of throughput. File systems with more than 1 TiB of Standard storage can burst to 100 MiB/s per TiB of data stored on EFS Standard. If you require higher amounts of throughput to EFS IA than your file system allows, use EFS Provisioned Throughput.

Data protection and availability

Q. How is Amazon EFS designed to provide high durability and availability?

Every file system object (i.e. directory, file, and link) is redundantly stored across multiple Availability Zones. In addition, a file system can be accessed concurrently from all Availability Zones in the region where it is located, which means that you can architect your application to failover from one AZ to other

AZs in the region in order to ensure the highest level of application availability. Mount targets themselves are designed to be highly available.

Q. Does Amazon EFS offer a Service Level Agreement (SLA)?

Yes. The Amazon EFS SLA provides for a service credit if a customer's monthly uptime percentage is below our service commitment in any billing cycle.

Q. How do I back up a file system?

Amazon EFS is designed to be highly durable. You can use AWS Backup to schedule automatic, incremental backups of your Amazon EFS file systems. For more information, please see the Amazon EFS Walkthrough: Backup Solutions for Amazon EFS File Systems.

Q. How do I access my file system from outside my VPC?

Amazon EC2 instances within your VPC can access your file system directly, and Amazon EC2 Classic instances outside your VPC can mount a file system via ClassicLink. Amazon EC2 instances in other VPCs can access your file system if connected using a VPC peering connection or VPC Transit Gateway. On-premises servers can mount your file systems via an AWS Direct Connect or AWS VPN connection to your VPC.

Scale and performance

Q. How much data can I store?

Amazon EFS file systems can store petabytes of data. Amazon EFS file systems are elastic, and automatically grow and shrink as you add and remove files. You do not provision file system size up front, and you pay only for what you use.

Q. How many Amazon EC2 instances can connect to a file system?

Amazon EFS supports one to thousands of Amazon EC2 instances connecting to a file system concurrently.

Q. How many file systems can I create?

You can create up to 1,000 file systems per region. For information on Amazon EFS limits, please visit the Amazon EFS Limits page.

Q. How does Amazon EFS performance compare to that of other storage solutions?

Amazon EFS file systems are distributed across an unconstrained number of storage servers, enabling file systems to grow elastically to petabyte-scale and allowing massively parallel access from Amazon EC2 instances to your data. Amazon EFS's distributed design avoids the bottlenecks and constraints inherent to traditional file servers.

This distributed data storage design means that multi-threaded applications and applications that concurrently access data from multiple Amazon EC2 instances can drive substantial levels of aggregate throughput and IOPS. Big Data and analytics workloads, media processing workflows, content management and web serving are examples of these applications.

The table below compares high-level performance and storage characteristics for AWS's file and block cloud storage offerings.

	Amazon EFS	Amazon EBS (io1)
Per-operation latency	Low, consistent	Lowest, consistent
Throughput scale	Multiple GBs per second	Single GB per second

Amazon EFS's distributed nature enables high levels of availability, durability, and scalability. This distributed architecture results in a small latency overhead for each file operation. Due to this per-operation latency, overall throughput generally increases as the average I/O size increases, since the overhead is amortized over a larger amount of data. Amazon EFS's support for highly parallelized workloads (i.e. with consistent operations from multiple threads and multiple EC2 instances) enables high levels of aggregate throughput and IOPS.

Q. What's the difference between "General Purpose" and "Max I/O" performance modes? Which one should I choose?

"General Purpose" performance mode is appropriate for most file systems, and is the mode selected by default when you create a file system. "Max I/O" performance mode is optimized for applications where tens, hundreds, or thousands of EC2 instances are accessing the file system — it scales to higher levels of aggregate throughput and operations per second with a tradeoff of slightly higher latencies for file operations. For more information, please see the documentation on File System Performance.

Q. How much throughput can a file system support?

The throughput available to a file system scales as a file system grows. Because file-based workloads are typically spiky — requiring high levels of throughput for periods of time and lower levels of throughput the rest of the time — Amazon EFS is designed to burst to allow high throughput levels for periods of time. All file systems deliver a consistent baseline performance of 50 MB/s per TB of Standard class storage, all file systems (regardless of size) can burst to 100 MB/s, and file systems with more than 1TB of Standard class storage can burst to 100 MB/s per TB. As you add data to your file system, the maximum throughput available to the file system scales linearly and automatically with your storage in the Amazon EFS Standard storage class.

File system throughput is shared across all Amazon EC2 instances connected to a file system. For example, a 1TB file system that can burst to 100 MB/s of throughput can drive 100 MB/s from a single Amazon EC2 instance, or 10 Amazon EC2 instances can collectively drive 100 MB/s. For more information, please see the documentation on File System Performance.

Provisioned Throughput

Q. What is Provisioned Throughput and when should I use it?

Provisioned Throughput enables Amazon EFS customers to provision their file system’s throughput independent of the amount of data stored, optimizing their file system throughput performance to match their application’s needs.

Amazon EFS Provisioned Throughput is available for applications with a high throughput to storage (MB/s per TB) ratio. For example, customers using Amazon EFS for development tools, web serving or content management applications, where the amount of data in their file system is low relative to throughput demands, are able to instantly get the high levels of throughput their applications require.

You can select your file system’s throughput mode via the AWS Console, AWS CLI, or AWS API. For more details, see the documentation on Provisioned Throughput.

Q. How does Amazon EFS Provisioned Throughput work?

When you select Provisioned Throughput for your file system, you can provision the throughput of your file system independently from the amount of data stored and pay for the storage and Provisioned Throughput separately. (ex. $0.30 per GB-Month for Standard storage and $6.00 per MB/s-Month for Provisioned Throughput in US-East (N. Virginia)).

When you select the default Bursting Throughput mode, the throughput of your file system scales linearly with the amount of data stored in the Amazon EFS Standard storage class. In the default Bursting Throughput mode, you get a baseline rate of 50 KB/s per GB of throughput included with the price of Standard storage.

Provisioned Throughput also includes 50 KB/s per GB (or 1 MB/s per 20 GB) of throughput in the price of Stanadrd storage. For example, if you store 20 GB for a month on Amazon EFS Stanadrd and configure a throughput of 5 MB/s for a month you will be billed for 20 GB-Month of storage and 4 (5-1) MB/s-Month of throughput.

Q. How will I be billed in Provisioned Throughput mode?

In the Provisioned Throughput mode, you are billed for storage you use and throughput you provisioned independently. You are billed hourly in the following dimensions:

- **Storage (per GB-Month)** - You are billed for the amount of storage you use in GB-Month.
- **Throughput (per MB/s-Month)** – You are billed for throughput you provision in MB/s-Month.

Q. How often can I change my file system's Provisioned Throughput?

If your file system is in Provisioned Throughput mode, you can increase the provisioned throughput of your file system as often as you want. You can decrease your file system throughput in Provisioned Throughput mode or change between Provisioned Throughput and the default Bursting Throughput modes as long as it's been more than 24 hours since the last decrease or throughput mode change.

Q. What is the throughput of my file system if the Provisioned Throughput mode is set less than the Baseline Throughput I am entitled to in the bursting mode?

In the default Bursting Throughput mode, the throughput of your file system scales with the amount of data stored. If your file system in the Provisioned Throughput mode grows in size after the initial configuration, it is possible that your file system has a higher baseline rate in the Bursting Throughput mode than the Provisioned Throughput mode.

In such cases, your file system throughput will be the throughput it is entitled to in the default Bursting Throughput mode and you will not incur any additional charge for the throughput beyond the bursting storage cost. You will also be able to burst according to the Amazon EFS throughput bursting model.

Access Control

Q. How do I control which Amazon EC2 instances can access my file system?

You control which EC2 instances can access your file system using VPC security group rules and AWS Identity and Access Management (IAM) policies. Use VPC security groups to control the network traffic to and from your file system. Attach an IAM policy to your file system to control which clients can mount your file system and with what permissions, and use EFS Access Points to manage application access. Control access to files and directories with POSIX-compliant user and group-level permissions.

Q. How can I use IAM policies to manage file system access?

Using the EFS console, you can apply common policies to your file system such as disabling root access, enforcing read-only access, or enforcing that all connections to your file system are encrypted. You can also apply more advanced policies such as granting access to specific IAM roles, including those in other AWS accounts.

Access Points

Q. What is an EFS Access Point?

EFS Access Points simplify providing applications access to shared data sets in an EFS file system. EFS Access Points work together with AWS IAM and enforce an operating system user and group, and a directory for every file system request made through the access point. You can create multiple access points per file system and use them to provide access to specific applications.

Q. Why should I use EFS Access Points?

EFS Access Points represent a flexible way to manage application access in NFS environments with increased scalability, security, and ease of use. Use cases that can benefit from EFS Access Points include container-based environments where developers build and deploy their own containers, data science applications that require access to production data, and sharing a specific directory in your file system with other AWS accounts.

Q. How do EFS Access Points work?

When you create an EFS Access Point, you can configure an operating system user and group, and a root directory for all connections that use it. If you specify the root directory's owner, EFS will automatically create it with the permissions you provide the first time a client connects to the access point. You can also update your file system's IAM policy to apply to your access points. For example, you can apply a policy that requires a specific IAM identity in order to connect to a given access point. For more information, see the EFS user guide.

Encryption

Q: What is Amazon EFS Encryption?

Amazon EFS offers the ability to encrypt data at rest and in transit.

Data encrypted at rest is transparently encrypted while being written, and transparently decrypted while being read, so you don't have to modify your applications. Encryption keys are managed by the AWS Key Management Service (KMS), eliminating the need to build and maintain a secure key management infrastructure.

Data encryption in transit uses industry standard Transport Layer Security (TLS) 1.2 to encrypt data sent between your clients and EFS file systems.

Encryption of data at rest and of data in transit can be configured together or separately to help meet your unique security requirements.

For more details, see the user documentation on Encryption.

Q: What is the AWS Key Management Service (KMS)?

AWS KMS manages the encryption keys for encrypted data at rest on EFS file systems. AWS KMS is a managed service that makes it easy for you to create and control the encryption keys used to encrypt your data. AWS Key Management Service is integrated with AWS services including Amazon EFS, Amazon EBS, and Amazon S3, to make it simple to encrypt your data with encryption keys that you manage. AWS Key Management Service is also integrated with AWS

CloudTrail to provide you with logs of all key usage to help meet your regulatory and compliance needs.

Q: How do I enable encryption for my Amazon EFS file system?

You can enable encryption at rest in the EFS console or by using the AWS CLI or SDKs. When creating a new file system in the EFS console, click “Create File System” and click the checkbox to enable encryption.

Data can be encrypted in transit between your Amazon EFS file system and its clients by using the EFS mount helper.

Encryption of data at rest and of data in transit can be configured together or separately to help meet your unique security requirements.

For more details, see the user documentation on Encryption.

Q: Does encryption impact Amazon EFS performance?

Encrypting your data has a minimal effect on I/O latency and throughput.

On-premises access

Q: How do I access an EFS file system from servers in my on-premises datacenter?

To access EFS file systems from on-premises, you must have an AWS Direct Connect or AWS VPN connection between your on-premises datacenter and your Amazon VPC.

You mount an EFS file system on your on-premises Linux server using the standard Linux mount command for mounting a file system via the NFSv4.1 protocol.

For more information about accessing EFS file systems from on-premises servers, please see the documentation.

Q: What can I do by enabling access to my EFS file systems from my on-premises servers?

You can mount your Amazon EFS file systems on your on-premises servers, and move file data to and from Amazon EFS using standard Linux tools and scripts or AWS DataSync. The ability to move file data to and from Amazon EFS file systems enables three use cases.

First, you can migrate data from on-premises datacenters to permanently reside in Amazon EFS file systems.

Second, you can support cloud bursting workloads to offload your application processing to the cloud. You can move data from your on-premises servers into your EFS file systems, analyze it on a cluster of EC2 instances in your Amazon VPC, and store the results permanently in your EFS file systems or move the results back to your on-premises servers.

Third, you can periodically copy your on-premises file data to EFS to support backup and disaster recovery scenarios.

Q: Can I access my Amazon EFS file system concurrently from my on-premises datacenter servers as well as Amazon EC2 instances?

Yes, you can access your Amazon EFS file system concurrently from servers in your on-premises datacenter as well as Amazon EC2 instances in your Amazon VPC. Amazon EFS provides the same file system access semantics, such as strong data consistency and file locking, across all EC2 instances and on-premises servers accessing a file system.

Q: What is the recommended best practice when moving file data to and from on-premises servers?

Because of the propagation delay tied to data traveling over long distances, the network latency of the network connection between your on-premises datacenter and your Amazon VPC can be tens of milliseconds. If your file operations are serialized, the latency of the network connection directly impacts your read and write throughput; in essence, the volume of data you can read or write during a period of time is bounded by the amount of time it takes for each read and write operation to complete. To maximize your throughput, parallelize your file operations so that multiple reads and writes are processed by EFS concurrently. Standard tools like GNU parallel enable you to parallelize the copying of file data. For more information, see the online documentation.

Q: How do I copy existing data from on-premises file storage to Amazon EFS?

There are a number of methods to copy existing on-premises data into Amazon EFS. AWS DataSync provides a fast and simple way to securely sync existing file systems into Amazon EFS, and works over any network, including AWS Direct Connect

AWS Direct Connect provides a high bandwidth and lower latency dedicated network connection over which you can mount your Amazon EFS file systems. Once mounted, you can use DataSync to copy data into Amazon EFS up to 10x faster than standard Linux copy tools.

For more information on AWS DataSync, please see the DataSync section of this FAQ.

AWS DataSync

Q. What is AWS DataSync?

AWS DataSync is an online data transfer service that makes it faster and simpler to move data between on-premises storage and Amazon EFS. DataSync uses a purpose-built protocol to accelerate and secure transfer over the Internet or AWS Direct Connect, at speeds up to 10 times faster than open-source tools. Using DataSync you can perform one-time data migrations, transfer on-premises data for timely in-cloud analysis, and automate replication to AWS for data protection and recovery. To learn more, visit the AWS DataSync page.

Q: How do I copy data into or out of my EFS file system with AWS DataSync?

To get started with AWS DataSync you first deploy a software agent that is available for download from the AWS Management Console. Once deployed, you can use the console or AWS Command Line Interface (CLI) to connect the agent to your on-premises or in-cloud file systems using the Network File System (NFS) protocol, select your Amazon EFS file system, and start copying data.

Q. Can EFS data be copied between regions with AWS DataSync?

Yes, you can use AWS DataSync to transfer files between two EFS file systems, including ones in different AWS Regions or ones belonging to different AWS accounts. To learn more, see the documentation.

Compatibility

Q. What interoperability and compatibility is there between existing AWS services and Amazon EFS?

Amazon EFS is integrated with a number of other AWS services, including Amazon CloudWatch, AWS CloudFormation, AWS CloudTrail, AWS IAM, and AWS Tagging services.

Amazon CloudWatch allows you to monitor file system activity using metrics. AWS CloudFormation allows you to create and manage file systems using templates.

AWS CloudTrail allows you to record all Amazon EFS API calls in log files.

AWS Identity and Access Management (IAM) allows you to control who can administer your file system. AWS Tagging services allows you to label your file systems with metadata that you define.

Q. What type of locking does Amazon EFS support?

Locking in Amazon EFS follows the NFSv4.1 protocol for advisory locking, and enables your applications to use both whole file and byte range locks.

Q. Are file system names global (like Amazon S3 bucket names)?

Every file system has an automatically generated ID number that is globally unique. You can tag your file system with a name, and these names do not need to be unique.

Pricing and billing

Q. How much does Amazon EFS cost?

With Amazon EFS, you pay only for what you use per month.

When using the Provisioned Throughput mode you pay for the throughput you provision per month. There is no minimum fee and there are no set-up charges.

EFS IA is priced based on the amount of storage used and the amount of data accessed. Until Lifecycle Management fully moves your file to EFS IA, it is stored on EFS Standard and billed at the Standard rate.

For more Amazon EFS pricing information, please visit the Amazon EFS Pricing page.

Q. Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Amazon FSx for Lustre FAQs

General

Q: What is Amazon FSx for Lustre?

A: Amazon FSx for Lustre makes it easy and cost effective to launch and run the world's most popular high-performance file system.

The open source Lustre file system is designed for applications that require fast storage – where you want your storage to keep up with your compute. Lustre was built to solve the problem of quickly and cheaply processing the world's ever-growing data sets, and it's the most widely used file system for the 500 fastest computers in the world.

As a fully managed service, Amazon FSx brings Lustre to the masses, allowing you to use it for any workload where storage speed matters. Amazon FSx eliminates the traditional complexity of setting up and managing high-performance Lustre file systems, allowing you in minutes to spin up and run a battle-tested high-performance file system. It also provides multiple deployment options so you can optimize cost for your needs.

Amazon FSx also integrates with Amazon S3, making it easy for you to process cloud data sets with the Lustre high-performance file system. When linked to an S3 bucket, an FSx for Lustre file system transparently presents S3 objects as files and allows you to write changed data back to S3.

Q: What use cases does Amazon FSx for Lustre support?

A: Use Amazon FSx for Lustre for workloads where speed matters, such as machine learning, high performance computing (HPC), video processing, financial modeling, genome sequencing, and electronic design automation (EDA).

Q: How do I get started with Amazon FSx for Lustre?

A: To use Amazon FSx for Lustre, you must have an AWS account. If you do not already have an AWS account, you can sign-up for an AWS account.

Once you have created an AWS account, you can create a file system via the AWS Management Console, the AWS Command Line Interface (AWS CLI), and Amazon FSx API (and various language-specific SDKs). To learn more, get started here.

Q: What is the difference between scratch and persistent deployment options?

A: Amazon FSx for Lustre provides two deployment options: scratch and persistent.

Scratch file systems are designed for temporary storage and shorter-term processing of data. Data is not replicated and does not persist if a file server fails.

Persistent file systems are designed for longer-term storage and workloads. The file servers are highly available, and data is automatically replicated within the AWS Availability Zone (AZ) that is associated with the file system. The data volumes attached to the file servers are replicated independently from the file servers to which they are attached.

Q: When and why should I use the persistent FSx for Lustre versus the scratch FSx for Lustre deployment option?

A: Use scratch file systems when you need cost-optimized storage for short-term, processing-heavy workloads.

Use persistent file systems for workloads that run for extended periods or indefinitely, and may be sensitive to disruptions in availability.

Q: What are the availability and durability characteristics of FSx for Lustre file systems?

A: : Amazon FSx for Lustre provides a parallel file system. In parallel file systems, data is stored across multiple network file servers to maximize performance and reduce bottlenecks, and each server has multiple disks. Larger file systems have more file servers and disks than smaller file systems.

On a persistent file system, if a file server becomes unavailable it is replaced automatically and within minutes. In the meantime, client requests for data on that server transparently retry and eventually succeed after the file server is replaced. With persistent file systems, data is replicated on disks and any failed disks are automatically replaced behind the scenes, transparently.

On a scratch file system, file servers are not replaced if they fail and data is not replicated. If a file server or a storage disk becomes unavailable, files stored on other servers are still accessible. If clients try to access files that are on the unavailable server, clients will get an I/O error. The following table provides the availability/durability for which scratch file systems of example sizes are designed. As larger file systems have more file servers and more disks, the probabilities of failure are increased.

Table: Availability/durability of scratch file systems of various example sizes

Scratch file system size (TiB)	*Number of file servers*	*Availability/durability during one day*	*Availability/durability during one week*
1.2	*2*	*99.9%*	*99.4%*
2.4	*2*	*99.9%*	*99.4%*
4.8	*3*	*99.8%*	*99.2%*
9.6	*5*	*99.8%*	*98.6%*
50.4	*22*	*99.1%*	*93.9%*

Please refer to the Amazon FSx for Lustre documentation for more information.

Q: How do I create a file system with Amazon FSx for Lustre?

A: Creating an Amazon FSx for Lustre file system from the Console, CLI, or API is a simple process. Within minutes, your file system is running and accessible to your compute instances. If you link your filesystem to an S3 data lake, your objects will appear as files and directories as soon as your file system is available.

Q: What instance types and AMIs work with Amazon FSx for Lustre?

A: FSx for Lustre is compatible with the most popular Linux-based AMIs, including Amazon Linux, Amazon Linux 2, Red Hat Enterprise Linux (RHEL), CentOS, SUSE Linux and Ubuntu. With FSx for Lustre, you can mix and match the instance types and Linux AMIs that are connected to a single file system.

Q: How do I access an FSx for Lustre file system from a compute instance?

A: To access your file system from a Linux instance, you first install the open-source Lustre client on that instance. Once it’s installed, you can mount your file system using standard Linux commands. Once mounted, you can work with the files and directories in your file system just like you would with a local file system.

The Lustre client is included with Amazon Linux 2 and Amazon Linux. For Red Hat Enterprise Linux, CentOS, and Ubuntu an AWS repository for the Lustre client is supported that provides clients compatible with these operating systems. See the FSx for Lustre documentation for details.

Q: Does Amazon FSx support shared VPCs?

A: Yes, with Amazon FSx, you can create and use file systems in shared Amazon Virtual Private Clouds (VPCs) from both owner accounts and participant accounts with which the VPC has been shared. VPC sharing enables you to reduce the number of VPCs that you need to create and manage, while you still benefit from using separate accounts for billing and access control.

Q: How do I manage an FSx for Lustre file system?

A: Amazon FSx is a fully managed service, so all of the file storage infrastructure is managed for you. When you use Amazon FSx, you avoid the complexity of deploying and maintaining complex file system infrastructure.

You can administer a file system via the AWS Management Console, the AWS command-line interface (CLI), or the Amazon FSx API (and various language-specific SDKs). The Console, API, and SDK provide the ability to create and delete file systems, create and edit file system tags, and display detailed information about file systems. If you link your filesystem to an S3 data lake, your objects will appear as files and directories as soon as your system is available.

Q: How does Amazon FS for Lustre work with long-term data repositories?

A: Amazon FSx for Lustre allows you to ingest and process large volumes of file data, while periodically writing intermediate results to your data repository. Doing so allows you to restart your workload at any time from the latest data you've stored in your data repository. When your workload is done, you can write final results from your file system to your data repository, and delete your file system.

Q: If I have data in S3, how do I access it from Amazon FSx for Lustre to process it?

A: If you have data in S3, you can seamlessly link your Amazon FSx for Lustre file system with a specified S3 bucket, making the data in your Amazon S3 data repository accessible to your file system. Once your file system is created, initially the S3 objects' names and prefixes will be visible as files and directories.

By default, Amazon S3 objects are only loaded into the file system when first accessed by your applications. If your applications access objects that haven't yet been loaded into your file system, Amazon FSx for Lustre automatically loads the corresponding objects from Amazon S3. Subsequent reads of these files are served directly out of the file system with low, consistent latencies. You can optionally batch hydrate your Amazon FSx for Lustre file system.

Amazon FSx for Lustre uses parallel data transfer techniques to transfer data from S3 at up to hundreds of GBs/s.

Q: How do I export files written to my Amazon FSx for Lustre file system back to Amazon S3?

A: At any time, you can export files from your file system back to your Amazon S3 bucket using the “Export to repository” action on the console, or the CreateDataRepositoryTask API. With this API, you can quickly initiate, monitor, and cancel writing new or changed files to S3. Since it’s an AWS-native API, you can use it to orchestrate data export tasks from cloud-native workflows such as Lambda-based serverless applications.

Q: How do I export POSIX metadata associated with files written to my Amazon FSx for Lustre file system back to Amazon S3?

A: You can use the “Export to repository” action on the console, or the CreateDataRepositoryTask API to transfer symbolic links, file ownership metadata, and file timestamps to S3. S3 stores file permissions and other file metadata in the same format used by AWS DataSync and AWS Storage Gateway.

Q: If I have data on-premises how do I make it available to Amazon FSx for Lustre to process it?

A: If you have high-performance or data processing workloads running on-premises and demand for computing capacity spikes, you can cloud burst your workloads to Amazon FSx for Lustre by using Amazon Direct Connect or VPN.

Q: How do I monitor my file system’s activity?

A: Amazon FSx for Lustre provides native CloudWatch integration, allowing you to monitor file system health and performance metrics in real time. Example metrics include storage consumed, number of compute instance connections, throughput, and number of file operations per second. You can log all Amazon FSx API calls using AWS CloudTrail.

Q: Is Amazon FSx for Lustre POSIX-compliant?

A: Yes.

Q: How is Amazon FSx for Lustre integrated with Amazon SageMaker?

A: Amazon FSx for Lustre can be an input data source for Amazon SageMaker. When FSx for Lustre is used as an input data source, Amazon SageMaker training jobs are accelerated by eliminating the initial S3 download step. SageMaker jobs can get started as soon as the FSx for Lustre file system linked with the S3 bucket is created without needing to download the full machine learning training data set from S3. Data is lazy loaded as needed from Amazon S3 for processing jobs. Another benefit is reduced TCO by avoiding the repeated download of common objects (saving S3 request costs) for iterative jobs on the same data set.

Q: How is Amazon FSx for Lustre integrated with Amazon Elastic Kubernetes Service (EKS)?

A: You can use persistent storage volumes backed by FSx for Lustre using the FSx for Lustre CSI driver from Amazon EKS or your self-managed Kubernetes on AWS. See Amazon EKS documentation for details.

Q: How is Amazon FSx for Lustre integrated with AWS ParallelCluster?

A: AWS ParallelCluster is an AWS-supported open-source cluster management tool that helps you to deploy and manage High Performance Computing (HPC) clusters in the AWS Cloud. AWS ParallelCluster supports automatic creation of a new Amazon FSx for Lustre file system or the ability to use an existing Amazon FSx for Lustre file system as part of the cluster creation process.

Q: What regions is Amazon FSx for Lustre available in?

A: Please refer to Regional Products and Services for details of Amazon FSx for Lustre service availability by region.

Q: Does Amazon FSx offer a Service Level Agreement (SLA)?

A: Yes. The Amazon FSx SLA provides for a service credit if a customer's monthly uptime percentage is below our service commitment in any billing cycle.

Scale and performance

Q: What performance can I expect from Amazon FSx for Lustre?

A: Amazon FSx for Lustre file systems scale to hundreds of GB/s of throughput and millions of IOPS. Amazon FSx for Lustre also supports concurrent access to the same file or directory from thousands of compute instances, enabling rapid data checkpointing from application memory to storage (a common technique in HPC). Amazon FSx for Lustre provides consistent, sub-millisecond latencies for file operations. Scratch file systems provide a baseline of 200 MB/s and a burst of 1200 MB/s of throughput per TiB of storage provisioned. With persistent file systems, you can choose from per unit throughput capacities of 50, 100, or 200 MB/s per TiB of storage provisioned.

Q: How does throughput scale with storage capacity?

A: Scratch file systems provides a baseline of 200 MB/s and a burst of 1200 MB/s of throughput per TiB of storage provisioned. Persistent file systems provide 50, 100, or 200 MB/s of baseline throughput per TiB of storage provisioned based on the option chosen.

Q: How many instances can connect to a file system?

A: An FSx for Lustre file system can be concurrently accessed by thousands of compute instances.

Q: What file system sizes are supported by FSx for Lustre & what is the increment granularity?

A: Scratch and persistent file systems can be created in sizes of 1.2 TiB or in increments of 2.4 TiB.

Q: How many file systems can I create?

A: There is a 100-file system limit per account which can be increased upon request.

Security and compliance

Q: Does Amazon FSx for Lustre support data encryption?

A: Yes, Amazon FSx for Lustre file systems support encryption at-rest and in-transit. Scratch file systems are encrypted at-rest with keys managed by the service and encrypted in-transit when accessed from supported EC2 instances. Persistent file systems are encrypted at-rest using either customer managed or AWS Managed Customer Master Key (CMK) in AWS KMS. Persistent file systems are encrypted in-transit when accessed from supported EC2 instances. See the Amazon FSx documentation for details on regions where in-transit encryption is supported for scratch and persistent file systems.

Q: What access control capabilities does Amazon FSx provide?

A: Every FSx for Lustre resource is owned by an AWS account, and permissions to create or access a resource are governed by permissions policies. You specify the Amazon Virtual Private Cloud (VPC) in which your file system is made accessible, and you control which resources within the VPC have access to your file system using VPC Security Groups. You control who can administer your file system and backup resources (create, delete, etc.) using AWS IAM.

Q: What compliance programs does Amazon FSx support?

A: AWS has the longest-running compliance program in the cloud and are committed to helping customers navigate their requirements. Amazon FSx has been assessed to meet global and industry security standards. It complies with PCI DSS, ISO 9001, 27001, 27017, and 27018), and SOC 1, 2, and 3, in addition to being HIPAA eligible. That makes it easier for you to verify our security and meet your own obligations. For more information and resources, visit our compliance pages. You can also go to the Services in Scope by Compliance Program page to see a full list of services and certifications.

Pricing and billing

Q: How will I be charged and billed for my use of Amazon FSx for Lustre?

A: You pay only for the resources you use. See the Amazon FSx for Lustre pricing page for details.

Q: Do your prices include taxes?

A: Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Amazon FSx for Windows File Server FAQs

General

Q: What is Amazon FSx for Windows File Server?

A: Amazon FSx for Windows File Server provides a fully managed native Microsoft Windows file system so you can easily move your Windows-based applications that require file storage to AWS.

Built on Windows Server, Amazon FSx provides shared file storage with the compatibility and features that your Windows-based applications rely on, including full support for the SMB protocol, Windows NTFS, and Active Directory (AD) integration. Amazon FSx uses SSD storage to provide the fast performance your Windows applications and users expect, with high levels of throughput and IOPS, and consistent sub-millisecond latencies.

With Amazon FSx, you can launch highly durable and available Windows file systems that can be accessed from up to thousands of compute instances using the industry-standard SMB protocol. Amazon FSx eliminates the typical administrative overhead of managing Windows file servers.

Q: What is an Amazon FSx for Windows File Server file system, and what is a file share?

A: A file system is the primary resource in Amazon FSx. It's where you store and access your files and folders. It is associated with a storage amount and a throughput capacity, as well as a DNS name for accessing it.

A file share is a specific folder (and its subfolders) within your file system that you make accessible to your compute instances – every file system comes with a default Windows file share, named “share” and you can create and manage as many other Windows file shares as you’d like.

Q: How do I get started with FSx for Windows File Server?

A: To use Amazon FSx, you must have an AWS account. If you do not already have an AWS account, you can sign up for an AWS account.

Once you have created an AWS account, you can create a file system via the AWS Management Console, the AWS Command Line Interface (AWS CLI), and Amazon FSx API (and various language-specific SDKs).

Q: What instance types and OS versions can I access my file system from?

A: By supporting the SMB protocol, Amazon FSx can connect your file system to Amazon EC2, VMware Cloud on AWS, Amazon WorkSpaces, and Amazon AppStream 2.0 instances. To ensure compatibility with your applications, Amazon FSx supports all Windows versions starting from Windows Server 2008 and Windows 7, and current versions of Linux (using the cifs-utils tool).

Q: How do I access data on my Amazon FSx file system?

A: From within Windows, use the “Map Network Drive” feature to map a drive letter (e.g., Z:) to a file share on your Amazon FSx file system. You can also access your file system from Linux using the cifs-utils tool to mount your file share. Once you've done this, you can work with the files and folders in your Amazon FSx file system just like you would with a local file system.

Q: How do I manage a file system?

A: Amazon FSx is a fully-managed service, so all of the file storage infrastructure is managed for you. When you use Amazon FSx, you avoid the complexity of deploying and maintaining complex file system infrastructure.

To create, view, tag, and delete file systems and backups, you can use the AWS Management Console, the AWS command-line interface (CLI), or the Amazon FSx API (and various language-specific SDKs). To administer file systems, including managing file shares, active user sessions and open files, shadow copies, user quotas, and Data Deduplication, you can use the Amazon FSx remote management CLI via PowerShell.

Q: How do I migrate my existing file data into an Amazon FSx file system?

A: To migrate your existing file data into Amazon FSx, use Windows's Robust File Copy (RoboCopy) to copy your files (both the data and the full set of metadata like ownership and Access Control Lists) directly to Amazon FSx. The RoboCopy tool performs the copy in entirety (including all metadata like ACLs, ownership, and time stamps), efficiently (with parallel copy) and reliably (by recovering from interruptions like network outages). Once you have moved your file and folder data, Amazon FSx offers programmatic share management support to help you easily migrate your file share configuration. Learn more in the documentation guide.

Q: How do I monitor my file system's activity?

A: You can monitor storage capacity and file system activity using Amazon CloudWatch, and monitor all Amazon FSx API calls using AWS CloudTrail.

Q: What workloads is Amazon FSx for Windows File Server designed for?

A: Amazon FSx was designed for a broad set of use cases that require Windows shared file storage, like CRM, ERP, custom or .NET applications, home directories, data analytics, media and entertainment workflows, web serving and content management, software build environments, and Microsoft SQL Server.

Q: How do I use Amazon FSx with Microsoft SQL Server?

High availability (HA) Microsoft SQL Server is typically deployed across multiple database nodes in a Windows Server Failover Cluster (WSFC), with each node having access to shared file storage. Amazon FSx for Windows File Server can be used as shared storage for High Availability (HA) Microsoft SQL Server deployments in two ways: as storage for active data files and as an SMB file share witness. For more information, see Using Amazon FSx with Microsoft SQL Server.

Q: When should I use Amazon FSx Windows File Servers vs. Amazon EFS vs. Amazon FSx for Lustre?

A: For Windows-based applications, Amazon FSx provides fully managed Windows file servers with features and performance optimized for "lift-and-shift" business-critical application workloads including home directories (user shares), media workflows, and ERP applications. It is accessible from Windows and Linux instances via the SMB protocol. If you have Linux-based applications, Amazon EFS is a cloud-native fully managed file system that provides simple, scalable, elastic file storage accessible from Linux instances via the NFS protocol.

For compute-intensive and fast processing workloads, like high performance computing (HPC), machine learning, EDA, and media processing, Amazon FSx for Lustre, provides a file system that’s optimized for performance, with input and output stored on Amazon S3.

Q. Does Amazon FSx support access from my on-premises environment?

A: Yes, you can access Amazon FSx file systems from your on-premises environment using an AWS Direct Connect or AWS VPN connection between your on-premises datacenter and your Amazon VPC. With support for AWS Direct Connect, Amazon FSx allows you to access your file system over a dedicated network connection from your on-premises environment. With support for AWS VPN, Amazon FSx allows you to access your file system from your on-premises devices over a secure and private tunnel.

With on-premises access, you can use Amazon FSx for hosting user shares accessible by on-premises end-users, for cloud bursting workloads to offload your application processing to the cloud, and for backup and disaster recovery solutions. For more information, see Accessing Amazon FSx from on-premises.

Q. Does Amazon FSx support access from multiple VPCs, accounts, and regions?

A: Yes, you can access your Amazon FSx file systems from multiple Amazon VPCs, AWS accounts, and AWS Regions using VPC Peering connections or AWS Transit Gateway. A VPC Peering connection is a networking connection between two VPCs that enables you to route traffic between them. A transit gateway is a network transit hub that you can use to interconnect your VPCs. With VPC Peering and AWS Transit Gateway, you can even interconnect VPCs across AWS accounts and AWS Regions.

With access to your file systems via VPC Peering and Transit Gateway, you can share your file data sets across users and applications in multiple VPCs, AWS accounts, and/or AWS Regions. For more information, see Accessing Amazon FSx from multiple VPCs, accounts or regions.

Q: Does Amazon FSx support shared VPCs?

A: Yes, with Amazon FSx, you can create and use file systems in shared Amazon Virtual Private Clouds (VPCs) from both owner accounts and participant accounts with which the VPC has been shared. VPC sharing enables you to reduce the number of VPCs that you need to create and manage, while you still benefit from using separate accounts for billing and access control.

Q: What regions is Amazon FSx for Windows File Server available in?

A: Please refer to Regional Products and Services for details of Amazon FSx for Windows File Server service availability by region.

Q: Does Amazon FSx offer a Service Level Agreement (SLA)?

A: Yes. The Amazon FSx SLA provides for a service credit if a customer's monthly uptime percentage is below our service commitment in any billing cycle.

Scale and performance

Q: What performance does FSx for Windows File Server provide?

A: FSx for Windows File Server uses high-performance SSD storage to provide consistent sub-millisecond latencies for file operations.

Every Amazon FSx file system has a throughput capacity that you configure when the file system is created. This throughput capacity determines the baseline and burst speeds at which the Windows file server hosting your file system can serve file data.

When creating a file system in the Amazon FSx console, it automatically recommends a throughput capacity for your file system.

If you want to select a specific throughput capacity, you can pick from the following.

Throughput capacity specified (MBps)	8	16	32	64	128	256	512	1,024	2,048
Baseline throughput (MBps)	8	16	32	64	128	256	512	1,024	2,048
Burst throughput (MBps)	192	192	192	256	438	438	-	-	-

Depending on your workload's access patterns, you may observe even higher levels of throughput (between 600 MBps and 3 GBps) by benefiting from in-memory caching on the Windows file server.

Q: How much data can I store on Amazon FSx for Windows File Server?

A: You can run up to thousands of Amazon FSx for Windows File Server file systems in your account, with each file system having up to 64 TB of data. To unify your data from multiple file systems into one common folder structure, Amazon FSx supports the use of Microsoft's Distributed File System (DFS) to organize shares into a single folder structure up to hundreds of PB in size.

Q: How do I scale out performance across multiple file systems?

A: Amazon FSx supports the use of Microsoft's Distributed File System (DFS) Namespaces to scale out performance across multiple file systems in the same namespace up to tens of GBps and millions of IOPs.

Security and compliance

Q: How does Amazon FSx integrate with Microsoft Active Directory (AD)?

A: Amazon FSx works with Microsoft Active Directory (AD) to integrate with your existing Windows environments. When creating a file system with Amazon FSx, you join it to your Microsoft AD -- either an AWS Managed Microsoft AD or your self-managed Microsoft AD. Your users can then use their existing AD-based user identities to authenticate themselves and access the Amazon FSx file system, and to control access to individual files and folders.

Q: What access control capabilities does Amazon FSx provide?

A: Amazon FSx provides standard Windows permissions (full support for Windows Access Controls ACLS) for files and folders.

You specify the Amazon Virtual Private Cloud (VPC) in which your file system is made accessible, and you control which resources within the VPC have access to your file system using VPC Security Groups.

You control who can administer your file system and backup resources (create, delete, etc.) using AWS IAM.

Q: Does Amazon FSx for Windows File Server support data encryption?

A: Yes. Amazon FSx for Windows File Server always encrypts your file system data and your backups at-rest using keys you manage through AWS Key Management Service (KMS). Amazon FSx encrypts data-in-transit using SMB Kerberos session keys, when you access your file system from clients that support SMB 3.0 (and higher). You can also choose to enforce in-transit encryption on all connections to your file system by limiting access to only those clients that support SMB 3.0 and higher to help meet compliance needs.

Q: What compliance programs does Amazon FSx support?

A: AWS has the longest-running compliance program in the cloud and is committed to helping customers navigate their requirements. Amazon FSx has been assessed to meet global and industry security standards. It complies with PCI DSS, ISO 9001, 27001, 27017, and 27018), and SOC 1, 2, and 3, in addition to being HIPAA eligible. That makes it easier for you to verify our security and meet your own obligations. For more information and resources, visit our compliance pages. You can also go to the Services in Scope by Compliance Program page to see a full list of services and certifications.

Availability and durability

Q: What does Amazon FSx for Windows File Server do to ensure high availability and durability?

A: To ensure high availability and durability, Amazon FSx automatically replicates your data within an Availability Zone (AZ) to protect it from component failure, continuously monitors for hardware failures, and automatically replaces infrastructure components in the event of a failure. You can also create a Multi-AZ file system, which provides redundancy across multiple AZs. Amazon FSx also takes highly durable backups (stored in S3) of your file system daily using Windows's Volume Shadow Copy Service, and allows you to take additional backups at any point.

Q: What is the difference between Single-AZ and Multi-AZ file systems?

Single-AZ file systems are designed to be highly available and durable within an AZ by replicating data within an AZ and automatically replacing infrastructure components in the event of a failure.

Multi-AZ file systems support all the availability and durability features of Single-AZ file systems, and in addition, are designed to provide continuous availability to data even in the event that an AZ is unavailable. In a Multi-AZ deployment, Amazon FSx automatically provisions and maintains a standby file server in a different Availability Zone. Any changes written to disk in your file system are synchronously replicated across AZs to the standby. Using Amazon FSx Multi-AZ deployments can enhance availability during planned system maintenance, and help protect your data against instance failure and Availability Zone disruption. In the event of planned file system maintenance or unplanned service disruption, Amazon FSx automatically fails over to the secondary file server, allowing you to continue accessing your data without manual intervention.

Q: What events would cause a Multi-AZ Amazon FSx file system to initiate a failover to the standby file server?

Amazon FSx automatically performs a failover in the event of a loss of availability to the active file server. This can be caused by a failure in the active Availability Zone, or by a failure of the active file server itself. Amazon FSx will also temporarily fail over to the standby file server during planned maintenance.

Q: What happens during a failover in a Multi-AZ file system and how long does a failover take?

A: Amazon FSx detects and automatically recovers from failures so that you can resume file system operations as quickly as possible without administrative intervention. When failing over from one file server to another, the new active file server will automatically begin serving all file system reads and write requests. Failovers, as defined by the interval between the detection of the failure on the active and promotion of the other file server to active, typically complete within 30 seconds. Failback will occur once the file server in the preferred subnet is fully recovered (typically under 20 minutes), and also completes within 30 seconds.

Windows clients accessing Multi-AZ file systems automatically fail over and fail back with the file system, meaning that users or applications running on Windows clients automatically benefit from the enhanced availability of Multi-AZ file systems. Linux clients don’t automatically connect to the standby file server upon a failover, and will automatically resume file system operations once failback to the preferred file server is complete.

Q: How does Amazon FSx keep Windows Server software up to date?

A: Amazon FSx performs routine software updates for the Windows Server software it manages. The maintenance window is your opportunity to control what day and time of the week this software patching occurs. Patching occurs infrequently, typically once every several weeks, and should require only a fraction of your 30-minute maintenance window.

Data protection

Q: How does Amazon FSx enable me to protect my data?

A: Beyond automatically replicating your file system's data to ensure high durability, Amazon FSx provides you with two options to further protect the data stored on your file systems: Windows shadow copies to enable your end-users to easily undo file changes and compare file versions by restoring files to previous versions, and backups to support your backup retention and compliance needs.

Q: How do my end-users restore files to previous versions?

A: Amazon FSx supports file- or folder-level restores to previous versions by supporting Windows shadow copies, which are snapshots of your file system at a point in time. With shadow copies enabled, your end-users can view and restore individual files or folders from a prior snapshot with the click of a button in Windows File Explorer. Storage administrators using Amazon FSx can easily schedule shadow copies to be taken periodically using Windows PowerShell commands.

Q: How do I take backups on Amazon FSx for Windows File Server?

A: Creating regular backups for your file system is a best practice that complements the replication that Amazon FSx performs for your file system. Working with Amazon FSx backups is easy, whether it's creating backups, restoring a file system from a backup, or deleting a backup.

Amazon FSx takes daily automatic backups of your file systems, and allows you to take additional backups at any point. Amazon FSx backups are incremental, which means that only the changes after your most recent backup are saved, thus saving on backup storage costs by not duplicating data.

Q: What durability and consistency does Amazon FSx provide for backups?

A: With Amazon FSx, backups are file-system-consistent and highly durable. To ensure file-system-consistency, Amazon FSx uses Windows's Volume Shadow Copy Service, allowing you to restore to a point in time snapshot of your file system. To ensure high durability, Amazon FSx stores backups in Amazon S3.

Q: What is the daily backup window?

A: The daily backup window is a 30-minute window that you specify when creating a file system. Amazon FSx takes the daily automatic backup of your file system during this window. At some point during the daily backup window, storage I/O might be suspended briefly while the backup process initializes (typically under a few seconds).

Q: What is the daily backup retention period?

A: The daily backup retention period specified for your file system (7 days by default) determines the number of days your daily automatic backups are kept.

Q: What happens to my backups if I delete my file system?

A: When you delete your file system, all automatic daily backups associated with the file system are deleted. Any user-initiated backups you created will remain.

Optimizing Total Cost of Ownership (TCO)

Q: What is Data Deduplication?

A: Large datasets often have redundant data. For example, with user file shares, multiple users tend to have files that are similar or identical. As another example, with software development shares, most binaries remain largely unchanged from build to build. Data Deduplication is a feature in Windows Server that reduces costs that are associated with redundant data by storing duplicated portions of files only once. It optimizes files transparently such that users and applications accessing the data are unaware of deduplication.

Q: How much storage savings can I expect with Data Deduplication?

A: The storage savings you can achieve with Data Deduplication depends on the nature of your data set, including how much duplication exists across files. Typical savings average 50-60% for general-purpose file shares, with savings ranging 30-50% for user documents, and 70-80% for software development data sets.

Q: How do I enable Data Deduplication on my file system?

A: You can enable Data Deduplication on your file system by running a single command (Enable-FSxDedup) on the Amazon FSx remote management CLI via PowerShell. Once enabled, Data Deduplication continually and automatically scans and optimizes your files in the background.

Q: How do I monitor and control individual users' storage consumption on my file system?

You can enable and configure user storage quotas on your file system to monitor usage and allocate storage costs to individual teams, and to impose restrictions at a user-level in order to prevent any one user from storing a lot of data.

Pricing and billing

Q: How will I be charged and billed for my use of Amazon FSx for Windows File Server?

A: You pay only for the resources you use. You are billed hourly for your file systems, based on their deployment type (Single-AZ or Multi-AZ), configured storage capacity (priced per GB-month) and throughput capacity (priced per MBps-month). You are also billed hourly for your backup storage (priced per GB-month). For pricing information, please visit the Amazon FSx pricing page.

Q: Do your prices include taxes?

A: Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese

billing address, use of AWS services is subject to Japanese Consumption Tax.
Learn more.

Amazon S3 Glacier FAQs

General

Q: Why is Amazon Glacier now called Amazon S3 Glacier?

Customers have long thought of Amazon Glacier, our backup and archival storage service, as a storage class of Amazon S3. In fact, a very high percentage of the data stored in Amazon Glacier today comes directly from customers using S3 Lifecycle policies to move cooler data into Amazon Glacier. Now, Amazon Glacier is officially part of S3 and will be known as Amazon S3 Glacier (S3 Glacier). All of the existing Glacier direct APIs continue to work just as they have, but we've now made it even easier to use the S3 APIs to store data in the S3 Glacier storage class.

Q: What is Amazon S3 Glacier?

Amazon S3 Glacier is an extremely low-cost storage service that provides secure, durable, and flexible storage for data backup and archival. With Amazon S3 Glacier, customers can reliably store their data for as little as $0.004 per gigabyte per month. Amazon S3 Glacier enables customers to offload the administrative burdens of operating and scaling storage to AWS, so that they don't have to worry about capacity planning, hardware provisioning, data replication, hardware failure detection and repair, or time-consuming hardware migrations.

Q: How can businesses, government and other organizations benefit from Amazon S3 Glacier?

Amazon S3 Glacier enables any business or organization to easily and cost effectively retain data for months, years, or decades. With Amazon S3 Glacier, customers can now cost effectively retain more of their data for future analysis or reference, and they can focus on their business rather than operating and maintaining their storage infrastructure. Customers seeking compliance storage can deploy compliance controls using Vault Lock to meet regulatory and compliance archiving requirements.

Q: How should I choose between Amazon S3 Glacier and Amazon Simple Storage Service (Amazon S3)?

Amazon S3 is a durable, secure, simple, and fast storage service designed to make web-scale computing easier for developers. Use Amazon S3 if you need low latency or frequent access to your data. Use Amazon S3 Glacier if low storage cost is paramount, and you do not require millisecond access to your data.

Q: What kind of data can I store?

You can store virtually any kind of data in any format. You can also deploy compliance storage controls with Vault Lock to store regulatory and compliance archives in an immutable, Write Once Read Many (WORM) format. Please refer to the Amazon Web Services Licensing Agreement for details.

Q: What does Amazon do with my data in Amazon S3 Glacier?

Amazon will store your data and track its associated usage for billing purposes. Amazon will not otherwise access your data for any purpose outside of the Amazon S3 Glacier offering, except if required to do so by law. Please refer to the Amazon Web Services Licensing Agreement for details.

Q: How do I use Amazon S3 Glacier?

Amazon S3 now supports four new features to reduce your storage costs by making it even easier to build archival applications using the Amazon S3 Glacier storage class and by enabling one-click data replication to S3 Glacier in another AWS Region. S3 PUT to Glacier, S3 Cross-Region Replication to Glacier, S3 Restore Notifications, and S3 Restore Speed Upgrade are available using the S3 APIs, AWS Software Development Kits (SDKs), and AWS Management Console for simpler integration with your archival workloads and applications. To learn more, visit our Amazon S3 Developer Guide.

Amazon S3 Glacier provides a simple, standards-based REST web services interface as well as Java and .NET SDKs. The AWS Management console can be used to quickly set up Amazon S3 Glacier. Data can then be uploaded and retrieved programmatically. View our documentation for more information on the Glacier direct APIs and SDKs.

Q: How durable is Amazon S3 Glacier?

Amazon S3 Glacier is designed to provide average annual durability of 99.999999999% for an archive. The service redundantly stores data in multiple facilities and on multiple devices within each facility. To increase durability, Amazon S3 Glacier synchronously stores your data across multiple facilities before returning SUCCESS on uploading archives. S3 Glacier performs regular, systematic data integrity checks and is built to be automatically self-healing.

Q: How reliable is Amazon S3 Glacier?

Amazon S3 Glacier gives any developer access to the same highly scalable, highly available, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites. Amazon S3 Glacier is designed for 99.99% availability and is backed by the Amazon S3 Service Level Agreement.

Q: Does Amazon S3 Glacier offer a Service Level Agreement (SLA)?

Yes. The Amazon S3 SLA provides for a service credit if a customer's monthly uptime percentage is below our service commitment in any billing cycle.

Q: What is the backend infrastructure supporting the S3 Glacier storage class?

In general, AWS does not disclose the backend infrastructure and architecture for our compute, networking, and storage services, as we are more focused on the customer outcomes of performance, durability, availability, and security. However, this question is often asked by our customers. We use a number of different technologies which allow us to offer the prices we do to our customers. Our services are built using common data storage technologies specifically assembled into purpose-built, cost-optimized systems using AWS-developed software. S3 Glacier benefits from our ability to optimize the sequence of inputs and outputs to maximize efficiency accessing the underlying storage.

Q: What is S3 Glacier Deep Archive?

S3 Glacier Deep Archive is a new Amazon S3 storage class that provides secure and durable object storage for long-term retention of data that is accessed once or twice in a year. From just $0.00099 per GB-month (less than one-tenth of one cent, or about $1 per TB-month), S3 Glacier Deep Archive offers the lowest cost storage in the cloud, at prices significantly lower than storing and maintaining data in on-premises magnetic tape libraries or archiving data off-site. To learn more about S3 Glacier Deep Archive, visit Amazon S3 FAQs.

Getting started

Q: How is data within Amazon S3 Glacier organized?

You store data in Amazon S3 Glacier as an archive. Each archive is assigned a unique archive ID that can later be used to retrieve the data. An archive can represent a single file or you may choose to combine several files to be uploaded as a single archive. You upload archives into vaults. Vaults are collections of archives that you use to organize your data.

Q: How much data can I store?

There is no maximum limit to the total amount of data that can be stored in Amazon S3 Glacier. Individual archives are limited to a maximum size of 40 terabytes.

Q: What is the minimum amount of data that I can store using Amazon S3 Glacier?

There is no minimum limit to the amount of data that can be stored in Amazon S3 Glacier and individual archives can be from 1 byte to 40 terabytes.

Q: Does the AWS Management Console support Amazon S3 Glacier?

Yes. The AWS Management Console allows you to create and configure vaults, allowing you to easily and quickly setup Glacier. Click here to go the AWS Management Console.

Billing

Q: How much does Amazon S3 Glacier cost?

With Amazon S3 Glacier, storage is priced from $0.004 per gigabyte per month, and you pay for what you use. There are no setup fees, and for most archive use cases your total costs will primarily be made up of your storage cost.

Upload requests are priced from $0.05 per 1,000 requests. In addition, archives stored in S3 Glacier have a minimum 90 days of storage, and archives deleted before 90 days incur a pro-rated charge equal to the storage charge for the remaining days. As Amazon S3 Glacier is designed to store data that is infrequently accessed and long lived, these charges will likely not apply to most of you.

We charge less where our costs are less. Some prices vary across Amazon S3 Glacier Regions and are based on the location of your vault. There is no Data Transfer charge for data transferred between Amazon EC2 and Amazon S3 Glacier within the same Region. Data transferred between Amazon EC2 and Amazon S3 Glacier across all other Regions (e.g. between the Amazon EC2 Northern California and Amazon S3 Glacier US East North Virginia Regions) will be charged at Internet Data Transfer rates on both sides of the transfer.

To learn more about AmazonS3 Glacier pricing, please visit the Amazon S3 Glacier pricing page.

Q: How is my storage charge calculated?

The volume of storage billed in a month is based on the average storage used throughout the month, measured in gigabyte-months (GB-Months). The size of each of your archives is calculated as the amount of data you upload plus an additional 32 kilobytes of data for indexing and metadata (e.g. your archive description). This extra data is necessary to identify and retrieve your archive. Here is an example of how to calculate your storage costs using US East (Northern Virginia) Region pricing:

If you upload 100,000 archives that are 1 gigabyte each, your total storage would be:

1.000032 gigabytes for each archive x 100,000 archives = 100,003.20 gigabytes

If you stored the archives for 1 month, you would be charged:

100,003.20 GB-Months x $0.004 = $400.01

If you upload 200,000 archives that are 0.5 gigabytes each, your total storage would be:

0.500032 gigabytes for each archive x 200,000 archives = 100,006.40 gigabytes

If you stored the archives for 1 month, you would be charged:

100,006.40 GB-Months x $0.004 = $400.03

Your storage is measured in “TimedStorage-ByteHrs,” which are added up at the end of the month to generate your monthly charges. For example, if you store an archive that is 1 gigabyte (inclusive of the 32 kilobyte overhead) for one day in the US East (Northern Virginia) Region, your storage usage would be:

1,073,741,824 bytes x 1 day x 24 hours = 25,769,803,776 Byte-Hours

Converting this to GB-Months (assuming a 30 day month) gives:

25,769,803,776 Byte-Hours x (1 GB / 1,073,741,824 bytes) x (1 month / 720 hours) = 0.03 GB-Months

So your storage charge for that day would be:

0.03 GB-Months x $0.004 = $0.00012

To learn more about Amazon S3 Glacier pricing and view prices for other regions, please visit the Amazon S3 Glacier pricing page.

Q: Why do prices vary depending on which Amazon S3 Glacier Region I choose?

We charge less where our costs are less. For example, our costs are lower in the US East (North Virginia) Region than in the US West (Northern California) Region.

Q: How will I be charged and billed for my use of Amazon S3 Glacier?

There are no setup fees to begin using the service. At the end of the month, your credit card will automatically be charged for that month's usage. You can view your charges for the current billing period at any time on the Amazon Web Services web site, by logging into your Amazon Web Services account, and clicking "Account Activity" under "Your Web Services Account".

Q: How much data can I retrieve for free?

Amazon S3 Glacier offers a 10 GB retrieval free tier. You can retrieve 10 GB of your Amazon S3 Glacier data per month for free. The free tier allowance can be used at any time during the month and applies to Standard retrievals.

Q: How much does Amazon S3 Glacier cost?

There are three ways to retrieve data from Amazon S3 Glacier and each has a different per-GB retrieval fee and per-archive request fee (i.e. requesting one archive counts as one request). Expedited retrievals cost $0.03 per GB and $0.01 per request. Standard retrievals cost $0.01 per GB and $0.05 per 1,000 requests. Bulk retrievals cost $0.0025 per GB and $0.025 per 1,000 requests.

For example, using Expedited retrievals, if you requested 10 archives with a size of 1 GB each, the cost would be 10 x $0.03 +10 x $0.01 = $0.40.

If you were using Standard retrievals to retrieve 500 archives that were 1 GB each, the cost would be 500GB x $0.01 + 500 x $0.05/1,000 = $5.025

Lastly, using Bulk retrievals, if you were to retrieve 500 archives that are 1 GB each, the cost would be 500GB x $0.0025 + 500 x $0.025/1,000 = $1.2625.

To learn more about Amazon S3 Glacier pricing, please visit the Amazon S3 Glacier pricing page.

Q: How will I be charged when retrieving only a range of an archive?

Range retrievals are priced in precisely the same way as regular retrievals from Amazon S3 Glacier. You are charged a per-GB fee for only the amount of data retrieved in the range you specify.

Q: How will I be charged for deleting data that is less than 3 months old?

Amazon S3 Glacier is designed for use cases where data is retained for months, years, or decades. Deleting data from Amazon S3 Glacier is free if the archive being deleted has been stored for three months or longer. If an archive is deleted within three months of being uploaded, you will be charged an early deletion fee. In the US East (Northern Virginia) Region, you would be charged a prorated early deletion fee of $0.012 per gigabyte deleted within three months. So if you deleted 1 gigabyte of data 1 month after uploading it, you would be charged a $0.008 early deletion fee. If, instead you deleted 1 gigabyte after 2 months, you would be charged a $0.004 early deletion fee.

To view prices for other regions, please visit the Amazon S3 Glacier pricing page.

Q: What can I expect the total cost of ownership (TCO) to be?

Amazon S3 Glacier is a secure, durable, and extremely low-cost cloud storage service for data archiving and long-term backup. Customers can reliably store large or small amounts of data for as little as $0.004 per gigabyte per month, a significant savings compared to on-premises solutions. To keep costs low yet suitable for varying retrieval needs, Amazon S3 Glacier provides three options for access to archives, from a few minutes to several hours. Your total cost of ownership (TCO) for your Amazon S3 Glacier storage will depend on your data access patterns. Below are several examples illustrating different use cases ranging from deep archives that are never retrieved to active workloads where large portions of data are accessed.

TCO example 1: Let’s assume that you upload 1 PB of data into Amazon S3 Glacier, that the average archive size is 1 GB and that you never retrieve any data. When you first upload the 1 PB, there are upload request fees of 1,048,576 GB x $0.05 / 1,000 = $52.43. Then the ongoing storage costs are 1,048,576 GB x $0.004 = $4,194.30 per month, or $50,331.65 per year.

TCO example 2: Now let’s assume the same storage as example 1 and also assume that you retrieve 3 TB (3,072 GB) a day on average using Bulk retrievals and that the average archive size was 1 GB for a total of 3,072 archives. That’s 90 TB retrieved per month or 8.8% of your data per month. The total retrieval fees per day would be 3,072 x $0.0025 + 3,072 * $0.025 / 1,000 = $7.76, which equates to $232.70 per month and $2,792.45 per year. Adding storage costs, your annual TCO is $50,331.65 + $2,792.45 = $53,124.10. In this example, retrieval fees make up just 5.3% of your total Glacier fees. Your total monthly cost per GB stored including retrieval fees is $0.004222/GB.

TCO example 2: Now let’s assume the same storage as example 1 and also assume that you retrieve 1 TB (1,024 GB) a day on average using Standard retrievals and that occasionally you use Expedited retrievals for urgent requests, averaging 10 GB per day. Here, we assume the average archive size is 1 GB. That’s 30.3 TB per month or 3% of your data per month. The total retrieval fees per day would be (1,024 x $0.01 + 1,024 x $0.05 / 1000) + (10 x $0.03 + 10 x $0.01) = $10.69, which equates to $320.74 per month and $3,848.83 per year. Adding storage costs, your annual TCO is $50,331.65 + $3,848.83 = $54,180.48. In this example, retrieval fees make up just 7.1% of your total Glacier fees. Your total monthly cost per GB stored including retrieval fees is $0.0043/GB.

To learn more about Amazon S3 Glacier pricing, please visit the Amazon S3 Glacier pricing page.

Q: Do your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Q: How will multipart upload requests to the S3 Glacier storage class appear on my bill?

For Initiate multipart and Upload part, you will be charged at S3 Standard PUT and POST request rates. For Complete multipart, you will be charged the S3 Glacier PUT and POST request rate.

Q: How will in-progress multipart uploads to the S3 Glacier storage class appear on my bill?

In-progress multipart parts for a PUT to the S3 Glacier storage class are billed as S3 Glacier Staging Storage at S3 Standard storage rates until the upload completes. Deleted in-progress multipart parts will not be subject to an S3 Glacier early delete fee. The 90 day early-delete window starts from the time the multipart upload is completed.

Security

Q: How do I control access to my data?

By default, only you can access your data. In addition, you can control access to your data in Amazon S3 Glacier by using the AWS Identity and Access Management (AWS IAM) service. You simply set up an AWS IAM policy that specifies which users within an account have rights to operations on a given vault.

Q: Is my data encrypted?

Yes, all data in the service will be encrypted on the server side. Amazon S3 Glacier handles key management and key protection for you. Amazon S3 Glacier uses one of the strongest block ciphers available, 256-bit Advanced Encryption Standard (AES-256). 256-bit is the largest key size defined for AES. Customers wishing to manage their own keys can encrypt data prior to uploading it.

Q: Does Amazon S3 Glacier support IAM permissions?

Yes, S3 Glacier will support API-level permissions through AWS Identity and Access Management (IAM) service integration

For more information about IAM, go to:

- AWS Identity and Access Management
- AWS Identity and Access Management Getting Started Guide

- Using AWS Identity and Access Management

Archives and vaults

Q: What is an archive?

An archive is a durably stored block of information. You store your data in Amazon S3 Glacier as archives. You may upload a single file as an archive, but your costs will be lower if you aggregate your data. TAR and ZIP are common formats that customers use to aggregate multiple files into a single file before uploading to Amazon S3 Glacier. The total volume of data and number of archives you can store are unlimited. Individual Amazon S3 Glacier archives can range in size from 1 byte to 40 terabytes. The largest archive that can be uploaded in a single Upload request is 4 gigabytes. For items larger than 100 megabytes, customers should consider using the Multipart upload capability. Archives stored in Amazon S3 Glacier are immutable, i.e. archives can be uploaded and deleted but cannot be edited or overwritten.

Q: How do I delete archives?

You can delete an archive at any time. You will stop being billed for your archive when your delete request succeeds at which point the archive itself will be inaccessible. Archives that are deleted within 3 months of being uploaded will be charged a deletion fee (see billing section for more details).

Q: How do I upload large archives?

When uploading large archives (100MB or larger), you can use multi-part upload to achieve higher throughput and reliability. Multi-part uploads allow you to break your large archive into smaller chunks that are uploaded individually. Once all the constituent parts are successfully uploaded, they are combined into a single archive.

Q: What is a vault?

A vault is a way to group archives together in Amazon S3 Glacier. You organize your data in Amazon S3 Glacier using vaults. Each archive is stored in a vault of your choice. You may control access to your data by setting vault-level access policies using the AWS Identity and Access Management (IAM) service. You can also attach notification policies to your vaults. These enable you or your application to be notified when data that you have requested for retrieval is ready for download. Click here to learn more about setting up notifications using the Amazon Simple Notification Service (Amazon SNS).

Q: How many vaults can I create?

You can create up to 1,000 vaults per account per region.

Q: How do I effectively manage my Amazon S3 Glacier vaults?

Amazon S3 Glacier allows you to tag your Glacier vaults for easier resource and cost management. Tags are labels that you can define and associate with your vaults, and using tags adds filtering capabilities to operations such as AWS cost reports. For example, you can use tags to allocate S3 Glacier costs and usage across multiple departments in your organization or by any other categorization. You can tag your vaults by using the S3 Glacier Console or the Amazon Glacier direct APIs. For more information see Tagging Your Amazon S3 Glacier Vaults.

Q: How do I delete a vault?

You may delete any S3 Glacier vault that does not contain any archives using the AWS Management Console, the Amazon Glacier direct APIs or the SDKs. Once a vault has been deleted, you can then re-create a vault with the same name. If your vault contains archives, you must delete all the archives before deleting the vault.

Vault access policies

Q: What is a vault access policy?

A vault access policy is a resource-based policy that you can attach directly to your S3 Glacier vault (the resource) to specify who has access to the vault and what actions they can perform on it. To learn more please read Managing Vault Access Policies in the Amazon S3 Glacier developer's guide.

Q: How are vault access policies different from access control based on AWS Identity and Access Management (IAM) policies?

Access permissions can be assigned in two ways: as user-based permissions or as resource-based permissions. Access control based on IAM policies is user-based where you would assign IAM policies to IAM users or groups to control the read, write, and delete permissions on your S3 Glacier vaults. Access control with vault access policies is resource-based where you would attach an access policy directly on a vault to govern access to all users. Vault access policies can make certain use cases simpler. For example, to protect information in a business-critical vault from unintended deletion, you can create a vault access policy that denies delete attempts from all users. This data protection procedure can be accomplished in a matter of minutes in the AWS Management Console without having to audit and revoke delete permissions assigned to users through IAM policies.

Q: Can I use vault access policies to manage cross-account access?

Yes you can. For example, you can grant read-only access on your vault to a business partner in a different AWS account by simply adding that account to the vault's access policy and specifying that only read activities are allowed.

Q: How does billing work in a cross-account access scenario?

The vault owner's account will be billed for the charges incurred during cross-account access. For example, Alice (account A) grants Bob (account B) access to Alice's "movies" vault and allows Bob to upload data. After Bob makes 1000 requests to upload 1GB of data, Alice's account (account A) will be billed for the 1000 requests as well as the 1GB of data until the data is deleted. Bob's account (account B) will not incur these charges.

Q: How do I create and manage vault access policies?

You can create and manage vault access policies in the AWS Glacier console or use the vault access APIs in the AWS SDK. To learn more please read Managing

Vault Access Policies in the Amazon S3 Glacier developer's guide.

Q: How many vault access policies can I have?

You can set one vault access policy for each vault. The vault access policy can be used as a single location to view the list of users with vault access and the allowed actions for each user.

Vault Lock

Q: What is Vault Lock?

Vault Lock allows you to easily deploy and enforce compliance controls on individual S3 Glacier vaults via a lockable policy (Vault Lock policy). Once locked, the Vault Lock policy becomes immutable and S3 Glacier will enforce the prescribed controls to help achieve your compliance objectives. To learn more, please read Amazon S3 Glacier Vault Lock in the Amazon S3 Glacier developer's guide.

Q: What type of compliance controls can I deploy with Vault Lock?

You can deploy a variety of compliance controls in a Vault Lock policy using the AWS Identity and Access Management (IAM) policy language. For example, you can easily set up "Write Once Read Many" (WORM) or time-based records retention for regulatory archives. To learn more, please read Amazon S3 Glacier Vault Lock in the Amazon S3 Glacier developer's guide.

Q: How does Vault Lock enforce my compliance controls?

Vault Lock enforces your compliance controls via a lockable policy (Vault Lock policy). Once locked, the Vault Lock policy becomes immutable and S3 Glacier will only allow operations on your data that are explicitly permitted by the compliance controls you specified. Vault Lock also ensures that a locked policy cannot be deleted or altered until there are no more archives to protect in the vault. Learn more about Locking a Vault for compliance in the Amazon S3 Glacier developer's guide.

Q: How is a Vault Lock policy different than a vault access policy?

Both policies govern access controls to your vault, however, a Vault Lock policy can be made immutable and provides strong enforcement for your compliance controls. You can use the Vault Lock policy to deploy regulatory and compliance controls that are typically restrictive and are “set and forget” in nature. In conjunction, you can use the vault access policy to implement access controls that are not compliance related, temporary, and subject to frequent modification. The two policies can be used in tandem to achieve governance and flexibility.

Q: What AWS electronic storage services have been assessed based on financial services regulations?

For customers in the financial services industry, Vault Lock provides added support for broker-dealers who must retain records in a non-erasable and non-rewritable format to satisfy regulatory requirements of SEC Rule 17a-4(f), FINRA Rule 4511, or CFTC Regulation 1.31. You can easily designate the records retention time frame to retain regulatory archives in the original form for the required duration, and also place legal holds to retain data indefinitely until the hold is removed.

Q: What AWS documentation supports the SEC 17a-4(f)(2)(i) and CFTC 1.31(c) requirement for notifying my regulator?

Provide notification to your regulator or “Designated Examining Authority (DEA)” of your choice to use Amazon S3 Glacier for electronic storage along with a copy of the Cohasset Assessment. For the purposes of these requirements, AWS is not a designated third party (D3P). Be sure to select a D3P and include this information in your notification to your DEA.

Q: What other controls can be applied with Amazon S3 Glacier Vault Lock?

In certain situations, you may be faced with the need to place a legal hold on your compliance archives for an indefinite period of time. A legal hold can be initiated on a S3 Glacier Vault by creating a vault access policy that denies the use of Glacier’s Delete functions if the vault is tagged in a particular way. In addition to time-based retention and legal hold, Glacier Vault Lock can be used to implement a variety of compliance controls which can be made immutable for strong governance, such as enforcing Multifactor Authentication on all data access/read activities to a vault with classified information.

Q: How do I set up Vault Lock?

You can set up Vault Lock in the Amazon S3 Glacier console or use the Vault Lock APIs in the AWS SDK. To learn more, please read Getting Started with Amazon S3 Glacier Vault Lock in the Amazon S3 Glacier developer's guide.

Data retrievals

Q: How can I retrieve data from the service? >>

Q: What are Standard retrievals? >>

Q: How do I use Standard retrievals? >>

Q: How much do Standard retrievals cost? >>

Q: When should I use Standard retrievals? >>

Q: What are Bulk retrievals? >>

Q: How do I use Bulk retrievals? >>

Q: How much do Bulk retrievals cost? >>

Q: When should I use Bulk retrievals? >>

Q: What are Expedited retrievals? >>

Q: What is a Provisioned capacity unit? >>

Q: When should I provision retrieval capacity? >>

Q: How do I purchase provisioned capacity? >>

Q: How much does provisioned capacity cost? >>

Q: How do I use Expedited retrievals? >>

Q: How much do Expedited retrievals cost? >>

Q: When should I use Expedited retrievals? >>

Q: Can I retrieve part of an archive? >>

Q: Why would I retrieve only a range of an archive? >>

Q: How do I view my jobs? >>

Q: Can I be notified when a job is completed? >>

Data retrieval policies

Q: What are data retrieval policies? >>

Q: How do I set up data retrieval policies? >>

Q: Are data retrieval policies specific to each AWS region? >>

Q: Can I use data retrieval policies to “slow down” my retrievals or spread them out? >>

Q: What impact does the change in the retrieval free tier to 10 GB per month have on my data retrieval policy? >>

Data inventories

Q: Can I see what archives I have stored in Amazon S3 Glacier? >>

Q: Can I obtain a real time list of my vaults? >>

Amazon S3 Glacier Select

Q: What is Amazon S3 Glacier Select? >>

Q: Why should I use Amazon S3 Glacier Select? >>

Q: How does the Amazon S3 Glacier Select compare to legacy archival solutions? >>

Q: What are some scenarios in which I can use Amazon S3 Glacier Select? >>

Q: What kind of latencies can I expect when querying against Amazon S3 Glacier? >>

Q: How do I get started using Amazon S3 Glacier Select? >>

AWS Backup FAQs

General

Q: What is AWS Backup?

A: AWS Backup is a centralized backup service that makes it easy and cost-effective for you to back up your application data across AWS services in the AWS Cloud and on premises, helping you meet your business and regulatory backup compliance requirements. AWS Backup makes protecting your AWS storage volumes, databases, and file systems simple by providing a central place where you can configure and audit the AWS resources you want to back up, automate backup scheduling, set retention policies, and monitor all recent backup and restore activity.

Q: How does AWS Backup work with other AWS services that have backup capabilities?

A: Today, several AWS services offer backup features that help you protect your data, such as EBS snapshots, RDS snapshots, DynamoDB backups, and Storage Gateway snapshots. All existing per-service backup capabilities remain unchanged. AWS Backup provides a new, common way to manage backups across AWS services both in the AWS Cloud and on premises. AWS Backup introduces a centralized backup console that offers backup scheduling, retention management, and backup monitoring. AWS Backup supports existing backup functionality provided by EBS, RDS, DynamoDB, and Storage Gateway. For AWS services that have backup functionality built on AWS Backup, such as Amazon EFS, AWS Backup provides you with backup management capabilities, such as backup scheduling, retention management and backup monitoring, as well as additional features, such as lifecycling backups to a low-cost storage tier, backup storage and encryption that is independent from its source data, and backup access policies.

Q: Why should I use AWS Backup?

A: Backing up your data is an important step towards protecting your application and ensuring that you meet your business and regulatory backup compliance requirements. Even durable resources are susceptible to threats like bugs in your application that could cause accidental deletions or corruption. Building and managing your own backup workflows across all your applications in a compliant and consistent manner can be complex and costly. AWS Backup removes the need for costly, custom solutions or manual processes by providing a fully managed, policy-based backup solution that provides automated backup scheduling and backup retention management.

Q: How does AWS Backup work?

A: To get started with AWS Backup, create a backup policy called a backup plan, which defines parameters such as how frequently to back up your resources and how long to store those backups. You can then assign resources to backup plans and AWS Backup will start automatically backing up these resources and managing backup retention on your behalf according to your backup plan. You can use AWS Backup's central console to view your AWS resources that are being protected, restore from a backup, and monitor backup and restore activity.

Q: What are the key features of AWS Backup?

A: AWS Backup provides a centralized console, automated backup scheduling, backup retention management, and backup monitoring and alerting. AWS Backup also offers advanced features such as lifecycling backups to a low-cost storage tier, backup storage and encryption that is independent from its source data, and backup access policies.

Q: What can I back up using AWS Backup?

A: You can use AWS Backup to manage the backups of EBS volumes, RDS databases, DynamoDB tables, EFS file systems, and Storage Gateway volumes.

Q: Can I use AWS Backup to back up on-premises data?

A: Yes. AWS Backup integrates with Storage Gateway to enable you to back up your on-premises Storage Gateway volumes, providing a common way to manage the backups of your application data both on premises and in the AWS cloud.

Q: Can I use AWS Backup to access backups created by services with existing backup capabilities?

A: Yes. Backups created using services with existing backup capabilities, such as EBS snapshots or DynamoDB backups, can be accessed using AWS Backup. Conversely, backups created by AWS Backup can be accessed using the source service, like EBS or DynamoDB.

Q: How does AWS Backup relate to Amazon Data Lifecycle Manager and when should I use one over the other?

A: Amazon Data Lifecycle Management (DLM) policies and backup plans created in AWS Backup work independently from each other and provide two ways to manage EBS snapshots. DLM provides a simple way to manage the lifecycle of EBS resources, such as volume snapshots. You should use DLM when you want to automate the creation, retention, and deletion of EBS snapshots. You should use AWS Backup to manage and monitor backups across the AWS services you use, including EBS volumes, from a single place.

Q: Does AWS Backup offer a Service Level Agreement (SLA)?

Yes. The AWS Backup SLA provides for a service credit if a customer's monthly uptime percentage is below our service commitment in any billing cycle.

Core Concepts

Q: What is a recovery point?

A: A recovery point represents the content of a resource at a specified time. Recovery points also include metadata such as information about the resource, restore parameters, and tags.

Q: What is a Backup Plan?

A: A backup plan is a policy expression that defines when and how you want to back up your AWS resources, such as DynamoDB tables or EFS file systems. You assign resources to backup plans and AWS Backup will then automatically backup and retain backups for those resources according to the backup plan. Backup plans are composed of one or more backup rules. Each backup rule is composed of 1) a backup schedule, which includes the backup frequency (Recovery Point Objective - RPO) and backup window, 2) a lifecycle rule that specifies when to transition a backup from one storage tier to another and when to expire the recovery point, 3) the Backup Vault in which to place the created recovery points in, and 4) the tags to be added to backups upon creation. For example, a backup plan might have a “daily backup rule” and a “monthly backup rule”. The daily rule backs up resources every day at midnight and retains the backups for one month. The monthly rule takes a backup once a month on the beginning of every month and retains the backups for one year.

Q: What is a Backup Vault?

A: A Backup Vault is a logical backup container for your recovery points that allows you to organize your backups.

Q: How does AWS Backup’s lifecycle feature work?

A: For AWS services that introduce backup functionality built on AWS Backup, such as Amazon EFS, AWS Backup provides a lifecycle feature that allows you to automatically transition your recovery points from a warm storage tier backed by Amazon S3 that provides millisecond access time to your backups to a lower-cost cold storage tier backed by Glacier that provides a restore time of 3-5 hours.

Q: How does encryption work in AWS Backup?

A: Backups from AWS services that introduce backup functionality built on AWS Backup, such as Amazon EFS, are encrypted in-transit and at-rest independently from the source services, giving your backups an additional layer of protection. Encryption is configured at the Backup Vault level. Backups from services with existing backup capabilities are encrypted using the source service’s backup encryption methodology. For example, EBS snapshots are encrypted using the encryption key of the volume the snapshot was created from.

Q: How do I use access policies in a Backup Vault to control access to backups

A: AWS Backup allows you to set resource-based policies on Backup Vaults, enabling you to control access to the Backup Vault and the backups in it.

Q: What services provide support for AWS Backup's advanced features?

A: Amazon EFS supports AWS Backup's advanced features with backup functionality integrated with AWS Backup.

Compliance

Q: Which compliance programs does AWS Backup support?

A: AWS has the longest-running compliance program in the cloud and are committed to helping customers navigate their requirements. AWS Backup has been assessed to meet global and industry security standards. It complies with PCI DSS, ISO 9001, 27001, 27017, and 27018), in addition to being HIPAA eligible. That makes it easier for you to verify our security and meet your own obligations. For more information and resources, visit our compliance pages. You can also go to the Services in Scope by Compliance Program page to see a full list of services and certifications.

Q: Is AWS Backup PCI compliant?

A: Yes. AWS Backup is PCI-DSS compliant, which means you can use it to transfer payment information. You can download the PCI Compliance Package in AWS Artifact to learn more about how to achieve PCI Compliance on AWS.

Q: Is AWS Backup HIPAA eligible?

A: Yes. AWS Backup is HIPAA eligible, which means if you have a HIPAA BAA in place with AWS, you can use AWS Backup to tranfer protected health information (PHI).

AWS Snow family FAQs

General

Q: What is the AWS Snow family?

The AWS Snow family are physical devices that help migrate large amounts of data into and out of the cloud without depending on networks. This helps you apply the wide variety of AWS services for analytics, file systems and archives to your data. Snowball is a suitcase-sized device, Snowball Edge is a rack mountable and clusterable suitcase sized device with compute capabilties, and Snowmobile is a shipping container moved with a tractor-trailer. These services can assist with data migration, disaster recovery, data center shutdown and remote data collection projects.

Q: What are some example use cases?

You can use AWS Snow family services for data transfer and occasional pre-processing on location. Some large data transfer examples include cloud migration, disaster recovery, data center relocation, and/or remote data collection projects. These projects typically require you to migrate large amounts of data in the shortest, and most cost-effective, amount of time.

Some example use cases for Snowball Edge computing capabilties include IoT sensor stream capture, on-the-fly media transcoding, image compression, aggregating metrics, and control signaling and alarming.

Q: Why do I need this service?

It can take a long time to transfer large amounts of data over the wire, and some locations don't have any connectivity at all. The Snow family helps expedite data transfers in a more secure and cost-effective way. Each service has a pre-set capacity level to make it easy to choose. Snowball Edge further helps bring computing applications closer to the data source to enhance analysis and deliver real-time results.

Q: Which Snow family service is best for me?

Th Snow family comes in multiple capacities and form factors to fit most data migration and/or remote data collection projects. First determine a capacity point then evaluate your need for on-site computing capabilities.

Please refer to the comparison table for more details.

Q: How quickly can I migrate data?

The Snowball and Snowball Edge services can typically transfer up to 100TBs in about a week. Snowmobile can transfer data at a rate of up to 1TB/s, which means 100PBs can be loaded in just a few weeks. In comparison, a dedicated T3 line at 50Mb/s takes years.

Q: How do I choose between Snow family and other AWS data migration services?

AWS Storage Gateway and Direct Connect services are good choices when network bandwidth limitations do not exist. For the most efficient means of data transfer, whether connected to a network or not, the Snow family of services provides good choices as well. Amazon offers a variety of tools to help you move data via networks, roads, and technology partners. See the pages below to find out which service is best for you.

Learn more about migrating to AWS here »

Learn more about which data migration tool is right for you »

Q: What AWS Regions are supported?

Snow family services are available for use in specific AWS regions. To discuss data transfer needs specific for your region, please follow up with AWS Sales, or see the Regional Service Availability pages, for more information.

AWS Storage Gateway FAQs

General

Q: What is the AWS Storage Gateway service?

A: The AWS Storage Gateway service enables hybrid cloud storage between on-premises environments and the AWS Cloud. It seamlessly integrates on-premises enterprise applications and workflows with Amazon's block and object cloud storage services through industry standard storage protocols. It provides low-latency performance by caching frequently accessed data on premises, while storing data securely and durably in Amazon cloud storage services. It provides an optimized data transfer mechanism and bandwidth management, which tolerates unreliable networks and minimizes the amount of data being transferred. It brings the security, manageability, durability, and scalability of AWS to existing enterprise environments through native integration with AWS encryption, identity management, monitoring, and storage services. Typical use cases include backup and archiving, disaster recovery, moving data to S3 for in-cloud workloads, and tiered storage.

AWS Storage Gateway supports three storage interfaces: file, tape, and volume. Each gateway you have can provide one type of interface.

The *file gateway* enables you to store and retrieve objects in Amazon S3 using file protocols, such as NFS. Objects written through file gateway can be directly accessed in S3.

The tape gateway provides your backup application with an iSCSI virtual tape library (VTL) interface, consisting of a virtual media changer, virtual tape drives, and virtual tapes. Virtual tape data is stored in Amazon S3 or can be archived to Amazon S3 Glacier

The *volume gateway* provides block storage to your applications using the iSCSI protocol. Data on the volumes is stored in Amazon S3. To access your iSCSI volumes in AWS, you can take EBS snapshots which can be used to create EBS volumes.

Q: How do I use the AWS Storage Gateway service?

A: You can have two touchpoints to use the service: the AWS Management Console and a gateway that is available as a virtual machine (VM) or as a physical hardware appliance.

You use the AWS Management Console to download the virtual appliance gateway or purchase the hardware appliance, configure storage, and manage and monitor the service. The gateway connects your applications to AWS storage by providing standard storage interfaces. It provides transparent caching, efficient data transfer, and integration with AWS monitoring and security services.

To get started, sign up for an AWS account and visit the AWS Storage Gateway Management Console to download a gateway VM appliance, or purchase the hardware appliance. Once you've installed your gateway, you associate it with your AWS Account through our activation process. After activation, you configure the gateway to connect to the appropriate storage type. For file gateway, you configure file shares that are mapped to selected S3 buckets, using IAM roles. For volume gateway, you create and mount volumes as iSCSI devices. For tape gateway, you connect your backup application to create and manage tapes. Once configured, you start using the gateway to write and read data to and from AWS storage. You can monitor the status of your data transfer and your storage interfaces through the AWS Management Console. Additionally, you can use the API or SDK to programmatically manage your application's interaction with the gateway.

Q: Where can I deploy a Storage Gateway appliance?

A: On-premises, you can deploy a virtual machine containing the Storage Gateway software on VMware ESXi, Microsoft Hyper-V, or Linux KVM, or you can deploy Storage Gateway as a hardware appliance. You can also deploy the Storage Gateway VM in VMware Cloud on AWS, or as an AMI in Amazon EC2.

Q: What is file gateway?

A: File gateway presents a file-based interface to Amazon S3, which appears as a network file share. It enables you to store and retrieve Amazon S3 objects through standard file storage protocols. File gateway allows your existing file-based applications or devices to use secure and durable cloud storage without needing to be modified. With file gateway, your configured S3 buckets will be available as Network File System (NFS) mount points or Server Message Block (SMB) file shares. Your applications read and write files and directories over NFS or SMB, interfacing to the gateway as a file server. In turn, the gateway translates these file operations into object requests on your S3 buckets. Your most recently used data is cached on the gateway for low-latency access, and data transfer between your data center and AWS is fully managed and optimized by the gateway. Once in S3, you can access the objects directly or manage them using features such as S3 Lifecycle Policies, object versioning, and cross-region replication. You can run file gateway on-premises or in EC2.

Q: What is tape gateway?

A: Tape gateway is a cloud-based Virtual Tape Library (VTL). It presents your backup application with a VTL interface, consisting of a media changer and tape drives. You can create virtual tapes in your virtual tape library using the AWS Management Console. Your backup application can read data from or write data to virtual tapes by mounting them to virtual tape drives using the virtual media changer. Virtual tapes are discovered by your backup application using its standard media inventory procedure. Virtual tapes are available for immediate access and are backed by Amazon S3. You can also archive tapes. Archived tapes are stored in Amazon S3 Glacier or Amazon S3 Glacier Deep Archive.

Q: What is volume gateway?

A: Volume gateway provides an iSCSI target, which enables you to create block storage volumes and mount them as iSCSI devices from your on-premises or EC2 application servers. The volume gateway runs in either a cached or stored mode.

- In the cached mode, your primary data is written to S3, while retaining your frequently accessed data locally in a cache for low-latency access.

- In the stored mode, your primary data is stored locally and your entire dataset is available for low-latency access while asynchronously backed up to AWS.

In either mode, you can take point-in-time snapshots of your volumes, which are stored as Amazon EBS Snapshots in AWS, enabling you to make space-efficient versioned copies of your volumes for data protection, recovery, migration and various other copy data needs.

Q: What benefits does AWS Storage Gateway provide?

A: AWS Storage Gateway provides a set of features that enable you to effectively leverage AWS storage within your existing applications and workflows. It provides a standard set of protocols such as iSCSI, SMB and NFS, which allow you to use your existing applications without any changes. Through its local cache, the gateway provides low-latency access to recently used data. The gateway optimizes data transfer to AWS storage, such as optimization of transfer through intelligent buffering, upload management to address network variations, and bandwidth management. The gateway provides you an effective mechanism to store data in AWS across the range of storage services most suitable for your use cases. The gateway is easy to deploy and can use your existing virtual infrastructure and hypervisor investments, or can be installed in your data center or remote offices as a hardware appliance. The gateway software running as a VM or on the hardware appliance is stateless, allowing you to easily create and manage new instances of your gateway as your storage needs evolve. Finally, the service integrates natively into AWS management services such as Amazon CloudWatch, AWS CloudTrail, AWS Key Management Service (KMS), and AWS Identity and Access Management (IAM).

Q: What sort of encryption does AWS Storage Gateway use to protect my data?

A: All data transferred between any type of gateway appliance and AWS storage is encrypted using SSL. By default, all data stored by AWS Storage Gateway in S3 is encrypted server-side with Amazon S3-Managed Encryption Keys (SSE-S3). Also, you can optionally configure different gateway types to encrypt stored data with AWS Key Management Service (KMS) via the Storage Gateway API.

See below for specifics on KMS support by File Gateway, Tape Gateway, and Volume Gateway.

Q: Is AWS Storage Gateway HIPAA eligible?

A: Yes. AWS Storage Gateway is HIPAA eligible. If you have an executed Business Associate Agreement (BAA) with AWS, you can use Storage Gateway to store, backup and archive protected health information (PHI) on scalable, cost-effective, and secure AWS storage services, including Amazon S3, Amazon S3 Glacier and Amazon EBS, which are also HIPAA eligible.

Information on HIPAA eligible services on AWS can be found at our HIPAA Compliance page, and you can enter into a BAA with AWS here. HIPAA eligibility for Storage Gateway applies to all gateway types (File, Volume and Tape).

Q: Is AWS Storage Gateway PCI compliant?

A: Yes, the AWS Storage Gateway in compliance with the Payment Card Industry Data Security Standard (PCI DSS) based on recent assessments. Existing customers can download the Attestation of Compliance (AOC) and PCI Responsibility Summary reports in the AWS Management Console with AWS Artifact, and prospective customers can request them through AWS sales.

Q: What AWS Storage Gateway types can I manage through AWS Backup?

A: You can manage backup and retention policies for cached and stored volume modes of Volume Gateway through AWS Backup.

Q: Can I deploy a Storage Gateway on my private non-routable network? Does Storage Gateway support AWS PrivateLink?

A: Yes, you can deploy a Storage Gateway on a private, non-routable network if that network is connected to your Amazon VPC via DX or VPN. Storage Gateway traffic will be routed via VPC endpoints powered AWS PrivateLink, a technology that enables private connectivity between AWS services using Elastic Network Interfaces (ENI) with private IPs in your VPCs. To learn more about PrivateLink, visit the PrivateLink documentation. To setup AWS PrivateLink for Storage Gateway, visit the AWS PrivateLink for Storage Gateway documentation.

File Gateway

Q: What is file gateway?

A: File Gateway is a configuration of the AWS Storage Gateway service that provides your applications a file interface to seamlessly store files as objects in Amazon S3, and access them using industry standard file protocols.

Q: What can I do with file gateway?

A: Use cases for file gateway include: (a) migrating on-premises file data to Amazon S3, while maintaining fast local access to recently accessed data, (b) Backing up on-premises file data as objects in Amazon S3 (including Microsoft SQL Server and Oracle databases and logs), with the ability to use S3 capabilities such as lifecycle management, versioning and cross region replication, and, (c) Hybrid cloud workflows using data generated by on-premises applications for processing by AWS services such as machine learning, big data analytics or serverless functions.

Q: What are the benefits of using file gateway to store data in S3?

A: File gateway enables your existing file-based applications, devices, and workflows to use Amazon S3, without modification. File gateway securely and durably stores both file contents and metadata as objects, while providing your on-premises applications low-latency access to cached data.

Q: Which Amazon S3 storage classes does File Gateway support?

A: File gateway supports Amazon S3 Standard, S3 Standard - Infrequent Access (S3 Standard - IA) and S3 One Zone - IA. For details on storage classes, refer to the Amazon S3 documentation. You configure the initial storage class for objects that the gateway creates, and then you can use bucket lifecycle policies to move files from Amazon S3 to Amazon S3 Glacier. If an application attempts to access a file/object stored through File Gateway that is now in Amazon S3 Glacier, you will receive a generic I/O error.

Q: What protocols does file gateway support?

A: File gateway supports Linux clients connecting to the gateway using Network File System (NFS) versions 3 and 4.1 for Linux clients, and supports Windows clients connecting to the gateway using Server Message Block (SMB) versions 2 and 3.

Q: How can I create and use a file share?

A: You can create an NFS or SMB file share using the AWS Management Console or service API and associate the file share with a new or existing Amazon S3 bucket. To access the file share from your applications, you mount it from your application using standard UNIX or Windows commands. For convenience, example command lines for each environment are shown in the management console.

Q: What options do I have to configure an NFS file share?

A: You can configure your NFS file share with administrative controls such as limiting access to specific NFS clients or networks, read-only or read-write, or enabling user permission squashing.

Q: What options do I have to configure an SMB file share?

A: You can configure your SMB file share to be accessed by Active Directory (AD) users only or provide authenticated guest access to users in your organization. You can further limit access to the file share as read-only or read-write, or to specific AD users and groups.

Q: Does file gateway support integration with on-premises Microsoft Active Directory (AD)?

A: Yes, file gateway integrates with Microsoft Active Directory on-premises as well as with in-cloud Active Directory solutions such as Managed Microsoft AD.

Q: Can I export an SMB file share without Active Directory?

A: Yes. You can export an SMB file shares using a guest username and password. You will need to change the default password using the Console or service API before setting up your file share for guest access.

Q: Can I export a mix of NFS and SMB file shares on the same gateway?

A: Yes.

Q: Can I export an NFS and SMB file share on the same bucket?

A: No, currently file metadata, such as ownership, stored as S3 object metadata cannot be mapped across different protocols.

Q: How does file gateway access my S3 bucket?

A: File gateway uses an AWS Identity and Access Management (IAM) role to access your S3 bucket. You can set up an IAM role yourself or have it automatically setup by the AWS Storage Gateway Management Console. For automatic setup, AWS Storage Gateway will create a new IAM role in your account and associate it with an IAM Access Policy to access your S3 bucket. The IAM role and IAM access policy are created in your account and you can fully manage them yourself.

Q: How does my application access my file share?

A: To use the file share, you mount it from your application using standard UNIX or Windows commands. For convenience, example command lines are shown in the management console

Q: What is the relationship between files and objects?

A: Files are stored as objects in your S3 buckets and you can configure the initial storage class for objects that file gateway creates. There is a one-to-one relationship between files and objects, and you can configure the initial storage class for objects that file gateway creates.

The object key is derived from the file path within the file system. For example, if you have a gateway with hostname *file.amazon.com* and have mapped *my-bucket*, then file gateway will expose a mount point called *file.amazon.com:/export/my-bucket*. If you then mount this locally on */mnt/my-bucket* and create a file named *file.html* in a directory */mnt/my-bucket/dir* this file will be stored as an object in the bucket *my-bucket* with a key of *dir/file.html*. Creating sparse files will result in a non-sparse zero-filled object in S3.

Q: What file system operations are supported by file gateway?

A: Your clients can create, read, update, and delete, files and directories. Files are stored as individual objects in Amazon S3. Directories are managed as folder objects in S3, using the same syntax as the S3 console. Symbolic links and hard links are not supported. Attempting to create a link will result in an error.

Rename operations will appear atomic to your clients, but S3 does not support renaming of objects. When you rename a file or directory the gateway performs copy-put requests to create a copy of the objects in S3 under the new keys and then deletes the originals. This avoids having to re-send large files over the network. Renaming directories containing a large number of files is not instantaneous and will result in 2 copies of your data being stored in S3 until the rename operation completes.

Q: What file system metadata can my client access and where is the metadata stored?

A: Your clients can access POSIX-style metadata including ownership, permissions, and timestamps that are durably stored in S3 in the user metadata of the object associated with the file. When you create a file share on an existing bucket, the stored metadata will be restored and made accessible to your clients.

Q: How do I set the Content-Type for files uploaded to S3?

A: For each file share, you can enable guessing of MIME types for uploaded objects upon creation or enable the feature later. If enabled, file gateway will use the filename extension to determine the MIME type for the file and set the S3 objects Content-Type accordingly. This is beneficial if you are using file gateway to manage objects in S3 which you access directly via URL or distribute through Amazon CloudFront.

Q: Can I directly access objects stored in S3 by using file gateway?

A: Yes. Once objects are stored in S3, you can access them directly in AWS for in-cloud workloads without requiring file gateway. Your objects inherit the properties of the S3 bucket in which they are stored, such as lifecycle management, and cross-region replication.

An object that needs to be accessed by using a file share should only be managed by the gateway. If you directly overwrite or update an object previously written by file gateway, it results in undefined behavior when the object is accessed through the file share.

Q: What if my bucket already contains objects?

A: If your bucket already contains objects when you configure it for use with file gateway, object keys will be used to present the objects as files to the NFS and SMB clients. The files are given default file system metadata.

To reduce latency and number of S3 requests, file gateway only scans the headers for file metadata associated with the objects when you explicitly list the files or directories. File metadata is collected as a part of that scan, file contents are downloaded only when the object is read.

Q: How are buckets accessed by the gateway? Are entire bucket or file contents downloaded?

A: The gateway does not automatically download full objects or all the data that exists in your bucket; data is only downloaded when it is explicitly accessed by your clients. Additionally, to reduce data transfer overhead, file gateway uses multipart uploads and copy put, so only changed data in your files is uploaded to S3.

Q: What metadata can my NFS client access for objects created outside of the gateway?

A: For objects uploaded to the S3 bucket directly, i.e. not using file gateway and an NFS share, you can configure default ownership and permissions.

Q: What metadata can my SMB client access for objects created outside of the gateway?

A: For objects uploaded to the S3 bucket directly, i.e. without using file gateway and an SMB share, metadata such as ownership and permissions will be inherited from the object's parent folder. Permissions at the root of the share are fixed and objects created directly under the root folder will inherit these fixed permissions. Refer to the documentation on metadata settings of objects created outside the gateway.

Q: Can I use multiple NFS clients with a single file gateway?

A: You can have multiple NFS clients accessing a single file gateway. However, as with any NFS server, concurrent modification from multiple NFS clients can lead to unpredictable behavior. Application level coordination is required to do this in a safe way.

Q: Can I have multiple writers to my S3 bucket?

A: No. We recommend a single writer to objects in your S3 bucket. If you directly overwrite or update an object previously written by file gateway, it results in undefined behavior when the object is accessed through the file share. Concurrent modification of the same object (e.g. via the S3 API and the file gateway) can lead to unpredictable results and we recommend against this configuration.

Q: Can I have two gateways writing independent data to the same bucket?

A: We do not recommend configuring multiple writers to a single bucket because it can lead to unpredictable results. You could enforce unique object names or prefixes through your application workflow. File gateway doesn't monitor or report on conflicts in such a setup.

Q: Can I have multiple gateways reading data from the same bucket?

A: Yes, you can have multiple readers on a bucket managed through a file gateway. You can configure a file share as read-only, and allow multiple gateways to read objects from the same bucket. Additionally, you can refresh the inventory of objects that your gateway knows about using the Storage Gateway Console or the RefreshCache API.

Note however that the if you do not configure a file share as read-only, file gateway does not monitor or restrict these readers from inadvertently writing to the bucket. It is up to you to maintain a single writer/multi reader configuration from your application.

Q: Can I monitor my file share using Amazon CloudWatch?

A: Yes, you can monitor usage of your file share using Amazon CloudWatch metrics and get notified on completion of file operations through CloudWatch Events. To learn more, visit Monitoring your File Share.

Q: How do I know when my file is uploaded?

A: When you write files to your file share with file gateway, the data is stored locally first and then asynchronously uploaded to your S3 bucket. You can request notifications through AWS CloudWatch Events when this upload completes. These notifications can be used to trigger additional workflows, such as invoking an AWS Lambda function or Amazon EC2 Systems Manager Automation, which is dependent upon the data that is now available in S3. To learn more, please refer to the documentation for File Upload Notification.

Q: Can I update my file gateway's view of a bucket to see objects created from an object-based workload or another file gateway?

A: Yes, you can refresh the inventory of objects that your file gateway knows about using the Console or the RefreshCache API. You will receive notifications through AWS CloudWatch Events when the refresh cache operation has completed. These notifications can be used to send emails using Amazon SNS, or trigger local processing using the updated contents. To learn more, please refer to the documentation.

Q: Can I use the gateway to update data in a bucket that belongs to another AWS account?

A: Yes, you can use the gateway for cross-account access to buckets. To learn more, please refer to the documentation for Using File Share for Cross-Account access.

Q: Can I use the gateway to access data in Requester Pays S3 buckets?

A: Yes, when creating your file share you can enable access to Requester Pays S3 buckets. As a requester, you will incur the charges associated with accessing data from Requester Pays buckets.

Q: How many files shares can I create per bucket?

A: There is a one-to-one mapping between a file share and a bucket. We do not limit the number of file shares per bucket. However, we recommend having a single writer to the bucket, either a file gateway or client accessing S3 directly.

Q: How many file shares can I create per gateway?

A: You can create up to 10 file shares per gateway.

Q: What is the maximum size of an individual file?

A: The maximum size of an individual file is 5 TB, which is the maximum size of an individual object in S3. If you write a file larger than 5 TB, you will get a "file too large" error message and only the first 5 TB of the file will be uploaded.

Q: My application checks storage size before copying data. What storage size does the gateway return?

A: The gateway returns a large number (8 EB) as your total capacity. Amazon S3 does not limit total storage.

Q: Can I use versioning, lifecycle, cross-region replication, and S3 event notification?

A: Yes. Your bucket policies for versioning, lifecycle management, cross-region replication, and S3 event notification, apply directly to objects stored in your bucket through AWS Storage Gateway.

You can use S3 lifecycle policies to change an object's storage tier or delete old objects or object versions. In the case of objects deleted by lifecycle policy, you will need to call the RefreshCache API to reflect these changes to your NFS clients.

When using an S3 bucket which is the target for cross-region replication, you may need to use the RefreshCache API to ensure the gateway cache and S3 bucket are in sync.

If using S3 event notifications you may receive events for partial files created by the gateway to ensure your data is durably stored in S3. Partial files may occur for a number of reasons, such as the gateway needing to free up cache space, or a high rate of writes to a file. These partial files may not be application consistent.

Q: Can I use file gateway with my backup application?

A: File gateway supports SMB versions 2 and 3 as well as NFS versions 3 and 4.1. We are continuing to do on-going testing with common backup apps. Please let us know of any specific apps with which you'd like to see compatibility tested.

Q: Can I use file gateway to write files to EFS?

A: No. File gateway allows you to store files as objects in S3.

Q: When should I use file gateway vs. the S3 API?

A: You can use file gateway when you want to access objects in S3 as files using standard filesystem operations. File gateway additionally provides low-latency local access and efficient data transfer. You can use the S3 API when your application doesn’t require file system operations and can manage data transfer directly.

Q: How does file gateway manage the local cache? What data gets stored locally?

A: Local disk storage on the gateway is used to temporarily hold changed data that needs to be transferred to AWS, and to locally cache data for low-latency read access. File gateway automatically manages the cache maintaining the most recently accessed data based on client read and write operations. Data is evicted from the cache only when space is needed to store more recently used data.

To maximize write performance, the gateway uses a write-back mechanism where data is first persisted to disk and then asynchronously uploaded to S3. The gateway serves data from the local cache to maximize read performance. If not present, data is efficiently synchronously fetched from Amazon S3 using byte-range gets.

The local cache should generally be sized for the working set of data that you need low-latency access to. If the cache is too small then read latencies will increase as data being requested must be fetched from S3, and writes could fail if there is no free cache space to store data locally pending upload to S3.

Q: What guidance should I use to provision size of the gateway's cache disk? What happens if I provision a smaller cache disk?

A: You should provision your cache based on:
1/ The size of your working dataset to which you need low-latency access, so you can reduce read latencies by decreasing the frequency with which data is requested from S3, and
2/ The size of files written to the gateway by your applications.

Smaller cache disks can result in poor performance and failures during writes if there is no free cache space to store data locally when pending upload to S3. To learn more about monitoring your cache usage, refer to Monitoring Your File Share in the documentation.

Q: When does data in the cache get evicted?

A: Data written to the cache from your applications or through retrieval from Amazon S3 is evicted from the cache only when space is needed to store more recently accessed data.

Q: Does file gateway perform data reduction (deduplication or compression)?

A: No. Files are mapped to objects one-to-one in your bucket without modification, enabling you to access your data directly in S3 without needing to use the gateway or deploy additional software to rehydrate your data.

File gateway uses multipart uploads and copy put, so only changed data is uploaded to S3 which can reduce data transfer. The gateway does not automatically download full objects or all the data that exists in your bucket; data is only downloaded when explicitly accessed by your NFS client.

Q: Can I use file gateway with Amazon S3 Transfer Acceleration?

A: File gateway will not use the accelerated endpoints even if your bucket is configured for S3 Transfer Acceleration.

Q: What sort of encryption does file gateway use to protect my data?

A: All data transferred between the gateway and AWS storage is encrypted using SSL. By default, all data stored in S3 is encrypted server-side with Amazon S3-Managed Encryption Keys (SSE-S3). For each file share you can optionally configure to have your objects encrypted with AWS KMS-Managed Keys using SSE-KMS. To learn more, please see "Encrypting Your Data Using AWS Key Management System," in the Storage Gateway User Guide, which includes critical details about usage of the feature.

Tape Gateway

Q: What are the benefits of storing virtual tapes in AWS compared to warehousing tapes offsite?

A: You get 11 9s of data durability, fixity checks by AWS on a regular basis, data encryption, right data when you restore, and cost savings, when storing virtual tapes in AWS using Tape Gateway with S3 Glacier Deep Archive compared to warehousing physical tapes offsite. First, all virtual tapes stored in S3 Glacier Deep Archive are replicated and stored across at least three geographically-dispersed Availability Zones, protected by 11 9s of durability. Second, AWS performs fixity checks on a regular basis to confirm your data can be read and no errors have been introduced. Third, all tapes stored in S3 Glacier Deep Archive are protected by S3 Server Side Encryption using default keys or your KMS keys. In addition, you also avoid physical security risk associated with tape portability. Fourth, compared to the experience of warehousing tapes offsite where you may receive an incorrect or broken tape during restore, with Tape

Gateway, you always get correct data. Finally, you can save in monthly storage costs when storing your data in S3 Glacier Deep Archive compared to warehousing tapes offsite.

Q: What Amazon S3 storage classes does Tape Gateway support?

A: Tape Gateway supports S3 Standard, S3 Glacier, and S3 Glacier Deep Archive storage classes. Data on your virtual tapes is stored in virtual tape library in Amazon S3 when backup application is writing data to tapes. After you eject tapes from backup application, your tapes are archived to S3 Glacier or S3 Glacier Deep Archive.

Q: How much data can I store on a virtual tape?

A: The minimum size and maximum size of a virtual tape you can create on a tape gateway is 100 GiB and 5 TiB respectively. Please note, you only pay for the amount of data stored on each tape, and not for the size of the tape.

Q: How many tapes can the virtual tape library (VTL) hold?

A: A single tape gateway can have up to 1,500 virtual tapes in the VTL with a maximum aggregate capacity of 1 PB, however there is no limit to the amount of data or number of virtual tapes you can archive. You can also deploy additional tape gateways to scale storage for virtual tapes that are not archived. For more information, please see our documentation on Storage Gateway limits.

Q: How much data can I store in tape archives?

A: There is no limit to the amount or size or virtual tapes that you can archive.

Q: Which S3 storage classes can I retrieve my archived virtual tape to?

A: You can retrieve a virtual tape archived in S3 Glacier or S3 Glacier Deep Archive to S3. A tape archived in S3 Glacier is retrieved to S3 using standard retrieval method typically within 3-5 hours. A tape archived in S3 Glacier Deep Archive is retrieved to S3 using standard retrieval method typically within 12 hours.

Q: How do I access my data on virtual tapes?

A: The virtual tape containing your data must be stored in a virtual tape library before it can be accessed. Access to virtual tapes in your virtual tape library is instantaneous. If the virtual tape containing your data is archived, you can retrieve the virtual tape using the AWS Management Console or API. First select the virtual tape, then choose the virtual tape library into which you want the virtual tape to be loaded.You can retrieve a tape archived in S3 Glacier and S3 Glacier Deep Archive to S3, typically within 3-5 hours and 12 hours, respectively. Once the virtual tape is available in the virtual tape library, you can use your backup application to make use of the virtual tape to restore data.

Q: Will I be able to access the virtual tapes in my virtual tape library using Amazon S3 or Amazon S3 Glacier APIs?

A: No. You cannot access virtual tape data using Amazon S3 or Amazon S3 Glacier APIs. However, you can use the tape gateway APIs to manage your virtual tape library and your virtual tape shelf.

Q: How do I use Tape Gateway with S3 Glacier Deep Archive storage class?

A: When creating new tapes through the Storage Gateway console or API, you can set archival storage target to S3 Glacier Deep Archive. When your backup software ejects the tapes, they will be archived to S3 Glacier Deep Archive. You can retrieve a virtual tape archived in S3 Glacier Deep Archive to S3 using standard retrieval method typically within 12 hours.

Q: Can I move my existing virtual tapes in S3 Glacier to S3 Glacier Deep Archive?

A: Yes. Tape Gateway supports moving your tapes in S3 Glacier to S3 Glacier Deep Archive. You can assign the tape placed in Glacier Pool to Deep Archive Pool using AWS Storage Gateway Console or API. Tape Gateway will then move the virtual tape to Deep Archive Pool associated with the S3 Glacier Deep Archive storage class. You will incur tape move charge for moving a tape from S3 Glacier to S3 Glacier Deep Archive and if applicable, an early deletion fee for S3 Glacier, if you move a tape from S3 Glacier to S3 Glacier Deep Archive prior to 90 days.

Q: Can I move a tape in S3 Glacier Deep Archive to S3 Glacier?

A: No, you cannot move a tape from S3 Glacier Deep Archive to S3 Glacier. You can retrieve a tape from S3 Glacier Deep Archive to S3 or delete a tape from S3 Glacier Deep Archive.

Q: What backup applications can I use with tape gateway?

A: The VTL interface is compatible with backup and archival applications that use the industry-standard iSCSI-based tape library interface. For a full list of the supported backup applications see the requirements section of the AWS Storage Gateway user guide.

Q: What sort of encryption does tape gateway use to protect my data?

A: All data transferred between the gateway and AWS storage is encrypted using SSL. By default, all data stored by tape gateway in S3 is encrypted server-side with Amazon S3-Managed Encryption Keys (SSE-S3).

You can optionally configure encryption on tapes using AWS KMS-Managed Keys via the Storage Gateway API. You will be able to specify one of the managed Customer Master Keys (CMKs) as the KMS key. The configured CMK used to encrypt tape data cannot be changed after creation. To learn more, please see "Encrypting Your Data Using AWS Key Management System," in the Storage Gateway User Guide, which includes critical details about usage of the feature.

Volume Gateway

Q: How much volume data can I manage per gateway? What is the maximum size of a volume?

A: Each *volume gateway* can support up to 32 volumes. In *cached mode*, each volume can be up to 32 TB for a maximum of 1 PB of data per gateway (32 volumes, each 32 TB in size). In *stored mode*, each volume can be up to 16 TB for a maximum of 512 TB of data per gateway (32 volumes, each 16 TB in size).

For more information, please refer to our documentation on Storage Gateway limits.

Volume gateways compress data before that data is transferred to AWS and while stored in AWS. This compression can reduce both data transfer and storage charges. Volume storage is not pre-provisioned; you will be billed for only the amount of data stored on the volume, not the size of the volume you create.

Q: When I look in Amazon S3 why can't I see my volume data?

A: Your volumes are stored in an Amazon S3 bucket maintained by the AWS Storage Gateway service. Your volumes are accessible for I/O operations through AWS Storage Gateway. You cannot directly access them using Amazon S3 API actions. You can take point-in-time snapshots of gateway volumes that are made available in the form of Amazon EBS snapshots, which can be turned into either Storage Gateway Volumes or EBS Volumes. Use the file gateway to work with your data natively in S3.

Q: What sort of encryption does volume gateway use to protect my data?

A: All data transferred between the gateway and AWS storage is encrypted using SSL. By default, all data stored by volume gateway in S3 is encrypted server-side with Amazon S3-Managed Encryption Keys (SSE-S3).

You can optionally configure encryption for data stored in AWS on volumes using AWS KMS managed keys via the Storage Gateway API. You will be able to specify one of the managed Customer Master Keys (CMKs) as the KMS key. The configured CMK used to encrypt a volume cannot be changed after creation. To learn more, please see "Encrypting Your Data Using AWS Key Management System," in the Storage Gateway User Guide, which includes critical details about usage of the feature.

Q: Can I create an EBS Snapshot from KMS-encrypted volume?

A: Yes. You can create an EBS snapshot from an AWS KMS-encrypted volume using the API. The EBS snapshot will be encrypted using the same key as the one used for volume encryption

Q: Can I create a volume from KMS-encrypted EBS snapshot?

A: Yes. You can create an encrypted volume from KMS-encrypted EBS snapshot using the API. The encrypted volume can use the same key that was used to encrypt the EBS snapshot, or you can specify a different encryption key for encrypting the volume.

Q: Why would I use snapshots?

A: You can take point-in-time snapshots of your volume gateway volumes in the form of Amazon EBS snapshots.You can use a snapshot of your volume as the starting point for a new Amazon EBS volume, which you can then attach to an Amazon EC2 instance. Using this approach, you can easily supply data from your on-premises applications to your applications running on Amazon EC2 if you require additional on-demand compute capacity for data processing or replacement capacity for disaster recovery purposes.

For cached volumes, where your volume data is already stored in Amazon S3, you can use snapshots to preserve versions of your data. Using this approach, you can revert to a prior version when required or repurpose a point-in-time version as a new volume. You can initiate snapshots on a scheduled or ad hoc basis. When taking a new snapshot, only the data that has changed since your last snapshot is stored. If you have a volume with 100 GB of data, but only 5 GB of data have changed since your last snapshot, only the 5 additional GB of snapshot data will be stored in Amazon S3. When you delete a snapshot, only the data not needed for any other snapshot is removed.

For stored volumes, where your volume data is stored on-premises, snapshots provide durable, off-site backups in Amazon S3. You can create a new volume from a snapshot if you need to recover a backup. You can also use a snapshot of your volume as the starting point for a new Amazon EBS volume which you can then attach to an Amazon EC2 instance.

Q: What data will my snapshot contain? How do I know when to take a snapshot to ensure my data is backed up?

A: Snapshots represent a point-in-time copy of the volume at the time the snapshot is requested. They contain all of the information needed to restore your data (from the time the snapshot was taken) to a new volume. Data written to the volume by your application prior to taking the snapshot, but not yet been uploaded to AWS, will be included in the snapshot.

In practical terms, the snapshot will be assigned an ID and visible in the AWS Management Console and AWS Command Line Interface (AWS CLI) immediately, but will initially be in a PENDING status. When all data written to the volume prior to the snapshot request has been uploaded from the gateway and into EBS, the status will change to AVAILABLE. At this point you can use the snapshot as the base for a new gateway or EBS volume.

Q: How do I restore a snapshot to a gateway?

A: Each snapshot is given a unique identifier that you can view using the AWS Management Console. You can create AWS Storage Gateway or Amazon EBS volumes based on any of your existing snapshots by specifying this unique identifier.

Using the AWS Management Console, you can create a new volume from a snapshot you've stored in Amazon S3. You can then mount this volume as an iSCSI device to your on-premises application server.

Because cached volumes store your primary data in Amazon S3, when creating a new volume from a snapshot, your gateway keeps the snapshot data in Amazon S3 where it becomes the primary data for your new volume.

Because stored volumes store your primary data locally, when creating a new volume from a snapshot, your gateway downloads the data contained within the snapshot to your local hardware. There it becomes the primary data for your new volume.

Q: Do the AWS Storage Gateway's volumes need to be un-mounted in order to take a snapshot? Does the snapshot need to complete before the volume can be used again?

A: No, taking snapshots does not require you to un-mount your volumes, nor does it impact your application's performance. However, snapshots only capture data that has been written to your AWS Storage Gateway volume, which may exclude any data that has been locally buffered by your application or OS.

Q: Can I schedule snapshots of my AWS Storage Gateway volumes?

A: Yes, you can create a snapshot schedule for each of your volumes. You can modify both the time the snapshot occurs each day, as well as the frequency (every 1, 2, 4, 8, 12, or 24 hours).

Q: How long does it take to complete a snapshot?

A: The time it takes to complete a snapshot is largely dependent upon the size of your volume and the speed of your Internet connection to AWS. The AWS Storage Gateway compresses all data prior to upload, reducing the time to take a snapshot.

Q: Will I be able to access my snapshot data using Amazon S3's APIs?

A: No, snapshots are only accessible from the AWS Storage Gateway and Amazon EBS and cannot be directly accessed using Amazon S3 APIs.

Q: What are the snapshot limits per gateway?

A: There are no limits to the number of snapshots or the amount of snapshot data a single gateway can produce.

Q: What are the benefits of using AWS Backup to protect my Volume Gateway volumes?

A: Using AWS Backup to backup Volume Gateway volumes simplifies and centralizes backup management, thus reducing operational burden and making it easier to meet compliance requirements across all your AWS resources. AWS Backup allows you to set customizable scheduled backup policies that meet your backup requirements. Using AWS Backup, you can set backup retention and expiration rules so you no longer need to develop custom scripts or manually manage the point-in-time backups of your Volume Gateway volumes. Finally, you can manage and monitor backups across multiple Volume Gateways, and other AWS resources such as EBS volumes and RDS databases, from a central view.

Q: How do I protect volumes on Volume Gateway using AWS Backup?

A: You can use AWS Backup to either take a one-time backup or define a backup schedule for Volume Gateway volumes. The volume backups are stored in Amazon S3 as Amazon EBS snapshots and visible in AWS Backup console or Amazon EBS console. The volume backups created by AWS Backup can manually or automatically be deleted from AWS Backup console.

Q: How do I use AWS Backup to manage backup and retention of my Volume Gateway volumes?

A: You can start from either the Storage Gateway console or the AWS Backup console to manage your backups. If you start from the Storage Gateway console, you have the ability to navigate to the AWS Backup console to complete your backup plan configuration or initiate an on-demand backup. Alternatively, you can start from the AWS Backup console to configure your backup plan or initiate an on-demand backup of Volume Gateway volumes.

Q: Does anything change with how I have been using Volume Gateway volumes today?

A: No. All existing Volume Gateway snapshot functionality and your existing Amazon EBS Snapshots remain available and unchanged. You can continue to use the Storage Gateway console to create volumes from your EBS Snapshots and use the Amazon EBS console to view or delete your snapshots.

Q: If I use AWS Backup, can I also continue to use Volume Gateway snapshot schedules and existing snapshots?

A: Yes. You can continue to use Volume Gateway’s existing snapshot capabilities to create Amazon EBS snapshots and use your previously created snapshots for restore purposes. AWS Backup’s backup schedule operates independently from the Volume Gateway scheduled snapshots, and provides you an additional way to centrally manage all your backup and retention policies.

Q: If I have a KMS-encrypted volume on Volume Gateway, will AWS Backup be able to backup that volume?

A: Yes. AWS Backup will backup KMS-encrypted volumes Volume Gateway with the same key as the one used for volume encryption.

Q: Can I use AWS Backup to create a backup of my Volume Gateway volume in a different region (e.g. cross region)?

A: AWS Backup supports backup of Volume Gateway volumes within the same region in which AWS Backup operates.

High Availability on VMware

Q: How does Storage Gateway provide high availability?

A: Storage Gateway achieves high availability by running a series of continuous health-checks against the operation of the gateway that connect to the VMware monitoring service. During a hardware, software, or network failure, VMware will trigger a gateway restart on a new host or on its existing host if the host is still operational. At a maximum, users and applications will experience up to 60 seconds of downtime during a restart. After a restart, connections to the gateway are automatically re-established, never needing manual intervention. On re-initialization the gateway will send metrics back to cloud to give customers a full view of the availability event.

Q: What environments are enabled for Storage Gateway high availability?

A: Storage Gateway high availability can currently be enabled in clustered VMware vSphere environments that have VMware HA enabled and have shared volume storage available.

Q: What does Storage Gateway with high availability cost?

A: There is no additional cost for running Storage Gateway with the high availability integration enabled.

Q: What types of failures are covered by Storage Gateway with high availability?

A: Storage Gateway with VMware HA enabled and application monitoring configured, will detect and recover from hardware failures, hypervisor failures, network failures, as well as software issues that lead to connection timeouts or file-share, volume, or virtual tape library unavailability.

Q: Will NFS and SMB sessions be maintained during a gateway restart?

A: Yes.

Q: Will gateway reads or writes fail during a gateway restart?

A: NFS clients connecting to File Gateways may hang for up to 60 seconds on a read or write operation while the gateway restarts and then will retry, given customers use the recommended mount settings. SMB clients may reject a file read or write during a restart depending on client settings. All iSCSI reads and writes for Volume Gateway and Tape Gateway will hang during a gateway restart and then automatically retry.

Q: Will Storage Gateway HA still have the ability to restart if its connection to AWS is broken?

A: Yes, gateways will be reinitialized using the same underlying shared storage, preserving local cache and upload queues

Q: Will I lose data during a gateway restart?

A: No, gateways will be reinitialized using the same underlying shared storage, preserving local cache and upload queues.

Q: Do I need to make any changes to my VMware environment to take advantage of the HA feature?

A: If the gateway is deployed to VMware with VMware HA enabled you will be able to configure the restart sensitivity of the Storage Gateway VM in the VMware vSphere control center. The Storage Gateway VM heartbeat will be available giving you the ability automatically restart the gateway on a specific timeout.

Q: What does Storage Gateway HA give me that I don't already have if I operate VMware HA?

A: VMware HA monitors underlying infrastructure, such as storage and networking. Storage Gateway provides a range of health checks such as file system availability, SMB endpoint availability, and NFS endpoint availability that monitor all of the critical operations of the gateway, ensuring the whole service and not just the underlying infrastructure is continuously available to your users and applications.

Q: Will this be available for VMware Cloud on AWS?

A: Yes. Storage Gateway High Availability can be used on VMware Cloud with no additional requirements. VMware Cloud on AWS has VMware HA enabled by default and shared volumes are available.

Q: How will I know if a gateway is capable of high availability and operating in HA-mode?

A: When setting up a new gateway for VMware, you will be given the option of testing HA. You may also test whether a deployed gateway is HA-capable by choosing the “Test VMware HA” action in console.

Q: What operational visibility will I have during a gateway restart?

A: AWS Storage Gateway console will show availability events in log tables and interruptions in performance graphs during a gateway restart.

Q: Will I see an availability event in CloudWatch when a gateway restart occurs?

A: Yes, if you have configured the integration with CloudWatch, availability events triggered from the gateway will be available through CloudWatch.

Q: How will I know when a gateway returns to operation?

A: If you have configured the integration with CloudWatch, a CloudWatch event will be triggered on re-initialization. Additionally, performance graphs will show the gateway’s operational metrics including number of active sessions.

Q: Will I be able to set a service timeout that triggers a gateway restart?

A: Yes, administrators will be able to set a timeout in the vSphere console that will restart the service if the gateway is unreachable for the specified number of seconds.

AWS PrivateLink support

Q: Does Storage Gateway support AWS PrivateLink for all types of gateways?

A: Yes, the service supports PrivateLink for all gateway types (File/Volume/Tape).

Q: What is the cost for using VPC endpoints with Storage Gateway?

A: You will be billed for each hour that your VPC endpoint remains provisioned. Data processing charges also apply for each Gigabyte processed through the VPC endpoint regardless of the traffic’s source or destination.

Q: How do I activate gateways that are connected to AWS via AWS PrivateLink?

A: PrivateLink enabled gateways can be activated through the AWS Console if your web browser has access to both the internet and your private network, or via the CLI in the region that they are based.

Q: How can I use PrivateLink with File Gateway?

A: To use File Gateway on-premises with PrivateLink and private virtual interfaces (VIFs) to access your Amazon S3 buckets, you will need to setup an Amazon EC2 based proxy server. In order to access Amazon S3 over a private network, you need to use S3's gateway endpoints, and these endpoints are not directly accessibly from on-premises environments. The proxy server will provide access through the VPC endpoint for S3, making it accessible to an on-premises File Gateway. We recommend using an EC2 instance family that is optimized for network bandwidth.

Q: Can a File Gateway use a VPC endpoint in one region and access an S3 bucket in another region?

A: No.

Q: How can I use PrivateLink with Volume Gateways and Tape Gateways?

A: Volume and Tape Gateways connect directly to AWS services through the Storage Gateway VPC endpoint without the need for a proxy to S3.

Q: Can I use AWS PrivateLink with my Storage Gateway hardware appliance from Dell EMC?

A: Yes, but the appliance must be activated before it is moved to the private network.

Hardware Appliance

Q: What is the Storage Gateway hardware appliance?

A: AWS Storage Gateway is available as a hardware appliance, which has Storage Gateway software pre-installed on a Dell EMC PowerEdge R640 server with a validated configuration. You manage the appliance from the AWS Management Console or API.

Q: What gateway types and storage interfaces are supported on the hardware appliance?

A: The hardware appliance supports File Gateway with NFS and SMB interfaces, Volume Gateway cached volumes with iSCSI, and Tape Gateway with iSCSI-VTL.

Q: Why might I need a hardware appliance?

A: The hardware appliance further simplifies procurement, deployment, and management of AWS Storage Gateway on-premises for IT environments such as remote offices and departments which lack existing virtual server infrastructure, adequate disk and memory resources, or staff with hypervisor management skills. It avoids having to procure additional infrastructure necessary for a virtual environment in order to operate the local Storage Gateway VM appliance.

Q: How many models of hardware appliances are available?

A: We are offering a single model at this time.

Q: What are the specifications of the hardware appliance?

A: The hardware appliance is based on a Dell EMC PowerEdge R640 server. Please refer to Storage Gateway hardware appliance for specifications.

Q: Where is the hardware appliance available? With which AWS regions does it work?

A: The hardware appliance can be shipped to US and Europe addresses. It can be used with the US East (Northern Virginia), US East (Ohio), US West (Oregon), US West (Northern California), EU (Frankfurt), EU (Ireland), EU (London), and EU (Paris) AWS regions. File Gateway file shares may be added against any global partition Amazon S3 bucket.

Q: Where do I buy the hardware appliance?

A: We recommend you start at AWS Storage Gateway management console to purchase the hardware appliance. In the Storage Gateway console when you create a new gateway, you have the option to use a hardware appliance in addition to using virtual platforms VMware ESXi, Microsoft Hyper-V, Linux KVM, and Amazon EC2. If you don't already own a hardware appliance, you are directed to Amazon US, Amazon UK, or Amazon Germany to purchase it. You can choose to purchase the appliance from these websites based on your delivery location preference using your Amazon or Amazon Business account.

Q: Who owns the hardware appliance?

A: After purchase, you own the hardware appliance.

Q: How do I use the hardware appliance?

A: Once you receive the hardware appliance, you configure IP address through the local hardware console, and use this IP address in the AWS Storage Gateway management console to activate your appliance. This associates your hardware appliance with your AWS account. Once the hardware appliance is activated, you select your desired gateway type from the console, either file, volume (cached), or tape. The selected type of gateway is then enabled on the appliance. Once activated, you manage and use your new hardware gateway appliance with the AWS Console, CLI, or SDK, similar to how you would with the virtual appliance today. For more information, please see the hardware appliance documentation.

Q: Can I run multiple gateways on a single hardware appliance?

A: No. Currently, the hardware appliance supports running only one gateway at a time.

Q: Can I change the type of gateway once it is installed on a hardware appliance?

A: Yes. To change the gateway type after it is installed on a hardware appliance, you choose *Remove Gateway* from the Storage Gateway console, which deletes the gateway and all associated resources. At that point, you are free to launch a new gateway on the hardware appliance.

Q: How can I purchase and use additional storage on the Storage Gateway hardware appliance?

A: You will have the option to purchase additional storage on Amazon US, Amazon UK, or Amazon Germany. You can purchase the base appliance which offers 5 TB usable storage and choose to add a package of 5 x 1.92 TB SSDs to bring the appliance’s usable storage capacity to 12 TB.

To expand your storage, simply insert the SSDs into the pre-configured appliance. The SSDs are hot pluggable, and the appliance will automatically recognize the extra storage upon adding SSDs to the appliance. See the documentation here for instructions.

Q: How can I purchase a fiber optic network card for the Storage Gateway hardware appliance?

A: You will have the option to purchase an Intel X710 4-port 10 Gigabit fiber optic network card on Amazon US, Amazon UK, or Amazon Germany for Storage Gateway hardware appliance. You can select the fiber optic network card option in addition to the base appliance when ordering the appliance. Upon receipt of the appliance and the fiber optic network card, you will swap out the 10 Gigabit copper network card with the fiber optic network card using instructions here.

Q: Can I add more storage to a Storage Gateway hardware appliance I had purchased before?

A: Yes. You can buy 5 x 1.92 TB SSDs available on Amazon US, Amazon UK, or Amazon Germany and add them to the appliance. If you have already activated the appliance and associated it with your AWS account, you will need to factory reset it before adding more storage.

Q: Can I add any SSD or hard drive to increase storage capacity for my Storage Gateway hardware appliance?

A: No. At this time you can only add the SSDs that are available for purchase on Amazon US, Amazon UK, or Amazon Germany. The SSDs available for purchase are qualified by Dell and AWS for the Storage Gateway hardware appliance.

Q: Does the Storage Gateway hardware appliance support RAID?

A: Yes. The hardware appliance uses software-based ZFS RAID and provides protection against storage drive failure. The base appliance offering 5 TB usable storage tolerates failure of 1 SSD and the 12 TB usable storage configuration tolerates failure of 2 SSDs.

Performance, monitoring, and maintenance

Q: What performance can I expect?

A: The AWS Storage Gateway sits between your applications and Amazon storage services. The performance you experience depends on what host platform (hardware appliance, virtual machine, Amazon EC2 instance) you are using to run Storage Gateway software and a number of other factors. These include the network bandwidth between your iSCSI initiator or NFS client and gateway, the speed and configuration of your underlying local disks, the configuration of your VM, the amount of local storage allocated to your gateway, and the bandwidth between your gateway and Amazon storage.

Our technical documentation provides guidance on how to optimize your AWS Storage Gateway environment for best performance.

Q: What are the minimum hardware and software requirements for the AWS Storage Gateway?

A: For running AWS Storage Gateway on a virtual machine or an Amazon EC2 instance, see the requirements section in the AWS Storage Gateway User Guide. AWS Storage Gateway is also available as a Hardware Appliance with pre-validated specifications.

Q: Can I use the AWS Storage Gateway with AWS Direct Connect?

A: Yes, you can use AWS Direct Connect to increase throughput and reduce your network costs by establishing a dedicated network connection between your on-premises gateway and AWS. Note that the AWS Storage Gateway efficiently uses your Internet bandwidth to help speed up the upload of your on-premises application data to AWS.

Q: Can I route my AWS Storage Gateway Internet traffic through a local proxy server?

A: Yes. Volume and tape gateways support configuration of a Socket Secure version 5 (SOCKS5) proxy between your on-premises gateway and AWS. File gateways support configuration of an HyperText Transfer Protocol (HTTP) proxy.

Q: What type of data reduction does AWS Storage Gateway perform?

A: Volume and tape gateways perform compression of data in-transit and at-rest which can reduce both data transfer and storage charges. The AWS Storage Gateway only uploads data that has changed, minimizing the amount of data sent over the Internet.

Q: Does the AWS Storage Gateway support bandwidth throttling?

A: Yes, using the AWS Management Console you can restrict the bandwidth between your tape and volume gateway and AWS based on a rate that you provide. You can specify individual rates for inbound and outbound traffic.

Q: How do I monitor my gateway?

A: You can use Amazon CloudWatch to monitor the performance metrics for your gateway, giving you insight into storage, bandwidth, throughput, and latency. These metrics are accessible directly from CloudWatch; or by following links in the AWS Storage Gateway Console, which take you directly to the CloudWatch metrics for the resource being viewed. Please refer to the CloudWatch details and pricing pages for additional information.

Q: How can I measure the cache performance of my gateway?

A: You can use Amazon CloudWatch metrics including CachePercentDirty, CacheHitPercent, CacheFree, and CachePercentUsed. These can be viewed by following the Monitoring link on the gateway details tab in the AWS Storage Gateway Console.

Q: How can I measure the bandwidth used by my gateway?

A: You can use Amazon CloudWatch metrics including CloudBytesUploaded and CloudBytesDownloaded.

Q: How does the AWS Storage Gateway manage updates?

A: AWS Storage Gateway periodically deploys important updates and software patches to your gateway virtual machine (VM). You can configure a weekly maintenance schedule allowing you to control when these updates will be applied to your gateway. Alternatively, you can apply updates manually when they are made available, either through the AWS Storage Gateway Console or API. Updates should take only a few minutes to complete. For more information, please visit the Managing Gateway Updates section of our documentation.

Billing

Q: How will I be billed for my use of AWS Storage Gateway?

A: There are 3 elements to how you will be billed for AWS Storage Gateway: Storage, requests, and data transfer. For detailed pricing information, please visit the AWS Storage Gateway Pricing page.

Q: How will I be charged for file storage when using a file gateway?

A: File gateways stores data directly in Amazon S3. You are billed by Amazon S3 for the objects stored and requests made by your file gateway. For more information, please visit the Amazon S3 Pricing page.

Q: How will I be charged for volume or virtual tape storage when using a volume or tape gateway?

A: You are billed for the amount of volume and virtual tape data you store in AWS. This fee is prorated daily and prices vary by region. You are only billed for the portion of volume or virtual tape capacity that you use, not for the provisioned size of the resource. All volume and virtual tape data is compressed before it is transferred to AWS by the gateway, which can reduce your storage charges. For detailed pricing information, please visit the AWS Storage Gateway Pricing page.

Q: How will I be charged for EBS snapshots taken from my AWS Storage Gateway volumes?

A: EBS snapshots taken from your Storage Gateway volumes are stored and billed by Amazon EBS. When taking a new snapshot only the data that has changed since your last snapshot is stored to reduce your storage charges. For more information, please visit the Amazon EBS Pricing page.

Q: How will I be charged for reading and writing data?

A: When your gateway writes data to AWS you will be charged at a flat rate of $0.01 per GB of data written to AWS up to a monthly maximum of no more than $125 per gateway. There is no charge for reading data from AWS. Since the gateway performs caching, bandwidth optimization, and, for volume and tape gateways, compression, the amount of data written to AWS may be less than the amount of data written to the gateway by your application. You can monitor the amount of data written by your gateway to AWS through the provided Amazon CloudWatch metrics and you can configure bandwidth limits on your gateway to manage your costs.

Q: How will I be charged when retrieving data on an archived virtual tape?

A: You are charged, when retrieving a virtual tape that has been archived in S3 Glacier, at a flat rate of $0.01 per GB of data stored on the tape. For example, retrieving 5 tapes that contain 100 GB each would cost 5 x 100GB x $0.01 = $5.00.

Q: How will I be charged for deleting an archived virtual tape?

A: If a virtual tape is deleted within three months of being archived in S3 Glacier or within six months of being archived S3 Glacier Deep Archive, you will be charged an early deletion fee. If the virtual tape has been stored for three months or longer in S3 Glacier or for six months or longer in S3 Glacier Deep Archive, there is no charge for deletion.

In the US East (Northern Virginia) Region, you would be charged a prorated early deletion fee of $0.012 per GB deleted within three months. For example, if you delete 1 virtual tape containing 1 GB of data 1 month after uploading archiving it in S3 Glacier, you would be charged a $0.008 early deletion fee. If, instead you delete the same virtual tape after 2 months, you would be charged a $0.004 early deletion fee.

Q: How am I charged for virtual tapes I store in S3 Glacier Deep Archive?

A: Virtual tapes stored in S3 Glacier Deep Archive will be charged S3 Glacier Deep Archive storage class rate. You can visit Storage Gateway pricing webpage to review Tape Gateway pricing.

Q: How will the virtual tapes I store in Deep Archive Pool, associated with S3 Glacier Deep Archive storage class, show up on my AWS bill and in the AWS Cost Management tool?

A: The usage and cost for virtual tapes you store in Deep Archive Pool will show up as an independent service line item on your monthly AWS bill under AWS Storage Gateway Deep Archive, separate from your AWS Storage Gateway and costs. However, if you are using the AWS Cost Management tool, usage and cost for virtual tapes you store in Deep Archive Pool will be included under AWS Storage Gateway in your detailed monthly spend reports, and not broken out as a separate service line item.

Q: How will I be charged for moving a virtual tape archived in S3 Glacier to S3 Glacier Deep Archive?

A: For AWS US East (N. Virginia) region, you are charged, when moving a virtual tape that has been archived in S3 Glacier to S3 Glacier Deep Archive, at a rate of $0.032 per GB of data stored on the tape. For example, moving a 100 GB tape archived in S3 Glacier to S3 Glacier Deep Archive will cost 100 GB x $0.032/GB = $3.2. If you move a tape that's archived for less than 90 days in S3 Glacier to S3 Glacier Deep Archive, you are also charged for early deletion fee for tape storage in S3 Glacier.

Q: How will I be charged for network data transfer to and from AWS when using AWS Storage Gateway?

A: You are billed for Internet data transfer for each GB downloaded from AWS to your gateway. All data transfer for uploading to AWS is free.

Q: How can I tell how much storage I am going to be billed for?

A: The Billing and Cost Management console shows an estimate of month-to-date usage for each service, including AWS Storage Gateway volumes and virtual tapes. For a breakdown of usage by individual volume or virtual tape Detailed Billing Reports enables you to see usage for each resource on a daily basis.

Q: When using file gateway, will I incur S3 request charges?

A: You will pay for the S3 requests made by file gateway on your behalf to store and retrieve your files in S3 as objects. The gateway caches data up to the capacity of the local disks you allocate, which can help reduce costs for data retrieval.

Q: When does each monthly billing cycle begin?

A: The billing system follows Coordinated Universal Time (UTC). The calendar month begins midnight UTC on the first day of every month.

Q: Do your prices include taxes?

A: Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of the Asia Pacific (Tokyo) Region is subject to Japanese Consumption Tax.

Q: How much does the hardware appliance cost?

A: Please refer to the hardware appliance listing on Amazon US, Amazon UK, or Amazon Germany for the current price.

Q: How do I pay for the hardware appliance?

A: You purchase the hardware appliance from Amazon US, Amazon UK, or Amazon Germany using an Amazon account, or an Amazon Business account with a purchase order.

Q: Can I lease or rent the hardware appliance?

A: No. You pay the full price at the time of purchase.

Support

Q: Does AWS Premium Support cover the AWS Storage Gateway?

A: Yes, AWS Premium Support covers issues related to your use of the AWS Storage Gateway. Please see the AWS Premium Support detail page for further information and pricing.

Q: What other support options are available?

A: You can tap into the breadth of existing AWS community knowledge through the AWS Storage Gateway discussion forum.

Q: Who do I call for support related to hardware appliance?

A: You contact AWS Support, who provides AWS Storage Gateway software and service support. AWS Support also coordinates and hands over any cases related to Dell EMC hardware to a fully trained Dell EMC support team. We recommend that you purchase AWS Premium Support.

Q: Where do I find the Dell EMC service tag for the hardware appliance (also known as serial number)?

A: The Dell EMC service tag for the hardware appliance can be found in the Hardware view of AWS Storage Gateway console.

Q: What if there is a hardware problem with the hardware appliance?

A: AWS Support works with Dell EMC for hardware support. Hardware support is included with your appliance purchase and includes 36 months of 7x24 phone support and next-business-day, on-site service for parts replacement.

Q: What are the warranty terms of the hardware appliance?

A: The hardware appliance comes with 3 years of warranty from Dell with next business day onsite service for parts replacement. You can find warranty information under *Product description* section of the hardware appliance listing on Amazon US, Amazon UK, or Amazon Germany.