**What is AWS (Amazon Web Services)? Did got chance to work on Amazon tools?**

AWS provides a set of flexible services designed to enable companies to create and deliver products with greater speed and reliability using AWS and DevOps practices. These services simplify commissioning and infrastructure management, application code deployment, automated software release process and monitoring of the application and infrastructure performance. Amazon used tools like AWS CodeCommit, AWS CodeDeploy, AWS CodePipeline etc. that helps to make DevOps easier.

**Name the several layers of Cloud Computing.**

Here is the list of layers of the cloud computing

* SaaS: Software as a Service (SaaS), it provides users access directly to the cloud application without installing anything on the system. Google Apps, Salesforce, Workday, Concur, Citrix GoToMeeting, Cisco WebEx.
* IaaS: Infrastructure as a service, it provides the infrastructure in terms of hardware like memory, processor speed etc. Rackspace, Amazon Web Services (AWS), Cisco Metapod, Microsoft Azure, Google Compute Engine (GCE)
* PaaS: Platform as a service, it provides cloud application platform for the developers, AWS Elastic Beanstalk,Windows Azure, Heroku, Force.com, Google App Engine, Apache Stratos.

**What are the advantages of using cloud computing?**

The advantages of using cloud computing are

* Data backup and storage of data
* Powerful server capabilities
* SaaS (Software as a service)
* Information technology sandboxing capabilities
* Increase in productivity
* Cost effective & Time saving

**List down the three basic clouds in cloud computing?**

* Professional cloud
* Personal cloud
* Performance cloud

**As an infrastructure as a service what are the resources that are provided by it?**

IAAS (Infrastructure as a Service) provides virtual and physical resources that are used to build a cloud. It deals with the complexities of deploying and maintaining of the services provided by this layer. Here the infrastructure is the servers, storage and other hardware systems.

**What are the business benefits involved in cloud architecture?**

The benefits involved in cloud architecture is

* Zero infrastructure investment
* Just in time infrastructure.
* More efficient resource utilization

**What are the of cloud architecture that separates it from traditional one?**

The characteristics that makes cloud architecture above traditional architecture is

* According to the demand cloud architecture provides the hardware requirement
* Cloud architecture can scale the resource on demand
* Cloud architecture can manage and handling dynamic workloads without failure

**Mention platforms which are used for large scale cloud computing?**

The platforms that are used for large scale cloud computing are

* Apache Hadoop
* MapReduce

**What are the security aspects provided with cloud?**

* Identity management: It authorizes the application services
* Access control: permission has to be provided to the users so that they can control the access of another user who is entering into the cloud environment.
* Authentication and Authorization: Allows only the authorized and authenticated user only to access the data and applications

**What are the different deployment models for cloud computing?**

The different models are:

* Private Cloud
* Public Cloud
* Hybrid Cloud
* Community Cloud

**List out different layers which define cloud architecture?**

There are five layers:

* Cloud Controller (CLC)
* Walrus
* Cluster Controller
* Storage Controller (SC)
* Node Controller (NC)

**Before going for cloud computing platform what are the essential things to be taken in concern by users?**

* Compliance
* Loss of data
* Data storage
* Business continuity
* Uptime
* Data integrity in cloud computing

**What are the security laws which are implemented to secure data in a cloud?**

The security laws which are implemented to secure data in cloud are

* Processing: Control the data that is being processed correctly and completely in an application
* File: It manages and control the data being manipulated in any of the file
* Output reconciliation:  It controls the data which has to be reconciled from input to output
* Input Validation: Control the input data
* Security and Backup: It provides security and backup it also controls the security breaches logs

**Explain the difference between cloud and traditional datacenters?**

* The cost of the traditional data center is higher due to heating and hardware/software issues
* Cloud gets scaled when the demand increases.  Majority of the expenses are spent on the maintenance of the data centers, while that is not the case with cloud computing

**Explain what are the different modes of software as a service (SaaS)?**

* Simple multi-tenancy:  In this each user has independent resources and are different from other users, it is an efficient mode.
* Fine grain multi-tenancy:  In this type, the resources can be shared by many but the functionality remains the same.

**How important is the platform as a service?**

Platform as a service or PAAS is an important layer in cloud computing.  It provides application platform for providers.  It is responsible for providing complete virtualization of the infrastructure layer and makes it work like a single server.

**What is a cloud service?**

Cloud service is used to build cloud applications using the server in a network through internet.  It provides the facility of using the cloud application without installing it on the computer. It also reduces the maintenance and support of the application which are developed using cloud service.

**What are system integrators in Cloud Computing?**

In Cloud Computing, systems integrator provides the strategy of the complicated process used to design a cloud platform. Integrator allows to create more accurate hybrid and private cloud network, as integrators have all the knowledge about the data center creation.

**What is “EUCALYPTUS” stands for?**

“EUCALYPTUS” stands for Elastic Utility Computing Architecture for Linking Your Programs To Useful Systems”

**Explain what is the use of “EUCALYPTUS” in cloud computing?**

“Eucalyptus” is an open source software infrastructure in cloud computing, which is used to implement clusters in cloud computing platform. It is used to build public, hybrid and private clouds. It can produce your own data center into a private cloud and allows you to use its functionality to many other organizations.

**What are the security laws which are implemented to secure data in a cloud?**

The security laws which are implemented to secure data in cloud are:

* Processing
* File
* Output reconciliation
* Input Validation
* Security and Backup

**Why API’s have in cloud services?**

Application Programming Interface (API) has the following uses:

* It eliminates the need to write fully fledged programs
* It provides the instructions to set up communication between one or more applications
* It allows easy creation of applications and links the cloud services with other systems

**How many data centers are deployed for cloud computing? What are they?**

There are two data centers in cloud computing:

* Containerized Data centers
* Low Density Data centers

**What is the difference in cloud computing and computing for mobiles?**

Mobile computing uses the same concept as cloud computing.   Cloud computing becomes active with the data with the help of internet rather than individual device. It provides users with the data which they have to retrieve on demand.  In mobile, the applications run on the remote server and gives user the access for storage and manage.

**What you know about Serverless model?**

* Serverless refers to a model where the existence of servers is hidden from developers. It means you no longer have to deal with capacity, deployments, scaling and fault tolerance and OS. It will essentially reduce maintenance efforts and allow developers to quickly focus on developing codes.
* Examples are Amazon AWS Lambda and Auth0 Serverless platform.

**What is AWS Certificate Manager?**



AWS Certificate Manager (ACM) handles the complexity of provisioning, deploying, and managing certificates provided by ACM (ACM Certificates) for your AWS-based websites and applications. You use ACM to request and manage the certificate and then use other AWS services to provision the ACM Certificate for your website or application. As shown by the following illustration, ACM Certificates are currently available for use with only Elastic Load Balancing and Amazon CloudFront. You cannot use ACM Certificates outside of AWS.

**What is IAM service?**

AWS Identity and Access Management (IAM) is a web service that helps you securely control access to AWS resources for your users. You use IAM to control who can use your AWS resources (authentication) and what resources they can use and in what ways (authorization).

**What is Redshift?**

Redshift is a fast, fully managed, petabyte-scale data warehouse service that makes it simple and cost-effective to efficiently analyze all your data using your existing business intelligence tools.

**What is VPC?**

A virtual private cloud (VPC) is a virtual network dedicated to your AWS account. You can configure or create your VPC as per requirement like select region, create subnets (IP- CIDR), configure route tables, security groups, Internet gateway etc. to your AWS account by which you can launch your AWS resources, such as Amazon EC2, RDS instances etc. into your VPC. So, basically you can say that Amazon VPC is the networking layer for AWS Infrastructure.

**What is VPC peering?**

A VPC peering connection is a networking connection between two VPCs that enables you to route traffic between them using private IP addresses. And instances which is in VPC can communicate with each other as if they are within the same network.

You can create a VPC peering connection between your own VPCs, or with a VPC in another AWS account within a single region.

If you have more than one AWS account within a same region and wants to share or transfer the data, you can peer the VPCs across those accounts to create a file sharing network. You can also use a VPC peering connection to allow other VPCs to access resources you have in one of your VPCs.

A VPC peering connection can help you to facilitate the transfer of data.

**What is VPC endpoints?**

A VPC endpoint enables you to create a private connection between your VPC with another AWS service without requiring access over the Internet, through a NAT device, a VPN connection, or AWS Direct Connect. They are horizontally scaled, redundant, and highly available VPC components that allow communication between instances in your VPC and AWS services without imposing availability risks or bandwidth constraints on your network traffic.

An endpoint enables instances in your VPC to use their private IP addresses to communicate with resources in other services. Don’t require public IP addresses to your instances, and you don’t need an Internet gateway, a NAT device, or a virtual private gateway in your VPC.

**What are microservices and why they have an impact on operations?**

Microservices are a product of software architecture and programming practices. Microservices architectures typically produce smaller, but more numerous artifacts that Operations is responsible for regularly deploying and managing. For this reason, microservices have an important impact on Operations. The term that describes the responsibilities of deploying microservices is micro deployments. So, what DevOps is really about is bridging the gap between microservices and micro deployments.

**What is an AMI?**

AMI stands for Amazon Machine Image. It is effectively a snapshot of theroot filesystem. AWS AMI provides the information required to launch an instance, which is a virtual server in the cloud. You specify an AMI when you launch an instance, and you can launch as many instances from the AMI as you need. You can also launch instances from as many different AMIs as you need.

An AMI includes the following:

* A template for the root volume for the instance (such as an operating system, an application server, and applications).
* Launch permissions that control which AWS accounts can use the AMI to launch instances.
* A block device mapping that specifies the volumes to attach to the instance when it’s launched
* Build a new AMI by first spinning up and instance from a trusted AMI.  Then adding packages and components as required.  Be wary of putting sensitive data onto an AMI.
* For instance, your access credentials should be added to an instance after spinning up.  With a database, mount an outside volume that holds your MySQL data after spinning up as well.

**What is the relation between instance and AMI?**

An Amazon Machine Image (AMI) is a template that contains a software configuration (for example, an operating system, an application server, and applications). From an AMI, you launch an instance, which is a copy of the AMI running as a virtual server in the cloud.

You can launch several types of instances from a single AMI. An instance type determines the hardware of the host computer used for your instance. Each instance type offers different compute and memory capabilities.

**What is instance profile? How do you create a role?**

It’s a container for IAM role and you can pass this role information to a EC2-instance when the instance starts. This role gives access permissions to S3 buckets and other repositories where your applications are stored.

* We can create IAM instance profile by using AWS CLI
* Command to create instance profile is ---
* Aws iam create -instance-profile - -instance-profile-name

**What are the key components of AWS (Amazon Web Service)?**  
The key components of AWS are:

* **Route 53:** A DNS web service
* **Simple E-mail Service:** It allows sending e-mail using RESTFUL API call or via regular SMTP
* **Identity and Access Management:** It provides enhanced security and identity management for your AWS account
* **Simple Storage Device or (S3):** It is a storage device and the most widely used AWS service
* **Elastic Compute Cloud (EC2):** It provides on-demand computing resources for hosting applications. It is very useful in case of unpredictable workloads
* **Elastic Block Store (EBS):** It provides persistent storage volumes that attach to EC2 to allow you to persist data past the lifespan of a single EC2
* **CloudWatch:** To monitor AWS resources. It allows administrators to view and collect key Also, one can set a notification alarm in case of trouble.

**What is AMI (Amazon Machine Image)?**

It’s a template that provides the information (an operating system, an application server and applications) required to launch an instance, which is a copy of the AMI running as a virtual server in the cloud. You can launch instances from as many different AMIs as you need.

**What is S3?**

S3 stands for Simple Storage Service. You can use S3 interface to store and retrieve any amount of data, at any time and from anywhere on the web. Also, we can host a website in Amazon S3. most of the companies storing the documents, images and other files to S3. For S3, the payment model is “pay as you go”.

**What Is Amazon EC2?**

Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) cloud. Using Amazon EC2 eliminates your need to invest in hardware up front, so you can develop and deploy applications faster. You can use Amazon EC2 to launch as many or as few virtual servers as you need, configure security and networking, and manage storage. Amazon EC2 enables you to scale up or down to handle changes in requirements or spikes in popularity, reducing your need to forecast traffic.

**What Is Amazon EC2 instance?**

An EC2 instance is a virtual server in Amazon’s Elastic Compute Cloud (EC2) for running applications on the Amazon Web Services (AWS) infrastructure.

**Explain some features of Amazon EC2?**

* Virtual computing environments, known as instances
* Preconfigured templates for your instances, known as Amazon Machine Images (AMIs), that package the bits you need for your server (including the operating system and additional software)
* Various configurations of CPU, memory, storage, and networking capacity for your instances, known as instance types
* Secure login information for your instances using key pairs (AWS stores the public key, and you store the private key in a secure place)
* Storage volumes for temporary data that’s deleted when you stop or terminate your instance, known as instance store volumes
* Persistent storage volumes for your data using Amazon Elastic Block Store (Amazon EBS), known as Amazon EBS volumes
* Multiple physical locations for your resources, such as instances and Amazon EBS volumes, known as regions and Availability Zones
* A firewall that enables you to specify the protocols, ports, and source IP ranges that can reach your instances using security groups
* Static IP addresses for dynamic cloud computing, known as Elastic IP addresses

**How you will find out the instance id from within an ec2 machine?**

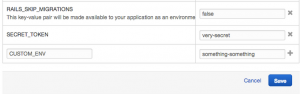
* wget -q -O – http://instance-data/latest/meta-data/instance-id
* If you need programmatic access to the instance ID from within a script  
  die() { status=$1; shift; echo “FATAL: $\*”; exit $status; }  
  EC2\_INSTANCE\_id=”`wget -q -O – http://instance-data/latest/meta-data/instance-id || die \”wget instance-id has failed: $?\”`”

**For a transport in cloud how you can secure your data?**

To secure your data while transporting them from one place to another, check that there is no leak with the encryption key implemented with the data you are sending.

**How do you pass custom environment variable on Amazon Elastic Beanstalk (AWS EBS)?**  
As a heads up to anyone who uses the .ebextensions/\*.config way: nowadays you can add, edit and remove environment variables in the Elastic Beanstalk web interface.

The variables are under Configuration? Software Configuration:



**What is DynamoDB?**

Amazon DynamoDB is a fully managed NoSQL database service that provides fast and predictable performance with seamless scalability. You can use Amazon DynamoDB to create a database table that can store and retrieve any amount of data, and serve any level of request traffic. Amazon DynamoDB automatically spreads the data and traffic for the table over a sufficient number of servers to handle the request capacity specified by the customer and the amount of data stored, while maintaining consistent and fast performance.

**What is ElastiCache?**

ElastiCache is a web service that makes it easy to set up, manage, and scale distributed in-memory cache environments in the cloud.

**What is Amazon Kinesis Firehose?**

Amazon Kinesis Firehose is a fully managed service for delivering real-time streaming data to destinations such as Amazon Simple Storage Service (Amazon S3) and Amazon Redshift.

**What Is Amazon CloudSearch and its features?**

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

You can use Amazon CloudSearch to index and search both structured data and plain text. Amazon CloudSearch features:

* Full text search with language-specific text processing
* Boolean search
* Prefix searches
* Range searches
* Term boosting
* Faceting
* Highlighting
* Autocomplete Suggestions

**What is the AWS Key Management Service?**

The AWS Key Management Service (AWS KMS) is a managed service that makes it easy for you to create and control the encryption keys used to encrypt your data.

**The following components of AWS Data Pipeline work together to manage your data:**

A pipeline definition specifies the business logic of your data management. For more information, see Pipeline Definition File Syntax.

A pipeline schedules and runs tasks. You upload your pipeline definition to the pipeline, and then activate the pipeline. You can edit the pipeline definition for a running pipeline and activate the pipeline again for it to take effect. You can deactivate the pipeline, modify a data source, and then activate the pipeline again. When you are finished with your pipeline, you can delete it.

Task Runner polls for tasks and then performs those tasks. For example, Task Runner could copy log files to Amazon S3 and launch Amazon EMR clusters. Task Runner is installed and runs automatically on resources created by your pipeline definitions. You can write a custom task runner application, or you can use the Task Runner application that is provided by AWS Data Pipeline. For more information, see Task Runners.

**What is Regions and Endpoints in AWS?**

* To reduce data latency in your applications, most Amazon Web Services products allow you to select a regional endpoint to make your requests. An endpoint is a URL that is the entry point for a web service. For example, https://dynamodb.us-west-2.amazonaws.com is an entry point for the Amazon DynamoDB service.
* Some services, such as IAM, do not support regions; their endpoints therefore do not include a region. A few services, such as Amazon EC2, let you specify an endpoint that does not include a specific region, for example, https://ec2.amazonaws.com. In that case, AWS routes the endpoint to us-east-1.

**What is AWS WAF? What are the potential benefits of using WAF?**

AWS WAF is a web application firewall that lets you monitor the HTTP and HTTPS requests that are forwarded to Amazon CloudFront and lets you control access to your content. Based on conditions that you specify, such as the IP addresses that requests originate from or the values of query strings, CloudFront responds to requests either with the requested content or with an HTTP 403 status code (Forbidden. You can also configure CloudFront to return a custom error page when a request is blocked.

Benefits of using WAF:

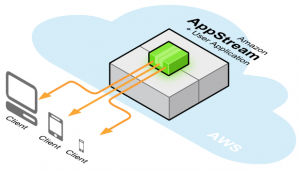
* Additional protection against web attacks using conditions that you specify. You can define conditions by using characteristics of web requests such as the IP address that the requests originate from, the values in headers, strings that appear in the requests, and the presence of malicious SQL code in the request, which is known as SQL injection.
* Rules that you can reuse for multiple web applications
* Real-time metrics and sampled web requests
* Automated administration using the AWS WAF API

**How to find your regions and Availability Zones using the Amazon EC2 CLI?**

Use the ec2-describe-regions command as follows to describe your regions.  
PROMPT> ec2-describe-regions  
REGION us-east-1 ec2.us-east-1.amazonaws.com  
REGION ap-northeast-1 ec2.ap-northeast-1.amazonaws.com  
REGION ap-southeast-1 ec2.ap-southeast-1.amazonaws.com

**What is Amazon AppStream and advantage of using AppStreaming?**

Amazon AppStream is an application streaming service that lets you stream your existing resource-intensive applications from the cloud without code modifications.



**Advantages of Streaming Your Application?**

Interactively streaming your application from the cloud provides several benefits:

* **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.
* **Support Multiple Platforms –** You can write your application once and stream it to multiple device platforms. To support a new device, just write a small client to connect to your streaming application.
* **Fast and Easy Updates –** Because your streaming application is centrally managed by Amazon AppStream, updating your application is as simple as providing a new version of your streaming application to Amazon AppStream. You can immediately upgrade all of your customers without any action on their part.
* **Instant On –** Streaming your application with Amazon AppStream lets your customers start using your application or game immediately, without the delays associated with large file downloads and time-consuming installations.
* **Improve Security –** Unlike traditional boxed software and digital downloads, where your application is available for theft or reverse engineering, Amazon AppStream stores your streaming application binary securely in AWS datacenters.
* **Automatic Scaling –** You can use Amazon AppStream to specify capacity needs, and then the service automatically scales your streamed application and connects customers’ devices to it.

**What are the benefits of EBS vs. instance-store?**

* EBS backed instances can be set so that they cannot be (accidentally) terminated through the API.
* EBS backed instances can be stopped when you’re not using them and resumed when you need them again (like pausing a Virtual PC), at least with my usage patterns saving much more money than I spend on a few dozen GB of EBS storage.
* EBS backed instances don’t lose their instance storage when they crash (not a requirement for all users, but makes recovery much faster)
* You can dynamically resize EBS instance storage.
* You can transfer the EBS instance storage to a brand-new instance (useful if the hardware at Amazon you were running on gets flaky or dies, which does happen from time to time)
* It is faster to launch an EBS backed instance because the image does not have to be fetched from S3.

**Which AWS responsible for managed email and calendaring?**

WorkMail is a managed email and calendaring service with strong security controls and support for existing desktop and mobile email clients. You can access their email, contacts, and calendars wherever you use Microsoft Outlook, your browser, or your iOS and Android mobile devices. You can integrate Amazon WorkMail with your existing corporate directory and control both the keys that encrypt your data and the location where your data is stored.

**What is auto-scaling? How does it work?**

* Horizontally Scaling
* Vertically Scaling

**Auto scaling is a feature of AWS** which allows you to configure and automatically provision and spinning up new instances without the need for your intervention. You can do this by setting thresholds and metrics to monitor. When those thresholds are crossed, a new instance of your choosing will be spun up, configured, and rolled into the load balancer pool. You’ve scaled horizontally without any operator intervention.

**Vertically Scaling:**This is an incredible feature of AWS and cloud virtualization. Spinning up a new larger instance than the one you are currently running.  Pause that instance and detach the root EBS volume from this server and discard. Then stop your live instance, detach its root volume. Note the unique device ID and attach that root volume to your new server. And the start it again. You have scaled vertically in-place.

## **Determining the ELB Option That’s Best for You**

Your mileage will vary depending on your exact situation, of course. But in general, the Classic Load Balancer is likely to be the best choice if your routing and load-balancing needs can all be handled based on IP addresses and TCP ports.

In contrast, the Application Load Balancer can address more complex load-balancing needs by managing traffic at the application level. This is especially advantageous for next-generation infrastructure, such as that based on containers, or if you are building complex web applications in which requests for certain components should be directed to one cluster, while others go to a different one.

**What automation tools can you use to spinning up the servers?**

Here below many types tools given any of the following tools can be used:

* Roll-your-own scripts, and use the AWS API tools. Such scripts could be written in bash, Perl or other language or your choice.
* Use a configuration management and provisioning tool like Ansible, puppet or its successor Opscode Chef etc.
* You might also look towards a tool like Scalr. Lastly you can go with a managed solution such as Rightscale.

**How you would simulate perimeter security using amazon web services model?**

Traditional perimeter security that we’re already familiar with using firewalls and so forth is not supported in the Amazon EC2 world.

AWS supports security groups.  One can create a security group for a jump box with ssh access – only port 22 open. From there a web server group and database group are created.

The web server group allows 80 and 443 from the world, but port 22 only from the jump box group.  Further the database group allows port 3306 from the web server group and port 22 from the jump box group. Add any machines to the web server group and they can all hit the database.

No one from the world can, and no one can directly SSH to any of your boxes.

Want to further lock this configuration down?  Only allow SSH access from specific IP addresses on your network, or allow just your subnet.

**What is S3? What is it used for? Should encryption be used in S3?**

Amazon S3 is stand for Simple storage service that is storage for the Internet. It as a, “simple storage service that offers software developers a highly-scalable, reliable, and low-latency data storage infrastructure at very low costs”.

Amazon S3 provides a simple web service interface which you can use to store and retrieve any amount of data, at any time, from anywhere on the web. Using this web service, developers can easily build applications that make use of Internet storage.

You can think of it like ftp storage, where you can move files to and from there, but not mount it like a file system. AWS automatically puts your snapshots there, as well as AMIs there. Encryption should be considered for sensitive data, as S3 is a proprietary technology developed by Amazon themselves, and as yet unproven vis-a-vis a security standpoint.

Encryption should be considered for sensitive data, as S3 is a proprietary technology developed by Amazon themselves, and yet to be proven from a security standpoint.

**How is buffer used in amazon web services?**

Buffer is used to make the system more resilient to burst of traffic or load by synchronizing different components. The components always receive and process the requests in an unbalanced way. Buffer keeps the balance between different components and makes them work at the same speed to provide faster services.

**What is the function of Amazon Elastic Compute Cloud?**

Amazon Elastic compute cloud also known as Amazon EC2 is an Amazon web service that provides scalable resources and makes the computing easier for developers. The main functions of Amazon EC2 are:

* It provides easy configurable options and allow user to configure the capacity.
* It provides the complete control of computing resources and let the user run the computing environment according to his requirements.
* It provides a fast way to run the instances and quickly book the system hence reducing the overall time.
* It provides scalability to the resources and changes its environment according to the requirement of the user.
* It provides varieties of tools to the developers to build failure resilient applications.

**What is ECS?**

* Amazon EC2 Container Service (ECS) is a highly scalable container management service and high performance that supports the Docker containers and allows you to easily run applications on a cluster managed by Amazon EC2 instances.
* The EC2 service is inseparable from the concept of Amazon Machine Image - AMI. The May is Indeed the image of a virtual machine That Will Be Executed. EC2 based on XEN virtualization, that's why it is quite easy to move XEN servers to EC2.

**What are the security for amazon EC2?**

There are several best practices for secure Amazon EC2. A few of them are given below:

* Use AWS Identity and Access Management (IAM) to control access to your AWS resources.
* Restrict access by only allowing trusted hosts or networks to access ports on your instance.
* Review the rules in your security groups regularly, and ensure that you apply the principle of least
* Privilege – only open up permissions that you require.
* Disable password-based logins for instances launched from your AMI. Passwords can be found or cracked, and are a security risk.

**What are the different components used in AWS?**

The components that are used in AWS are:

* Amazon S3: it is used to retrieve input data sets that are involved in making a cloud architecture and also used to store the output data sets that is the result of the input.
* Amazon SQS: it is used for buffering requests that is received by the controller of the Amazon. It is the component that is used for communication between different controllers.
* Amazon Simple DB: it is used to store intermediate status log and the tasks that are performed by the user
* Amazon EC2: it is used to run a large distributed processing on the Hadoop cluster. It provides automatic parallelization and job scheduling.

**Mention what are the differences between Amazon S3 and EC2?**

**S3:** Amazon S3 is just a storage service, typically used to store large binary files. Amazon also has other storage and database services, like RDS for relational databases and DynamoDB for NoSQL.

**EC2:** An EC2 instance is like a remote computer running Windows or Linux and on which you can install whatever software you want, including a Web server running PHP code and a database server.

**Is it possible to use AWS as a web host? What are the way of using AWS as a web host?**

Yes, it is completely possible to host websites on AWS in 2 ways

* Easy – S3 (Simple Storage Solution) is a bucket storage solution that lets you serve static content e.g. images but has recently been upgraded so you can use it to host flat .html files and your site will get served by a default Apache installation with very little configuration on your part (but also little control).
* Trickier – You can use EC2 (Elastic Compute Cloud) and create a virtual Linux instance then install Apache/Nginx (or whatever) on that to give you complete control over serving whatever/however you want. You use SecurityGroups to enable/disable ports for individual machines or groups of them.

**How step you follow to make 10,000 files as public in S3?**

I will generate a bucket policy which gives access to all the files in the bucket. The bucket policy can be added to a bucket through AWS console.  
{  
“Id”: “…”,  
“Statement”: [ {  
“Sid”: “…”,  
“Action”: [  
“s3:GetObject”  
],  
“Effect”: “Allow”,  
“Resource”: “arn:aws:s3:::bucket/\*”,  
“Principal”: {  
“AWS”: [ “\*” ]  
}  
} ]  
}

**How to delete files recursively from an S3 bucket?**

* aws s3 rm –recursive s3://your\_bucket\_name/foo/
* Or delete everything under the bucket:  
  aws s3 rm –recursive s3://your\_bucket\_name
* If what you want is to actually delete the bucket, there is one-step shortcut:  
  aws s3 rb –force s3://your\_bucket\_name

**Write down the command you will use to copy all files from one S3 bucket to another with s3cmd?**

s3cmd sync s3://from/this/bucket/ s3://to/this/bucket/

**How do you see how much disk space is using by S3 bucket?**

s3cmd can show you this by running s3cmd du, optionally passing the bucket name as an argument.

**How to access/ping a server located on AWS?**

Using UI:  
**In your security group:**

* Click the inbound tab
* Create a custom ICMP rule
* Select echo request
* Use range 0.0.0.0/0 for everyone or lock it down to specific IPs
* Apply the changes and you’ll be able to ping.

**Using cmd: To do this on the command line you can run:**

* ec2-authorize -P icmp -t -1:-1 -s 0.0.0.0/0

**What happens when I reboot an EC2 instance?**

Rebooting an instance is like rebooting a PC. The hard disk isn’t affected. You don’t return to the image’s original state, but the contents of the hard disks are those before the reboot.  
Rebooting isn’t associated with billing. Billing starts when you instantiate an image and stops when you terminate it. Rebooting in between hasn’t any effect.

**What is the difference between Amazon SNS and Amazon SQS?**

* 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—without requiring each component to be concurrently available.

**How you will change the root EBS device of my amazon EC2 instance?**

* Stop the instance.
* Detach the root EBS volume.
* Attach the alternate EBS volume (as the root e.g. /dev/sda1)
* Start the instance.
* This presupposes that your alternate EBS volume is bootable, of course – it has to contain the bootable OS image.

**Can I vertically scale an Amazon instance?  How?**

Yes.  This is an incredible feature of AWS and cloud virtualization.  Spin up a new larger instance than the one you are currently running.  Pause that instance and detach the root EBS volume from this server and discard.  Then stop your live instance, detach its root volume.  Note the unique device ID and attach that root volume to your new server.   And the start it again.  Voila you have scaled vertically in-place

**What is the difference between NACL and SG?**

* NACL are applicable at the subnet level, if any instance in the subnet associated with NACL has to follow the NACL rules whereas security groups are applicable at the instance level.
* In NACL we can set both allow and deny rules for instance, whereas in SG we can only set allow rules for instance, by default everything is denied.
* SG evaluates everything before allowing the traffic. But it’s not the same case while coming to NACL. It has to first check the deny rules and the check the allow rules.

**What is EBS (Elastic Block Storage)?  What type of performance can you expect?  How do you back it up?  How do you improve performance?**

* EBS is a virtualized SAN or storage area network. Elastic Block Store (Amazon EBS) provides persistence block level storage volumes for use with EC2 instances. EBS volumes are highly available and reliable storage volumes that can be attached to any running instance that is in the same Availability Zone.
* **Performance that we can expect:**Performance on EBS can exhibit variability. That is it can go above the SLA performance level, then drop below it. The SLA provides you with an average disk I/O rate you can expect. This can frustrate some folks especially performance experts who expect reliable and consistent disk throughput on a server. Traditional physically hosted servers behave that way. Virtual AWS instances do not.

**Amazon EBS offering high availability & durability. And it offers the consistent & low-latency performance needed to run your workloads.**

* **EBS Magnetic volumes:**You can create EBS Magnetic volumes from 1 GB to 1 TB in size
* **EBS General Purpose SSD (gp2):** You can create **EBS General Purpose SSD (1G–16TB )**
* **Provisioned IOPS SSD (io1**)**:**Highest-performance SSD volume designed for mission-critical applications (4 GB – 16 TB)
* **Cold HDD (sc1):**Lowest cost HDD volume designed for less frequently accessed workloads (**500 GB – 16 TB**)
* **Amazon EBS Encryption:** You can use encrypted EBS volumes to meet a wide range of data-at-rest encryption requirements for regulated/audited data and applications.
* **Amazon EBS Snapshots:**You can create point-in-time snapshots of EBS volumes, which are persisted to Amazon S3. Snapshots protect data for long-term durability, and they can be used as the starting point for new EBS volumes. The same snapshot can be used to instantiate as many volumes as you wish. These snapshots can be copied across AWS regions.
* Performance metrics, such as bandwidth, throughput, latency, and average queue length, are available through the AWS Management Console. These**metrics, provided by AmazonCloudWatch,** allow you to monitor the performance of your volumes to make sure that you are providing enough performance for your applications without paying for resources you don’t need.

**What is cfn-init?**

It is a script that reads the template metadata from AWS::CloudFormation::init key and used to

1. Fetch data from cloud formation
2. Install packages
3. Write files to disk
4. Enable/disable, start/stop service

If you use cfn-init to update an existing file it creates a backup file with .bak extension.

**Elastic Search:**

Is a free open source engine used to search logs in their general, Amazon made it easier for us as service to create cloud, we connect to it cloud watch and can monitor,

After creating ES, we can add several ways to add data or connect to logs, we can use it by API,

We use Elastic Bean stalk to quickly deploy some applications on AWS, and the Elastic Bean stalk will create the own infrastructure itself. It will setup the load balancing and auto scaling, and configures monitoring.

First, we have to create infrastructure, and then we have to create a pipeline to deploy our code to that infrastructure. And new code can be deployed to a testing or QA environment before going to production. For these developers have to learn how to deploy resources and how-to setup those pipeline, instead of creating the features, The EBS aims to solve that problem using EBS we can deploy our code and service will automatically provision our capacity such as Auto Scaling, Load Balancing, and also monitors the environment. And we can customize this auto generated environment according to the requirement.

If we have application with shorter life cycles which don’t require more flexibility, if we are working with longer lifecycle applications which requires more configuration, customization it is best to use cloud formation template.

Languages:

* Java, .Net, PHP, Node.js, Python, Ruby,

Web Servers:

* Apache, Tomcat, Docker, Nginx, Java SE, IIS,

Supported Deployment Platforms

* Git, IDEs (Eclipse, Visual Studio), Manual upload (eg WAR files, ZIP)

**We use Elastic Beanstalk**

For spending minimal time to set up infrastructure, and testing, maintain some flexibility and control over the resources used to power applications.

**How EBS Work**

**Deployment Methods in EBS:**

**All at once**: It updates all instances at the same time for the reason it causes downtime, and this method is the fastest method among all. While deploying an application, instead of taking batches and updating, this method will update all instances all at same time.

**Rolling:** Rolling Deployment prevent the downtime, because of not updating all at once, and will update the application using batch method. Batch is a group of instances which are fixed in number, during a deployment each batch will take out of service, rest of them in service in serving traffic. While deployment, ELB will detach the batch and will serve the traffic with other batches, and then EBS will create a new batch for the updated version and the will reattach to the ELB. And ELB will check the traffic and instance health, and once the health check reaches the min threshold, then it will start serve the traffic.

**Rolling with additional batch:** It Is same as the rolling deployment method, instead of detaching the batching directly, we will add a new batch before detaching and then serve the traffic to the batch, and the detached batch will update the version and will reattach the batch.

**Immutable:** EBS will create a second Auto scaling group with a new application version and puts I our environment, then EBS launch a single instance with the updated application/configuration. That Instance start serving the traffic, once instance passes health checks EBS starts to launch more instances with the new application configuration until the no of instances in the new auto scaling group matches the no in the original.

**Blue**/**Green:** It follows the Immutable deployment method, we will create new resources like Immutable. We need to clone our current environment, or launch a new environment with different configurations, then we deploy the new application version to that new environment, once we test the new version in the new environment and we verify that is good to go we swap environment URLs which is basically modifies route 53 DNS configurations to point your green deployment to become original blue deployment,

**I have some private servers on my premises, also I have distributed some of my workload on the public cloud, what is this architecture called?**

Hybrid Cloud

**Explanation:** This type of architecture would be a hybrid cloud. Why? Because we are using both, the public cloud, and you’re on premises servers i.e. the private cloud. To make this hybrid architecture easy to use, wouldn’t it be better if your private and public cloud were all on the same network(virtually). This is established by including your public cloud servers in a virtual private cloud, and connecting this virtual cloud with your on-premise servers using a VPN (Virtual Private Network).

**What does the following command do with respect to the Amazon EC2 security groups?**

**ec2-create-group Create Security Group**

Creates a new security group for use with your account.

**Explanation:**A Security group is just like a firewall, it controls the traffic in and out of your instance. In AWS terms, the inbound and outbound traffic. The command mentioned is pretty straight forward, it says create security group, and does the same. Moving along, once your security group is created, you can add different rules in it. For example, you have an RDS instance, to access it, you must add the public IP address of the machine from which you want access the instance in its security group.

**You have a video trans-coding application. The videos are processed according to a queue. If the processing of a video is interrupted in one instance, it is resumed in another instance. Currently there is a huge back-log of videos which needs to be processed, for this you need to add more instances, but you need these instances only until your backlog is reduced. Which of these would be an efficient way to do it?**

You should be using an **On-Demand** instance for the same. Why? First of all, the workload has to be processed now, meaning it is urgent, secondly you don’t need them once your backlog is cleared, therefore Reserved Instance is out of the picture, and since the work is urgent, you cannot stop the work on your instance just because the spot price spiked, therefore Spot Instances shall also not be used. Hence On-Demand instances shall be the right choice in this case.

**You have a distributed application that periodically processes large volumes of data across multiple Amazon EC2 Instances. The application is designed to recover gracefully from Amazon EC2 instance failures. You are required to accomplish this task in the most cost-effective way.**

**Which of the following will meet your requirements?**

Spot Instances

**Explanation:** Since the work we are addressing here is not continuous, a reserved instance shall be idle at times, same goes with On-Demand instances. Also, it does not make sense to launch an On-Demand instance whenever work comes up, since it is expensive. Hence Spot Instances will be the right fit because of their low rates and no long-term commitments.

**How is stopping and terminating an instance different from each other?**

Starting, stopping and terminating are the three states in an EC2 instance, let’s discuss them in detail:

* **Stopping and Starting** an instance: When an instance is stopped, the instance performs a normal shutdown and then transitions to a stopped state. All of its Amazon EBS volumes remain attached, and you can start the instance again at a later time. You are not charged for additional instance hours while the instance is in a stopped state.
* **Terminating** an instance: When an instance is terminated, the instance performs a normal shutdown, then the attached Amazon EBS volumes are deleted unless the volume’s *deleteOnTermination* attribute is set to false. The instance itself is also deleted, and you can’t start the instance again at a later time.

**If I want my instance to run on a single-tenant hardware, which value do I have to set the instance’s tenancy attribute to?**

Dedicated

**Explanation:** The Instance tenancy attribute should be set to Dedicated Instance. The rest of the values are invalid.

**When will you incur costs with an Elastic IP address (EIP)?**

When it is allocated and associated with a stopped instance.

**Explanation:** You are not charged, if only one Elastic IP address is attached with your running instance. But you do get charged in the following conditions:

* When you use more than one Elastic IPs with your instance.
* When your Elastic IP is attached to a stopped instance.
* When your Elastic IP is not attached to any instance.

**How is a Spot instance different from an On-Demand instance or Reserved Instance?**

First of all, let’s understand that Spot Instance, On-Demand instance and Reserved Instances are all models for pricing. Moving along, spot instances provide the ability for customers to purchase compute capacity with no upfront commitment, at hourly rates usually lower than the On-Demand rate in each region. Spot instances are just like bidding, the bidding price is called Spot Price. The Spot Price fluctuates based on supply and demand for instances, but customers will never pay more than the maximum price they have specified. If the Spot Price moves higher than a customer’s maximum price, the customer’s EC2 instance will be shut down automatically. But the reverse is not true, if the Spot prices come down again, your EC2 instance will not be launched automatically, one has to do that manually.  In Spot and On-demand instance, there is no commitment for the duration from the user side, however in reserved instances one has to stick to the time period that he has chosen.

**Are the Reserved Instances available for Multi-AZ Deployments?**

Available for all instance types

**Explanation:** Reserved Instances is a pricing model, which is available for all instance types in EC2.

**How to use the processor state control feature available on the c4.8xlarge instance?**

The processor state control consists of 2 states:

* The C state – Sleep state varying from c0 to c6. C6 being the deepest sleep state for a processor
* The P state – Performance state p0 being the highest and p15 being the lowest possible frequency.

Now, why the C state and P state. Processors have cores, these cores need thermal headroom to boost their performance. Now since all the cores are on the processor the temperature should be kept at an optimal state so that all the cores can perform at the highest performance.

Now how will these states help in that? If a core is put into sleep state it will reduce the overall temperature of the processor and hence other cores can perform better. Now the same can be synchronized with other cores, so that the processor can boost as many cores it can by timely putting other cores to sleep, and thus get an overall performance boost.

Concluding, the C and P state can be customized in some EC2 instances like the c4.8xlarge instance and thus you can customize the processor according to your workload.

**What kind of network performance parameters can you expect when you launch instances in cluster placement group?**

The network performance depends on the instance type and network performance specification, if launched in a placement group you can expect up to

* 10 Gbps in a single-flow,
* 20 Gbps in multiflow i.e., full duplex
* Network traffic outside the placement group will be limited to 5 Gbps (full duplex).

**To deploy a 4-node cluster of Hadoop in AWS which instance type can be used?**

First let’s understand what happens in a Hadoop cluster, the Hadoop cluster follows a master slave concept. The master machine processes all the data, slave machines store the data and act as data nodes. Since all the storage happens at the slave, a higher capacity hard disk would be recommended and since master does all the processing, a higher RAM and a much better CPU is required. Therefore, you can select the configuration of your machine depending on your workload. For e.g. – In this case c4.8xlarge will be preferred for master machine whereas for slave machine we can select i2.large instance. If you don’t want to deal with configuring your instance and installing Hadoop cluster manually, you can straight away launch an Amazon EMR (Elastic Map Reduce) instance which automatically configures the servers for you. You dump your data to be processed in S3, EMR picks it from there, processes it, and dumps it back into S3.

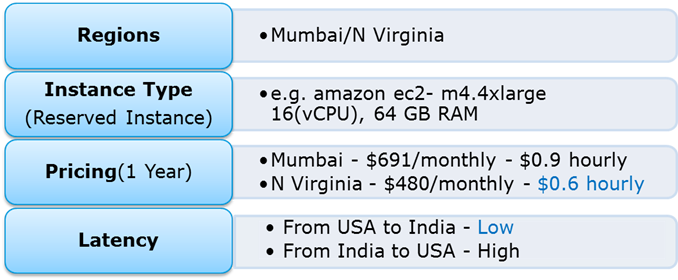
**Where do you think an AMI fits, when you are designing an architecture for a solution?**

AMIs (Amazon Machine Images) are like templates of virtual machines and an instance is derived from an AMI. AWS offers pre-baked AMIs which you can choose while you are launching an instance, some AMIs are not free, therefore can be bought from the AWS Marketplace. You can also choose to create your own custom AMI which would help you save space on AWS. For example, if you don’t need a set of software on your installation, you can customize your AMI to do that. This makes it cost efficient, since you are removing the unwanted things.

**How do you choose an Availability Zone?**

Let’s understand this through an example, consider there’s a company which has user base in India as well as in the US.

Let us see how we will choose the region for this use case:



So, with reference to the above figure the regions to choose between are, Mumbai and North Virginia. Now let us first compare the pricing, you have hourly prices, which can be converted to your per month figure. Here North Virginia emerges as a winner. But, pricing cannot be the only parameter to consider. Performance should also be kept in mind hence, let’s look at latency as well. Latency basically is the time that a server takes to respond to your requests i.e. the response time. North Virginia wins again!

So, concluding, North Virginia should be chosen for this use case.

**Is one Elastic IP address enough for every instance that I have running?**

Depends! Every instance comes with its own private and public address. The private address is associated exclusively with the instance and is returned to Amazon EC2 only when it is stopped or terminated. Similarly, the public address is associated exclusively with the instance until it is stopped or terminated. However, this can be replaced by the Elastic IP address, which stays with the instance if the user doesn’t manually detach it. But what if you are hosting multiple websites on your EC2 server, in that case you may require more than one Elastic IP address.

**What are the best practices for Security in Amazon EC2?**

There are several best practices to secure Amazon EC2. A few of them are given below:

* Use AWS Identity and Access Management (IAM) to control access to your AWS resources.
* Restrict access by only allowing trusted hosts or networks to access ports on your instance.
* Review the rules in your security groups regularly, and ensure that you apply the principle of least
* Privilege – only open up permissions that you require.
* Disable password-based logins for instances launched from your AMI. Passwords can be found or cracked, and are a security risk.

**You need to configure an Amazon S3 bucket to serve static assets for your public-facing web application. Which method will ensure that all objects uploaded to the bucket are set to public read?**

Configure the bucket policy to set all objects to public read.

**Explanation:** Rather than making changes to every object, it’s better to set the policy for the whole bucket. IAM is used to give more granular permissions, since this is a website, all objects would be public by default.

**A customer wants to leverage Amazon Simple Storage Service (S3) and Amazon Glacier as part of their backup and archive infrastructure. The customer plans to use third-party software to support this integration. Which approach will limit the access of the third-party software to only the Amazon S3 bucket named “company-backup”?**

A custom IAM user policy limited to the Amazon S3 API in “company-backup”.

**Explanation:** Taking queue from the previous questions, this use case involves more granular permissions, hence IAM would be used here.

**Can S3 be used with EC2 instances, if yes, how?**

Yes, it can be used 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 Amazon Machine Images (AMIs) into Amazon S3 and to move them between Amazon S3 and Amazon EC2.

Another use case could be for websites hosted on EC2 to load their static content from S3.

**A customer implemented AWS Storage Gateway with a gateway-cached volume at their head office. An event takes the link between the main and branch office offline. Which methods will enable the branch office to access their data?**

Launch a new AWS Storage Gateway instance AMI in Amazon EC2, and restore from a gateway snapshot.

**Explanation:** The fastest way to do it would be launching a new storage gateway instance. Why? Since time is the key factor which drives every business, troubleshooting this problem will take more time. Rather than we can just restore the previous working state of the storage gateway on a new instance.

**When you need to move data over long distances using the internet, for instance across countries or continents to your Amazon S3 bucket, which method or service will you use?**

Amazon Transfer Acceleration

**Explanation:** You would not use Snowball, because for now, the snowball service does not support cross region data transfer, and since, we are transferring across countries, Snowball cannot be used. Transfer Acceleration shall be the right choice here as it throttles your data transfer with the use of optimized network paths and Amazon’s content delivery network up to 300% compared to normal data transfer speed.

**How can you speed up data transfer in Snowball?**

The data transfer can be increased in the following way:

* By performing multiple copy operations at one time i.e. if the workstation is powerful enough, you can initiate multiple cp commands each from different terminals, on the same Snowball device.
* Copying from multiple workstations to the same snowball.
* Transferring large files or by creating a batch of small file, this will reduce the encryption overhead.
* Eliminating unnecessary hops i.e. make a setup where the source machine(s) and the snowball are the only machines active on the switch being used, this can hugely improve performance.

**If you want to launch Amazon Elastic Compute Cloud (EC2) instances and assign each instance a predetermined private IP address you should:**

Launch the instances in the Amazon Virtual Private Cloud (VPC).

**Explanation:** The best way of connecting to your cloud resources (for ex- ec2 instances) from your own data center (for ex.- private cloud) is a VPC. Once you connect your datacenter to the VPC in which your instances are present, each instance is assigned a private IP address which can be accessed from your datacenter. Hence, you can access your public cloud resources, as if they were on your own network.

**Can I connect my corporate datacenter to the Amazon Cloud?**

Yes, you can do this by establishing a VPN (Virtual Private Network) connection between your company’s network and your VPC (Virtual Private Cloud), this will allow you to interact with your EC2 instances as if they were within your existing network.

**Is it possible to change the private IP addresses of an EC2 while it is running/stopped in a VPC?**

Primary private IP address is attached with the instance throughout its lifetime and cannot be changed, however secondary private addresses can be unassigned, assigned or moved between interfaces or instances at any point.

**Why do you make subnets?**

To efficiently utilize networks that have a large no. of hosts.

**Explanation:** If there is a network which has a large no. of hosts, managing all these hosts can be a tedious job. Therefore, we divide this network into subnets (sub-networks) so that managing these hosts becomes simpler.

**Which of the following is true?**

You can attach multiple subnets to a route table

**Explanation:** Route Tables are used to route network packets, therefore in a subnet having multiple route tables will lead to confusion as to where the packet has to go. Therefore, there is only one route table in a subnet, and since a route table can have any no. of records or information, hence attaching multiple subnets to a route table is possible.

**In CloudFront what happens when content is NOT present at an Edge location and a request is made to it?**

CloudFront delivers the content directly from the origin server and stores it in the cache of the edge location

**Explanation:** CloudFront is a content delivery system, which caches data to the nearest edge location from the user, to reduce latency. If data is not present at an edge location, the first time the data may get transferred from the original server, but from the next time, it will be served from the cached edge.

**If I’m using Amazon CloudFront, can I use Direct Connect to transfer objects from my own data center?**

Yes. Amazon CloudFront supports custom origins including origins from outside of AWS. With AWS Direct Connect, you will be charged with the respective data transfer rates.

**If my AWS Direct Connect fails, will I lose my connectivity?**

If a backup AWS Direct connect has been configured, in the event of a failure it will switch over to the second one. It is recommended to enable Bidirectional Forwarding Detection (BFD) when configuring your connections to ensure faster detection and failover. On the other hand, if you have configured a backup IPsec VPN connection instead, all VPC traffic will failover to the backup 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.

**If I launch a standby RDS instance, will it be in the same Availability Zone as my primary?**

No

**Explanation:** No, since the purpose of having a standby instance is to avoid an infrastructure failure (if it happens), therefore the standby instance is stored in a different availability zone, which is a physically different independent infrastructure.

**When would I prefer Provisioned IOPS over Standard RDS storage?**

If you have batch-oriented workloads

**Explanation:**  Provisioned IOPS deliver high IO rates but on the other hand it is expensive as well. Batch processing workloads do not require manual intervention they enable full utilization of systems, therefore a provisioned IOPS will be preferred for batch oriented workload.

**How is Amazon RDS, DynamoDB and Redshift different?**

* Amazon RDS is a database management service for relational databases, it manages patching, upgrading, backing up of data etc. of databases for you without your intervention. RDS is a Db management service for structured data only.
* DynamoDB, on the other hand, is a NoSQL database service, NoSQL deals with unstructured data.
* Redshift, is an entirely different service, it is a data warehouse product and is used in data analysis.

**If I am running my DB Instance as a Multi-AZ deployment, can I use the standby DB Instance for read or write operations along with primary DB instance?**

No

**Explanation:** No,Standby DB instance cannot be used with primary DB instance in parallel, as the former is solely used for standby purposes, it cannot be used unless the primary instance goes down.

**Your company’s branch offices are all over the world, they use a software with a multi-regional deployment on AWS, they use MySQL 5.6 for data persistence.**

**The task is to run an hourly batch process and read data from every region to compute cross-regional reports which will be distributed to all the branches. This should be done in the shortest time possible. How will you build the DB architecture in order to meet the requirements?**

For each regional deployment, use RDS MySQL with a master in the region and a read replica in the HQ region

**Explanation:** For this we will take an RDS instance as a master, because it will manage our database for us and since we should read from every region, we’ll put a read replica of this instance in every region where the data has to be read from. It is not correct since putting a read replica would be more efficient than putting a snapshot, a read replica can be promoted if needed to an independent DB instance, but with a Db snapshot it becomes mandatory to launch a separate DB Instance.

**Can I run more than one DB instance for Amazon RDS for free?**

Yes. You can run more than one Single-AZ Micro database instance, that too for free! 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.

**Which AWS services will you use to collect and process e-commerce data for near real-time analysis?**

Amazon DynamoDB and Amazon Redshift

**Explanation:** DynamoDB is a fully managed NoSQL database service. DynamoDB, therefore can be fed any type of unstructured data, which can be data from e-commerce websites as well, and later, an analysis can be done on them using Amazon Redshift. We are not using Elastic MapReduce, since a near real time analyses is needed.

**Can I retrieve only a specific element of the data, if I have a nested JSON data in DynamoDB?**

Yes. When using the GetItem, BatchGetItem, Query or Scan APIs, you can define a Projection Expression to determine which attributes should be retrieved from the table. Those attributes can include scalars, sets, or elements of a JSON document.

**A company is deploying a new two-tier web application in AWS. The company has limited staff and requires high availability, and the application requires complex queries and table joins. Which configuration provides the solution for the company’s requirements?**

Amazon DynamoDB

**Explanation:** DynamoDB can scale more than RDS or any other relational database service, therefore DynamoDB would be the apt choice.

**What happens to my backups and DB Snapshots if I delete my DB Instance?**

When you delete a DB instance, you have an option of creating a final DB snapshot, if you do that you can restore your database from that snapshot. RDS retains this user-created DB snapshot along with all other manually created DB snapshots after the instance is deleted, also automated backups are deleted and only manually created DB Snapshots are retained.

**Which of the following use cases are suitable for Amazon DynamoDB? Choose 2 answers**

Storing metadata for Amazon S3 objects and Running relational joins and complex updates.

**Explanation:** If all your JSON data have the same fields ex: [id, name, age] then it would be better to store it in a relational database, the metadata on the other hand is unstructured, also running relational joins or complex updates would work on DynamoDB as well.

**How can I load my data to Amazon Redshift from different data sources like Amazon RDS, Amazon DynamoDB and Amazon EC2?**

You can load the data in the following two ways:

* You can use the COPY command to load data in parallel directly to Amazon Redshift from Amazon EMR, Amazon DynamoDB, or any SSH-enabled host.
* 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.

**Your application must retrieve data from your user’s mobile every 5 minutes and the data is stored in DynamoDB, later every day at a time the data is extracted into S3 on a per user basis and then your application is later used to visualize the data to the user. You are asked to optimize the architecture of the backend system to lower cost, what would you recommend?**

Introduce Amazon ElastiCache to cache reads from the Amazon DynamoDB table and reduce provisioned read throughput.

**Explanation:** Since our work requires the data to be extracted and analyzed, to optimize this process a person would use provisioned IO, but since it is expensive, using a ElastiCache memoryinsread to cache the results in the memory can reduce the provisioned read throughput and hence reduce cost without affecting the performance.

**You are running a website on EC2 instances deployed across multiple Availability Zones with a Multi-AZ RDS MySQL Extra Large DB Instance. The site performs a high number of small reads and writes per second and relies on an eventual consistency model. After comprehensive tests, you discover that there is read contention on RDS MySQL. Which are the best approaches to meet these requirements? (Choose 2 answers)**

Deploy ElastiCache in-memory cache running in each availability zone

Increase the RDS MySQL Instance size and Implement provisioned IOPS

**Explanation:**Since it does a lot of read writes, provisioned IO may become expensive. But we need high performance as well, therefore the data can be cached using ElastiCache which can be used for frequently reading the data. As for RDS since read contention is happening, the instance size should be increased and provisioned IO should be introduced to increase the performance.

**A startup is running a pilot deployment of around 100 sensors to measure street noise and air quality in urban areas for 3 months. It was noted that every month around 4GB of sensor data is generated. The company uses a load balanced auto scaled layer of EC2 instances and a RDS database with 500 GB standard storage. The pilot was a success and now they want to deploy at least 100K sensors which need to be supported by the backend. You need to store the data for at least 2 years to analyze it. Which setup of the following would you prefer?**

Replace the RDS instance with a 6 node Redshift cluster with 96TB of storage

**Explanation:** A Redshift cluster would be preferred because it easy to scale, also the work would be done in parallel through the nodes, therefore is perfect for a bigger workload like our use case. Since each month 4 GB of data is generated, therefore in 2 years, it should be around 96 GB. And since the servers will be increased to 100K in number, 96 GB will approximately become 96TB. Hence option C is the right answer.

**Suppose you have an application where you have to render images and also do some general computing. From the following services which service will best fit your need?**

Application Load Balancer

**Explanation:** You will choose an application load balancer, since it supports path based routing, which means it can take decisions based on the URL, therefore if your task needs image rendering it will route it to a different instance, and for general computing it will route it to a different instance.

**What is the difference between Scalability and Elasticity?**

Scalability is the ability of a system to increase its hardware resources to handle the increase in demand. It can be done by increasing the hardware specifications or increasing the processing nodes.

Elasticity is the ability of a system to handle increase in the workload by adding additional hardware resources when the demand increases (same as scaling) but also rolling back the scaled resources, when the resources are no longer needed. This is particularly helpful in Cloud environments, where a pay per use model is followed.

**How will you change the instance type for instances which are running in your application tier and are using Auto Scaling. Where will you change it from the following areas?**

Auto Scaling launch configuration

**Explanation:** Auto scaling tags configuration, is used to attach metadata to your instances, to change the instance type you have to use auto scaling launch configuration.

**You have a content management system running on an Amazon EC2 instance that is approaching 100% CPU utilization. Which option will reduce load on the Amazon EC2 instance?**

Create a load balancer, and register the Amazon EC2 instance with it

**Explanation:** Creating alone an auto-scaling group will not solve the issue, until you attach a load balancer to it. Once you attach a load balancer to an auto-scaling group, it will efficiently distribute the load among all the instances. Option B – CloudFront is a CDN, it is a data transfer tool therefore will not help reduce load on the EC2 instance. Similarly, the other option – Launch configuration is a template for configuration which has no connection with reducing loads.

**When should I use a Classic Load Balancer and when should I use an Application load balancer?**

A Classic Load Balancer is ideal for simple load balancing of traffic across multiple EC2 instances, while an Application Load Balancer is ideal for microservices or container-based architectures where there is a need to route traffic to multiple services or load balance across multiple ports on the same EC2 instance.

**What does Connection draining do?**

 Re-routes traffic from instances which are to be updated or failed a health check.

**Explanation:** Connection draining is a service under ELB which constantly monitors the health of the instances. If any instance fails a health check or if any instance has to be patched with a software update, it  pulls all the traffic from that instance and re-routes them to other instances.

**When an instance is unhealthy, it is terminated and replaced with a new one, which of the following services does that?**

Fault Tolerance

**Explanation:** When ELB detects that an instance is unhealthy, it starts routing incoming traffic to other healthy instances in the region. If all the instances in a region becomes unhealthy, and if you have instances in some other availability zone/region, your traffic is directed to them. Once your instances become healthy again, they are re-routed back to the original instances.

**What are lifecycle hooks used for in Auto-Scaling?**

 They are used to put an additional wait time to a scale in or scale out event.

**Explanation:** Lifecycle hooks are used for putting wait time before any lifecycle action i.e., launching or terminating an instance happens. The purpose of this wait time, can be anything from extracting log files before terminating an instance or installing the necessary software’s in an instance before launching it.

**A user has setup an Auto Scaling group. Due to some issue, the group has failed to launch a single instance for more than 24 hours. What will happen to Auto Scaling in this condition?**

Auto Scaling will suspend the scaling process

**Explanation:** Auto Scaling allows you to suspend and then resume one or more of the Auto Scaling processes in your Auto Scaling group. This can be very useful when you want to investigate a configuration problem or other issue with your web application, and then make changes to your application, without triggering the Auto Scaling process.

**You have an EC2 Security Group with several running EC2 instances. You changed the Security Group rules to allow inbound traffic on a new port and protocol, and then launched several new instances in the same Security Group. The new rules apply:**

Immediately to all instances in the security group.

**Explanation:** Any rule specified in an EC2 Security Group applies immediately to all the instances, irrespective of when they are launched before or after adding a rule.

**To create a mirror image of your environment in another region for disaster recovery, which of the following AWS resources do not need to be recreated in the second region? ( Choose 2 answers )**

Route 53 Record Sets and Elastic IP Addresses (EIP)

**Explanation:** Elastic IPs and Route 53 record sets are common assets therefore there is no need to replicate them, since Elastic IPs and Route 53 are valid across regions

**A customer wants to capture all client connection information from his load balancer at an interval of 5 minutes, which of the following options should he choose for his application?**

Enable AWS CloudTrail for the load-balancer.

**Explanation:** AWS CloudTrail provides inexpensive logging information for load balancer and other AWS resources. This logging information can be used for analyses and other administrative work, therefore is perfect for this use case.

**A customer wants to track access to their Amazon Simple Storage Service (S3) buckets and also use this information for their internal security and access audits. Which of the following will meet the Customer requirement?**

Enable AWS CloudTrail to audit all Amazon S3 bucket access.

**Explanation:** AWS CloudTrail has been designed for logging and tracking API calls. Also, this service is available for storage, therefore should be used in this use case.

**Which of the following are true regarding AWS CloudTrail? (Choose 2 answers)**

CloudTrail is enabled on a per-region and service basis

Logs can be delivered to a single Amazon S3 bucket for aggregation.

**Explanation:** CloudTrail is not enabled for all the services and is also not available for all the regions. Therefore option B is correct, also the logs can be delivered to your S3 bucket, hence C is also correct.

**What happens if CloudTrail is turned on for my account but my Amazon S3 bucket is not configured with the correct policy?**

CloudTrail files are delivered according to S3 bucket policies. If the bucket is not configured or is misconfigured, CloudTrail might not be able to deliver the log files.

**How do I transfer my existing domain name registration to Amazon Route 53 without disrupting my existing web traffic?**

You will need to get a list of the DNS record data for your domain name first, it is generally available in the form of a “zone file” that you can get from your existing DNS provider. Once you receive the DNS record data, you can use Route 53’s Management Console or simple web-services interface to create a hosted zone that will store your DNS records for your domain name and follow its transfer process. It also includes steps such as updating the nameservers for your domain name to the ones associated with your hosted zone. For completing the process, you have to contact the registrar with whom you registered your domain name and follow the transfer process. As soon as your registrar propagates the new name server delegations, your DNS queries will start to get answered.

**Which of the following services you would not use to deploy an app?**

Lambda

**Explanation:** Lambda is used for running server-less applications. It can be used to deploy functions triggered by events. When we say server-less, we mean without you worrying about the computing resources running in the background. It is not designed for creating applications which are publicly accessed.

**How does Elastic Beanstalk apply updates?**

By having a duplicate ready with updates before swapping.

**Explanation:** Elastic Beanstalk prepares a duplicate copy of the instance, before updating the original instance, and routes your traffic to the duplicate instance, so that, in case your updated application fails, it will switch back to the original instance, and there will be no downtime experienced by the users who are using your application.

**How is AWS Elastic Beanstalk different than AWS OpsWorks?**

AWS Elastic Beanstalk is an application management platform while OpsWorks is a configuration management platform. BeanStalk is an easy to use service which is used for deploying and scaling web applications developed with Java, .Net, PHP, Node.js, Python, Ruby, Go and Docker. Customers upload their code and Elastic Beanstalk automatically handles the deployment. The application will be ready to use without any infrastructure or resource configuration.

In contrast, AWS OpsWorks is an integrated configuration management platform for IT administrators or DevOps engineers who want a high degree of customization and control over operations.

**What happens if my application stops responding to requests in beanstalk?**

AWS Beanstalk applications have a system in place for avoiding failures in the underlying infrastructure. If an Amazon EC2 instance fails for any reason, Beanstalk will use Auto Scaling to automatically launch a new instance. Beanstalk can also detect if your application is not responding on the custom link, even though the infrastructure appears healthy, it will be logged as an environmental event ( ex: a bad version was deployed) so you can take an appropriate action.

**How is AWS OpsWorks different than AWS CloudFormation?**

OpsWorks and CloudFormation both support application modelling, deployment, configuration, management and related activities. Both support a wide variety of architectural patterns, from simple web applications to highly complex applications. AWS OpsWorks and AWS CloudFormation differ in abstraction level and areas of focus.

AWS CloudFormation is a building block service which enables customer to manage almost any AWS resource via JSON-based domain specific language. It provides foundational capabilities for the full breadth of AWS, without prescribing a particular model for development and operations. Customers define templates and use them to provision and manage AWS resources, operating systems and application code.

In contrast, AWS OpsWorks is a higher-level service that focuses on providing highly productive and reliable DevOps experiences for IT administrators and ops-minded developers. To do this, AWS OpsWorks employs a configuration management model based on concepts such as stacks and layers, and provides integrated experiences for key activities like deployment, monitoring, auto-scaling, and automation. Compared to AWS CloudFormation, AWS OpsWorks supports a narrower range of application-oriented AWS resource types including Amazon EC2 instances, Amazon EBS volumes, Elastic IPs, and Amazon CloudWatch metrics.

**I created a key in Oregon region to encrypt my data in North Virginia region for security purposes. I added two users to the key and an external AWS account. I wanted to encrypt an object in S3, so when I tried, the key that I just created was not listed.  What could be the reason?**

The Key should be in the same region.

**Explanation:** The key created and the data to be encrypted should be in the same region. Hence the approach taken here to secure the data is incorrect.

**A company needs to monitor the read and write IOPS for their AWS MySQL RDS instance and send real-time alerts to their operations team. Which AWS services can accomplish this?**

Amazon CloudWatch

**Explanation:** Amazon CloudWatch is a cloud monitoring tool and hence this is the right service for the mentioned use case. The other options listed here are used for other purposes for example route 53 is used for DNS services, therefore CloudWatch will be the apt choice.

**What happens when one of the resources in a stack cannot be created successfully in AWS OpsWorks?**

When an event like this occurs, the “automatic rollback on error” feature is enabled, which causes all the AWS resources which were created successfully till the point where the error occurred to be deleted. This is helpful since it does not leave behind any erroneous data, it ensures the fact that stacks are either created fully or not created at all. It is useful in events where you may accidentally exceed your limit of the no. of Elastic IP addresses or maybe you may not have access to an EC2 AMI that you are trying to run etc.

**What automation tools can you use to spinning up the servers?**

Any of the following tools can be used:

* Roll-your-own scripts, and use the AWS API tools.  Such scripts could be written in bash, Perl or other language of your choice.
* Use a configuration management and provisioning tool like puppet or its successor Opscode Chef.  You can also use a tool like Scalr.
* Use a managed solution such as Rightscale.

**What is OpenStack?**

* OpenStack is often called Cloud Operating System, and that is not far from the truth. It is the complete environment for deploying IaaS which gives you possibility of making your own cloud similar to AWS. It is highly modular and consists of many sub-projects so you can pick and choose which functionality you need. OpenStack distribution are available from Red Hat, Mirantis, HPE, Oracle, Canonical and many others. It is completely open source project but some vendors make proprietary distributions.

**Classify Cloud Platforms anategory?**

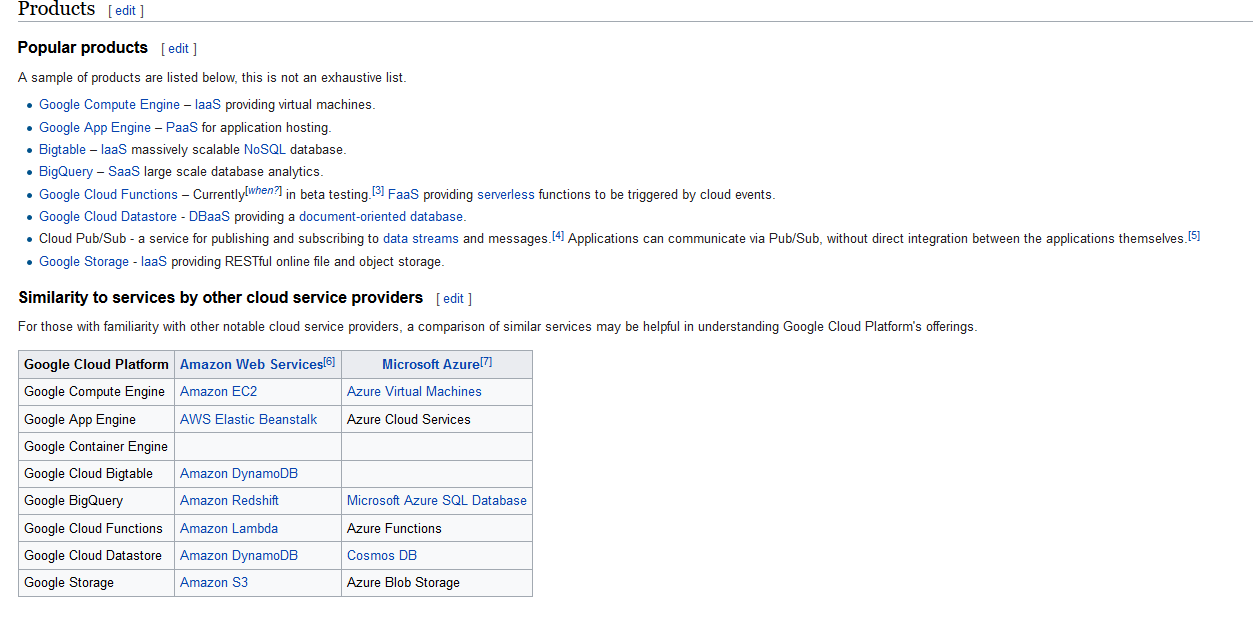
* Cloud Computing software can be classified as Software as a Service or SaaS, Infrastructure as a Service or IaaS and Platform as a Service or PaaS.
* SaaS is peace of software that runs over network on remote server and has only user interface exposed to users, usually in web browser. For example, salesforce.com.
* Infrastructure as a service is a cloud environment that exposes VM to user to use as entire OS or container where you could install anything you would install on your server.
* Example for this would-be OpenStack, AWS, Eucalyptus.

PaaS allows users to deploy their own application on the preinstalled platform, usually framework of application server and suite of developer tools. Examples for this would be OpenShHeroku.

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Google Cloud\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

Benefits of google cloud:

* Future proof infrastructure
* Powerful data & analytics
* Serverless, fully managed computing
* Customer friendly pricing
* Security at scale
* Data center innovation

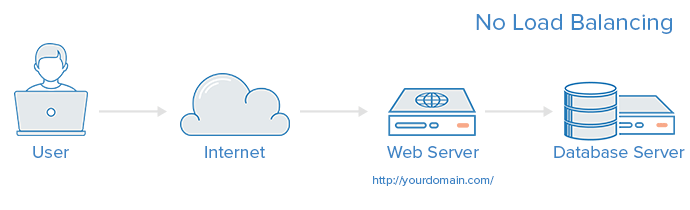


**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*AWS\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Load balancers**

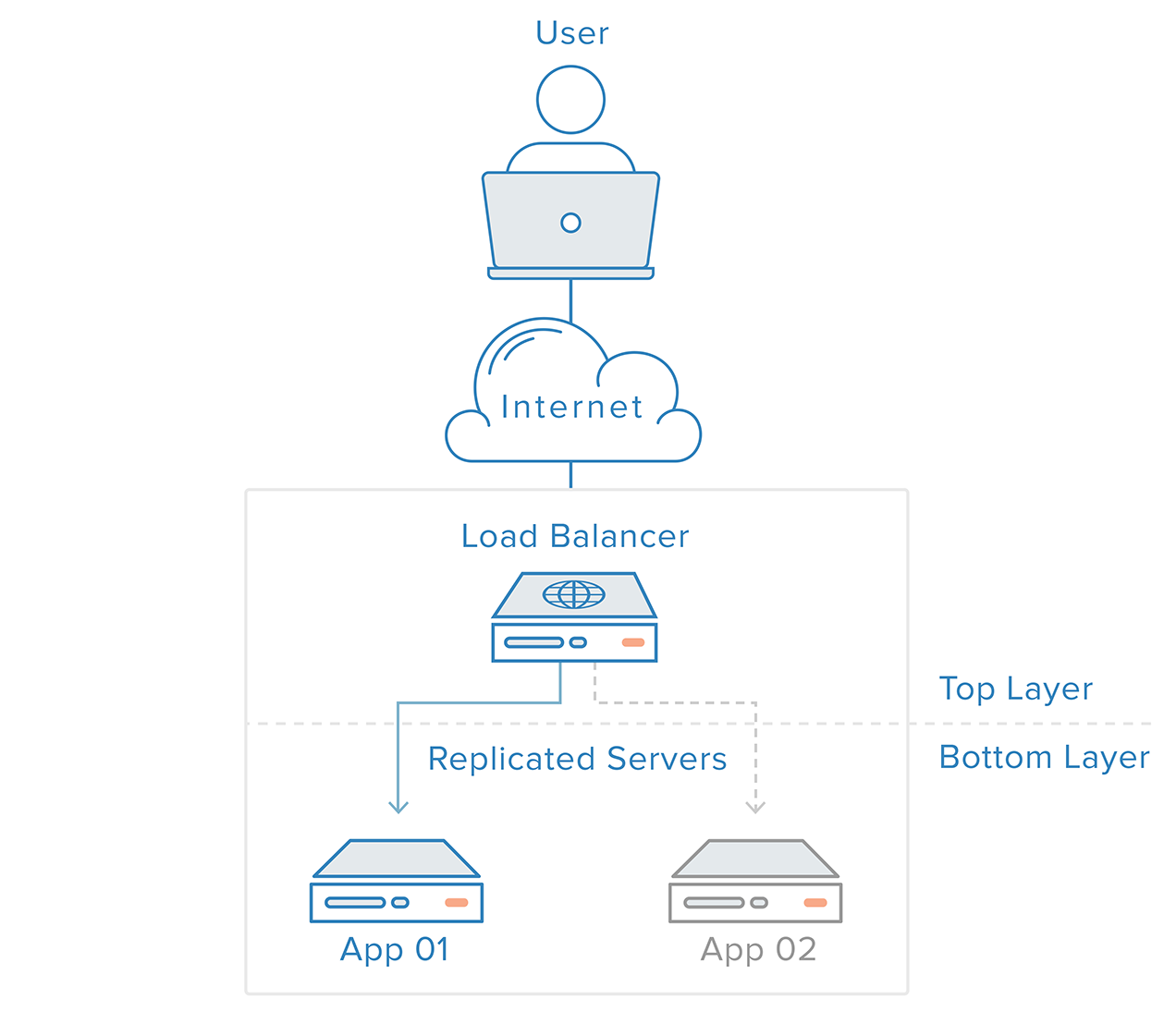
Load balancing is a key component of highly-available infrastructures commonly used to improve the performance and reliability of web sites, applications, databases and other services by distributing the workload across multiple servers.

A web infrastructure with no load balancing might look something like the following:



In this example, the user connects directly to the web server, at yourdomain.com. If this single web server goes down, the user will no longer be able to access the website. In addition, if many users try to access the server simultaneously and it is unable to handle the load, they may experience slow load times or may be unable to connect at all.

This single point of failure can be mitigated by introducing a load balancer and at least one additional web server on the backend. Typically, all of the backend servers will supply identical content so that users receive consistent content regardless of which server responds.



In the example illustrated above, the user accesses the load balancer, which forwards the user's request to a backend server, which then responds directly to the user's request. In this scenario, the single point of failure is now the load balancer itself. This can be mitigated by introducing a second load balancer, but before we discuss that, let's explore how load balancers work.

## **What kind of traffic can load balancers handle?**

Load balancer administrators create forwarding rules for four main types of traffic:

* **HTTP** — Standard HTTP balancing directs requests based on standard HTTP mechanisms. The Load Balancer sets the X-Forwarded-For, X-Forwarded-Proto, and X-Forwarded-Port headers to give the backends information about the original request.
* **HTTPS** — HTTPS balancing functions the same as HTTP balancing, with the addition of encryption. Encryption is handled in one of two ways: either with **SSL passthrough** which maintains encryption all the way to the backend or with **SSL termination** which places the decryption burden on the load balancer but sends the traffic unencrypted to the back end.
* **TCP** — For applications that do not use HTTP or HTTPS, TCP traffic can also be balanced. For example, traffic to a database cluster could be spread across all of the servers.
* **UDP** — More recently, some load balancers have added support for load balancing core internet protocols like DNS and syslogd that use UDP.

These forwarding rules will define the protocol and port on the load balancer itself and map them to the protocol and port the load balancer will use to route the traffic to on the backend.

## **How does the load balancer choose the backend server?**

Load balancers choose which server to forward a request to based on a combination of two factors. They will first ensure that any server they can choose is actually responding appropriately to requests and then use a pre-configured rule to select from among that healthy pool.

### Health Checks

Load balancers should only forward traffic to "healthy" backend servers. To monitor the health of a backend server, health checks regularly attempt to connect to backend servers using the protocol and port defined by the forwarding rules to ensure that servers are listening. If a server fails a health check, and therefore is unable to serve requests, it is automatically removed from the pool, and traffic will not be forwarded to it until it responds to the health checks again.

### Load Balancing Algorithms

The load balancing algorithm that is used determines which of the healthy servers on the backend will be selected. A few of the commonly used algorithms are:

**Round Robin** — Round Robin means servers will be selected sequentially. The load balancer will select the first server on its list for the first request, then move down the list in order, starting over at the top when it reaches the end.

**Least Connections** — Least Connections means the load balancer will select the server with the least connections and is recommended when traffic results in longer sessions.

**Source** — With the Source algorithm, the load balancer will select which server to use based on a hash of the source IP of the request, such as the visitor's IP address. This method ensures that a particular user will consistently connect to the same server.

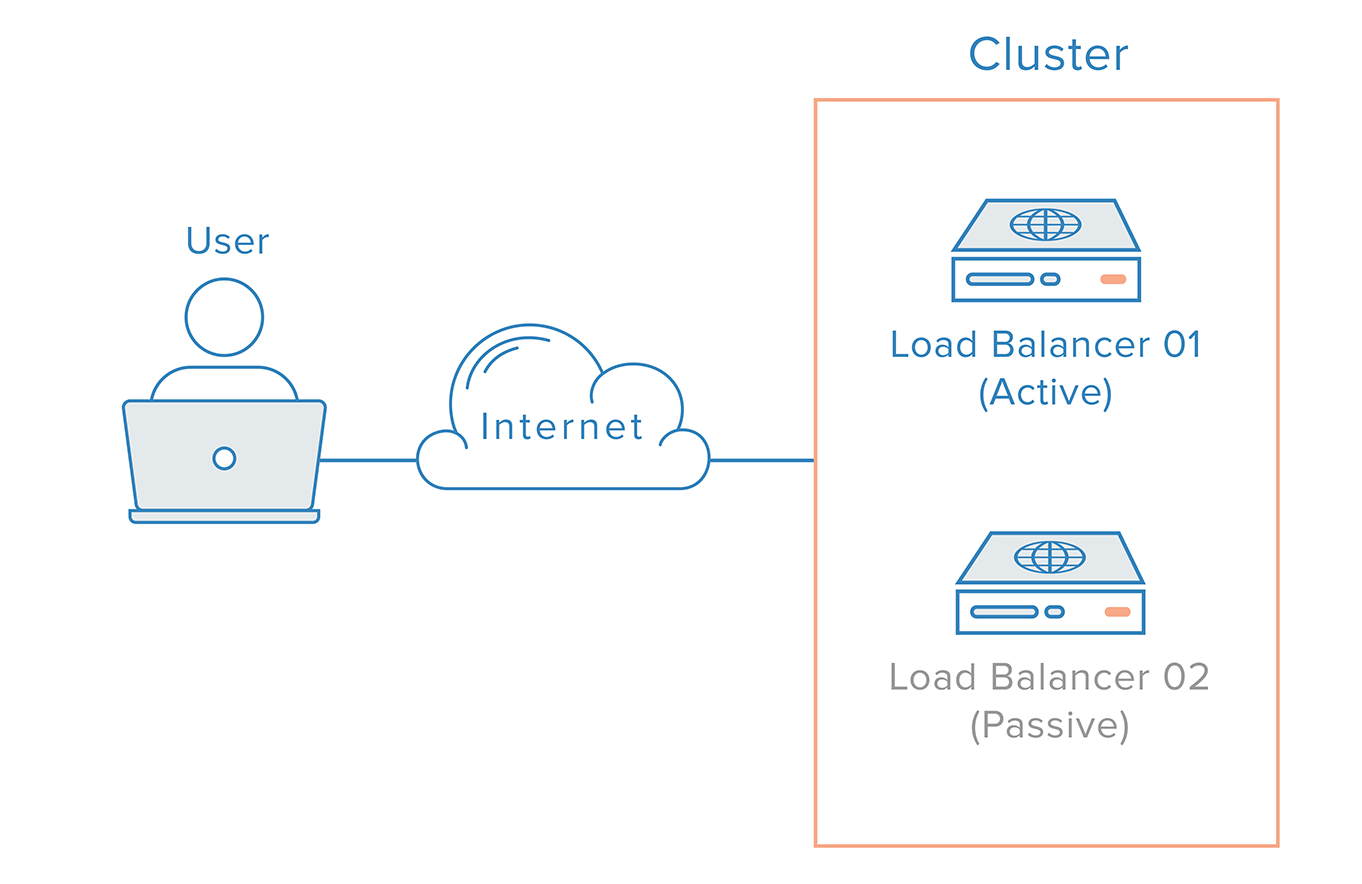
The algorithms available to administrators vary depending on the specific load balancing technology in use.

### How do load balancers handle state?

Some applications require that a user continues to connect to the same backend server. A Source algorithm creates an affinity based on client IP information. Another way to achieve this at the web application level is through **sticky sessions**, where the load balancer sets a cookie and all of the requests from that session are directed to the same physical server.

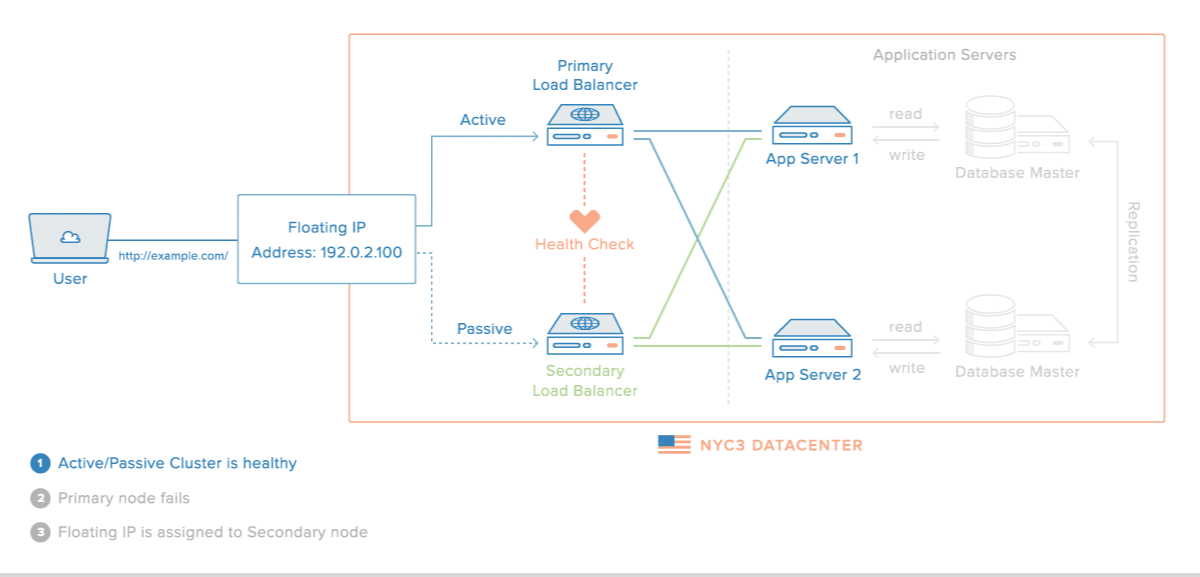
## **Redundant Load Balancers**

To remove the load balancer as a single point of failure, a second load balancer can be connected to the first to form a cluster, where each one monitors the others’ health. Each one is equally capable of failure detection and recovery.



In the event the main load balancer fails, DNS must take users to the to the second load balancer. Because DNS changes can take a considerable amount of time to be propagated on the Internet and to make this failover automatic, many administrators will use systems that allow for flexible IP address remapping, such as floating IPs. On demand IP address remapping eliminates the propagation and caching issues inherent in DNS changes by providing a static IP address that can be easily remapped when needed. The domain name can remain associated with the same IP address, while the IP address itself is moved between servers.

This is how a highly available infrastructure using Floating IPs might look:



<https://f5.com/glossary/load-balancer>

<https://en.wikipedia.org/wiki/Load_balancing_(computing)>

**AIP Management and Gateways**

* Application program interface (**API**) is a set of routines, protocols, and tools for building software applications. An **API** specifies how software components should interact. Additionally, **APIs** are used when programming graphical user interface (GUI) components.
* **API management** is the process of creating and publishing web APIs, enforcing their usage policies, controlling access, nurturing the subscriber community, collecting and analyzing usage statistics, and reporting on performance.
* **Gateway**: a server that act as an API front-end, receives API requests, enforces throttling and security policies, passes requests to the back-end service and then passes the response back to the requester. A gateway often includes a transformation engine to orchestrate and modify the requests and responses on the fly. A gateway can also provide functionality such collecting analytics data and providing caching. The gateway can provide functionality to support authentication, authorization, security, audit and regulatory compliance.

**REST and SOAP API’s**

**Amazon API Gateway:**

Amazon API Gateway is a fully managed service that makes it easy for developers to create, publish, maintain, monitor, and secure APIs at any scale.

With a few clicks in the AWS Management Console, you can create an API that acts as a “front door” for applications to access data, business logic, or functionality from your back-end services, such as workloads running on Amazon Elastic Compute Cloud (Amazon EC2), code running on AWS Lambda, or any Web application.

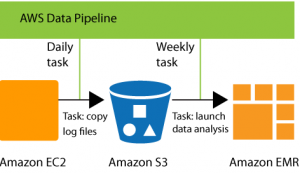
Amazon API Gateway handles all the tasks involved in accepting and processing up to hundreds of thousands of concurrent API calls, including traffic management, authorization and access control, monitoring, and API version management.

Amazon API Gateway has no minimum fees or startup costs. You pay only for the API calls you receive and the amount of data transferred out.

**Benefits:**

* Low-cost efficient
* Performance at any scale
* Easily monitor API activity
* Streamline API Development
* Flexible Security Controls
* Run your API’s without servers
* Create RESTful endpoints for existing servers

**Data Pipeline in AWS**



AWS Data Pipeline is a web service that helps you reliably process and move data between different AWS compute and storage services, as well as on-premise data sources, at specified intervals.

With AWS Data Pipeline, you can regularly access your data where it’s stored, transform and process it at scale, and efficiently transfer the results to AWS services such as Amazon S3, Amazon RDS, Amazon DynamoDB, and Amazon EMR.

AWS Data Pipeline helps you easily create complex data processing workloads that are fault tolerant, repeatable, and highly available.

You don’t have to worry about ensuring resource availability, managing inter-task dependencies, retrying transient failures or timeouts in individual tasks, or creating a failure notification system.

AWS Data Pipeline also allows you to move and process data that was previously locked up in on-premise data silos.

# Processor State Control for Your EC2 Instance

C-states control the sleep levels that a core can enter when it is idle. C-states are numbered starting with C0 (the shallowest state where the core is totally awake and executing instructions) and go to C6 (the deepest idle state where a core is powered off).

P-states control the desired performance (in CPU frequency) from a core. P-states are numbered starting from P0 (the highest performance setting where the core is allowed to use Intel Turbo Boost Technology to increase frequency if possible), and they go from P1 (the P-state that requests the maximum baseline frequency) to P15 (the lowest possible frequency).

The following instance types provide the ability for an operating system to control processor C-states and P-states:

* c4.8xlarge
* d2.8xlarge
* f1.16xlarge
* g3.16xlarge
* i3.16xlarge
* m4.10xlarge
* m4.16xlarge
* p2.16xlarge
* r4.8xlarge
* r4.16xlarge
* x1.16xlarge
* x1.32xlarge

You might want to change the C-state or P-state settings to increase processor performance consistency, reduce latency, or tune your instance for a specific workload.

The default C-state and P-state settings provide maximum performance, which is optimal for most workloads.

However, if your application would benefit from reduced latency at the cost of higher single- or dual-core frequencies, or from consistent performance at lower frequencies as opposed to bursty Turbo Boost frequencies, consider experimenting with the C-state or P-state settings that are available to these instances.

# Placement Groups

A placement group is a logical grouping of instances within a single Availability Zone. Placement groups are recommended for applications that benefit from low network latency, high network throughput, or both. To provide the lowest latency, and the highest packet-per-second network performance for your placement group, choose an instance type that supports enhanced networking. For more information, see Enhanced Networking.

First, you create a placement group and then you launch multiple instances into the placement group. We recommend that you launch the number of instances that you need in the placement group in a single launch request and that you use the same instance type for all instances in the placement group. If you try to add more instances to the placement group later, or if you try to launch more than one instance type in the placement group, you increase your chances of getting an insufficient capacity error.

There is no charge for creating a placement group.

If you stop an instance in a placement group and then start it again, it still runs in the placement group. However, the start fails if there isn't enough capacity for the instance.

If you receive a capacity error when launching an instance in a placement group that already has running instances, stop and start all of the instances in the placement group, and try the launch again. Restarting the instances may migrate them to hardware that has capacity for all the requested instances.

## **Placement Group Limitations**

Placement groups have the following limitations:

* A placement group can't span multiple Availability Zones.
* The name you specify for a placement group must be unique within your AWS account.
* The following are the only instance types that you can use when you launch an instance into a placement group:
  + General purpose: m4.large | m4.xlarge | m4.2xlarge | m4.4xlarge | m4.10xlarge | m4.16xlarge
  + Compute optimized: c4.large | c4.xlarge | c4.2xlarge | c4.4xlarge | c4.8xlarge | c3.large | c3.xlarge | c3.2xlarge | c3.4xlarge | c3.8xlarge | cc2.8xlarge
  + Memory optimized: cr1.8xlarge | r3.large | r3.xlarge | r3.2xlarge | r3.4xlarge | r3.8xlarge | r4.large | r4.xlarge | r4.2xlarge | r4.4xlarge | r4.8xlarge | r4.16xlarge | x1.16xlarge | x1.32xlarge
  + Storage optimized: d2.xlarge | d2.2xlarge | d2.4xlarge | d2.8xlarge | hi1.4xlarge | hs1.8xlarge | i2.xlarge | i2.2xlarge | i2.4xlarge | i2.8xlarge | i3.large | i3.xlarge | i3.2xlarge | i3.4xlarge | i3.8xlarge | i3.16xlarge
  + Accelerated computing: cg1.4xlarge | f1.2xlarge | f1.16xlarge | g2.2xlarge | g2.8xlarge | g3.4xlarge | g3.8xlarge | g3.16xlarge | p2.xlarge | p2.8xlarge | p2.16xlarge
* The maximum network throughput speed of traffic between two instances in a placement group is limited by the slower of the two instances. For applications with high-throughput requirements, choose an instance type with 10 Gbps or 20 Gbps network connectivity.
* Although launching multiple instance types into a placement group is possible, this reduces the likelihood that the required capacity will be available for your launch to succeed. We recommend using the same instance type for all instances in a placement group.
* You can't merge placement groups. Instead, you must terminate the instances in one placement group, and then relaunch those instances into the other placement group.
* A placement group can span peered VPCs; however, you will not get full-bisection bandwidth between instances in peered VPCs.
* You can't move an existing instance into a placement group. You can create an AMI from your existing instance, and then launch a new instance from the AMI into a placement group.
* Reserved Instances provide a capacity reservation for EC2 instances in an Availability Zone. The capacity reservation can be used by instances in a placement group that are assigned to the same Availability Zone. However, it is not possible to explicitly reserve capacity for a placement group.
* To ensure that network traffic remains within the placement group, members of the placement group must address each other via their private IPv4 addresses or IPv6 addresses (if applicable). If members address each other using their public IPv4 addresses, throughput drops to 5 Gbps or less.
* Network traffic to and from resources outside the placement group is limited to 5 Gbps.

## **Launching Instances into a Placement Group**

We suggest that you create an AMI specifically for the instances that you'll launch into a placement group.

**To launch instances into a placement group using the console**

1. Open the Amazon EC2 console at https://console.aws.amazon.com/ec2/.
2. Create an AMI for your instances.
   1. From the Amazon EC2 dashboard, choose **Launch Instance**. After you complete the wizard, choose **Launch**.
   2. Connect to your instance. (For more information, see Connect to Your Linux Instance.)
   3. Install software and applications on the instance, copy data, or attach additional Amazon EBS volumes.
   4. (Optional) If your instance type supports enhanced networking, ensure that this feature is enabled by following the procedures in Enhanced Networking on Linux.
   5. In the navigation pane, choose **Instances**, select your instance, choose **Actions**, **Image**, **Create Image**. Provide the information requested by the **Create Image** dialog box, and then choose **Create Image**.
   6. (Optional) You can terminate this instance if you have no further use for it.
3. Create a placement group.
   1. In the navigation pane, choose **Placement Groups**.
   2. Choose **Create Placement Group**.
   3. In the **Create Placement Group** dialog box, provide a name for the placement group that is unique in the AWS account you're using, and then choose **Create**.

When the status of the placement group is available, you can launch instances into the placement group.

1. Launch instances into your placement group.
   1. In the navigation pane, choose **Instances**.
   2. Choose **Launch Instance**. Complete the wizard as directed, taking care to do the following:

* On the **Choose an Amazon Machine Image (AMI)** page, select the **My AMIs** tab, and then select the AMI that you created.
* On the **Choose an Instance Type** page, select an instance type that can be launched into a placement group.
* On the **Configure Instance Details** page, enter the total number of instances that you'll need in this placement group, as you might not be able to add instances to the placement group later on.
* On the **Configure Instance Details** page, select the placement group that you created from **Placement group**. If you do not see the **Placement group** list on this page, verify that you have selected an instance type that can be launched into a placement group, as this option is not available otherwise.

**To launch instances into a placement group using the command line**

1. Create an AMI for your instances using one of the following commands:
   * create-image (AWS CLI)
   * New-EC2Image (AWS Tools for Windows PowerShell)
2. Create a placement group using one of the following commands:
   * create-placement-group (AWS CLI)
   * New-EC2PlacementGroup (AWS Tools for Windows PowerShell)
3. Launch instances into your placement group using one of the following options:
   * --placement with run-instances (AWS CLI)
   * -PlacementGroup with New-EC2Instance (AWS Tools for Windows PowerShell)

## **Deleting a Placement Group**

You can delete a placement group if you need to replace it or no longer need a placement group. Before you can delete your placement group, you must terminate all instances that you launched into the placement group.

**To delete a placement group using the console**

1. Open the Amazon EC2 console at https://console.aws.amazon.com/ec2/.
2. In the navigation pane, choose **Instances**.
3. Select and terminate all instances in the placement group. (You can verify that the instance is in a placement group before you terminate it by checking the value of **Placement Group** in the details pane.)
4. In the navigation pane, choose **Placement Groups**.
5. Select the placement group, and then choose **Delete Placement Group**.
6. When prompted for confirmation, choose **Yes, Delete**.

**To delete a placement group using the command line**

You can use one of the following sets of commands.

* terminate-instances and delete-placement-group (AWS CLI)
* Stop-EC2Instance and Remove-EC2PlacementGroup(AWS Tools for Windows PowerShell)

**Amazon EMR**

Amazon EMR provides a managed Hadoop framework that makes it easy, fast, and cost-effective to process vast amounts of data across dynamically scalable Amazon EC2 instances. You can also run other popular distributed frameworks such as Apache Spark, HBase, Presto, and Flink in Amazon EMR, and interact with data in other AWS data stores such as Amazon S3 and Amazon DynamoDB.

Amazon EMR securely and reliably handles a broad set of big data use cases, including log analysis, web indexing, data transformations (ETL), machine learning, financial analysis, scientific simulation, and bioinformatics.

**AWS Snowball**

AWS Snowball is a service that accelerates transferring large amounts of data into and out of AWS using physical storage appliances, bypassing the Internet. Each AWS Snowball appliance type can transport data at faster-than internet speeds. This transport is done by shipping the data in the appliances through a regional carrier. The appliances are rugged shipping containers, complete with E Ink shipping labels.

With a Snowball, you can transfer hundreds of terabytes or petabytes of data between your on-premises data centers and Amazon Simple Storage Service (Amazon S3). AWS Snowball uses Snowball appliances and provides powerful interfaces that you can use to create jobs, transfer data, and track the status of your jobs through to completion. By shipping your data in Snowballs, you can transfer large amounts of data at a significantly faster rate than if you were transferring that data over the Internet, saving you time and money.

There are many options for transferring your data into AWS. Snowball is intended for transferring large amounts of data. If you want to transfer less than 10 terabytes of data between your on-premises data centers and Amazon S3, Snowball might not be your most economical choice.

Snowball uses Snowball appliances shipped through your region's carrier. Each Snowball is protected by AWS Key Management Service (AWS KMS) and made physically rugged to secure and protect your data while the Snowball is in transit. In the US regions, Snowballs come in two sizes: 50 TB and 80 TB. All other regions have 80 TB Snowballs only.

## **Snowball Features**

Snowball with the Snowball appliance has the following features:

* You can import and export data between your on-premises data storage locations and Amazon S3.
* Snowball has an 80 TB model available in all regions, and a 50 TB model only available in the US regions.
* Encryption is enforced, protecting your data at rest and in physical transit.
* You don't have to buy or maintain your own hardware devices.
* You can manage your jobs through the AWS Snowball Management Console, or programmatically with the job management API.
* You can perform local data transfers between your on-premises data center and a Snowball can be done through the Snowball client, a standalone downloadable client, or programmatically using Amazon S3 REST API calls with the downloadable Amazon S3 Adapter for Snowball. For more information, see Transferring Data with a Snowball.
* The Snowball is its own shipping container, and its E Ink display changes to show your shipping label when the Snowball is ready to ship. For more information, see Shipping Considerations for AWS Snowball.
* For a list of regions where the Snowball appliance is available,

**AWS Cloud Trail**

AWS CloudTrail is an AWS service that helps you enable governance, compliance, and operational and risk auditing of your AWS account. Actions taken by a user, role, or an AWS service are recorded as events in CloudTrail. Events include actions taken in the AWS Management Console, AWS Command Line Interface, and AWS SDKs and APIs.

Visibility into your AWS account activity is a key aspect of security and operational best practices. You can use CloudTrail to view, search, download, archive, analyze, and respond to account activity across your AWS infrastructure. You can identify who or what took which action, what resources were acted upon, when the event occurred, and other details to help you analyze and respond to activity in your AWS account.

You can integrate CloudTrail into applications using the API, automate trail creation for your organization, check the status of trails you create, and control how users view CloudTrail events.

CloudTrail is enabled on your AWS account when you create it. When activity occurs in your AWS account, that activity is recorded in a CloudTrail event. You can easily view events in the CloudTrail console by going to **Event history**.

Event history allows you to view, search, and download the past seven days of supported activity in your AWS account. In addition, you can create a CloudTrail trail to further archive, analyze, and respond to changes in your AWS resources. A trail is a configuration that enables delivery of events to an Amazon S3 bucket that you specify. You can also deliver and analyze events in a trail with Amazon CloudWatch Logs and Amazon CloudWatch Events. You can create a trail with the CloudTrail console, the AWS CLI, or the CloudTrail API.

You can create two types of trails:

**A trail that applies to all regions**

When you create a trail that applies to all regions, CloudTrail creates the same trail in each region. It then records events in each region and delivers the CloudTrail event log files to an S3 bucket that you specify. This is the default option when you create a trail in the CloudTrail console.

**A trail that applies to one region**

When you create a trail that applies to one region, CloudTrail records the log files in that region only. It then delivers the CloudTrail event log files log to an S3 bucket that you specify. If you create additional single trails, you can have those trails deliver CloudTrail event log files to the same Amazon S3 bucket or to separate buckets.

For both types of trails, you can specify an Amazon S3 bucket from any region.

By default, CloudTrail event log files are encrypted using Amazon S3 server-side encryption (SSE). You can also choose to encrypt your log files with an AWS Key Management Service (AWS KMS) key. You can store your log files in your bucket for as long as you want. You can also define Amazon S3 lifecycle rules to archive or delete log files automatically. If you want notifications about log file delivery and validation, you can set up Amazon SNS notifications.

CloudTrail typically delivers log files within 15 minutes of account activity. In addition, CloudTrail publishes log files multiple times an hour, about every five minutes. These log files contain API calls from services in the account that support CloudTrail.

**What Is Auto Scaling?**

Auto Scaling helps you ensure that you have the correct number of Amazon EC2 instances available to handle the load for your application. You create collections of EC2 instances, called *Auto Scaling groups*. You can specify the minimum number of instances in each Auto Scaling group, and Auto Scaling ensures that your group never goes below this size. You can specify the maximum number of instances in each Auto Scaling group, and Auto Scaling ensures that your group never goes above this size. If you specify the desired capacity, either when you create the group or at any time thereafter, Auto Scaling ensures that your group has these many instances. If you specify scaling policies, then Auto Scaling can launch or terminate instances as demand on your application increases or decreases.

For example, the following Auto Scaling group has a minimum size of 1 instance, a desired capacity of 2 instances, and a maximum size of 4 instances. The scaling policies that you define adjust the number of instances, within your minimum and maximum number of instances, based on the criteria that you specify.


   An illustration of a basic Auto Scaling group.
  

## Auto Scaling Components

The following table describes the key components of Auto Scaling.

|  |  |
| --- | --- |
| A graphic representing an Auto Scaling group. | **Groups**  Your EC2 instances are organized into groups so that they can be treated as a logical unit for the purposes of scaling and management. When you create a group, you can specify its minimum, maximum, and, desired number of EC2 instances. |
| A graphic representing a launch configuration. | **Launch configurations**  Your group uses a launch configuration as a template for its EC2 instances. When you create a launch configuration, you can specify information such as the AMI ID, instance type, key pair, security groups, and block device mapping for your instances. |
| A graphic representing a launch configuration. | **Scaling plans**  A scaling plan tells Auto Scaling when and how to scale. For example, you can base a scaling plan on the occurrence of specified conditions (dynamic scaling) or on a schedule. |

**Minimum, Maximum and Desired values:**

The ASG will always try to maintain the Desired Capacity. If you scale up or down, and the new number of instances is different than the Desired Capacity, the ASG will add or remove an instance to go back to the desired capacity.

If you use scaling policies, and the policy condition is met, the ASG will change the Desired Capacity to match the result of your scaling policy. E.g., you have a Desired Capacity of 2, and a policy that says to scale up if the CPU utilization goes over a threshold.

If that policy is fulfilled, then the Desired Capacity will increase to 3, and so on.

So manually scaling up and down will result in your ASG restoring the number of instances to the Desired Capacity.

If you want to manually scale up and down, you could set your Max and Min to a wide value, and move Desired Capacity within it.

So, you could do Max=10, Min=1, Desired=3. Then you could scale up or down just by changing the Desired Capacity. (This is how we use Auto Scaling, and I think why I gave you a bad answer before.)

If you want to terminate an instance and change the Desired Capacity at the same time, the CLI can do that.

**Route 53**

The **name** is a reference to TCP or UDP port **53**, where DNS server requests are addressed.

# Choosing a Routing Policy

When you create a resource record set, you choose a routing policy, which determines how Amazon Route 53 responds to queries:

* **Simple routing policy** – Use for a single resource that performs a given function for your domain, for example, a web server that serves content for the example.com website.
* **Failover routing policy** – Use when you want to configure active-passive failover.
* **Geolocation routing policy** – Use when you want to route traffic based on the location of your users.
* **Latency routing policy** – Use when you have resources in multiple locations and you want to route traffic to the resource that provides the best latency.
* **Multivalue answer routing policy** – Use when you want Amazon Route 53 to respond to DNS queries with up to eight healthy records selected at random.
* **Weighted routing policy** – Use to route traffic to multiple resources in proportions that you specify.

**Reserved IP Addresses in AWS**

* 10.0.0.0: Network address.
* 10.0.0.1: Reserved by AWS for the VPC router.
* 10.0.0.2: Reserved by AWS. The IP address of the DNS server is always the base of the VPC network range plus two; however, we also reserve the base of each subnet range plus two. For more information, see Amazon DNS Server.
* 10.0.0.3: Reserved by AWS for future use.
* 10.0.0.255: Network broadcast address. We do not support broadcast in a VPC, therefore we reserve this address.

# VPC Peering

A VPC peering connection is a networking connection between two VPCs that enables you to route traffic between them using private IPv4 addresses or IPv6 addresses. Instances in either VPC can communicate with each other as if they are within the same network. You can create a VPC peering connection between your own VPCs, or with a VPC in another AWS account. In both cases, the VPCs must be in the same region.

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.

**Connection Draining:**

To ensure that a Classic Load Balancer stops sending requests to instances that are de-registering or unhealthy, while keeping the existing connections open, use *connection draining*. This enables the load balancer to complete in-flight requests made to instances that are de-registering or unhealthy.

When you enable connection draining, you can specify a maximum time for the load balancer to keep connections alive before reporting the instance as de-registered. The maximum timeout value can be set between 1 and 3,600 seconds (the default is 300 seconds). When the maximum time limit is reached, the load balancer forcibly closes connections to the de-registering instance.

While in-flight requests are being served, the load balancer reports the state of a de-registering instance as InService: Instance deregistration currently in progress. When the de-registering instance is finished serving all in-flight requests, or when the maximum timeout limit is reached, the load balancer reports the instance state as OutOfService: Instance is not currently registered with the LoadBalancer.

If an instance becomes unhealthy, the load balancer reports the instance state as OutOfService. If there are in-flight requests made to the unhealthy instance, they are completed. The maximum timeout limit does not apply to connections to unhealthy instances.

If your instances are part of an Auto Scaling group and connection draining is enabled for your load balancer, Auto Scaling waits for the in-flight requests to complete, or for the maximum timeout to expire, before terminating instances due to a scaling event or health check replacement.

You can disable connection draining if you want your load balancer to immediately close connections to the instances that are de-registering or have become unhealthy. When connection draining is disabled, any in-flight requests made to instances that are de-registering or unhealthy are not completed.

**Terms in Load Balancing**

**LTM:** Local Traffic Manager; **GTM:** Global Traffic Manager

**Difference between GTM and LTM**

The **Global Traffic Manager**, aka **GTM**, and now referred to **DNS**, is one of the cutting-edge modules offered on F5 Networks BIG-IP® platform. “Global” is the right word for this module because it has the ability to make name resolution load balancing decisions for systems located anywhere in the World, not just the US. You can think of the GTM as an intelligent DNS that is security minded. In other words, its’ logic can make informed decisions on correlating a hostname to an IP address while keeping security in check.

Most things you do on the Internet or Private networks start with name resolution – so it makes sense if you’re going to load balance an application it would start at this layer – resolving names to IPs based on availability, Performance, and even Persistence. It’s important to note, traffic does not “route” through the GTM, the GTM simply tells you the best IP to route to based on metrics for the URL in question. That IP can be almost anything really, but usually it’s an actual server, or a virtual IP that fronts multiple servers. Like a traditional DNS architecture, you usually have multiple GTMs in your architecture, this is for redundancy/availability.

The main configuration element in a GTM is called a Wide IP or WIP for short. There are many configuration elements that work in concert with a WIP, but at the base of it all is the Wide IP. A WIP equates to the common URL you’re load balancing, for example www.yourcompany.com. A pool or pools are usually attached to a WIP which contain the IPs it’s intelligently resolving. Like your run of the mill DNS server, the GTM does not tell the requester any information about ports. Though, the monitors associated with the pool members can indeed monitor availability or performance on ports.

## **LTM – Local Traffic Manager Overview**

The **Local Traffic Manager**, aka **LTM**, is the most popular module offered on F5 Networks BiG-IP® platform. The real power of the LTM is it’s a Full Proxy, allowing you to augment client and server-side connections. All while making informed load balancing decisions on availability, performance, and persistence. “Local” in the name is important, opposed to the GTM, traffic actually flows through the LTM to the servers it balances traffic to. Usually the servers it’s load balancing sit “locally” in the same data center as the LTM, though that is not a requirement. With SNAT (secure network address translation) configured on the VIP, if you can route to it you can load balance it – so it’s possible to have servers in different data centers be a part of the same pool in an LTM VIP.

The main configuration element on an LTM is the Virtual IP or VIP for short. There is a plethora of configuration elements that work with VIPs, but at the heart of the technology it’s a VIP they are all a part of. Like a WIP, VIPs equate to the URL you’re load balancing, but at its lowest level. Like a WIP it usually contains a pool with the servers it’s load balancing & monitor(s) to measure availability / performance.

## **Some of the Key differences of the GTM vs LTM**

* The biggest difference between the GTM and LTM, as mentioned earlier, is traffic doesn’t actually flow through the GTM to your servers.
* The GTM is an intelligent name resolver, intelligently resolving names to IP addresses.
* Once the GTM provides you with an IP to route to you’re done with the GTM until you ask it to resolve another name for you.
* Similar to a usual DNS server, the GTM does not provide any port information in its resolution.
* The LTM doesn’t do any name resolution and assumes a DNS decision has already been made.
* When traffic is directed to the LTM traffic flows directly through its’ full proxy architecture to the servers it’s load balancing.
* Since the LTM is a full proxy it’s easy for it to listen on one port but direct traffic to multiple hosts listening on any port specified.

## **How do the GTM & LTM work together?**

The GTM and LTM can work together or they can be totally independent. If your organization owns both modules it’s usually using them together, and that’s where the real power comes in. They do this via a proprietary protocol called iQuery. iQuery, functioning on TCP port 4353, reports VIP availability / performance back to the GTMs. The GTMs can then dynamically resolve VIPs that live on an LTM(s).

When a GTM has LTMs as servers in its’ configuration there is no need to monitor the actual VIP(s) with application monitors, as the LTM is doing that & iQuery reports the information back to the GTM.

As you can see the GTM & LTM modules are powerful tools in the world of Application Delivery / Load Balancing. Together the GTM & LTM make one mean lean Application Delivery machine.

# Amazon EC2:

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](https://aws.amazon.com/what-is-cloud-computing/). 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](http://developer.amazonwebservices.com/connect/kbcategory.jspa?categoryID=87), 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](https://aws.amazon.com/contact-us/ec2-request/).

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](https://aws.amazon.com/ec2/purchasing-options/spot-instances/) work.

If you prefer, you can also perform all these actions from the [AWS Management Console](https://aws.amazon.com/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 storage (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 is dependant 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 Simple Storage Service (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 Amazon Machine Images (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?](https://aws.amazon.com/ec2/faqs/#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 up to at total of 20 On-Demand instances across the instance family, purchasing 20 Reserved Instances, and requesting Spot Instances per your [dynamic Spot limit](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-spot-limits.html) per region. New AWS accounts may start with limits that are lower than the limits described here. Certain instance types are further limited per region as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Instance Type | On-Demand Limit | Reserved Limit | Spot Limit |
| m4.4xlarge | 10 | 20 | Dynamic Spot Limit |
| m4.10xlarge | 5 | 20 | Dynamic Spot Limit |
| m4.16xlarge | 5 | 20 | Dynamic Spot Limit |
| c4.4xlarge | 10 | 20 | Dynamic Spot Limit |
| c4.8xlarge | 5 | 20 | Dynamic Spot Limit |
| cg1.4xlarge | 2 | 20 | Dynamic Spot Limit |
| hi1.4xlarge | 2 | 20 | Dynamic Spot Limit |
| hs1.8xlarge | 2 | 20 | Not offered |
| cr1.8xlarge | 2 | 20 | Dynamic Spot Limit |
| p2.xlarge | 1 | 20 | Dynamic Spot Limit |
| p2.8xlarge | 1 | 20 | Dynamic Spot Limit |
| p2.16xlarge | 1 | 20 | Dynamic Spot Limit |
| g3.4xlarge | 1 | 20 | Dynamic Spot Limit |
| g3.8xlarge | 1 | 20 | Dynamic Spot Limit |
| g3.16xlarge | 1 | 20 | Dynamic Spot Limit |
| r4.large | 20 | 20 | Dynamic Spot Limit |
| r4.xlarge | 20 | 20 | Dynamic Spot Limit |
| r4.2xlarge | 20 | 20 | Dynamic Spot Limit |
| r4.4xlarge | 10 | 20 | Dynamic Spot Limit |
| r4.8xlarge | 5 | 20 | Dynamic Spot Limit |
| r4.16xlarge | 1 | 20 | Dynamic Spot Limit |
| r3.4xlarge | 10 | 20 | Dynamic Spot Limit |
| r3.8xlarge | 5 | 20 | Dynamic Spot Limit |
| i3.large | 2 | 20 | Dynamic Spot limit |
| i3.xlarge | 2 | 20 | Dynamic Spot limit |
| i3.2xlarge | 2 | 20 | Dynamic Spot limit |
| i3.4xlarge | 2 | 20 | Dynamic Spot limit |
| i3.8xlarge | 2 | 20 | Dynamic Spot limit |
| i3.8xlarge | 2 | 20 | Dynamic Spot limit |
| i3.16xlarge | 2 | 20 | Dynamic Spot limit |
| i2.2xlarge | 8 | 20 | Dynamic Spot Limit |
| i2.4xlarge | 4 | 20 | Dynamic Spot Limit |
| i2.8xlarge | 2 | 20 | Dynamic Spot Limit |
| d2.4xlarge | 10 | 20 | Dynamic Spot Limit |
| d2.8xlarge | 5 | 20 | Dynamic Spot Limit |
| t2.nano | 20 | 20 | Not offered |
| t2.micro | 20 | 20 | Not offered |
| t2.small | 20 | 20 | Not offered |
| t2.medium | 20 | 20 | Not offered |
| t2.large | 20 | 20 | Not offered |
| t2.xlarge | 20 | 20 | Not offered |
| t2.2xlarge | 20 | 20 | Not offered |
| All Other Instance Types | 20 | 20 | Dynamic Spot Limit |

*Note that cc2.8xlarge, cg1.4xlarge, hi1.4xlarge, hs1.8xlarge, cr1.8xlarge, G2, D2, and I2 instances are not available in all regions.*

If you need more instances, complete the [Amazon EC2 instance request form](https://aws.amazon.com/support/createCase?type=service_limit_increase&serviceLimitIncreaseType=ec2-instances) with your use case and your instance increase will be considered. Limit increases are tied to the region they were requested for.

Q: Are there any limitations in sending email from EC2 instances?

Yes. In order to maintain the quality of 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](https://portal.aws.amazon.com/gp/aws/html-forms-controller/contactus/ec2-email-limit-rdns-request).

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.

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](https://console.aws.amazon.com/cloudtrail/home).  For more information, visit the [CloudTrail](https://aws.amazon.com/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](https://d1.awsstatic.com/whitepapers/aws-security-whitepaper.8a79838e616a8c3914a452a39d765c6caf6bd9a3.pdf) white paper and to our [Amazon EC2 running Windows Security Guide](http://developer.amazonwebservices.com/connect/entry.jspa?externalID=1767).

## **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](https://aws.amazon.com/contact-us/eip_limit_request/). 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 address is associated exclusively with the instance and is only returned to Amazon EC2 when the instance is stopped or 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?

Yes, you can configure the reverse DNS record of your Elastic IP address by [filling out this form](https://portal.aws.amazon.com/gp/aws/html-forms-controller/contactus/ec2-email-limit-rdns-request). Note that a corresponding forward DNS record pointing to that Elastic IP address must exist before we can create the reverse DNS record.

## **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](http://aws.amazon.com/about-aws/global-infrastructure/regional-product-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.

## **Amazon Elastic Block Storage (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](https://aws.amazon.com/ebs/details/), and for additional information on performance, see the [Amazon EC2 User Guide's EBS Performance section](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSPerformance.html).

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](https://aws.amazon.com/ebs/details/#ebsoptimized).

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](https://aws.amazon.com/ebs/details/).

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](http://developer.amazonwebservices.com/connect/kbcategory.jspa?categoryID=243). 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](https://aws.amazon.com/ebs/faqs/).

## **Amazon Elastic File Storage (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](http://docs.aws.amazon.com/efs/latest/ug/gs-mount-fs-on-ec2instance-and-test.html).

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](http://docs.aws.amazon.com/efs/latest/ug/gs-mount-fs-on-ec2instance-and-test.html).

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](https://aws.amazon.com/efs/faq/#on-premises) of the Amazon EFS FAQ.

For more information about moving data to the Amazon cloud, please see the [Cloud Data Migration page](https://aws.amazon.com/cloud-data-migration/).

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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/vpc-classiclink.html). On-premises servers can mount your file systems via an [AWS Direct Connect](https://aws.amazon.com/directconnect/) 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](https://aws.amazon.com/efs/).

## **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.

## **Auto Scaling**

Q: Can I scale up my Amazon EC2 capacity fast but scale it down slowly?

Yes. For example, you can define a scale up condition to increase your Amazon EC2 capacity by 10% and a scale down condition to decrease it by 5%.

Q: What happens if a scaling activity causes me to reach my Amazon EC2 limit of instances?

Auto Scaling Service 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](https://aws.amazon.com/contact-us/ec2-request/).

Q: What happens to my Amazon EC2 instances if I delete my Auto Scaling Group?

If you have an Auto Scaling group with running instances and you choose to delete the Auto Scaling group, the instances will be terminated and the Auto Scaling group will be deleted.

## **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](https://aws.amazon.com/elasticloadbalancing/classicloadbalancer/) that routes traffic based on either application or network level information, and the [Application Load Balancer](https://aws.amazon.com/elasticloadbalancing/applicationloadbalancer/) 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](https://aws.amazon.com/elasticloadbalancing/) for more information.

## **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](https://aws.amazon.com/ec2/faqs/#convertible) 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](https://aws.amazon.com/ec2/faqs/#convertible).

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 m4.2xlarge Linux/Unix regional RI with default tenancy in US East (N.Virginia). Then this RI’s discounted rate can automatically apply to two m4.xlarge instances in us-east-1a or four m4.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.

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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts-reserved-instances-application.html#apply_ri) 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 hourly billing, and how it applies to RIs, see [Billing Benefits and Payment Options](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts-reserved-instances-application.html).

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 m4.2xlarge RI that is scoped to a region, then your discounted rate could apply towards the usage of 1 m4.2xlarge or 2 m4.xlarge instances.

[Click here](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts-reserved-instances-application.html#apply_ri) to learn more about how instance size flexibility of RIs applies to your EC2 usage. And [click here](http://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/billing-reports.html#enhanced-RI) 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 |
| 10xlarge | 80 |
| 16xlarge | 128 |
| 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](http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/concepts-reserved-instances-tiers.html) portion of the [Amazon EC2 User Guide](http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/Welcome.html).

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](http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/concepts-reserved-instances-tiers.html) portion of the [Amazon EC2 User Guide](http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/Welcome.html).

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](http://aws.amazon.com/contact-us/aws-sales/) about receiving discounts beyond those that are automatically provided.

## **Spot Instances**

Q. What is a Spot Instance?

Spot instances are a new way to purchase and consume Amazon EC2 Instances. They allow customers to bid on unused EC2 capacity and run those instances for as long as their bid exceeds the current Spot Price. The Spot Price changes periodically based on supply and demand, and customers whose bids meet or exceed it gain access to the available Spot instances. Spot instances are complementary to On-Demand instances and Reserved Instances, providing another option for obtaining compute capacity.

Q. How is a Spot instance different than an On-Demand instance or Reserved Instance?

Spot instances provide the ability for customers to purchase compute capacity with no upfront commitment, at hourly rates usually lower than the On-Demand rate. Spot instances allow you to specify the maximum hourly price that you are willing to pay to run a particular instance type. Amazon EC2 sets a Spot Price for each instance type in each availability zone, which is the hourly price all customers will pay to run a Spot instance for that given period. The Spot Price fluctuates based on supply and demand for instances, but customers will never pay more than the maximum price they have specified. If the Spot Price moves higher than a customer’s maximum price, the customer’s instance will be shut down by Amazon EC2. Other than those differences, Spot instances perform exactly the same as On-Demand or Reserved Instances. See [here](https://aws.amazon.com/ec2/purchasing-options/spot-instances/) for more details on Spot instances.

Q. How do I purchase and start up a Spot instance?

Spot instances can be requested using the [EC2 Management Console](https://aws.amazon.com/console/) or Amazon EC2 APIs. To start with the EC2 Management Console:

1. Log in to the EC2 Management Console.
2. Choose "Spot Requests" in the left navigation pane.
3. Choose "Request Spot Instances".
4. Complete the Launch Instance Wizard process, choosing an AMI, region and instance size and type.
5. Enter the number of Spot instances you would like to request, your maximum price, and whether the request is persistent or not.
6. After choosing your key pair and security group(s), you are ready to submit your Spot instance request.

For detail on how to request Spot instances through the Amazon EC2 API, see the [Amazon EC2 API Reference](http://awsdocs.s3.amazonaws.com/EC2/latest/ec2-api.pdf).

For a more detailed walk-through of using Spot instances and more information on how to get the most out of Spot instances, see [Introduction to Spot Instances](http://developer.amazonwebservices.com/connect/entry.jspa?externalID=3260).

Q. How many Spot instances can I request?

You are limited to requesting Spot instances per your dynamic Spot limit for each region. Note that not all instance types are available on Spot, and new AWS accounts might start with a lower limit. To learn more about Spot instance limits, please refer to the Amazon EC2 [User Guide](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-spot-limits.html).  
  
If you would like a higher limit, complete the[Amazon EC2 instance request form](https://aws.amazon.com/support/createCase?serviceLimitIncreaseType=ec2-instances&type=service_limit_increase&isauthcode=true&code=dzkwDcZ52oii6pr9iPYylIvuM78ObPVkQug42ntPetVgfanCcFjKifdSEPXRx2_tIQp6hLDGGq_GSvJ7t1wKy_3x97YfW6IheoEZJhc9I_rifBs7i_76ZG8f5qZ8EU10kPjq1Bm6L-vOdh3iTfdFaObJ_irgWqaWTsMcZrt0Je9t9HV3272554QNYYjLS0PdDnBFbZlalCUf4p37KTyoifrZPgMGdVVZZuRE0KKJIOvEZo3GIw-MpYkcA0MlUqT3jenFQ9V8CYtkM0T7BCfdWmdXN59U4fmS9NKp75gpAvub4PTi1J5GqzS6QRE_UsKy6BXqguDhyQs57Unu93COwLZdwEmqMqo6WBhAi90ZY10XfleVigCgLvcp1Nh7_dPZ2W6Dn4oldC3odsTS4ZrX7tfsa13LsysDgEdgI6TzNeZ3rV1UpRlUwQe8W7tVYLmKQNPHz4DdEKynfj7Zas-RSTAUOJcbGmylTsraI4Vx9mpnZgs8vnFf-3sAzgmuchdoZLx0qmTrJSZWNm7qOLwxUVbNX2kaGocCzioUn1wxcN2AFcyanjFUHuVlxRoSYc4WYGncCZGuCS-bnoNZqYmTxLZniybzAixVS3UecOgOFG8QX8LJkVFkB-I2LUdfp4Cvq1UqT9rB00VKRNW2NPauDVxb2zYpcOabNGylt9dZiitfTsCy3mrULMRPOZQiBRqHyCIQ6Th-UZfT9uelnfpCXG5WD2lEEIN0W7ZZEpkoYbmVrFv23y2ahqXkzXLs3_YqOwDw9ODQt8jpc9_v-lPXaHLwvwOJNEYLzbFDdWo07E6w6dBiXoOBjBgOtFLC8lUU4UmrrMEahS32I-WdShTpBrQAPfx1TVGDAFhPbhEG-DG199LowkOlVAPWVWXZrHmKHxapnpflHWOm6ZNrA_dh4a5xho-ca101el-3AG2CEwIPZw8s0L5f2t9clLrwzuVH6OD5NbH23PptJs3iGVN20bjwPgpzlxThzxKMHziT4wXAcrTOAzFUa7t0v5wtbLNJSPURrXYKWYX7SJmzZELL7z9YKqRYUGCyiVmvyIhOHHjSLhsmvqfWO3lgjeBkNBTeKEbsLzMHwGNlFZKGqnyD8SCGwdrIBvLxanjR2XsAmPX-CrWAJU9KaRfXDZli-OusDX-T1DWcUKoLtJe1hnOiqZgzCgrb7wABkRRpCG8hgmvVaeKPSp-_YvlFbg) with your use case and your instance increase will be considered. Limit increases are tied to the region they were requested for.

Q. How can I determine the status of my Spot request?

You can determine the status of your Spot request in the instance provisioning lifecycle by inspecting its Spot Bid Status code and message. By reviewing Spot bid statuses, you can see why your Spot requests state has or has not changed and you can learn how to optimize your Spot requests to get them fulfilled. You can access Spot Bid Status information on the Spot Instance page of the EC2 console of the AWS Management Console, as well as through the DescribeSpotInstanceRequests API action and the ec2-describe-spot-instance-requests CLI command. For more information, please visit the [Amazon EC2 Developer guide](https://aws.amazon.com/documentation/ec2/).

Q. Are Spot instances available for all instance families and sizes and in all regions?

Instance types supported in each region are listed [here](https://aws.amazon.com/ec2/spot/pricing/). Spot instance APIs are available in all regions except the US GovCloud region.

Q. Which operating systems are available as Spot instances?

Linux/UNIX and Windows Server are available. Windows Server with SQL Server is not currently available.

Q. Are there any features or services of Amazon Web Services that are not supported for use with Spot instances?

Amazon DevPay is not supported for use with Spot instances.

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. Will I be charged if my Spot instance is terminated by Amazon EC2 before the hour is up?

No. If the Spot instance is terminated by Amazon EC2, you will not be charged for a partial hour of usage. However, if you terminate the instance yourself, you will be charged for any hour in which the instance ran.

Q. How often should I expect the Spot price to change?

Amazon EC2 will change the Spot price periodically as new requests are received and as available Spot capacity changes (e.g., due to instance terminations). While the Spot price may change anytime, in general it will change once per hour and in many cases less frequently. We publish the current Spot price and historical prices for Spot instances through the API, and they can also be viewed using the AWS Management Console. This can help you assess the levels and timing of fluctuations in the Spot price over time.

Q. Will all Spot instances started at the same time be charged the same price?

Yes.

Q. Will the price I’m charged for a running Spot instance change during its instance-hour as the Spot price changes?

No. The price per instance-hour for a Spot instance is set at the beginning of each instance-hour for the entire hour. Any changes to the Spot price will not be reflected until the next instance-hour begins.

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 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. Why do Spot prices differ across accounts for the same instance type, operating system, and Availability Zone?

To ensure that resources are distributed across Availability Zones for a region, Availability Zones are independently mapped to identifiers for each account. For example, your Availability Zone us-east-1a might not be the same location as us-east-1a for another account. So, Spot prices for the same Availability Zone identifier may be different in different accounts. Note that there's no way for you to coordinate Availability Zones between accounts.

Q. What is a Spot fleet?

A Spot fleet allows you to automatically bid on 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, and define a target capacity based on your application needs (in units including instances, vCPUs, memory, storage, or network throughput). 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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-fleet.html).

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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-spot-limits.html#spot-fleet-limitations) 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 bid 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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-fleet-requests.html#manage-spot-fleet) section of the Amazon EC2 User Guide to learn how to describe your Spot fleet request's history.

Q. What happens if my Spot fleet request bid price exceeds my Spot bid price limit for one of the instance types I am requesting?

If your Spot fleet request bid price exceeds your Spot bid price limits, we will submit Spot requests for that instance type at your current Spot bid price limit. Your Spot fleet request’s history will show if any of your fleet’s instances were affected by your Spot bid price limit. Visit the [Monitoring Your Spot Fleet](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-fleet-requests.html#manage-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 bids simultaneously, and are subject to the same availability and prices as a single Spot instance request. For example, if no resources are available at your Spot fleet request bid price, we may be unable to fulfill your request partially or in full.

Q. Can I submit a multi-Availability Zone fleet request?

Yes, visit the [Spot Fleet Examples](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-fleet-examples.html) 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 two allocation strategies: lowestPrice and diversified. 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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-fleet.html#spot-fleet-allocation-strategy) to learn more.

Q. Can I tag a Spot fleet request?

We currently do not support tagging Spot fleet requests.

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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-fleet-requests.html#manage-spot-fleet) to learn how to describe your Spot fleet request.

Q. Can I modify my Spot fleet request?

Currently, you can only 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 Spot fleet with Auto Scaling. You cannot use Elastic Load Balancing or Elastic MapReduce to trigger Spot fleet requests.

Q. Does a Spot fleet request terminate Spot instances when they are no longer running in the lowest priced Spot pools and relaunch them in the lowest priced pools?

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.

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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/spot-interruptions.html#spot-instance-termination-notices)), and you will not be charged for the affected instance(s).

## **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](http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/concepts_micro_instances.html) 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.

## **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 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](https://aws.amazon.com/ibm/).

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 ECS:**

Q: What is Amazon EC2 Container Service?  
Amazon EC2 Container Service (ECS) is a highly scalable, high performance container management service that supports Docker containers and allows you to easily run applications on a managed cluster of Amazon EC2 instances. Amazon ECS eliminates the need for you to install, operate, and scale your own cluster management infrastructure. With simple API calls, you can launch and stop container-enabled applications, query the complete state of your cluster, and access many familiar features like security groups, Elastic Load Balancing, EBS volumes and IAM roles. You can use Amazon ECS to schedule the placement of containers across your cluster based on your resource needs and availability requirements. You can also integrate your own scheduler or third-party schedulers to meet business or application specific requirements.

Q: Why should I use Amazon ECS?  
Amazon ECS makes it easy to use containers as a building block for your applications by eliminating the need for you to install, operate, and scale your own cluster management infrastructure. Amazon ECS lets you schedule long-running applications, services, and batch processes using Docker containers. Amazon ECS maintains application availability and allows you to scale your containers up or down to meet your application's capacity requirements. Amazon ECS is integrated with familiar features like Elastic Load Balancing, EBS volumes, VPC, and IAM. Simple APIs let you integrate and use your own schedulers or connect Amazon ECS into your existing software delivery process.

Q: What is the pricing for Amazon ECS?  
There is no additional charge for Amazon ECS. You pay for AWS resources (e.g. EC2 instances or EBS volumes) you create to store and run your application. You only pay for what you use, as you use it; there are no minimum fees and no upfront commitments.

Q: How is Amazon ECS different from AWS Elastic Beanstalk?  
AWS Elastic Beanstalk is an application management platform that helps customers easily deploy and scale web applications and services. It keeps the provisioning of building blocks (e.g., EC2, RDS, Elastic Load Balancing, Auto Scaling, CloudWatch), deployment of applications, and health monitoring abstracted from the user so they can just focus on writing code. You simply specify which container images are to be deployed, the CPU and memory requirements, the port mappings, and the container links. Elastic Beanstalk will automatically handle all the details such as provisioning an Amazon ECS cluster, balancing load, auto-scaling, monitoring, and placing your containers across your cluster.

Elastic Beanstalk is ideal if you want to leverage the benefits of containers but just want the simplicity of deploying applications from development to production by uploading a container image. You can work with Amazon ECS directly if you want more fine-grained control for custom application architectures.

Q: How is Amazon ECS different from AWS Lambda?  
Amazon EC2 Container Service is a highly scalable Docker container management service that allows you to run and manage distributed applications that run in Docker containers. AWS Lambda is an event-driven task compute service that runs your code in response to “events” such as changes in data, website clicks, or messages from other AWS services without you having to manage any compute infrastructure.

## **Using Amazon EC2 Container Service**

Q: How do I get started using Amazon ECS?  
Visit our [Getting Started](https://aws.amazon.com/ecs/getting-started/) page for more information on how to start.

Q: Does EC2 Container Service support any other container types?  
No. Docker is the only container platform supported by EC2 Container Service at this time.

Q: I want to launch containers. Why do I have to launch Tasks?  
Docker encourages you to split your applications up into their individual components, and EC2 Container Service is optimized for this pattern. Tasks allow you to define a set of containers that you would like to be placed together (or part of the same placement decision), their properties, and how they may be linked. Tasks include all the information that EC2 Container Service needs to make the placement decision. To launch a single container, your Task Definition should only include one container definition.

Q: Does Amazon ECS support applications and services?  
Yes. The Amazon ECS Service scheduler can manage long-running applications and services. The Service scheduler helps you maintain application availability and allows you to scale your containers up or down to meet your application's capacity requirements. The Service scheduler allows you to distribute traffic across your containers using Elastic Load Balancing. Amazon ECS will automatically register and deregister your containers from the associated load balancer. The Service scheduler will also automatically recover containers that become unhealthy (fail ELB health checks) or stop running to ensure you have the desired number of healthy containers supporting your application. You can scale your application up and down by changing the number of containers you want the service to run. You can update your application by changing its definition or using a new image. The scheduler will automatically start new containers using the new definition and stop containers running the previous version (waiting for the ELB connections to drain if ELB is used).

Q: Does Amazon ECS support dynamic port mapping?  
Yes. It is possible to associate a service on Amazon EC2 Container Service (ECS) to an Application Load Balancer for the Elastic Load Balancing (ELB) service. The Application Load Balancer supports a target group that contains a set of instance:ports. You can specify a dynamic port in the ECS task definition which gives the container an unused port when it is scheduled on the EC2 instance. The ECS scheduler will automatically add the task to the Application Load Balancer’s target group using this port.

Q: Does Amazon ECS support batch jobs?  
Yes. You can use Amazon ECS Run task to run one or more tasks once. Run task starts the task on an instance that meets the task’s requirements including CPU, memory and ports.

Q: Can I use my own scheduler with Amazon ECS?  
ECS provides Blox, a collection of open source projects for container management and orchestration. Blox makes it easy to consume events from Amazon ECS, store the cluster state locally and query the local data store through APIs. Blox also includes a daemon scheduler that can be used as a reference for how to use the cluster state server. See the [Blox GitHub page](https://blox.github.io/) to learn more.

Q: Can I use my own AMI?  
Yes. You can use any AMI that meets the Amazon ECS AMI specification. We recommend starting from the Amazon ECS-enabled Amazon Linux AMI. Partner AMIs compatible with Amazon ECS are also available. You can review the Amazon ECS AMI specification in the documentation.

Q: How can I configure my container instances to pull from Amazon EC2 Container Registry?  
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 attach the [AmazonEC2ContainerServiceforEC2Role](http://docs.aws.amazon.com/AmazonECS/latest/developerguide/instance_IAM_role.html) to your instances. Then Amazon ECS will retrieve the appropriate images for your applications.

## **Security**

Q: How does Amazon ECS isolate containers belonging to different customers?  
Amazon ECS schedules containers for execution on customer-controlled Amazon EC2 instances and builds on the same isolation controls and compliance that are available for EC2 customers.

* Your compute instances are located in a Virtual Private Cloud (VPC) with an IP range that you specify. You decide which instances are exposed to the Internet and which remain private.
* Your EC2 instances use an IAM role to access the ECS service.
* Your ECS tasks use an IAM role to access services and resources.
* Security Groups and networks ACLs allow you to control inbound and outbound network access to and from your instances.
* You can connect your existing IT infrastructure to resources in your VPC using industry-standard encrypted IPsec VPN connections.
* You can provision your EC2 resources as Dedicated Instances. Dedicated Instances are Amazon EC2 Instances that run on hardware dedicated to a single customer for additional isolation.

Q: Can I apply additional security configuration and isolation frameworks to my container instances?  
Yes. As an Amazon EC2 customer, you have root access to the operating system of your container instances, enabling you to take ownership of the operating system’s security settings as well as load and configure additional software components for security capabilities such as monitoring, patch management, log management and host intrusion detection.

Q: Can I operate container instances with different security settings or segregate different tasks across different environments?  
Yes. You can configure your different container instances using the tooling of your choice. Amazon ECS allows you to control the placement of tasks in different container instances through the construct of clusters and targeted launches.

Q: Does Amazon ECS support retrieving Docker images from a private or internal source?  
Yes. Customers can configure their container instances to access a private Docker image registry within a VPC or a registry that’s accessible outside a VPC such as the [Amazon EC2 Container Registry](https://aws.amazon.com/ecr/getting-started/).

Q: How do I configure IAM roles for ECS tasks?   
You first need to create an IAM role for your task, using the 'Amazon EC2 Container Service Task Role’ service role and attaching a policy with the required permissions. When you create a new task definition or a task definition revision you can then specify a role by selecting it form the ’Task Role’ drop-down or using the ‘taskRoleArn’ filed in the JSON format.

# Amazon VPC:

Q. What is Amazon Virtual Private Cloud (Amazon VPC)?

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.

#### **Manage Your AWS Resources**

[Sign in to the Console](https://console.aws.amazon.com/console/home)

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 (VPC): 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.
* Hardware VPN Connection: A hardware-based VPN connection between your Amazon VPC and your datacenter, home network, or co-location facility.
* Virtual Private Gateway: The Amazon VPC side of a VPN connection.
* Customer Gateway: Your side of a VPN connection.
* Router: Routers interconnect subnets and direct traffic between Internet gateways, virtual private gateways, NAT gateways, and subnets.
* Peering Connection: A peering connection enables you to route traffic via private IP addresses between two peered VPCs.
* VPC Endpoint: Enables private connectivity to Amazon services  from within your VPC without using an Internet gateway or NAT.
* 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, is exposed to the Internet. You can also leverage the greatly enhanced security options in Amazon VPC to provide more granular access both 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. VPC with a Single Public Subnet Only
2. VPC with Public and Private Subnets
3. VPC with Public and Private Subnets and Hardware VPN Access
4. VPC with a Private Subnet Only and Hardware VPN Access

## **Connectivity**

Q. What are the connectivity options for my VPC?

You may connect your VPC to:

* The Internet (via an Internet gateway)
* Your corporate data center using a Hardware 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 VPCs (via VPC peering connections)

Q. What are the different types of VPC Endpoints available on Amazon VPC?

VPC endpoints enable you to privately connect your VPC to an AWS service without requiring an Internet gateway, a NAT device, 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 for S3 and DynamoDB. These endpoints will add an entry to you’re the route table you selected and route the traffic to the supported services through Amazon’s private network.

Interface type endpoints are available for Kinesis, EC2 Systems Manager (SSM), EC2 API and ELB API. These endpoints provide private connectivity to services powered by PrivateLink, and provide connectivity over Direct Connect. More services will be supported by these endpoints in the future. Please refer to [Amazon VPC pricing](https://aws.amazon.com/vpc/pricing/) for interface type endpoints' pricing.

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. How does a hardware VPN connection work with Amazon VPC?

A hardware 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 hardware VPN connection.

Q. What is IPsec?

[IPsec](http://en.wikipedia.org/wiki/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?

There are two types of VPN connections that you can create: 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 VPN connections must be able to:

* Establish Border Gateway Protocol (BGP) peerings
* Bind tunnels to logical interfaces (route-based VPN)
* Utilize IPsec Dead Peer Detection

Q. Which Diffie-Hellman Groups do you support?

We support the following Diffie-Hellman (DH) groups in Phase1 and Phase2.

* Phase1 DH groups 2, 14-18, 22, 23, 24
* Phase2 DH groups 2, 5, 14-18, 22, 23, 24

Q. What customer gateway devices are known to work with Amazon VPC?

The following devices meeting the aforementioned requirements are known to work with hardware VPN connections, and have support in the command line tools for automatic generation of configuration files appropriate for your device:

* Statically-routed VPN connections
  + [Cisco ASA 5500 Series](http://www.cisco.com/cisco/web/solutions/small_business/products/security/ASA_5500_series/index.html) version 8.2 (or later) software
  + [Cisco ISR](http://www.cisco.com/en/US/products/ps10906/Products_Sub_Category_Home.html) running Cisco IOS 12.4 (or later) software
  + [Dell SonicWALL Next Generation Firewalls (TZ, NSA, SuperMassive Series) running SonicOS5.8 (or later)](http://www.sonicwall.com/us/en/products/network-security.html)
  + [Juniper J-Series Service Router](http://www.juniper.net/us/en/products-services/routing/j-series/) running JunOS 9.5 (or later) software
  + [Juniper SRX-Series Services Gateway](http://www.juniper.net/us/en/products-services/security/srx-series/) running JunOS 9.5 (or later) software
  + [Juniper SSG](http://www.juniper.net/us/en/products-services/security/ssg-series/) running ScreenOS 6.1, or 6.2 (or later) software
  + [Juniper ISG](http://www.juniper.net/us/en/products-services/security/isg-series/) running ScreenOS 6.1, or 6.2 (or later) software
  + [Microsoft Windows Server 2008 R2](http://www.microsoft.com/en-us/server-cloud/windows-server/) (or later) software
  + [Yamaha RTX1200 router](http://www.rtpro.yamaha.co.jp/RT/docs/amazon-vpc/index.html)
* Dynamically-routed VPN connections (requires BGP)
  + [Astaro Security Gateway](http://www.astaro.com/products/hardware-appliances) running version 8.3 (or later)
  + [Astaro Security Gateway Essential Firewall Edition](http://www.astaro.com/landingpages/en-worldwide-essential-firewall) running version 8.3 (or later)
  + [Cisco ISR](http://www.cisco.com/en/US/products/ps10906/Products_Sub_Category_Home.html) running Cisco IOS 12.4 (or later) software
  + [Dell SonicWALL Next Generation Firewalls (TZ, NSA, SuperMassive Series) running SonicOS5.9 (or later)](http://www.sonicwall.com/us/en/products/network-security.html)
  + [Juniper J-Series Service Router](http://www.juniper.net/us/en/products-services/routing/j-series/) running JunOS 9.5 (or later) software
  + [Juniper SRX-Series Services Gateway](http://www.juniper.net/us/en/products-services/security/srx-series/) running JunOS 9.5 (or later) software
  + [Juniper SSG](http://www.juniper.net/us/en/products-services/security/ssg-series/) running ScreenOS 6.1, or 6.2 (or later) software
  + [Juniper ISG](http://www.juniper.net/us/en/products-services/security/isg-series/) running ScreenOS 6.1, or 6.2 (or later) software
  + [Palo Alto Networks PA Series](https://www.paloaltonetworks.com/products/platforms/firewalls.html) running PANOS 4.1.2 (or later) software
  + [Vyatta Network OS](http://www.vyatta.com/product/vyatta-network-os/specs) 6.5 (or later) software
  + [Yamaha RTX1200 router](http://www.rtpro.yamaha.co.jp/RT/docs/amazon-vpc/index.html)

Please note, these sample configurations are for the minimum requirement of AES128, SHA1, and DH Group 2. You will need to modify these sample configuration files to take advantage of AES256, SHA256, or other DH groups.

Q. If my device is not listed, where can I go for more information about using it with Amazon VPC?

We recommend checking the [Amazon VPC forum](http://developer.amazonwebservices.com/connect/forum.jspa?forumID=58) as other customers may be already using your device.

Q. Are there any VPN connection throughput limitations?

VPN connection throughput can depend on multiple factors, such as the capability of your Customer Gateway (CGW), the capacity of your connection, average packet size, the protocol being used (TCP vs. UDP), and the network latency between your CGW and the Virtual Private Gateway (VGW).

Q. What tools are available to me to help troubleshoot my Hardware VPN configuration?

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?

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)](http://en.wikipedia.org/wiki/Network_address_translation) on Amazon EC2 instances within a VPC accessed via a hardware VPN connection.

Q. Can I NAT my CGW behind a router or firewall?

Yes, you will need to enable NAT-T and open UDP port 4500 on your NAT device.

Q. What IP address do I use for my CGW address?

You will use the public IP address of your NAT device.

Q. How does my connection decide to use NAT-T?

If your device has NAT-T enabled on the tunnel, AWS will use it by default. You will need to open UDP port 4500 or else the tunnel will not establish.

Q. How do I disable NAT-T on my connection?

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 CGWs behind a NAT, what do I need to do to configure that?

You will use the public IP address of your NAT device for the CGW for each of your connections. You will also need to make sure UDP port 4500 is open.

Q. How many IPsec security associations can be established concurrently per tunnel?

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.

## **IP Addressing**

Q. What IP address ranges can I use within my VPC?

You can use any [IPv4](http://en.wikipedia.org/wiki/IPv4) address range, including [RFC 1918](https://tools.ietf.org/html/rfc1918) or publicly routable IP ranges for the primary CIDR block. For the secondary CIDR blocks certain [restrictions](http://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/VPC_Subnets.html#add-cidr-block-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 VPCs?

You assign a single [Classless Internet Domain Routing (CIDR)](http://en.wikipedia.org/wiki/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 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 hardware VPN, and to my VPC?

Yes, you can route traffic via the hardware VPN connection and advertise the address range from your home network.

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 a VPC's size?

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](https://aws.amazon.com/contact-us/vpc-request/).

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 the [EC2 User Guide](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html) 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.

## **Routing & Topology**

Q. What does an Amazon VPC router do?

An Amazon VPC router enables Amazon EC2 instances within subnets to communicate with Amazon EC2 instances in other subnets within the same VPC. The VPC router also enables subnets, Internet gateways, and virtual private gateways to communicate with each other. Network usage data is not available from the router; however, you can obtain network usage statistics from your instances using Amazon CloudWatch.

Q. Can I modify the VPC route tables?

Yes. You can create route rules to specify which subnets are routed to the Internet gateway, the virtual private gateway, or other instances.

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.

Q. Does Amazon VPC support [multicast](http://en.wikipedia.org/wiki/IP_multicast) or [broadcast](http://en.wikipedia.org/wiki/Broadcast_address#IP_network_broadcasting)?

No.

## **Security & 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, they can communicate using 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. Why can’t I ping the router, or my default gateway, that connects my subnets?

Ping (ICMP Echo Request and Echo Reply) requests to the router in your VPC is not supported. Ping between Amazon EC2 instances within VPC is supported as long as your operating system's firewalls, VPC security groups, and network ACLs permit such traffic.

Q. Can I monitor the network traffic in my VPC?

Yes. You can use the Amazon VPC Flow Logs feature to monitor the network traffic in your VPC.

## **Amazon VPC & EC2**

Q. Within which Amazon EC2 region(s) is Amazon VPC available?

Amazon VPC is currently available in multiple [Availability Zones](http://developer.amazonwebservices.com/connect/entry.jspa?externalID=1347) 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](http://aws.amazon.com/contact-us/vpc-request/).

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](http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/FAQ_Regions_Availability_Zones.html).

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](http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/FAQ_Regions_Availability_Zones.html).

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](http://aws.amazon.com/ec2/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](https://forums.aws.amazon.com/ann.jspa?annID=1875) 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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-vpc.html) 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. If you delete your default VPC, you can create a new default VPC later from the VPC Console or by using the CreateDefaultVPC API.This will create a new default VPC in the region, rather than restore the previous one.

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, contact [AWS Support](https://aws.amazon.com/contact-us/). We will review your request and your existing AWS services and EC2-Classic presence to determine if you are eligible for a default VPC.

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?

No. Peering connections are only available between VPCs in the same region.

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](http://aws.amazon.com/ec2/pricing/) 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](http://docs.aws.amazon.com/AmazonVPC/latest/PeeringGuide/) 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.

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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/placement-groups.html).

## **PrivateLink**

Q. What is PrivateLink?

PrivateLink is a purpose-built technology designed for customers to access Amazon services in a highly available and scalable manner, while keeping all the network traffic within the AWS network. Customers can use this to privately access Amazon services from their Amazon Virtual Private Cloud (VPC) or their own data center, without using public IPs, and without requiring the traffic to traverse across the Internet.

Q. How can I use PrivateLink?

In order to use PrivateLink, you will need to create interface type VPC endpoints for Amazon services that are powered by PrivateLink. These service endpoints will appear as Elastic Network Interface (ENI) with private IPs in your VPCs. Once these endpoints are created, any traffic destined to these IPs will get privately routed to the corresponding Amazon services.

Q. Can I privately access services powered by PrivateLink over Direct Connect?

Yes. The application in your own data center can connect to the service endpoints in Amazon VPC over AWS Direct Connect.&nbsp; The service endpoints will automatically direct the traffic to Amazon Services powered by PrivateLink.

Q. Which services are currently available on PrivateLink?

Elastic Compute Cloud (EC2), Elastic Load Balancer (ELB), Kinesis, and EC2 Systems Manager are currently available on PrivateLink.

## **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](https://aws.amazon.com/ec2/pricing/).

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.

## **Virtual Private Gateway - Bring your own Autonomous System Number**

Q. What is this feature?

For any new VGWs, 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?   
There is no additional charge for this feature.

Q. How can I configure/assign my ASN to be advertised as Amazon side ASN?

You can configure/assign an ASN to be advertised as the Amazon side ASN during creation of the new Virtual Private Gateway (VGW). You can create a VGW using the VPC console or a EC2/CreateVpnGateway API call.

Q. What ASN did Amazon assign prior to this feature?

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?

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?

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.

Amazon will provide a default ASN for the VGW 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?

We will ask you to re-enter a private ASN once you attempt to create the VGW, 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?

Amazon will provide an ASN for the VGW 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?

You can view the Amazon side ASN in the VGW 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?

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 VGW with desired ASN, and create a new VIF with the newly created VGW. 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?

You will need to create a new VGW with the desired ASN, and recreate your VPN connections between your Customer Gateways and the newly created VGW.

Q. I already have a VGW 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 VGW, what Amazon side ASN will I be assigned?

Amazon will assign 64512 to the Amazon side ASN for the new VGW.

Q. I have a VGW 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?

You can configure/assign an ASN to be advertised as the Amazon side ASN during creation of the new Virtual Private Gateway (VGW). You can create VGW using console or EC2/CreateVpnGateway API call. As noted earlier, we will allow the use of the “legacy public ASN” for your newly created VGW.

Q. I have a VGW 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 VGW, what Amazon side ASN will I be assigned?

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 VGW and defaults to that ASN.

Q. I’m attaching multiple private VIFs to a single VGW. Can each VIF have a separate Amazon side ASN?

No, you can assign/configure separate Amazon side ASN for each VGW, not each VIF. Amazon side ASN for VIF is inherited from the Amazon side ASN of the attached VGW.

Q. I’m creating multiple VPN connections to a single VGW. Can each VPN connection have a separate Amazon side ASN?

No, you can assign/configure separate Amazon side ASN for each VGW, not each VPN connection. Amazon side ASN for VPN connection is inherited from the Amazon side ASN of the VGW.

Q. Where can I select my own ASN?

When creating a VGW 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 VGW is configured with Amazon side ASN, the private VIFs or VPN connections created using the VGW will use your Amazon side ASN.

Q. I use CloudHub today. Will I have to adjust my configurations 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 VGW 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 VGW and recreate a new VGW with the desired ASN.

Q. Is there a new API to configure/assign the Amazon side ASN?

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?

No. You can view the Amazon side ASN with the same EC2/DescribeVpnGateways API. We just added a new parameter (amazonSideAsn) to this API.

## **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, Internet gateways, customer gateways, virtual private gateways, and VPN connections 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 VPC
* Five virtual private gateways per AWS account per region
* Fifty customer gateways per AWS account per region
* Ten IPsec VPN Connections per virtual private gateway

See the [VPC User Guide](http://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/VPC_Appendix_Limits.html) for more information on VPC limits.

Q. Does the Amazon VPC VPN Connection have a Service Level Agreement (SLA)?

Not currently.

Q. Can I obtain AWS Support with Amazon VPC?

Yes. [Click here](http://aws.amazon.com/premiumsupport/) for more information on AWS Support.

Q. Can I use [ElasticFox](http://sourceforge.net/projects/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.

**ELASTIC beanstalk:**

: What is AWS Elastic Beanstalk?  
AWS Elastic Beanstalk makes it even easier for developers to quickly deploy and manage applications in the AWS Cloud. Developers simply upload their application, and Elastic Beanstalk automatically handles the deployment details of capacity provisioning, load balancing, auto-scaling, and application health monitoring.

Q: Who should use AWS Elastic Beanstalk?  
Those who want to deploy and manage their applications within minutes in the AWS Cloud. You don’t need experience with [cloud computing](https://aws.amazon.com/what-is-cloud-computing/) to get started. AWS Elastic Beanstalk supports Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker web applications.

Q: Which languages and development stacks does AWS Elastic Beanstalk support?  
AWS Elastic Beanstalk supports the following languages and development stacks:

* Apache Tomcat for Java applications
* Apache HTTP Server for PHP applications
* Apache HTTP Server for Python applications
* Nginx or Apache HTTP Server for Node.js applications
* Passenger or Puma for Ruby applications
* Microsoft IIS 7.5, 8.0, and 8.5 for .NET applications
* Java SE
* Docker
* Go

See [Supported Platforms](https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/concepts.platforms.html) for a complete, up-to-date list of supported language and development stacks.

Q: Will AWS Elastic Beanstalk support other languages?  
Yes. AWS Elastic Beanstalk is designed so that it can be extended to support multiple development stacks and programming languages in the future. AWS is working with solution providers on the APIs and capabilities needed to create additional Elastic Beanstalk offerings.

Q: What can developers now do with AWS Elastic Beanstalk that they could not before?  
AWS Elastic Beanstalk automates the details of capacity provisioning, load balancing, auto scaling, and application deployment, creating an environment that runs a version of your application. You can simply upload your deployable code (e.g., WAR file), and AWS Elastic Beanstalk does the rest. The AWS Toolkit for Visual Studio and the AWS Toolkit for Eclipse allow you to deploy your application to AWS Elastic Beanstalk and manage it without leaving your IDE. Once your application is running, Elastic Beanstalk automates management tasks–such as monitoring, application version deployment, a basic health check–and facilitates log file access. By using Elastic Beanstalk, developers can focus on developing their application and are freed from deployment-oriented tasks, such as provisioning servers, setting up load balancing, or managing scaling.

Q: How is AWS Elastic Beanstalk different from existing application containers or platform-as-a-service solutions?  
Most existing application containers or platform-as-a-service solutions, while reducing the amount of programming required, significantly diminish developers’ flexibility and control. Developers are forced to live with all the decisions predetermined by the vendor–with little to no opportunity to take back control over various parts of their application’s infrastructure. However, with AWS Elastic Beanstalk, developers retain full control over the AWS resources powering their application. If developers decide they want to manage some (or all) of the elements of their infrastructure, they can do so seamlessly by using Elastic Beanstalk’s management capabilities.

Q: What elements of my application can I control when using AWS Elastic Beanstalk?  
With AWS Elastic Beanstalk, you can:

* Select the operating system that matches your application requirements (e.g., Amazon Linux or Windows Server 2012 R2)
* Choose from several available database and storage options
* Enable login access to Amazon EC2 instances for immediate and direct troubleshooting
* Quickly improve application reliability by running in more than one Availability Zone
* Enhance application security by enabling HTTPS protocol on the load balancer
* Access built-in Amazon CloudWatch monitoring and getting notifications on application health and other important events
* Adjust application server settings (e.g., JVM settings) and pass environment variables
* Run other application components, such as a memory caching service, side-by-side in Amazon EC2
* Access log files without logging in to the application servers

Q: What are the Cloud resources powering my AWS Elastic Beanstalk application?  
AWS Elastic Beanstalk uses proven AWS features and services, such as Amazon EC2, Amazon RDS, Elastic Load Balancing, Auto Scaling, Amazon S3, and Amazon SNS, to create an environment that runs your application. The current version of AWS Elastic Beanstalk uses the Amazon Linux AMI or the Windows Server 2012 R2 AMI.

Q: What kinds of applications are supported by AWS Elastic Beanstalk?  
AWS Elastic Beanstalk supports Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker, and is ideal for web applications. However, due to Elastic Beanstalk’s open architecture, non-web applications can also be deployed using Elastic Beanstalk. We expect to support additional application types and programming languages in the future. See [Supported Platforms](https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/concepts.platforms.html) to learn more.

Q: Which operating systems does AWS Elastic Beanstalk use?  
AWS Elastic Beanstalk runs on the Amazon Linux AMI and the Windows Server 2012 R2 AMI. Both AMIs are supported and maintained by Amazon Web Services and are designed to provide a stable, secure, and high-performance execution environment for Amazon EC2 Cloud computing.

Q: How quickly will my application start running?  
It typically takes a few minutes to create the AWS resources to run your application, measured from the time you upload your application version (e.g., WAR file, ASP.NET files, Node.js files, PHP files, Python files, or Ruby files) to when it is fully deployed and accessible to your users. This time is dependent on a number of factors, including the size of your deployable code and the number of application servers you are deploying.

Q: How quickly will my application get updated?  
Deploying new application versions to existing resources (e.g., environments) happens much faster (typically under a minute) and is mostly dependent on the size of the new application version.

Q: How quickly can my application scale up and down?  
AWS Elastic Beanstalk provides a truly elastic environment using Auto Scaling. Your application can be configured to automatically scale tens or even hundreds of times based on thresholds, such as CPU utilization or network bandwidth. These thresholds can be easily configured for your specific application using the Elastic Beanstalk console. With Elastic Beanstalk, you don’t have to worry if you will be able to scale quickly to handle peaks in traffic or users, nor if you will be forced to pay for resources that you don’t need.

Q: Can I have multiple versions of my application running at the same time?  
Yes. AWS Elastic Beanstalk is designed to support multiple running environments, such as one for integration testing, one for pre-production, and one for production. Each environment is independently configured and runs on its own separate AWS resources. Elastic Beanstalk also stores and tracks application versions over time, so an existing environment can be easily rolled back to a prior version or a new environment can be launched using an older version to try and reproduce a customer problem.

Q: How many applications can I run with AWS Elastic Beanstalk?  
You can create up to 75 applications and 1,000 application versions. By default, you can run up to 200 environments across all of your applications. If you are also using AWS outside of AWS Elastic Beanstalk, you may not be able to create 10 environments since other limits may be hit sooner. For example, the default AWS account limits allow you to launch up to 20 EC2 instances and create up to 10 elastic load balancers. If you need more resources, complete the AWS Elastic Beanstalk request form, and your request will be promptly evaluated.

Q: Can I use AWS Elastic Beanstalk to deploy applications that must be highly available?  
Yes. To do this, you edit your environment configuration settings, select 2 or more instances for Auto Scaling minimum, and set Multiple Availability Zones to “Any 2”. AWS Availability Zones are designed to be physically distinct, fail independently, and be reliable.

Q: What happens if my application stops responding to requests?  
AWS Elastic Beanstalk applications are protected against failures in the underlying infrastructure. If an Amazon EC2 instance fails for any reason, AWS Elastic Beanstalk will use Auto Scaling to automatically launch a new instance. Elastic Beanstalk can also detect if your application is not responding on the custom URL even though the underlying infrastructure appears healthy, and will log that as an environment event (e.g., a bad version was deployed) so you can take appropriate action.

Q: Which AWS Regions is AWS Elastic Beanstalk available in?

Please refer to [Regional Products and Services](https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/) for details of Elastic Beanstalk availability by Region.

Q: How do I access AWS Elastic Beanstalk?  
You can use the AWS Management Console, the AWS Elastic Beanstalk command line interface (CLI), the AWS Toolkit for Visual Studio, the AWS Toolkit for Eclipse, the AWS Elastic Beanstalk API, or AWS SDKs.

Q: Can I use an integrated development environment, like Eclipse or Microsoft Visual Studio?  
Yes. You can use Eclipse and Visual Studio to deploy your application to AWS Elastic Beanstalk. You can use the AWS Toolkit for Eclipse for Java applications and the AWS Toolkit for Visual Studio for .NET applications. The toolkits allow you to develop your application, deploy it to Elastic Beanstalk, and even test it out without having to switch your focus away from your IDE.

## **Getting Started**

Q: How do I sign up for AWS Elastic Beanstalk?

To sign up for AWS Elastic Beanstalk, choose the Sign Up Now button on the Elastic Beanstalk 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 Elastic Beanstalk process. After signing up, please refer to the [AWS Elastic Beanstalk Getting Started Guide](http://docs.amazonwebservices.com/elasticbeanstalk/latest/dg/GettingStarted.html).

Q: Why am I asked to verify my phone number when signing up for AWS Elastic Beanstalk?

AWS Elastic Beanstalk 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 Elastic Beanstalk is to work through the [AWS Elastic Beanstalk Getting Started Guide](http://docs.amazonwebservices.com/elasticbeanstalk/latest/dg/GettingStarted.html), part of our technical documentation. Within a few minutes, you will be able to deploy and use a sample application or upload your own application.

Q: Is there a sample application that I can use to check out AWS Elastic Beanstalk?

Yes. AWS Elastic Beanstalk includes a sample application that you can use to test drive the offering and explore its functionality.

## **Databases and Storage**

Q: Does AWS Elastic Beanstalk store anything in Amazon S3?

Yes. AWS Elastic Beanstalk stores your application files and, optionally, server log files in Amazon S3. If you are using the AWS Management Console, the AWS Toolkit for Visual Studio, or AWS Toolkit for Eclipse, an Amazon S3 bucket will be created in your account for you and the files you upload will be automatically copied from your local client to Amazon S3. Optionally, you may configure Elastic Beanstalk to copy your server log files every hour to Amazon S3. You do this by editing the environment configuration settings.

Q: Can I use Amazon S3 to store application data, like images?

Yes. You can use Amazon S3 for application storage. The easiest way to do this is by including the AWS SDK as part of your application’s deployable file. For example, you can include the AWS SDK for Java as part of your application's WAR file.

Q: What database solutions can I use with AWS Elastic Beanstalk?

AWS Elastic Beanstalk does not restrict you to any specific data persistence technology. You can choose to use Amazon Relational Database Service (Amazon RDS) or Amazon DynamoDB, or use Microsoft SQL Server, Oracle, or other relational databases running on Amazon EC2.

Q: How do I set up a database for use with AWS Elastic Beanstalk?

Elastic Beanstalk can automatically provision an Amazon RDS DB instance. The information about connectivity to the DB instance is exposed to your application by environment variables. To learn more about how to configure RDS DB instances for your environment, see the [Elastic Beanstalk Developer Guide](http://docs.amazonwebservices.com/elasticbeanstalk/latest/dg/AWSHowTo.RDS.html).

Q: Does this mean I need to modify the application code when moving from test to production?

Not with AWS Elastic Beanstalk. With Elastic Beanstalk, you can specify the connection information in the environment configuration. By extracting the connection string from the application code, you can easily configure different Elastic Beanstalk environments to use different databases.

## **Security**

Q: How do I make my application private?

By default, your application is available publicly at myapp.elasticbeanstalk.com for anyone to access. You can use [Amazon VPC](https://aws.amazon.com/vpc/) to provision a private, isolated section of your application in a virtual network that you define. This virtual network can be made private through specific security group rules, network ACLs, and custom route tables. You can also easily control what other incoming traffic, such as SSH, is delivered or not to your application servers by changing the EC2 security group settings.

Q: Can I run my application inside a Virtual Private Cloud (VPC)?

Yes, you can run your applications in a VPC. For more details, see the [AWS Elastic Beanstalk Developer Guide](http://docs.amazonwebservices.com/elasticbeanstalk/latest/dg/AWSHowTo-vpc.html).

Q: Where can I find more information about security and running applications on AWS?

For more information about security on AWS, please refer to our [Amazon Web Services: Overview of Security Processes](https://d0.awsstatic.com/whitepapers/Security/AWS_Security_Whitepaper.pdf) document and visit our [Security Center](https://aws.amazon.com/security/).

Q: Is it possible to use Identity & Access Management (IAM) with AWS Elastic Beanstalk?

Yes. IAM users with the appropriate permissions can now interact with AWS Elastic Beanstalk.

Q: Why should I use IAM with AWS Elastic Beanstalk?

IAM allows you to manage users and groups in a centralized manner. You can control which IAM users have access to AWS Elastic Beanstalk, and limit permissions to read-only access to Elastic Beanstalk for operators who should not be able to perform actions against Elastic Beanstalk resources. All user activity within your account will be aggregated under a single AWS bill.

Q: How do I create IAM users?

You can use the [IAM console](http://console.aws.amazon.com/iam/), IAM command line interface (CLI), or IAM API to provision IAM users. By default, IAM users have no access to AWS services until permissions are granted.

Q: How do I grant an IAM user access to AWS Elastic Beanstalk?

You can grant IAM users access to services by using policies. To simplify the process of granting access to AWS Elastic Beanstalk, you can use one of the policy templates in the IAM console to help you get started. Elastic Beanstalk offers two templates: a read-only access template and a full-access template. The read-only template grants read access to Elastic Beanstalk resources. The full-access template grants full access to all Elastic Beanstalk operations, as well as permissions to manage dependent resources, such as Elastic Load Balancing, Auto Scaling, and Amazon S3. You can also use the AWS Policy Generator to create custom policies. For more details, see the [AWS Elastic Beanstalk Developer Guide](http://docs.amazonwebservices.com/elasticbeanstalk/latest/dg/AWSHowTo.iam.html).

Q: Can I restrict access to specific AWS Elastic Beanstalk resources?

Yes. You can allow or deny permissions to specific AWS Elastic Beanstalk resources, such as applications, application versions, and environments.

Q: Who gets billed for the AWS resources that an IAM user creates?

All resources created by IAM users under a root account are owned and billed to the root account.

Q: Who has access to an AWS Elastic Beanstalk environment launched by an IAM user?

The root account has full access to all AWS Elastic Beanstalk environments launched by any IAM user under that account. If you use the Elastic Beanstalk template to grant read-only access to an IAM user, that user will be able to view all applications, application versions, environments, and any associated resources in that account. If you use the Elastic Beanstalk template to grant full access to an IAM user, that user will be able to create, modify, and terminate any Elastic Beanstalk resources under that account.

Q: Can an IAM user access the AWS Elastic Beanstalk console?

Yes. An IAM user can access the AWS Elastic Beanstalk console using their username and password.

Q: Can an IAM user call the AWS Elastic Beanstalk API?

Yes. An IAM user can use their access key and secret key to perform operations using the Elastic Beanstalk API.

Q: Can an IAM user use the AWS Elastic Beanstalk command line interface?

Yes. An IAM user can use their access key and secret key to perform operations using the AWS Elastic Beanstalk command line interface (CLI).

## **Managed Platform Updates**

Q: How can I keep the underlying platform of the environment running my application automatically up-to-date?

You can opt-in to having your AWS Elastic Beanstalk environments automatically updated to the latest version of the underlying platform running your application during a specified maintenance window. Elastic Beanstalk regularly releases new versions of supported platforms (Java, PHP, Ruby, Node.js, Python, .NET, Go, and Docker) with operating system, web and application server, and language and framework updates.

Q: How can I get started with managed platform updates?

To let Elastic Beanstalk automatically manage your platform updates, you must enable managed platform updates in the Configuration tab of the Elastic Beanstalk console or use the EB CLI or API. After you have enabled the feature, you can configure which types of updates to allow and when updates can occur.

Q: What kinds of platform version updates will managed platform updates apply?

AWS Elastic Beanstalk can automatically perform platform updates for new patch and minor platform versions. Elastic Beanstalk will not automatically perform major platform version updates (e.g., Java 7 Tomcat 7 to Java 8 Tomcat 8) because they include backwards incompatible changes and require additional testing. In these cases, you must manually initiate the update.

Q: How does AWS Elastic Beanstalk distinguish between “major,” “minor,” and “patch” version releases?

AWS Elastic Beanstalk platforms are versioned using this pattern: MAJOR.MINOR.PATCH (e.g., 2.0.0). Each portion is incremented as follows:

* MAJOR version when there are incompatible changes.
* MINOR version when there is additional functionality added in a backward-compatible manner.
* PATCH version when there are backward-compatible bug fixes.

Q: When and how can I perform major version updates?

You can perform major version updates at any time using the AWS Elastic Beanstalk management console, API, or CLI. You have the following options to perform a major version update:

* Apply the update in-place on an existing environment. See [Updating Your Elastic Beanstalk Environment's Platform Version](https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/using-features.platform.upgrade.html).
* Create a clone of an existing environment with the new platform version. See [Clone an Environment](https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/using-features.managing.clone.html) to learn more.

Q: How does Elastic Beanstalk apply managed platform updates?

The updates are applied using an immutable deployment mechanism that ensures that no changes are made to the existing environment until a parallel fleet of Amazon EC2 instances, with the updates installed, is ready to be swapped with the existing instances, which are then terminated. In addition, if the Elastic Beanstalk health system detects any issues during the update, traffic is redirected to the existing fleet of instances, ensuring minimal impact to end users of your application.

Q: Will my application be available during the maintenance windows?

Since managed platform updates use an immutable deployment mechanism to perform the updates, your application will be available during the maintenance window and consumers of your application will not be impacted by the update.

Q: What does it cost to use managed platform updates?

There is no additional charge for the managed platform updates feature. You simply pay for the additional EC2 instances necessary to perform the update for the duration of the update.

Q: What is a maintenance window?

A maintenance window is a weekly two-hour-long time slot during which AWS Elastic Beanstalk will initiate platform updates if managed platform updates is enabled and a new version of the platform is available. For example, if you select a maintenance window that begins every Sunday at 2 AM, AWS Elastic Beanstalk will initiate the platform update sometime between 2-4 AM every Sunday. It is important to note that, depending on the configuration of your applications, updates could complete outside of the maintenance window.

The maintenance window is set on a per-environment basis, providing you the option to set different maintenance windows for your various application components or applications. This allows environment updates to be staggered if you do not want multiple pieces of your application to be updated at the same time. If you enable managed platform updates but do not specify a maintenance window, a default weekly 2-hour window will be assigned for your environment. If you want to change when maintenance is performed on your behalf, you can do so by modifying the managed update configuration in the AWS Management Console or by using the UpdateEnvironment API.

Q: How will I be notified of the availability of new platform versions?

You will be notified about the availability of new platform versions through the AWS Management Console, forum announcements, and release notes.

Q: Where can I find details of changes between platform versions?

Details on changes between platform versions can be found on the AWS Elastic Beanstalk [Release Notes](https://aws.amazon.com/releasenotes/AWS-Elastic-Beanstalk) page.

Q: What operations can I perform on the environment while a managed update is in progress?

The only action available to you while a managed platform update is in-progress is ‘abort’. This will allow you to stop the update immediately and roll back to the previous version.

Q: Which platform version will my environment be updated to if there are multiple new versions released in between maintenance windows?

Your environment will always be updated to the latest version available based on the level (minor plus patch or patch only) you have selected.

Q: Where can I find details of all the managed platform updates that have been performed on my environment?

Details for every managed platform update are available on the events page and are tagged with an event type of “MAINTENANCE.”

Q: How often are platform version updates released?

The number of version releases in a given year varies based on the frequency and content of releases and patches from the language/framework’s vendor or core team, and the outcome of a thorough vetting of these releases and patches by our platform engineering team.

## **Billing**

Q: How much does AWS Elastic Beanstalk cost?

There is no additional charge for AWS Elastic Beanstalk–you pay only for the AWS resources actually used to store and run your application. New AWS customers who are eligible for the AWS Free Tier may deploy an application that runs within the Free Tier using the default settings of Elastic Beanstalk.

Q: How much do the AWS resources powering my application on AWS Elastic Beanstalk cost?

You pay only for what you use, and there is no minimum fee for the use of any AWS resources. For Amazon EC2 pricing information, please visit the pricing section on the EC2 detail page. For Amazon S3 pricing information, please visit the pricing section on the S3 detail page. You can use the AWS simple calculator to estimate your bill for different application sizes.

Q: How do I check how many AWS resources have been used by my application and access my bill?

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 choosing Account Activity under Your Web Services Account.

## **Support**

Q: Does AWS Support cover AWS Elastic Beanstalk?

Yes. AWS Support covers issues related to your use of AWS Elastic Beanstalk. For further details and pricing, see the AWS Support page.

Q: What other support options are available?

You can tap into the breadth of existing AWS community knowledge to help you with your development through the [AWS Elastic Beanstalk discussion forum](https://forums.aws.amazon.com/forum.jspa?forumID=86).

# AWS Lambda

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. Learn more about serverless computing by visiting [here](https://aws.amazon.com/serverless/).

Q: What events can trigger an AWS Lambda function?

Please see our [documentation](http://docs.aws.amazon.com/lambda/latest/dg/intro-core-components.html#intro-core-components-event-sources) 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](https://aws.amazon.com/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](https://aws.amazon.com/elasticbeanstalk/) 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](https://aws.amazon.com/ecs/) 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 supports code written in Node.js (JavaScript), Python, Java (Java 8 compatible), and C# (.NET Core). Your code can include existing libraries, even native ones. Please read our documentation on using [Node.js](http://docs.aws.amazon.com/lambda/latest/dg/authoring-function-in-nodejs.html), [Python](http://docs.aws.amazon.com/lambda/latest/dg/python-lambda.html), [Java](http://docs.aws.amazon.com/lambda/latest/dg/java-lambda.html), and [C#](http://docs.aws.amazon.com/lambda/latest/dg/current-supported-versions.html).

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](https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/).

## **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](http://docs.aws.amazon.com/lambda/latest/dg/lambda-introduction.html). 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 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 the inline editor in the AWS Lambda console. [Go to the console to get started.](https://aws.amazon.com/console/) 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](http://docs.aws.amazon.com/lambda/latest/dg/getting-started.html) 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](http://docs.aws.amazon.com/lambda/latest/dg/env_variables.html).

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](http://docs.aws.amazon.com/kms/latest/developerguide/overview.html) 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 Developers Guide](http://docs.aws.amazon.com/lambda/latest/dg/welcome.html) to learn more.

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, 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](http://docs.aws.amazon.com/lambda/latest/dg/monitoring-functions.html) 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](http://docs.aws.amazon.com/lambda/latest/dg/monitoring-functions.html) 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. You can set your memory in 64MB increments from 128MB to 1.5GB.

Q: How long can an AWS Lambda function execute?

All calls made to AWS Lambda must complete execution within 300 seconds. The default timeout is 3 seconds, but you can set the timeout to any value between 1 and 300 seconds.

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](https://aws.amazon.com/lambda/pricing/) for details.

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](http://docs.aws.amazon.com/lambda/latest/dg/versioning-aliases.html).

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 a Kinesis stream and execute a Lambda function for each message in the stream.

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](http://docs.aws.amazon.com/lambda/latest/dg/intro-core-components.html#intro-core-components-event-sources) 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 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](http://docs.aws.amazon.com/AmazonS3/latest/dev/NotificationHowTo.html). To learn more about Amazon DynamoDB Streams visit the [DynamoDB Stream Developers Guide](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Streams.html). To learn more about invoking Lambda functions using Amazon SNS, visit the [Amazon SNS Developers Guide](http://docs.aws.amazon.com/sns/latest/dg/sns-lambda.html). For more information on Amazon Cognito events, visit [Amazon Cognito](https://aws.amazon.com/cognito/). For more information on AWS CloudTrail logs and auditing API calls across AWS services, see [AWS CloudTrail](https://aws.amazon.com/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](http://docs.aws.amazon.com/AmazonCloudWatch/latest/DeveloperGuide/WhatIsCloudWatch.html) 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](https://aws.amazon.com/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](http://docs.aws.amazon.com/lambda/latest/dg/welcome.html) 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](https://aws.amazon.com/mobile/sdk/) page. To learn more about Amazon API Gateway, visit the [Amazon API Gateway](https://aws.amazon.com/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](https://developer.amazon.com/appsandservices/solutions/alexa/alexa-skills-kit/docs/developing-an-alexa-skill-as-a-lambda-function) 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 do I automate deployment for a serverless application?  
You can automate your serverless application’s release process using CodePipeline. CodePipeline is a continuous delivery service that enables you to model, visualize and automate the steps required to release your serverless application. To learn more about serverless CI/CD, visit our [documentation](http://docs.aws.amazon.com/lambda/latest/dg/automating-deployment.html).

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](http://docs.aws.amazon.com/lambda/latest/dg/deploying-lambda-apps.html).

Q: How do I coordinate calls between multiple AWS Lambda functions?  
You can use [AWS Step Functions](https://aws.amazon.com/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](https://aws.amazon.com/xray/) 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](http://docs.aws.amazon.com/lambda/latest/dg/lambda-x-ray.html) to learn more. AWS X-Ray rates will apply.

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](https://github.com/awslabs/serverless-application-specification).

## **Lambda@Edge**

Q: What is Lambda@Edge?   
[Lambda@Edge](https://aws.amazon.com/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 code to AWS Lambda and configure your function to be triggered in response to [Amazon CloudFront](https://aws.amazon.com/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](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/lambda-at-the-edge.html).

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 only supports Node.js for global invocation by CloudFront events at this time. Learn more in our [documentation](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/lambda-at-the-edge.html).

Q: When should I use Lambda@Edge?  
Lambda@Edge is optimized for latency sensitive use cases where your end viewers are distributed globally. Ideally, all the information you need to make a decision is 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 right from the edge in Node.js-6.10 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 Node.js-6.10 Lambda functions with CloudFront events for global invocation if the function satisfies the Lambda@Edge service limits. Read more [here](http://docs.aws.amazon.com/lambda/latest/dg/API_UpdateFunctionConfiguration.html) 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](https://aws.amazon.com/lambda/edge/) and [Amazon CloudFront](https://aws.amazon.com/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](http://docs.aws.amazon.com/lambda/latest/dg/concurrent-executions.html#concurrent-execution-safety-limit) for info on default safety throttle limits). If you wish to submit a request to increase the throttle limit you can visit our [Support Center](https://aws.amazon.com/support), 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](http://docs.aws.amazon.com/lambda/latest/dg/setting-up.html) 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](http://docs.aws.amazon.com/lambda/latest/dg/welcome.html) 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](http://docs.aws.amazon.com/lambda/latest/dg/welcome.html) to learn more.

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](http://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/vpc-nat-gateway.html) 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](http://docs.aws.amazon.com/lambda/latest/dg/java-lambda.html).

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](https://aws.amazon.com/blogs/compute/nodejs-packages-in-lambda/).

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](https://aws.amazon.com/blogs/compute/running-executables-in-aws-lambda/).

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](https://aws.amazon.com/blogs/compute/nodejs-packages-in-lambda/).

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](https://aws.amazon.com/blogs/compute/running-executables-in-aws-lambda/).

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](http://docs.aws.amazon.com/lambda/latest/dg/authoring-function-in-nodejs.html).

## **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”

## **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](http://docs.aws.amazon.com/lambda/latest/dg/current-supported-versions.html).

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 language runtime to all users of the service.

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](http://docs.aws.amazon.com/step-functions/latest/dg/hello-lambda.html) for more details.

# Amazon Simple Storage Service (S3)

Q: What is Amazon S3?

Amazon S3 is storage for the Internet. It’s a simple storage service that offers software developers a highly-scalable, reliable, and low-latency 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, developers can easily build applications that make use of Internet storage. Since Amazon S3 is highly scalable and you only pay for what you use, developers can start small and grow their application as they 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, not figuring out how to store their data.

Q: How can I get started using Amazon S3?

To sign up for Amazon S3, click the “Sign up for This Web Service” button on the [Amazon S3](https://aws.amazon.com/s3/details/) 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 S3 sign-up process. After signing up, please refer to the Amazon S3 documentation and sample code in the [Resource Center](http://docs.aws.amazon.com/AmazonS3/latest/API/RelatedResources.html) to begin using Amazon S3.

Q: What are the technical benefits of Amazon S3?

Amazon S3 was carefully engineered to meet the requirements for scalability, reliability, speed, low-cost, and simplicity that must be met for Amazon’s internal developers. Amazon S3 passes these same benefits onto any external developer. More information about the Amazon S3 design requirements is available on the [Amazon S3](https://aws.amazon.com/s3/details/) detail page.

Q: What can developers do now that they could not before?

Until now, a sophisticated and scalable data storage infrastructure like Amazon’s has been beyond the reach of small developers. 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?

You can store virtually any kind of data in any format. Please refer to the [Amazon Web Services Licensing Agreement](http://aws.amazon.com/agreement) for details.

Q: How much data can I store?

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](http://docs.amazonwebservices.com/AmazonS3/latest/dev/UploadingObjects.html) capability.

Q: What storage classes does Amazon S3 offer?

Amazon S3 offers a range of storage classes designed for different use cases. There are three highly durable storage classes including Amazon S3 Standard for general-purpose storage of frequently accessed data, Amazon S3 Standard - Infrequent Access for long-lived, but less frequently accessed data, and Amazon Glacier for long-term archive. You can learn more about those three storage classes on the [Amazon S3 Storage Classes page](https://aws.amazon.com/s3/storage-classes/).

Reduced Redundancy Storage (RRS) is an Amazon S3 storage option that enables customers to reduce their costs by storing noncritical, reproducible data at lower levels of redundancy than Amazon S3’s standard storage. You can learn more about Reduced Redundancy Storage on the [Reduced Redundancy detail page.](https://aws.amazon.com/s3/reduced-redundancy/)

Q: How can I delete large numbers of objects?

You can use [Multi-Object Delete](http://docs.amazonwebservices.com/AmazonS3/latest/dev/DeletingObjects.html) to delete large numbers of objects from Amazon S3. This feature allows you to send multiple object keys in a single request to speed up your deletes. Amazon does not charge you for using Multi-Object Delete.

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](http://aws.amazon.com/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 can be constructed to mimic hierarchical attributes.

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, reliable, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites. S3 Standard is designed for 99.99% availability and Standard - IA is designed for 99.9% availability. Both are backed by the [Amazon S3 Service Level Agreement](https://aws.amazon.com/s3/sla/).

Q: What data consistency model does Amazon S3 employ?

Amazon S3 buckets in all Regions provide read-after-write consistency for PUTS of new objects and eventual consistency for overwrite PUTS and DELETES.

[Learn more](http://docs.aws.amazon.com/AmazonS3/latest/dev/Introduction.html#ConsistencyModel)

Q: What happens 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: What is the BitTorrent™ protocol, and how do I use it with Amazon S3?

BitTorrent is an open source Internet distribution protocol. Amazon S3’s bandwidth rates are inexpensive, but BitTorrent allows developers to further save on bandwidth costs for a popular piece of data by letting users download from Amazon and other users simultaneously. Any publicly available data in Amazon S3 can be downloaded via the BitTorrent protocol, in addition to the default client/server delivery mechanism. Simply add the ?torrent parameter at the end of your GET request in the REST API.

Q: Does Amazon S3 offer a Service Level Agreement (SLA)?

Yes. The [Amazon S3 SLA](https://aws.amazon.com/s3/sla/) provides for a service credit if a customer’s monthly uptime percentage is below our service commitment in any billing cycle.

Q: How can I Increase the number of Amazon S3 buckets that I can provision?

By default, customers can provision up to 100 buckets per AWS account. However, you can increase your Amazon S3 bucket limit by visiting [AWS Service Limits](http://docs.aws.amazon.com/general/latest/gr/aws_service_limits.html).

## **Regions**

Q: Where is my data stored?

You specify a region when you create your Amazon S3 bucket. Within that region, your objects are redundantly stored on multiple devices across multiple facilities. Please refer to [Regional Products and Services](https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/) for details of Amazon S3 service availability by region.

Q: How do I decide which 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](https://aws.amazon.com/s3/pricing/).

Q: I’m not in the US or Europe; can I use Amazon S3?

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.

Q. Wasn’t there a US Standard region?

We renamed the US Standard Region to US East (Northern Virginia) Region to be consistent with AWS regional naming conventions. There is no change to the endpoint and you do not need to make any changes to your application.

## **Security**

Q: How secure is my data?

Amazon S3 is secure by default. Only the bucket and object owners originally 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. 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 or the Server Side Encryption with Customer-Provide Keys (SSE-C) option to encrypt data stored-at-rest. Amazon S3 provides the encryption technology for both SSE and SSE-C. 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, companies 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, companies 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](http://docs.amazonwebservices.com/AmazonS3/latest/dev/index.html?UsingAuthAccess.html) topic in the [Amazon S3 Developer Guide](http://docs.aws.amazon.com/AmazonS3/latest/dev/Welcome.html).

Q: Does Amazon S3 support data access auditing?

Yes, customers can optionally configure Amazon S3 buckets to create access log records for all requests made against it. 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](http://docs.amazonwebservices.com/AWSJavaSDK/latest/javadoc/com/amazonaws/services/s3/AmazonS3EncryptionClient.html). 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](https://aws.amazon.com/kms/) (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](http://docs.amazonwebservices.com/AWSJavaSDK/latest/javadoc/com/amazonaws/services/s3/AmazonS3EncryptionClient.html), 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](http://docs.amazonwebservices.com/AmazonS3/latest/dev/UsingEncryption.html) in the [Amazon S3 Developer Guide](http://docs.amazonwebservices.com/AmazonS3/latest/dev/).

Q: How does Amazon protect SSE encryption keys?

With SSE, every protected object is encrypted with a unique key. This object key is itself encrypted by a separate master key. A new master key is issued at least monthly. Encrypted data, encryption keys and master keys are stored and secured on separate hosts for multiple layers of protection.

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 (Ireland) or EU (Frankfurt) region. It is your responsibility to ensure that you comply with EU privacy laws.

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 document.](https://aws.amazon.com/whitepapers/overview-of-security-processes/)

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](http://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/vpc-endpoints.html).

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](http://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/vpc-endpoints.html).

Q: What is Amazon Macie?

[Amazon Macie](https://aws.amazon.com/macie/faq/) 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](http://docs.aws.amazon.com/macie/latest/userguide).

## **Data Protection**

Q: How durable is Amazon S3?

Amazon S3 Standard and Standard - IA are 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 objects with Amazon S3, you can on average expect to incur a loss of a single object once every 10,000,000 years. In addition, Amazon S3 is designed to sustain the concurrent loss of data in two facilities.

As with any environments, the best practice is to have a backup and to put in place safeguards against malicious or accidental users errors. For S3 data, that best practice includes secure access permissions, Cross-Region Replication, versioning and a functioning, regularly tested backup.

Q: How is Amazon S3 designed to achieve 99.999999999% durability?

Amazon S3 Standard and Standard - IA redundantly stores your objects on multiple devices across multiple facilities in an Amazon S3 Region. The service is designed to sustain concurrent device failures by quickly detecting and repairing any lost redundancy. When processing a request to store data, the service will redundantly store your object across multiple facilities before returning SUCCESS. Amazon S3 also regularly verifies 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](http://docs.aws.amazon.com/AmazonS3/latest/dev/ObjectVersioning.html).

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](http://docs.aws.amazon.com/AmazonS3/latest/dev/object-lifecycle-mgmt.html) 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](http://docs.aws.amazon.com/AmazonS3/latest/dev/object-lifecycle-mgmt.html) along with [Versioning](http://docs.aws.amazon.com/AmazonS3/latest/dev/ObjectVersioning.html) 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 MFA Delete capability, which uses [multi-factor authentication](http://aws.amazon.com/mfa/), 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](http://docs.amazonwebservices.com/AmazonS3/latest/dev/Versioning.html).

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](https://aws.amazon.com/s3/pricing/).

## **S3 Standard - Infrequent Access**

Q: What is S3 Standard - Infrequent Access?

Amazon S3 Standard - Infrequent Access (Standard - IA) is an Amazon S3 storage class for data that is accessed less frequently, but requires rapid access when needed. Standard - IA offers the high durability, throughput, and low latency of Amazon S3 Standard, with a low per GB storage price and per GB retrieval fee. This combination of low cost and high performance make Standard - IA ideal for long-term storage, backups, and as a data store for disaster recovery. The Standard - IA storage class is set at the object level and can exist in the same bucket as Standard, allowing you to use lifecycle policies to automatically transition objects between storage classes without any application changes.

Q: Why would I choose to use Standard - IA?

Standard - IA is ideal for data that is accessed less frequently, but requires rapid access when needed. Standard - IA is ideally suited for long-term file storage, older data from sync and share, backup data, and disaster recovery files.

Q: What performance does S3 Standard - Infrequent Access offer?

S3 Standard - Infrequent Access provide the same performance as S3 Standard storage.

Q: How durable is Standard - IA?

S3 Standard - IA is designed for the same 99.999999999% durability as Standard and Amazon Glacier. Standard - IA is designed for 99.9% availability, and carries a [service level agreement](https://aws.amazon.com/s3/sla/) providing service credits if availability is less than our service commitment in any billing cycle.

Q: How available is Standard - IA?

Designed for 99.9% availability, Standard - IA has a thinner front end that provides nine one-hundredths of a percent less availability than S3 Standard. Standard - IA carries a [service level agreement](https://aws.amazon.com/s3/sla/) providing service credits if availability is less than our service commitment in any billing cycle.

Q: How do I get my data into Standard - IA?

There are two ways to get data into Standard – IA from within S3. You can directly PUT into Standard – IA by specifying STANDARD\_IA in the x-amz-storage-class header. You can also set lifecycle policies to transition objects from Standard to Standard - IA.

Q: Are my Standard - IA objects backed with the Amazon S3 Service Level Agreement?

Yes, Standard - IA is backed with the [Amazon S3 Service Level Agreement](https://aws.amazon.com/s3/sla/), 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 Standard - IA?

You should expect the same latency and throughput performance as Amazon S3 Standard when using Standard - IA.

Q: How am I charged for using Standard - IA?

Please see the [Amazon S3 pricing page](https://aws.amazon.com/s3/pricing/) for general information about Standard - IA pricing.

Q. What charges will I incur if I change storage class of an object from Standard-IA to Standard with a copy request?

You will incur charges for an Standard-IA copy request and a Standard-IA data retrieval.

Q: Is there a minimum duration for Standard - IA?

Standard - IA is designed for long-lived, but infrequently accessed data that is retained for months or years. Data that is deleted from Standard - IA within 30 days will be charged for a full 30 days. Please see the [Amazon S3 pricing page](https://aws.amazon.com/s3/pricing/) for information about Standard - IA pricing.

Q: Is there a minimum object size for Standard - IA?

Standard - IA is designed for larger objects and 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 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](https://aws.amazon.com/s3/pricing/) for information about Standard - IA pricing.

Q: Can I tier objects from Standard - IA to Amazon Glacier?

Yes. In addition to using lifecycle policies to migrate objects from Standard to Standard - IA, you can also set up lifecycle policies to tier objects from Standard - IA to Amazon Glacier.

## **Amazon Glacier**

Q: Does Amazon S3 provide capabilities for archiving objects to lower cost storage options?

Yes, Amazon S3 enables you to utilize [Amazon Glacier’s](https://aws.amazon.com/glacier/) extremely low-cost storage service as storage for data archival.  Amazon Glacier stores data for as little as $0.004 per gigabyte per month. To keep costs low yet suitable for varying retrieval needs, Amazon Glacier provides three options for access to archives, 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 Glacier option?

You can use [Lifecycle rules](http://docs.aws.amazon.com/AmazonS3/latest/dev/object-lifecycle-mgmt.html) to automatically archive sets of Amazon S3 objects to Amazon Glacier based on lifetime. 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 Amazon S3 Standard or Amazon S3 Standard - IA objects which have names beginning with the specified prefix and which have aged past the specified time period are archived to Amazon Glacier. To retrieve Amazon S3 data stored in Amazon Glacier, initiate a retrieval job via the Amazon S3 APIs or Management Console. Once the job is complete, you can access your data through an Amazon S3 GET object request.

For more information on using Lifecycle rules for archival, please refer to the [Object Archival](http://docs.aws.amazon.com/AmazonS3/latest/dev/object-archival.html) 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 Glacier?

Yes, like Amazon S3’s other storage options (Standard or Standard - IA), Amazon 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 Amazon Glacier option, using the Amazon S3 LIST API.

Q: Can I use Amazon Glacier APIs to access objects that I’ve archived to Amazon Glacier?

Because Amazon S3 maintains the mapping between your user-defined object name and Amazon Glacier’s system-defined identifier, Amazon S3 objects that are stored using the Amazon Glacier option 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 Glacier?

To retrieve Amazon S3 data stored in Amazon Glacier, 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 RRS while leaving the archived data intact in Amazon Glacier. You can specify the amount of time in days for which the temporary copy is stored in RRS. You can then access your temporary copy from RRS through an Amazon S3 GET request on the archived object.

Q: How long will it take to retrieve my objects archived in Amazon Glacier?

When processing a retrieval job, Amazon S3 first retrieves the requested data from Amazon Glacier, and then creates a temporary copy of the requested data in RRS (which typically takes on the order of 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. Lastly, Bulk retrievals typically complete within 5 – 12 hours. For more information about the retrieval options, please refer to the [Glacier FAQ](https://aws.amazon.com/glacier/faqs/#dataretrievals).

Q: What am I charged for archiving objects in Amazon Glacier?

Amazon Glacier storage is priced from $0.004 per gigabyte per month. Lifecycle transition requests into Amazon Glacier cost $0.05 per 1,000 requests. Objects that are archived to 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.

Q: How is my storage charge calculated for Amazon S3 objects archived to Amazon 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 32 kilobytes of Glacier data plus an additional 8 KB of S3 standard storage data. Amazon Glacier requires an additional 32 KB 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 Glacier. This enables you to get a real-time list of all of your Amazon S3 objects, including those stored using the Amazon Glacier option, using the Amazon S3 LIST API. 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 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 region on the [Amazon S3 Pricing Page](https://aws.amazon.com/s3/pricing/).

Q: How much data can I retrieve for free?

You can retrieve 10 GB of your Amazon 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 am I charged for deleting objects from Amazon Glacier that are less than 3 months old?

Amazon Glacier is designed for use cases where data is retained for months, years, or decades. Deleting data that is archived to Amazon Glacier is free if the objects being deleted have been archived in Amazon Glacier for three months or longer. If an object archived in Amazon Glacier is deleted or overwritten within three months of being archived then there will be an early deletion fee. This fee is prorated. If you delete 1GB of data 1 month after uploading it, you will be charged an early deletion fee for 2 months of Amazon Glacier storage. If you delete 1 GB after 2 months, you will be charged for 1 month of Amazon Glacier storage.

Q: How much does it cost to retrieve data from Glacier?

There are three ways to retrieve data from 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 costs start at $0.03 per GB and $0.01 per request. Standard retrievals costs start at $0.01 per GB and $0.05 per 1,000 requests. Bulk retrievals costs start at $0.0025 per GB and $0.025 per 1,000 requests.

For example, using Expedited retrievals in the US East (Northern Virginia) region, 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 in the US East (Northern Virginia) region to retrieve 500 archives that were 1 GB each, the cost would be 500GB x $0.01 + 500 x $0.05/1,000 = $5.25

Lastly, using Bulk retrievals in the US East (Northern Virginia) region, 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.

When an archived object is retrieved, it resides in both RRS and Glacier.

To learn more about Glaicer pricing, please visit the [Glacier pricing page](https://aws.amazon.com/glacier/pricing/).

## **Event Notification**

Q1: 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](https://aws.amazon.com/sns/), [Amazon SQS](https://aws.amazon.com/sqs/), or directly to [AWS Lambda](https://aws.amazon.com/lambda/).

Q2: 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 Amazon S3. You can use Amazon 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 Amazon 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/."

Q3: 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](http://docs.aws.amazon.com/AmazonS3/latest/dev/NotificationHowTo.html) topic in the [Amazon S3 Developer Guide](http://docs.aws.amazon.com/AmazonS3/latest/dev/Welcome.html).

Q4: 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](http://docs.aws.amazon.com/AmazonS3/latest/dev/NotificationHowTo.html) topic in the [Amazon S3 Developer Guide](http://docs.aws.amazon.com/AmazonS3/latest/dev/Welcome.html). You can learn more about the AWS messaging services in the [Amazon SNS Documentation](http://docs.aws.amazon.com/sns/latest/dg/welcome.html) and the [Amazon SQS Documentation](http://docs.aws.amazon.com/AWSSimpleQueueService/latest/SQSGettingStartedGuide/Introduction.html).

Q5: What does it cost to use Amazon S3 event notifications?

There are no additional charges from 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 the AWS Lambda function. Visit the Amazon SNS, Amazon SQS, or AWS Lambda pricing pages to view the pricing details for these services.

## **Static Website Hosting**

Q: Can I host my static website on Amazon S3?

Yes, you can host your entire static website on Amazon S3 for an inexpensive, highly available hosting solution that scales automatically to meet traffic demands. Amazon S3 gives you access to the same highly scalable, reliable, fast, inexpensive infrastructure that Amazon uses to run its own global network of web sites. Service availability corresponds to storage class and the service level agreement provides service credits if a customer’s availability falls below our service commitment in any billing cycle. To learn more about hosting your website on Amazon S3, please see our [walkthrough](http://docs.aws.amazon.com/AmazonS3/latest/dev/website-hosting-custom-domain-walkthrough.html) on setting up an Amazon S3 hosted website.

Q: What kinds of websites should I host using Amazon S3 static website hosting?

Amazon S3 is ideal for hosting websites that contain only static content, including html files, images, videos, and client-side scripts such as JavaScript. Amazon EC2 is recommended for websites with server-side scripting and database interaction.

Q: Can I use my own host name with my Amazon S3 hosted website?

Yes, you can easily and durably store your content in an Amazon S3 bucket and map your domain name (e.g. “example.com”) to this bucket. Visitors to your website can then access this content by typing in your website’s URL (e.g., “http://example.com”) in their browser.

Q: Does Amazon S3 support website redirects?

Yes, Amazon S3 provides multiple ways to enable redirection of web content for your static websites. Redirects enable you to change the Uniform Resource Locator (URL) of a web page on your Amazon S3 hosted website (e.g. from www.example.com/oldpage to www.example.com/newpage) without breaking links or bookmarks pointing to the old URL. You can set rules on your bucket to enable automatic redirection. You can also configure a redirect on an individual S3 object.

Q: Is there an additional charge for hosting static websites on Amazon S3?

There is no additional charge for hosting static websites on Amazon S3. The same pricing dimensions of storage, requests, and data transfer apply to your website objects.

Refer to the [S3 Pricing](http://aws.amazon.com/s3/pricing/) page for more information.

## **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 new 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 lifecycle policies to manage transitions to the S3 Standard – Infrequent Access and Glacier storage tiers.

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 regions using Cross-Region Replication. For more information about setting up Cross-Region Replication, please visit [How to Set Up Cross-Region Replication](http://docs.aws.amazon.com/AmazonS3/latest/dev/crr-how-setup.html) in the [Amazon S3 Developer Guide](http://docs.aws.amazon.com/AmazonS3/latest/dev/crr.html).

For customers with Cross-Region Replication already enabled, new permissions are required in order for tags to replicate. For more information on the policies required, please visit [How to Set Up Cross-Region Replication](http://docs.aws.amazon.com/AmazonS3/latest/dev/crr-how-setup.html) in the [Amazon S3 Developer Guide](http://docs.aws.amazon.com/AmazonS3/latest/dev/crr.html).

Q. How much do Object Tags cost?

Object Tags are priced at $0.01 per 10,000 tags per month. The requests associated with adding and updating Object Tags are priced the same as existing request prices, please see the [Amazon S3 pricing page](https://aws.amazon.com/s3/pricing/effective-february-2014/) for more information.

### **S3 Analytics - Storage Class Analysis**

Q. What is S3 Analytics – 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 Analytics feature automatically identifies infrequent access patterns to help you transition storage to Standard-IA. You can configure a storage class analysis policy to monitor an entire bucket, a prefix, or object tag. Once an infrequent access pattern is observed, you can easily create a new 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 a S3 bucket to analyze using business intelligence tools of your choice such as Amazon QuickSight.

Q. How do I get started with S3 Analytics – 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 Standard-IA or archived to Glacier. You can navigate to the “Management” tab in the S3 Console to manage S3 Analytics, S3 Inventory, and S3 CloudWatch metrics.

Q. How am I charged for using S3 Analytics – Storage Class Analysis?

Please see the [Amazon S3 pricing page](https://aws.amazon.com/s3/pricing/) for general information about S3 Analytics – Storage Class Analysis pricing.

Q. How often is the Storage Class Analysis updated?

Storage Class Analysis is updated on a daily basis on the S3 Management Console. Additionally, you can configure S3 Analytics to export your daily storage class analysis to a S3 bucket of your choice.

### **S3 Inventory**

Q. What is S3 Inventory?

You can simplify and speed up business workflows and big data jobs using S3 Inventory which provides a scheduled alternative to Amazon S3’s synchronous List API. S3 Inventory provides a CSV (Comma Separated Values) flat-file output of your objects and their corresponding metadata on a daily or weekly basis for an S3 bucket or a shared prefix.

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 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 inventory, the output file output format (CSV), 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, or replication status.

Q. Can S3 Inventory improve the performance for my big data jobs and business workflow applications?

Yes, S3 Inventory can be used as a ready-made input into a big data job or workflow application instead of the synchronous S3 LIST API, saving the time and compute resources it takes to call and process the LIST API response.

Q. How am I charged for using S3 Inventory?

Please see the [Amazon S3 pricing page](https://aws.amazon.com/s3/pricing/) for general information about S3 Inventory pricing.

### **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 metrics for your S3 bucket or configure filters for the metrics using a prefix or object tag. Alternately, you can call the S3 PUT Bucket Metrics API to enable and configure publication of S3 storage metrics. Storage metrics will be available in CloudWatch within 15 minutes of being enabled.

Q. Can I align storage metrics to my applications or business organizations?

Yes, you can configure S3 CloudWatch metrics to generate metrics for your S3 bucket or configure filters for the metrics using a prefix or object tag. For example, you can monitor a spark application that accesses data under the prefix “/Bucket01/BigData/SparkCluster” as metrics filter 1 and define a second metrics filter with the tag “Dept, 1234” as metrics filter 2. An object can be a member of multiple filters, e.g., an object within the prefix “/Bucket01/BigData/SparkCluster” and with the tag “Dept,1234” will be in both metrics filter 1 and 2. In this way, metrics filters can be aligned to business applications, team structures or organizational budgets, allowing you to monitor and alert on multiple workloads separately within the same S3 bucket.

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 fire 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 fire a CloudWatch alarm to alert a Dev Ops engineer.

Q. How am I charged for using S3 CloudWatch Metrics?

S3 CloudWatch Metrics are priced as custom metrics for Amazon CloudWatch. Please see [Amazon CloudWatch pricing page](https://aws.amazon.com/cloudwatch/pricing/) for general information about S3 CloudWatch metrics pricing.

### **Lifecycle Management Policies**

Q. What is 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 lifecycle transition policy to automatically migrate Amazon S3 objects to Standard - Infrequent Access (Standard - IA) and/or Amazon Glacier 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 upload based on the age of the upload.

Q. How do I set up a 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 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](http://aws.amazon.com/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. For example, you can set infrequently accessed objects to move into lower cost storage tier (like Standard-Infrequent Access) after a period of time. After another period, it can be moved into Amazon Glacier for archive and compliance, and eventually deleted. These rules can invisibly lower storage costs and simplify management efforts, and may be leveraged across the Amazon family of storage services. 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’s lifecycle policy to help lower my Amazon S3 storage costs?

With Amazon S3’s lifecycle policies, you can configure your objects to be migrated to Standard - Infrequent Access (Standard - IA), archived to Amazon Glacier, or deleted 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 Standard - IA or Amazon Glacier, and/or an expiration. For example, you could create a rule that archives into Amazon 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. 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.

[Learn more.](http://docs.aws.amazon.com/AmazonS3/latest/dev/object-lifecycle-mgmt.html)

Q: How can I configure my objects to be deleted after a specific time period?

You can set a 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. Each Object Expiration rule allows you to specify a prefix and an expiration period. 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/”). For expiration period, provide the number of days from creation date (i.e. age) after which you want your objects removed. You may create multiple rules for different prefixes. For example, you could create a rule that removes all objects with the prefix “logs/” 30 days from creation, and a separate rule that removes all objects with the prefix “backups/” 90 days from creation.

After an Object Expiration rule is added, the rule is applied to objects that already exist in the bucket as well as new objects added to the bucket. Once objects are past their expiration date, they are identified and queued for removal. You will not be billed for storage for objects on or after their expiration date, though you may still be able to access those objects while they are in queue before they are removed. As with standard delete requests, Amazon S3 doesn’t charge you for removing objects using Object Expiration. You can set Expiration rules for your versioning-enabled or versioning-suspended buckets as well.

[Learn more.](http://docs.aws.amazon.com/AmazonS3/latest/dev/object-lifecycle-mgmt.html)

Q. Why would I use a lifecycle policy to expire incomplete multipart uploads?

The 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](http://docs.aws.amazon.com/AmazonS3/latest/dev/mpuoverview.html#mpu-abort-incomplete-mpu-lifecycle-config).

### **Cross-Region Replication**

Q: What is Amazon S3 Cross-Region Replication (CRR)?

CRR is an Amazon S3 feature that automatically replicates data across AWS regions. With CRR, every object uploaded to an S3 bucket is automatically replicated to a destination bucket in a different AWS region that you choose. 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.

Q: How do I enable CRR?

CRR is a bucket-level configuration. You enable a CRR configuration on your source bucket by specifying a destination bucket in a different region for replication. You can use either the AWS Management Console, the REST API, the AWS CLI, or the AWS SDKs to enable CRR. Versioning must be turned on for both the source and destination buckets to enable CRR. To learn more, please visit [How to Set Up Cross-Region Replication](https://docs.aws.amazon.com/AmazonS3/latest/dev/crr-how-setup.html) in the [Amazon S3 Developer Guide](https://docs.aws.amazon.com/AmazonS3/latest/dev/crr.html).

Q: What does CRR replicate to the target bucket?

CRR replicates every object-level upload that you directly make to your source bucket. The metadata and ACLs associated with the object are also part of the replication. Any change to the underlying data, metadata, or ACLs on the object would trigger a new replication to the destination bucket. You can either choose to replicate all objects uploaded to a source bucket or just a subset of objects uploaded by specifying prefixes. Existing data in the bucket prior to enabling CRR is not replicated. You can use S3’s COPY API to copy the existing data into your destination bucket. To learn more about CRR please visit [How to Set Up Cross-Region Replication](https://docs.aws.amazon.com/AmazonS3/latest/dev/crr-how-setup.html) in the [Amazon S3 Developer Guide](https://docs.aws.amazon.com/AmazonS3/latest/dev/crr.html).

Q: Can I use CRR with lifecycle rules?

Yes, you can configure separate lifecycle rules on the source and destination buckets. For example, you can configure a lifecycle rule to migrate data from Standard to Standard - IA on the destination bucket or configure a lifecycle rule to archive data into Amazon Glacier.

Q: What is the pricing for CRR?

You pay the Amazon S3 charges for storage, requests, and inter-region data transfer for the replicated copy of data. For example, if you replicate 1,000 1 GB objects (1,000 GB) between regions you will incur a request charge of $0.005 (1,000 requests x $0.005 per 1,000 requests) for replicating 1,000 objects and a charge of $20 ($0.020 per GB transferred x 1,000 GB) for inter-region data transfer. After replication, the 1,000 GB will incur storage charges based on the destination region.

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 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.  
  
Please visit the [S3 pricing](https://aws.amazon.com/s3/pricing/) page for more information.

## **Amazon S3 Transfer Acceleration**

Q. What is 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. 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 Transfer Acceleration?

It's easy to get started with Transfer Acceleration. First, [enable](http://docs.aws.amazon.com/AmazonS3/latest/dev/transfer-acceleration.html#transfer-acceleration-getting-started) Transfer Acceleration on an S3 bucket using the Amazon S3 console, the Amazon S3 API, or the AWS CLI. After 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: <bucketname>.s3-accelerate.amazonaws.com or <bucketname>.s3-accelerate.dualstack.amazonaws.com for the “[dual-stack](http://docs.aws.amazon.com/AmazonS3/latest/dev/dual-stack-endpoints.html)” endpoint. If you want to use standard data transfer, you can continue to use the regular endpoints.

There are certain restrictions on which bucket will work with transfer acceleration. For details, please refer the Amazon S3 developer guide [here](http://docs.aws.amazon.com/AmazonS3/latest/dev/transfer-acceleration.html#transfer-acceleration-requirements).

Q. How fast is Transfer Acceleration?

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. 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](http://s3-accelerate-speedtest.s3-accelerate.amazonaws.com/en/accelerate-speed-comparsion.html) to get a preview of the performance benefit from your location!

Q. Who should use Transfer Acceleration?

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.

Q. How secure is Transfer Acceleration?

Transfer Acceleration provides the same security as regular transfers to Amazon S3. All Amazon S3 security features, such as restricting access based on a client’s IP address, are supported as well. 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 Transfer Acceleration isn't faster?

Each time you use Transfer Acceleration to upload an object, we will check whether Transfer Acceleration is likely to be faster than a regular Amazon S3 transfer. If we determine that 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 that use of Transfer Acceleration for that transfer, and may bypass the Transfer Acceleration system for that upload.

Q. Can I use Transfer Acceleration with multipart uploads?

Yes, Transfer Acceleration supports all bucket level features including multipart upload.

Q. How should I choose between Transfer Acceleration and Amazon CloudFront’s PUT/POST?

Transfer Acceleration optimizes the TCP protocol and adds additional intelligence between the client and the S3 bucket, making 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 Transfer Acceleration and AWS Snowball?

The AWS Import/Export Snowball 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, Transfer Acceleration over a fully-utilized 1 Gbps line can transfer up to 75 TBs in the same time. 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, 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 Transfer Acceleration.

Q. Can Transfer Acceleration complement AWS Direct Connect?

AWS Direct Connect is a good choice for customers with a private networking requirement or have access to AWS Direct Connect exchanges. 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 Transfer Acceleration to help with remote office transfers, where they may suffer from poor Internet performance.

Q. Can 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](https://aws.amazon.com/storagegateway/faqs/#file) to learn more about the AWS implementation.

Q. Can Transfer Acceleration complement 3rd party integrated software?

Yes. Software packages that connect directly into Amazon S3 (read more about storage partner solutions [here](https://aws.amazon.com/backup-recovery/partner-solutions/)) can take advantage of Transfer Acceleration when they send their jobs to Amazon S3.

Q: Is 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 enables fast, easy, and secure transfers of files including protected health information (PHI) over long distances between your client and your Amazon S3 bucket. For more information, see [HIPAA Compliance](https://aws.amazon.com/compliance/hipaa-compliance).

## **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 needing 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](http://docs.aws.amazon.com/AmazonS3/latest/dev/dual-stack-endpoints.html), 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. Your applications may continue to access data through the existing APIs and virtual hosted style (e.g. http://bucket.s3.dualstack.aws-region.amazonaws.com) or path style (e.g. http://s3.dualstack.aws-region.amazonaws.com/bucket) URLs without code changes. When using Amazon S3 Transfer Acceleration, the “dual-stack” endpoint must be of the form http(s)://bucket.s3-accelerate.dualstack.amazonaws.com. However, you must also evaluate your bucket and Identity and Access Management (IAM) policies to ensure you have the appropriate access configured for your new IPv6 addresses. For more information about getting started accessing Amazon S3 over IPv6, see [Making Requests to Amazon S3 over IPv6](http://docs.aws.amazon.com/AmazonS3/latest/dev/ipv6-access.html).

Q. If I point to Amazon S3's "dual-stack" endpoint, will I still be able to access Amazon S3's APIs over IPv4?

Yes, you can continue to access Amazon S3 APIs using both IPv6 and IPv4 addresses when connecting to the Amazon S3 “dual-stack” endpoints. You will need to configure your client to prefer IPv4 addresses, which can be an application-level or host-level configuration option for many application runtime languages. Please consult the documentation for the language you are using for your runtime platform for the specific configuration option that prefers IPv4 connections.

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. Will existing VPC Endpoints continue to work if I point to Amazon S3's "dual-stack" endpoint?

Yes, you can continue using VPC Endpoint to access Amazon S3 over IPv4. If you use the dual-stack endpoint in an IPv4-only VPC, the VPC instances will drop the AAAA record and always access Amazon S3 over IPv4.

Q. If I enable IPv6, will the IPv6 address appear in the Server Access Log?

Yes, IPv6 addresses will now be shown in the Server Access logs if you have the Amazon S3 Server Access logs feature enabled. Any customer tool or software that parses the logs should be updated to handle the new IPv6 address format. Please contact [Developer Support](https://aws.amazon.com/contact-us/) if you have any issues with IPv6 traffic impacting your tool or software’s ability to handle IPv6 addresses in Server Access logs.

Q. Do I need to update my bucket and IAM policies?

Yes, if you use policies to grant or restrict access via IP addresses, you will need to update those policies to include the associated IPv6 ranges before you switch to the “dual-stack” endpoint. If your bucket grants or restricts access to specific IAM users, you will also need to have the IAM policy administrator review those users’ IAM policies to ensure they have appropriate access to the associated IPv6 ranges before you switch to the “dual-stack” endpoint. Failure to do so may result in clients incorrectly losing or gaining access to the bucket when they start using IPv6.

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 regions?

You can use IPv6 with Amazon S3 in all commercial AWS Regions except China (Beijing). You can also use IPv6 in the AWS GovCloud (US) region.

# Amazon EBS

Q: Are Amazon EBS volume and snapshot ID lengths changing in 2016?

Yes, please visit [the EC2 FAQ](https://aws.amazon.com/ec2/faqs-old-convertible/#longer-ids) 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. For more performance information see the [EBS product details page](https://aws.amazon.com/ebs/details/).

For more information about Amazon EBS performance guidelines, see [Increasing EBS Performance](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSPerformance.html).

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](https://aws.amazon.com/ebs/details/#elasticvolumes) 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](https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-expand-volume.html).

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](https://aws.amazon.com/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](https://aws.amazon.com/ebs/details/#ebsoptimized) deliver dedicated throughput between Amazon EC2 and Amazon EBS, with options between 62.5 MB/s and 1,750 MB/s depending on the instance type used.

### **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 process your application reads and writes in I/O sizes of 256KB or less. Every increase in I/O size above 256KB increases linearly the resources you need to achieve the same IOPS rate. For example, if you have provisioned a volume with 500 IOPS, that means that it can handle 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](https://aws.amazon.com/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 500 provisioned IOPS in a minute. For example, for a volume provisioned with 1500 IOPS, the queue depth average must be 3. For more information about ensuring consistent performance of your volumes, see [Increasing EBS Performance](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSPerformance.html).

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.](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSPerformance.html)

Q: Can I stripe multiple volumes together to get better performance?

Yes. You can stripe multiple volumes together to achieve up to 75,000 IOPS or 1,750 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.

### **Snapshots**

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](https://console.aws.amazon.com/ec2/v2/home#Snapshots:visibility=public;sort=startTime). All information on public datasets is available in our [AWS Public Datasets](https://aws.amazon.com/public-datasets/) resource center.

### **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](https://aws.amazon.com/kms/) (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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSEncryption.html).

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](https://aws.amazon.com/kms/) 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](https://aws.amazon.com/what-is-cloud-computing/). 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.

### **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](http://aws.amazon.com/support/trustedadvisor/).  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](https://aws.amazon.com/c-tax-faqs/).

# Amazon Elastic File System (EFS)

. What is Amazon Elastic File System?

Amazon EFS is a fully-managed service that makes it easy to set up and scale [file storage](https://aws.amazon.com/what-is-cloud-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 that 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 the storage you use.

Q. What use cases is Amazon EFS intended for?

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 Simple Storage Service (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](https://aws.amazon.com/efs/) is a [file storage](https://aws.amazon.com/what-is-cloud-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](https://aws.amazon.com/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](https://aws.amazon.com/s3/) is an object storage service. Amazon S3 makes data available through an Internet API that can be accessed anywhere.

Q. Where is my data stored?

Please refer to [Regional Products and Services](https://aws.amazon.com/about-aws/global-infrastructure/regional-product-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 Amazon Web Services account. If you do not already have an AWS account, you can create one by clicking the “Try the Free Tier” button on the [Amazon EFS detail page](https://aws.amazon.com/efs/details/).

Once you have created an AWS account, please refer to the Amazon EFS [Getting Started](http://docs.aws.amazon.com/efs/latest/ug/gs-mount-fs-on-ec2instance-and-test.html) 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. 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](http://docs.aws.amazon.com/efs/latest/ug/gs-mount-fs-on-ec2instance-and-test.html) guide.

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 Amazon EFS [Getting Started](http://docs.aws.amazon.com/efs/latest/ug/gs-mount-fs-on-ec2instance-and-test.html) guide.

Q. How do I manage a file system?

Amazon EFS is a fully-managed service, so all of the [file storage](https://aws.amazon.com/what-is-cloud-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, and display detailed information about file systems.

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](https://aws.amazon.com/efs/faq/#on-premises) section of this FAQ.

For more information about moving data to the Amazon cloud, please see the [Cloud Data Migration](https://aws.amazon.com/cloud-data-migration/) page.

## **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. How do I back up a file system?

Amazon EFS is designed to be highly durable. If you want to be able to revert to earlier versions of files to undo changes, you can use standard 3rd party backup software.

You can also use [AWS Data Pipeline](https://aws.amazon.com/datapipeline/) to create regular, automated copies of your file system based on a schedule that you define. For more information and to access an AWS Data Pipeline template provided by Amazon EFS, please see the Amazon EFS [Walkthrough: Back Up an EFS File System](http://docs.aws.amazon.com/efs/latest/ug/efs-backup.html).

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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/vpc-classiclink.html). On-premises servers can mount your file systems via an [AWS Direct Connect](https://aws.amazon.com/directconnect/) 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 or specify a size up front and you pay only for the storage 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?

By default, you can create up to 10 file systems per AWS account per region. You can request to increase your file system limit by visiting [AWS Service Limits](http://docs.aws.amazon.com/general/latest/gr/aws_service_limits.html).

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 Amazon’s file and block cloud storage offerings.

|  |  |  |
| --- | --- | --- |
|  | Amazon EFS | Amazon EBS PIOPS |
| 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](http://docs.aws.amazon.com/efs/latest/ug/limits-throughput.html).

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 storage, all file systems (regardless of size) can burst to 100 MB/s, and file systems larger than 1TB can burst to 100 MB/s per TB of storage. As you add data to your file system, the maximum throughput available to the file system scales linearly and automatically with your storage.

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 100MB/s of throughput can drive 100MB/s from a single Amazon EC2 instance, or 10 Amazon EC2 instances can collectively drive 100MB/s. For more information, please see the documentation on [File System Performance](http://docs.aws.amazon.com/efs/latest/ug/limits-throughput.html).

## **Access Control**

Q. How do I control which Amazon EC2 instances can access my file system?

When you create a file system, you create endpoints in your VPC called “mount targets.” When mounting from an EC2 instance, your file system’s DNS name, which you provide in your mount command, resolves to a mount target’s IP address. Only resources that can access a mount target can access your file system. You can control the network traffic to and from your file system mount targets using VPC [security groups](https://docs.aws.amazon.com/efs/latest/ug/security-considerations.html#network-access).

Q. How do I control who can access my file system?

You can control who can administer your file system using [AWS Identity and Access Management (IAM)](https://docs.aws.amazon.com/efs/latest/ug/auth-and-access-control.html#access-control). You can control access to files and directories with POSIX-compliant [user and group-level permissions](https://docs.aws.amazon.com/efs/latest/ug/accessing-fs-nfs-permissions.html).

## **Encryption**

Q: What is Amazon EFS Encryption?

Amazon EFS seamlessly offers encryption of EFS file systems. Data 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.

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 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](https://aws.amazon.com/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](https://aws.amazon.com/kms/) product page.

Q: How do I enable encryption for my Amazon EFS file system?

You can enable encryption for your EFS file system 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. For more details, see the user documentation on [Encryption](http://docs.aws.amazon.com/efs/latest/ug/encryption.html).

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 connection between your on-premises datacenter and your Amazon VPC. AWS Direct Connect establishes a private network connection between your on-premises datacenter and AWS. If you do not already have an AWS Direct Connect connection, you can create one by following the instructions in the [AWS Direct Connect user guide](http://docs.aws.amazon.com/directconnect/latest/UserGuide/Welcome.html).

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 via AWS Direct Connect, please see the [documentation](http://docs.aws.amazon.com/efs/latest/ug/how-it-works.html#how-it-works-direct-connect).

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. 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 use an AWS VPN to access an EFS file system from on-premises?

No, Amazon EFS does not support access over AWS VPN.

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 via AWS Direct Connect?

Because of the propagation delay tied to data traveling over long distances, the network latency of a Direct Connect 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 Direct Connect 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](https://www.gnu.org/software/parallel/).

## **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 the amount of file system storage you use per month in GB. There is no minimum fee and no set-up costs. There are no additional costs for bandwidth or requests. For Amazon EFS pricing information, please visit the pricing section on the [Amazon EFS Pricing page](https://aws.amazon.com/efs/pricing/).

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](http://aws.amazon.com/c-tax-faqs/).

# Amazon Glacier

Q: What is Amazon Glacier?

Amazon Glacier is an extremely low-cost storage service that provides secure, durable, and flexible storage for data backup and archival. With Amazon Glacier, customers can reliably store their data for as little as $0.004 per gigabyte per month. Amazon 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 Glacier?

Amazon Glacier enables any business or organization to easily and cost effectively retain data for months, years, or decades. With Amazon 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](http://docs.aws.amazon.com/amazonglacier/latest/dev/vault-lock.html) to meet regulatory and compliance archiving requirements.

Q: How should I choose between Amazon 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 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](http://docs.aws.amazon.com/amazonglacier/latest/dev/vault-lock.html) to store regulatory and compliance archives in an immutable, Write Once Read Many (WORM) format. Please refer to the [Amazon Web Services Licensing Agreement](https://aws.amazon.com/agreement/) for details.

Q: What does Amazon do with my data in Amazon 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 Glacier offering, except if required to do so by law. Please refer to the [Amazon Web Services Licensing Agreement](https://aws.amazon.com/agreement/) for details.

Q: How do I use Amazon Glacier?

Amazon 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 Glacier. Data can then be uploaded and retrieved programmatically. View our documentation for more information on the Glacier APIs and SDKs.

Q: How durable is Amazon Glacier?

Amazon 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 Glacier synchronously stores your data across multiple facilities before returning SUCCESS on uploading archives. Glacier performs regular, systematic data integrity checks and is built to be automatically self-healing.

## **Getting Started**

Q: How is data within Amazon Glacier organized?

You store data in Amazon 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 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 Glacier?

There is no minimum limit to the amount of data that can be stored in Amazon Glacier and individual archives can be from 1 byte to 40 terabytes.

Q: Does the AWS Management Console support Amazon Glacier?

Yes. The AWS Management Console allows you to create and configure vaults, allowing you to easily and quickly setup Glacier. [Click here](https://aws.amazon.com/console/) to go the AWS Management Console.

## **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 Glacier by using the [AWS Identity and Access Management](https://aws.amazon.com/iam) (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 Glacier handles key management and key protection for you. Amazon 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 Glacier support IAM permissions?

Yes, 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](https://aws.amazon.com/iam)
* [AWS Identity and Access Management Getting Started Guide](http://docs.amazonwebservices.com/IAM/latest/GettingStartedGuide/)
* [Using AWS Identity and Access Management](http://docs.amazonwebservices.com/IAM/latest/UserGuide/)

## **Archives and Vaults**

Q: What is an archive?

An archive is a durably stored block of information. You store your data in Amazon 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 Glacier. The total volume of data and number of archives you can store are unlimited. Individual Amazon 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 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 Glacier. You organize your data in Amazon 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)](https://aws.amazon.com/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](https://aws.amazon.com/sns/) 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 Glacier vaults?

Amazon 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 Glacier costs and usage across multiple departments in your organization or by any other categorization. You can tag your vaults by using the [Glacier Console](https://console.aws.amazon.com/glacier) or the [Glacier APIs](http://docs.aws.amazon.com/amazonglacier/latest/dev/vault-operations.html). For more information see [Tagging Your Amazon Glacier Vaults](http://docs.aws.amazon.com/amazonglacier/latest/dev/tagging.html).

Q: How do I delete a vault?

You may delete any Glacier vault that does not contain any archives using the AWS Management Console, the Amazon Glacier 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](http://docs.aws.amazon.com/IAM/latest/UserGuide/PermissionsOverview.html#TypesPermissions) that you can attach directly to your 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](http://docs.aws.amazon.com/amazonglacier/latest/dev/vault-access-policy.html) in the Amazon 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](http://docs.aws.amazon.com/IAM/latest/UserGuide/PermissionsOverview.html#TypesPermissions) or as [resource-based permissions](http://docs.aws.amazon.com/IAM/latest/UserGuide/PermissionsOverview.html#TypesPermissions). 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 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](http://docs.aws.amazon.com/amazonglacier/latest/dev/vault-access-policy.html) in the Amazon 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 Glacier vaults via a lockable policy (Vault Lock policy). Once locked, the Vault Lock policy becomes immutable and Glacier will enforce the prescribed controls to help achieve your compliance objectives. To learn more, please read [Amazon Glacier Vault Lock](http://docs.aws.amazon.com/amazonglacier/latest/dev/vault-lock.html) in the Amazon 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 Glacier Vault Lock](http://docs.aws.amazon.com/amazonglacier/latest/dev/vault-lock.html) in the Amazon 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 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](http://docs.aws.amazon.com/amazonglacier/latest/dev/vault-lock.html) in the Amazon 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 AWS Glacier for electronic storage along with a copy of the [Cohasset Assessment](https://d0.awsstatic.com/whitepapers/Amazon-GlacierVaultLock_CohassetAssessmentReport.pdf). 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 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 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 AWS Glacier console or use the Vault Lock APIs in the AWS SDK. To learn more, please read [Getting Started with Amazon Glacier Vault Lock](http://docs.aws.amazon.com/amazonglacier/latest/dev/vault-lock.html) in the Amazon Glacier developer’s guide.

## **Data Retrievals**

Q: How can I retrieve data from the service?

When you make a request to retrieve data from Glacier, you initiate a retrieval job for an archive. Once the retrieval job completes, your data will be available to download or access it using [Amazon Elastic Compute Cloud](https://aws.amazon.com/ec2) (Amazon EC2) for 24 hours. There are three options for retrieving data with varying access times and cost: Expedited, Standard, and Bulk retrievals.

Q: What are Standard retrievals?

Standard retrievals allow you to access any of your archives within several hours. Standard retrievals typically complete within 3 – 5 hours.

Q: How do I use Standard retrievals?

To make a Standard retrieval, set the “Tier” parameter in the InitiateJob API request to “Standard”. If no tier is specified, the request will default to Standard.

Q: How much do Standard retrievals cost?

Standard retrievals are priced at a flat rate of $0.01 per GB and $0.05 per 1,000 requests. For example, retrieving 500 archives that are 1 GB each would cost 500GB x $0.01 + 500 x $0.05/1,000 = $5.025

Q: When should I use Standard retrievals?

Standard retrievals are a low-cost way to access your data within just a few hours. For example, you can use Standard retrievals to restore backup data, retrieve archived media content for same-day editing or distribution, or pull and analyze logs to drive business decisions within hours.

Q: What are Bulk retrievals?

Bulk retrievals are Glacier’s lowest-cost retrieval option, enabling you to retrieve large amounts, even petabytes, of data inexpensively in a day. Bulk retrievals typically complete within 5 – 12 hours.

Q: How do I use Bulk retrievals?

To make a Bulk retrieval, set the “Tier” parameter in the InitiateJob API request to Bulk.

Q: How much do Bulk retrievals cost?

Bulk retrievals are priced at a flat rate of just $0.0025 per GB and $0.025 per 1,000 requests. For example, retrieving 500 archives that are 1 GB each would cost 500GB x $0.0025 + 500 x $0.025/1,000 = $1.2625.

Q: When should I use Bulk retrievals?

Bulk retrievals are designed to enable customers to cost-effectively pull large amounts of data for non-urgent use cases such as transcoding petabytes of raw video content or analyzing large genomics sequences.

Q: What are Expedited retrievals?

Expedited retrievals allow you to quickly access your data when occasional urgent requests for a subset of archives are required. For all but the largest archives (250MB+), data accessed using Expedited retrievals are typically made available within 1 – 5 minutes. There are two types of Expedited retrievals: On-Demand and Provisioned. On-Demand requests are like EC2 On-Demand instances and are available the vast majority of the time. Provisioned requests are guaranteed to be available when you need them.

Q: What is a Provisioned capacity unit?

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.

Q: When should I provision retrieval capacity?

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.

Q: How do I purchase provisioned capacity?

You can purchase provisioned capacity using the console, SDK, or the CLI.

Q: How much does provisioned capacity cost?

Each unit of provisioned capacity costs $100 per month from the date of purchase.

Q: How do I use Expedited retrievals?

To make an Expedited retrieval, set the “Tier” parameter in the InitiateJob API request to Expedited. There is no need to designate whether an Expedited retrieval is On-Demand or Provisioned. If you have purchased provision capacity, then all Expedited retrievals will be automatically be served via your Provisioned capacity.

Q: How much do Expedited retrievals cost?

Expedited retrievals are priced at a flat rate of $0.03 per GB and $0.01 per request. For example, retrieving 10 objects with a size of 1GB each, the cost would be 10 x $0.03 +10 x $0.01 = $0.40.

Q: When should I use Expedited retrievals?

Expedited retrievals are optimized for the occasional urgent request for a small number of archives. For all but the largest archives (250MB+), data accessed using Expedited retrievals are typically made available within 1 – 5 minutes. If your application or workload requires a guarantee that your Expedited retrievals will be available when you need it, then you should consider using Provisioned capacity.

Q: Can I retrieve part of an archive?

Yes, range retrievals enable you to retrieve a specific range of an archive. Range retrievals are similar to regular retrievals in Amazon Glacier. Both require the initiation of a retrieval job (See [How can I retrieve data?](https://aws.amazon.com/glacier/faqs/#How_can_I_retrieve_data) for more information). You can use range retrievals to reduce or eliminate your retrieval fees (See [How much data can I retrieve for free?](https://aws.amazon.com/glacier/faqs/#How_much_data_can_I_retrieve_for_free))

When initiating a retrieval job using range retrievals, you provide a byte range that can start at zero (which would be the beginning of your archive), or at any 1MB interval thereafter (e.g. 1MB, 2MB, 3MB, etc). The end of the range can either be the end of your archive or any 1MB interval greater than the beginning of your range.

Q: Why would I retrieve only a range of an archive?

There are several reasons why you might choose to perform a range retrieval. For example, you may have aggregated several files and uploaded them as a single archive. You may then need to retrieve a small selection of those files, in which case you could retrieve only the ranges of the archive that contained the required files. Another reason you could choose to perform a range retrieval is to manage how much data you download from Amazon Glacier in a given period. When you make a request to retrieve data from Glacier, you initiate a retrieval job for an archive. Once the retrieval job completes, your data will be available to download or access using [Amazon Elastic Compute Cloud](https://aws.amazon.com/ec2) (Amazon EC2) for 24 hours. The data retrieved is then available for download for 24 hours. You could therefore retrieve an archive in parts in order to manage the schedule of your downloads.

Q: How do I view my jobs?

You can list your ongoing jobs for any of your vaults by calling the ListJobs API. The list of jobs provides information including the job’s creation time and date and the job’s status (e.g. in-progress, completed successfully, or not in which case reasons for the job not succeeding are provided). The progress of a single job can be tracked by calling the DescribeJob API and providing the corresponding job ID. The status of the job will be returned immediately.

Q: Can I be notified when a job is completed?

Yes. You can optionally configure vaults to send notifications to you or your application when jobs complete. Notifications will be delivered via the Amazon Simple Notification Service ([Click here](http://docs.aws.amazon.com/amazonglacier/latest/dev/configuring-notifications.html) to learn more about Amazon SNS).

## **Data Retrieval Policies**

Q: What are data retrieval policies?

Amazon Glacier data retrieval policies let you define your own data retrieval limits with a few clicks in the AWS console. You can limit retrievals to “Free Tier Only”, or if you wish to retrieve more than the free tier, you can specify a “Max Retrieval Rate” to limit your retrieval speed and establish a retrieval cost ceiling. In both cases, Amazon Glacier will not accept retrieval requests that would exceed the retrieval limits you defined. Retrieval policies apply to Standard retrievals.

To learn more please read [Configuring Data Retrieval Policies](http://docs.aws.amazon.com/amazonglacier/latest/dev/data-retrieval-policy.html) in the Amazon Glacier developer’s guide.

Q: How do I set up data retrieval policies?

You can set up data retrieval policies in the Amazon Glacier console or via the Amazon Glacier APIs. To learn more please read [Configuring Data Retrieval Policies](http://docs.aws.amazon.com/amazonglacier/latest/dev/data-retrieval-policy.html) in the Amazon Glacier developer’s guide.

Q: Are data retrieval policies specific to each AWS region?

Yes. You can set one data retrieval policy for each AWS region which will govern all data retrieval activities in the region under your account. Data retrieval policies are region-specific because data retrieval costs vary across AWS regions.

Please visit [Amazon Glacier Pricing](https://aws.amazon.com/glacier/pricing/) for more information.

Q: Can I use data retrieval policies to “slow down” my retrievals or spread them out?

No, data retrieval policies such as “Free Tier Only” and “Max Retrieval Rate” will not accept a data retrieval request which would exceed your predefined data retrieval limit to help you manage data retrieval cost. Data retrieval policies will not change the 3 to 5 hour data retrieval latency or spread out your retrievals. You can leverage Amazon Glacier’s range retrieval feature to spread out retrievals and lower the peak retrieval speed. [Learn more](http://docs.aws.amazon.com/amazonglacier/latest/dev/downloading-an-archive.html#downloading-an-archive-range).

Q: What impact does the change in the retrieval free tier to 10 GB per month have on my data retrieval policy?

There is no impact to your policy in terms of the rate of data retrieval. If your retrieval policy was set to the previous free tier of 5% of your average monthly storage prior to the change in the retrieval free tier on November 21, 2016, your policy will remain the same GB-per-hour retrieval rate as your previous 5% free tier rate as of November 21, 2016. For example, if on that day your average monthly storage was 14,400 GB, your retrieval rate limit was 14,400 GB x 5% / 30 day / 24 hours = 1 GB per hour. Your new policy will remain at 1 GB per hour, but will be a “Max Retrieval Rate” rather than a “Free Tier Only” policy.

## **Data Inventories**

Q: Can I see what archives I have stored in Amazon Glacier?

Yes. Although you will need to maintain your own index of data you upload to Amazon Glacier, an inventory of all archives in each of your vaults is maintained for disaster recovery or occasional reconciliation purposes. The vault inventory is updated approximately once a day. You can request a vault inventory as either a JSON or CSV file and will contain details about the archives within your vault including the size, creation date and the archive description (if you provided one during upload). The inventory will represent the state of the vault at the time of the most recent inventory update.

Q: Can I obtain a real time list of my vaults?

Yes, you can list your vaults stored in Amazon Glacier using either the AWS Management Console or by calling the ListVaults API. As well as a list of vault names, you will also be able to see when the vault’s inventory was last updated and a summary of the vault’s contents at that time, as well as the vault’s creation date and creator.

# Amazon Aurora

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 with MySQL-compatibility delivers up to five times the performance of MySQL without requiring any changes to most MySQL applications. Amazon RDS manages your Amazon Aurora database, 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. Amazon Aurora with PostgreSQL compatibility is now available in preview. FAQs on PostgreSQL compatibility for Amazon Aurora are available [here](https://aws.amazon.com/rds/aurora/faqs/#postgresql).

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 using the InnoDB storage engine. Certain MySQL features like the MyISAM storage engine are not available with Amazon Aurora.

Q: How do I try Amazon Aurora?

To try Amazon Aurora, sign in to the [AWS console](https://console.aws.amazon.com/), 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](http://aws.amazon.com/rds/aurora/pricing/) 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](https://aws.amazon.com/rds/aurora/pricing/) 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 Amazon Aurora’s [Data Export and Import guide](https://d1.awsstatic.com/product-marketing/Aurora/Aurora_Export_Import_Best_Practices_v1-3.d19863e4e1350cc38a457a6b0bf7de2c0fd67a90.pdf).

Q: Does Amazon Aurora participate in the AWS Free Tier?

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](http://aws.amazon.com/rds/aurora/pricing/) 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.

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.

## **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 Performance Benchmarking Guide](https://d1.awsstatic.com/product-marketing/Aurora/RDS_Aurora_Performance_Assessment_Benchmarking_v1-2.3d5390614fec6ddf9c47d6a10a5beff7f3062c95.pdf).

Q: How do I optimize my database workload for Amazon Aurora?

Amazon Aurora is designed to be compatible with MySQL 5.6, 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.

## **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, up to 32 vCPUs and 244 GiB Memory, in the AWS Management Console (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?

Aurora gives you the ability to create snapshots of your databases, which you can use later to restore a database. You can share this 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](https://aws.amazon.com/rds/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?

No. Sharing an encrypted Aurora snapshot is not supported at his time.

## **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 supports two kinds of replicas. Amazon Aurora Replicas share the same underlying volume as the primary instance. Updates made by the primary are visible to all Amazon Aurora Replicas. You can also create 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 |
| 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 |

Q. Can I have cross-region replicas with Amazon Aurora?

Yes, you can setup a cross-region Aurora Replica from the RDS console. The cross-region replication is based on single threaded MySQL binlog replication and the replication lag will be influenced by the change/apply rate and delays in network communication between the specific regions selected.

Q. Can I create Aurora Read Replicas on the cross-region replica cluster?  
Yes, you can add Aurora Replicas on the cluster that will share the same underlying storage as the cross-region replica. The 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 failover 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. The promotion process typically takes a few minutes depending on your workload. The cross-region replication will stop once you initiate the promotion process.

Q: Can I prioritize certain replicas as failover targets over others?

A: 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?

A: 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?

A: 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. Amazon Aurora Replicas share the same underlying storage as the primary instance. Any Amazon 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.

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, Amazon Aurora flips the canonical name record (CNAME) for your DB Instance to point at the healthy replica, which is in turn is promoted to become the new primary. Start-to-finish, failover typically completes within a minute.
* If you do not have an Amazon Aurora Replica (i.e. single instance), Aurora will first attempt to create a new DB Instance in the same Availability Zone as the original instance. If unable to do so, Aurora will attempt to create a new DB Instance in a different Availability Zone. From start to finish, failover typically completes in under 15 minutes.

Your application should retry database connections in the event of connection loss.

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 begin routing your read/write traffic to an Amazon Aurora Replica. On average, this failover will take less than a minute. In addition, the read traffic that your Amazon Aurora Replicas were serving will be briefly interrupted.

Q: How far behind the primary will my replicas to be?

Since Amazon Aurora Replicas share the same data volume as the primary, 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.

## **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 data in transit. 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Overview.Encryption.html).

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](https://d1.awsstatic.com/product-marketing/Aurora/RDS_Aurora_Connectivity_Guide_v5-2.afff3b74d3758e2bf177282180312ec752667477.pdf).

## **PostgreSQL Compatibility (in preview)**

Amazon Aurora with PostgreSQL compatibility is now available in preview, and delivers up to twice the performance of PostgreSQL running on the same hardware without requiring any modification to your existing PostgreSQL applications. This section of the Amazon Aurora FAQs applies specifically to the preview of Amazon Aurora with PostgreSQL compatibility. [Sign up for the preview.](https://pages.awscloud.com/amazon-aurora-with-postgresql-compatibility-preview-form.html)

Q: What does "PostgreSQL-compatible" mean?

It means that all of the code, applications, drivers and tools you use today with your PostgreSQL databases can be used with Amazon Aurora with no change. Amazon Aurora is designed to be wire-compatible with PostgreSQL 9.6.

Q: Once I sign up for the preview, how long it will take for me to get access and how will I know?

We admit people to the preview when they apply. When you are admitted, you will receive an email with details on how to access the preview and the private forum to provide feedback.

Q: Will I be charged for my usage during the preview?

No, there is no charge for your preview database instances or storage used by your preview databases during the preview period.

Q: What are the unique features of the PostgreSQL-compatible edition of Amazon Aurora that are not available with the community edition of PostgreSQL?

• Consistently High Throughput: Amazon Aurora uses a variety of software and hardware techniques to ensure the database engine is able to fully leverage available compute, memory and networking. I/O operations use distributed systems techniques such as quorums to improve performance consistency. Testing on standard benchmarks such as pgbench has shown up to a 3X increase over stock PostgreSQL on the same hardware.

• Fault-tolerant and Self-healing Storage: Your data is replicated six ways, across three Availability Zones. Amazon Aurora storage is fault-tolerant, transparently handling 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 replaced automatically.

• Storage Auto-scaling: Amazon Aurora will automatically grow the size of your database volume as your database storage needs grow. Your volume will grow in increments of 10 GB up to a maximum of 64 TB or a maximum volume size you define. You don't need to provision excess storage for your database to handle future growth; you only pay for the storage you actually use.

• Instance Monitoring and Repair: Amazon RDS continuously monitors the health of your Amazon Aurora database and underlying EC2 instance. If a database issue occurs, Amazon RDS will automatically restart the database and associated processes. Amazon Aurora does not require crash recovery replay of database redo logs, greatly reducing restart times. Amazon Aurora also isolates the database buffer cache from the database processes, allowing the cache to survive a database restart. If there are issues with the underlying instance, Amazon RDS uses [RDS Multi-AZ](https://aws.amazon.com/rds/maz/) technology to automate failover to one of up to 15 Amazon Aurora Replicas you have created in any of three Availability Zones. If no Amazon Aurora Replicas have been provisioned, in the case of a failure, Amazon RDS will attempt to create a new Amazon Aurora DB instance for you automatically.

Q. Amazon Aurora replicates my database six ways across three Availability Zones. Does that mean that my effective storage price will be more than what is shown on the pricing page?

No. Amazon Aurora’s replication is included in the price. You are charged based only on the storage your database uses.

Q: How can I migrate from PostgreSQL to Amazon Aurora and vice versa?

To migrate from PostgreSQL on premises, PostgreSQL on EC2, or RDS for PostgreSQL, you can use the standard pg\_dump utility to export data from PostgreSQL and the pg\_restore utility to import data to Amazon Aurora, and vice versa. You can also move data from PostgreSQL into Amazon Aurora using the [AWS Database Migration Service](https://aws.amazon.com/dms/), which can easily and transparently migrate data without application downtime from any Oracle, SQL Server, MySQL, or PostgreSQL database, either running on-premise or within AWS, directly into a MySQL-compatible or PostgreSQL-compatible instance of Amazon Aurora. Database schemas and database code (such as code written in Oracle PL/SQL) can also be easily migrated from Oracle and SQL Server to Amazon Aurora using the [AWS Schema Conversion Tool](https://aws.amazon.com/dms/#sct), a self-service database schema and code migration tool.

Q: Do I need to change client drivers to use Amazon Aurora with PostgreSQL compatibility?

No, Amazon Aurora will work with standard PostgreSQL database drivers.

Q: Do PostgreSQL extensions work with Amazon Aurora?

Amazon Aurora supports the same popular extensions that are available for Amazon RDS for PostgreSQL including PostGIS, dblink, and many data type, index, search, and other useful extensions. We will continue to make additional extensions available based on customer need.

Q: What does twice 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.

Q: Will you continue to support Amazon RDS for PostgreSQL (Community Edition)?

Yes, we will continue to support and enhance future releases of PostgreSQL on Amazon RDS.

Q. What will the pricing be for Amazon Aurora with PostgreSQL compatibility?

The pricing for PostgreSQL compatibility will be the same as the pricing for MySQL compatibility, which you can find on the [pricing page](https://aws.amazon.com/rds/aurora/pricing/).

Q: Can I use Amazon Aurora with PostgreSQL compatibility in Amazon Virtual Private Cloud (Amazon VPC)?

Yes, all Amazon Aurora database 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: What security features does Amazon Aurora with PostgreSQL compatibility provide?

Amazon Aurora supports SSL to protect data in transit and transparent database encryption to protect data at rest. You can choose to use AWS KMS to manage your encryption key or supply your own encryption keys using a hardware security module (HSM). Amazon Aurora integrates with AWS CloudTrail and AWS Identity and Access Management (IAM), enabling tracking of all API calls and integration via federation with your LDAP and Active Directory systems.

Q: Will all Amazon Aurora data be encrypted at rest?

Yes, at your option, all data stored on PostgreSQL-compatible Amazon Aurora database instances can be encrypted at rest, using keys you can manage on the AWS Key Management System (KMS).

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 enable backups for my DB Instance?

Automated backups are taken continuously and are always enabled on Amazon Aurora database instances. Backups do not impact database performance.

Q: Can I take DB Snapshots and keep them around as long as I want?

Yes. There is no performance impact when taking snapshots. Note that restoring data from database snapshots requires creating a new database instance.

Q: Will I have data loss if my Amazon Aurora instance fails?

Amazon Aurora automatically maintains 6 copies of your data across 3 Availability Zones (AZ) and will 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 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 as recent as 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. Automated backups are deleted when you delete your DB Instance.

Q: What kind of replicas does Aurora support?

Amazon Aurora Replicas share the same underlying volumes as the primary instances. Updates made by the primary are visible to all Amazon Aurora Replicas with a short replication lag, on the order of milliseconds in most cases.

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.

# Amazon RDS

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](https://aws.amazon.com/dms/) in [the cloud](https://aws.amazon.com/what-is-cloud-computing/). 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](https://aws.amazon.com/rds/faqs/#36), 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](https://aws.amazon.com/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](https://aws.amazon.com/running_databases/) 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](https://portal.aws.amazon.com/gp/aws/developer/registration/index.html) if you do not already have one. After you are signed up, please refer to the [Amazon RDS documentation](http://aws.amazon.com/documentation/rds/), which includes our [Getting Started Guide](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_GettingStarted.html).

Amazon RDS is part of the [AWS Free Tier](https://aws.amazon.com/free/) so that new AWS customers can get started with a managed database service in the cloud for free.

## **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](https://console.aws.amazon.com/), [Amazon RDS APIs](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/Welcome.html), and [AWS Command Line Interface](http://docs.aws.amazon.com/cli/latest/reference/rds/index.html). 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](https://console.aws.amazon.com/), [Amazon RDS APIs](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/Welcome.html), or [AWS Command Line Interface](http://docs.aws.amazon.com/cli/latest/reference/rds/index.html). 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](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/API_CreateDBInstance.html) or [create-db-instance command](http://docs.aws.amazon.com/cli/latest/reference/rds/create-db-instance.html).

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](https://console.aws.amazon.com/), [DescribeDBInstances API](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/API_DescribeDBInstances.html) or [describe-db-instances command](http://docs.aws.amazon.com/cli/latest/reference/rds/describe-db-instances.html). 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_GettingStarted.html).

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 or SQL Server under the "BYOL" model. If your application requires more DB instances, you can request additional DB instances via [this request form](https://aws.amazon.com/contact-us/request-to-increase-the-amazon-rds-db-instance-limit/).

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: 30 databases per instance
* 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/MySQL.Procedural.Importing.html) or the [Data Import Guide for Oracle](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Oracle.Procedural.Importing.html) or the [Data Import Guide for SQL Server](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/SQLServer.Procedural.Importing.html) or the [Data Import Guide for PostgreSQL](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/PostgreSQL.Procedural.Importing.html).

In addition, [AWS Database Migration Service](https://aws.amazon.com/dms/) 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 (such as scaling DB instance class) and software patching occur, in the event they 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 maintenance window you identify. Maintenance windows are 30 minutes in duration.

The only 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) or required software patching. Required 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](https://console.aws.amazon.com/), the [ModifyDBInstance API](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/API_ModifyDBInstance.html) or the [modify-db-instance command](http://docs.aws.amazon.com/cli/latest/reference/rds/modify-db-instance.html). Each of your DB instances can have different preferred maintenance windows, if you so choose.

Running your DB instance as a [Multi-AZ deployment](https://aws.amazon.com/rds/faqs/#36) can further reduce the impact of a maintenance event. Please refer to the [Amazon RDS User Guide](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_UpgradeDBInstance.Maintenance.html) for more information on maintenance operations.

Q: What should I do if my queries seem to be running slow?

For production databases we encourage you to enable [Enhanced Monitoring](https://aws.amazon.com/rds/faqs/#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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_Monitoring.html).

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](http://docs.amazonwebservices.com/AmazonRDS/latest/UserGuide/Appendix.MySQL.CommonDBATasks.html) 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](http://docs.amazonwebservices.com/AmazonRDS/latest/UserGuide/Appendix.Oracle.CommonDBATasks.html#Appendix.Oracle.CommonDBATasks.WorkingWithTracefiles).

If you are 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](http://docs.amazonwebservices.com/AmazonRDS/latest/UserGuide/Appendix.SQLServer.CommonDBATasks.html#Appendix.SQLServer.CommonDBATasks.WorkingWithTracefiles).

## **Database Engine Versions**

Q: Which relational database engine versions does Amazon RDS support?

* [Amazon RDS for MySQL](https://aws.amazon.com/rds/mysql/faqs/) currently supports MySQL Community Edition 5.5, 5.6 and 5.7.
* [Amazon RDS for MariaDB](https://aws.amazon.com/rds/mariadb/faqs/) currently supports MariaDB Server 10.0 and 10.1.
* [Amazon RDS for PostgreSQL](https://aws.amazon.com/rds/postgresql/faqs/) currently supports PostgreSQL 9.3, 9.4, 9.5, and 9.6.
* [Amazon RDS for Oracle](https://aws.amazon.com/rds/oracle/faqs/) currently supports Oracle Database 11gR2 and 12c.
* [Amazon RDS for SQL Server](https://aws.amazon.com/rds/sqlserver/faqs/) currently supports Microsoft SQL Server 2008 R2,  2012, 2014 and 2016.

See the [Amazon Aurora User Guide](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Aurora.DatabaseEngineUpdates.html) for details about Amazon Aurora database versions.

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 version releases supported in a given year 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 Instanceoperation in the AWS Management Console or the [CreateDBInstance](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/API_CreateDBInstance.html) 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](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/API_ModifyDBInstance.html) 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_UpgradeDBInstance.Maintenance.html#Concepts.DBMaintenance). 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 will announce scheduled upgrades on the Amazon RDS Forum and send customer e-mail notifications at least 30 days in advance. 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_UpgradeDBInstance.Upgrading.html).

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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_RestoreFromSnapshot.html).

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 5.6.21, 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 schedule automatic upgrades for instances with the Auto Minor Version Upgrade setting to occur at least 30 days after announcing the deprecation on the forum and sending e-mail notifications to customers. We will also disable the creation of new instances for this version. After a minimum three month grace period after announcement, all instances still running the deprecated minor version will be scheduled for automatic upgrade to a supported minor version during the specified maintenance window.

When a major version of database engine is deprecated in Amazon RDS, we will provide a minimum six month grace period after the announcement of a deprecation for you to initiate an upgrade to a supported major version. At the end of this grace period, an automatic upgrade 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.

## **Free Tier**

Q: What does the AWS Free Tier for Amazon RDS offer?

The [AWS Free Tier for Amazon RDS](https://aws.amazon.com/rds/free/) 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](https://aws.amazon.com/free/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](https://aws.amazon.com/rds/pricing/).

Q: Is the AWS Free Tier for Amazon RDS available in all AWS Regions?

The AWS Free Tier for Amazon RDS is available in all AWS Regions except GovCloud (US).

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](https://aws.amazon.com/rds/pricing/)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](https://aws.amazon.com/contact-us/request-to-increase-the-amazon-rds-db-instance-limit/).

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](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/API_DescribeReservedDBInstances.html) or [describe-reserved-db-instances command](http://docs.aws.amazon.com/cli/latest/reference/rds/describe-reserved-db-instances.html). 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_WorkingWithReservedDBInstances.html#USER_WorkingWithReservedDBInstances.SizeFlexible).

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 call the [DescribeReservedDBInstancesOfferings API](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/API_DescribeReservedDBInstancesOfferings.html) or [describe-reserved-db-instances-offerings command](http://docs.aws.amazon.com/cli/latest/reference/rds/describe-reserved-db-instances-offerings.html), simply look for the Multi-AZ options listed among the DB Instance configurations available for purchase. If you want to purchase a reservation for a DB instance with synchronous replication across multiple Availability Zones, specify one of these offerings in your PurchaseReservedDBInstancesOffering call.

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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Concepts.DBInstanceClass.html).

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](https://console.aws.amazon.com/) (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.

Monitor the compute and storage resource utilization of your DB instance, for no additional charge, via Amazon CloudWatch. You can access metrics such as CPU utilization, storage utilization, and network traffic by clicking the “Monitoring” tab for your DB Instance in the [AWS Management Console](https://console.aws.amazon.com/) or using the Amazon CloudWatch APIs. To learn more about monitoring your active DB Instances, read the [Amazon RDS User Guide](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_Monitoring.html).

Please note that for SQL Server, because of the extensibility limitations of striped storage attached to a Windows Server environment, Amazon RDS does not currently support increasing storage. While we plan to support this functionality in the future, we recommend you to provision storage based on anticipated future storage growth. In the interim, if you need to increase the storage of a SQL Server DB instance, you will need to export the data, create a new DB Instance with increased storage, and import the data into it. Please refer to the [data import guide for SQL Server](http://docs.amazonwebservices.com/AmazonRDS/latest/UserGuide/ImportData.SQLServer.html) 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](https://console.aws.amazon.com/), the [ModifyDBInstance API](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/API_ModifyDBInstance.html), or the [modify-db-instance command](http://docs.aws.amazon.com/cli/latest/reference/rds/modify-db-instance.html).

However, for SQL Server, because of the extensibility limitations of striped storage attached to a Windows Server environment, Amazon RDS does not currently support increasing storage.

For more information, see [Storage for Amazon RDS](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_Storage.html).

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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_Storage.html).

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](http://docs.amazonwebservices.com/AmazonRDS/latest/UserGuide/USER_PIOPS.html).

## **Automated 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](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/API_DescribeDBInstances.html) 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](https://console.aws.amazon.com/) 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](https://console.aws.amazon.com/), [CreateDBSnapshot API](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/API_CreateDBSnapshot.html), or [create-db-snapshot command](http://docs.aws.amazon.com/cli/latest/reference/rds/create-db-snapshot.html) 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](http://docs.aws.amazon.com/cli/latest/reference/rds/copy-db-snapshot.html)) 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 extend your backup retention period beyond one day, you can do so using the CreateDBInstance API (when creating a new DB Instance) or ModifyDBInstance API (for an existing DB Instance). You can use these APIs to change the RetentionPeriod parameter from 1 to the desired number of days. For more information on automated backups, please refer to the [Amazon RDS User Guide](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_CommonTasks.BackupRestore.html).

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](https://console.aws.amazon.com/), the [ModifyDBInstance API](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/API_ModifyDBInstance.html), or the [modify-db-instance command](http://docs.aws.amazon.com/cli/latest/reference/rds/modify-db-instance.html) 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](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/API_DescribeDBSnapshots.html) or [describe-db-snapshots command](http://docs.aws.amazon.com/cli/latest/reference/rds/describe-db-snapshots.html) and delete snapshots with the [DeleteDBSnapshot API](http://docs.aws.amazon.com/AmazonRDS/latest/APIReference/API_DeleteDBSnapshot.html) or [delete-db-snapshot command](http://docs.aws.amazon.com/cli/latest/reference/rds/delete-db-snapshot.html).

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](https://aws.amazon.com/rds/pricing/) 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](http://docs.amazonwebservices.com/AmazonVPC/latest/UserGuide/).

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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-vpc.html#differences).

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 a 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_VPC.WorkingWithRDSInstanceinaVPC.html#USER_VPC.InstanceInVPC) in the Amazon RDS User Guide.

Q: How do I control network access to my DB Instance(s)?

Visit the [Security Groups](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Overview.RDSSecurityGroups.html) 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](http://docs.amazonwebservices.com/AmazonVPC/latest/UserGuide) 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_VPC.WorkingWithRDSInstanceinaVPC.html#USER_VPC.Non-VPC2VPC) 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 a 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_VPC.WorkingWithRDSInstanceinaVPC.html#USER_VPC.VPC2VPC) 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](http://docs.amazonwebservices.com/AmazonRDS/latest/UserGuide/) 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](http://docs.amazonwebservices.com/AmazonRDS/latest/UserGuide/RDSFAQ.SQLServer.html) 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Overview.RDSSecurityGroups.html). 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?

Yes, this option is currently supported for the MySQL, MariaDB, SQL Server, PostgreSQL, and Oracle engines.

Amazon RDS generates an SSL 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 encryption is a compute-intensive operation and will increase the latency of your database connection. SSL 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_MySQL.html#MySQL.Concepts.SSLSupport), [MariaDB User Guide](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_MariaDB.html#MariaDB.Concepts.SSLSupport), [SQL Server User Guide](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_SQLServer.html#SQLServer.Concepts.General.SSL), [PostgreSQL User Guide](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_PostgreSQL.html#PostgreSQL.Concepts.General.SSL) or [Oracle User Guide](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_Oracle.html#Oracle.Concepts.SSL). To learn more about how SSL works with these engines, you can refer directly to the [MySQL documentation](http://dev.mysql.com/doc/refman/5.6/en/ssl-connections.html), the [MariaDB documentation](https://mariadb.com/kb/en/mariadb/secure-connections-overview/), the [MSDN SQL Server documentation](http://msdn.microsoft.com/en-us/library/ms189067.aspx), the [PostgreSQL documentation](http://www.postgresql.org/docs/9.3/static/ssl-tcp.html), or the [Oracle Documentation](https://docs.oracle.com/database/121/DBSEG/asossl.htm).

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)](https://aws.amazon.com/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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Overview.Encryption.html).

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 technologies. Transparent Data Encryption in Oracle is integrated with [AWS CloudHSM](https://aws.amazon.com/cloudhsm/), which allows you to securely generate, store, and manage your cryptographic keys in single-tenant Hardware Security Module (HSM) appliances within the AWS cloud. For more information, see the Amazon RDS User's Guide sections on [Oracle](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Appendix.Oracle.Options.html#Appendix.Oracle.Options.AdvSecurity) and [SQL Server](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Appendix.SQLServer.Options.html#Appendix.SQLServer.Options.TDE).

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](https://aws.amazon.com/iam) users and groups can take on RDS resources. You do this by referencing the RDS resources in the [AWS IAM policies](http://docs.aws.amazon.com/IAM/latest/UserGuide/PermissionsAndPolicies.html) 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, "Test" DB Instances etc), and write AWS IAM policies that list the permissions (i.e. actions) that can taken on resources with the same tags. For more information, refer to [Managing Access to Your Amazon RDS Resources and Databases](https://aws.amazon.com/iam/) and [Tagging Amazon RDS Resources](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_Tagging.html)

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](https://aws.amazon.com/cloudtrail/), and turn it on via [CloudTrail's AWS Management Console home page](https://console.aws.amazon.com/cloudtrail/home).

## **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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_WorkingWithParamGroups.html).

Q: How can I monitor the configuration of my Amazon RDS resources?

You can use [AWS Config](https://aws.amazon.com/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)](https://aws.amazon.com/sns/). You can also create AWS Config Rules to evaluate whether these RDS resources have the desired configurations.

## **Multi-AZ Deployments and Read Replicas**

Q: What types of replication does Amazon RDS support and when should I use each?

Amazon RDS provides two distinct replication options to serve different purposes.

If you are looking to use replication to increase database availability while protecting your latest database updates against unplanned outages, consider running your DB instance as a [Multi-AZ deployment](https://aws.amazon.com/rds/details/multi-az/). When you create or modify your DB instance to run as a Multi-AZ deployment, Amazon RDS will automatically provision and manage a “standby” replica in a different Availability Zone (independent infrastructure in a physically separate location). In the event of planned database maintenance, DB instance failure, or an Availability Zone failure, Amazon RDS will automatically failover to the standby so that database operations can resume quickly without administrative intervention. Multi-AZ deployments utilize synchronous replication, making database writes concurrently on both the primary and standby so that the standby will be up-to-date in the event a failover occurs. While our technological implementation for Multi-AZ DB Instances maximizes data durability in failure scenarios, it precludes the standby from being accessed directly or used for read operations. The fault tolerance offered by Multi-AZ deployments make them a natural fit for production environments.

To help you to scale beyond the capacity constraints of a single DB Instance for read-heavy database workloads, Amazon RDS offers [Read Replicas](https://aws.amazon.com/rds/details/read-replicas/). You can create a Read Replica of a given source DB Instance using the AWS Management Console, the RDS API, or the AWS Command Line Interface. Once the Read Replica is created, database updates on the source DB instance will be propagated to the Read Replica. You can create multiple Read Replicas for a given source DB Instance and distribute your application’s read traffic amongst them.

Read Replicas are supported by Amazon Aurora, Amazon RDS for MySQL, MariaDB and PostgreSQL. Unlike Multi-AZ deployments, Read Replicas for these engines use each's built-in replication technology and 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 (“asynchronous” replication), and replication lag can vary significantly. This means recent database updates made to a standard (non Multi-AZ) source DB instance may not be present on associated Read Replicas in the event of an unplanned outage on the source DB instance. As such, Read Replicas do not offer the same data durability benefits as Multi-AZ deployments. While Read Replicas can provide some read availability benefits, they and are not designed to improve write availability.

You can use Multi-AZ deployments and Read Replicas in conjunction to enjoy the complementary benefits of each. You can simply specify that a given Multi-AZ deployment is the source DB instance for your Read Replica(s). That way you gain both the data durability and availability benefits of Multi-AZ deployments and the read scaling benefits of Read Replicas.

### **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 failover 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Concepts.MultiAZ.html).

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](https://aws.amazon.com/rds/faqs/#87).

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 deployments?

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, the standby replica 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](https://aws.amazon.com/rds/faqs/#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](https://console.aws.amazon.com/). 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 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.

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 an 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 failover 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_Events.html) 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 failover without user intervention under a variety of [failure conditions](https://aws.amazon.com/rds/faqs/#43). 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.](https://aws.amazon.com/rds/details/multi-az/)

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](https://console.aws.amazon.com/) 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](https://aws.amazon.com/rds/faqs/#36).

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.

Q: Do I need to enable automatic backups on my DB Instance before I can create read replicas?

Yes. Enable automatic backups on your 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 (MySQL)*: All DB Instances.

*Amazon RDS for MySQL*: DB Instances with MySQL version 5.5 or newer support creation of Read Replicas. Automatic backups must be and remain enabled ou the source DB Instance for Read Replica operations. Automatic backups 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*: DB Instances with MariaDB 10.0 or newer support creation of Read Replicas. Automatic backups must be and remain enabled on the source DB Instance for Read Replica operations.

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 Aurora (MySQL) allows you to create up to 15 Read Replicas for a given source DB Instance.

Amazon RDS for MySQL, MariaDB and PostgreSQL currently 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, RDS supports cross-region Read Replicas.

Q: Do Amazon RDS Read Replicas support synchronous replication?

No. Read Replicas in Amazon RDS for MySQL, MariaDB and PostgreSQL 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 a 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 make my Amazon RDS Read Replicas themselves Multi-AZ?

Amazon RDS for MySQL, MariaDB and PostgreSQL do not presently support this.

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 a Multi-AZ failover, any associated and available Read Replicas should 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:* 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](http://docs.amazonwebservices.com/AmazonRDS/latest/UserGuide/Overview.ReadReplica.html) 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”. Click [here](https://aws.amazon.com/rds/faqs/#98) for guidance on what to do if your Read Replica(s) fall significantly behind its source.

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 "DB 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](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_ReadRepl.html#USER_ReadRepl.Troubleshooting) or [PostgreSQL](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_ReadRepl.html#USER_ReadRepl.TroubleshootingPostgreSQL).

If a replication error is fixed, the Replication State changes to Replicating.

Q: My Read Replica has fallen significantly behind its source DB Instance. What should I do?

As discussed in the previous questions, “inconsistency” or lag between a Read Replica and its source DB Instance is common with asynchronous replication. If an existing Read Replica has fallen too far behind to meet your requirements, you can delete it and create a new one with the same endpoint by using the same DB Instance Identifier and Source DB Instance Identifier as the deleted Read Replica. Keep in mind that the re-creation process will be counter-productive at small lag levels (e.g. under five minutes of lag), and should be used with prudence (i.e. only when the Read Replica is significantly behind its source DB Instance). Also keep in mind that replica lag may naturally grow and shrink over time, depending on your source DB Instance’s steady-state usage pattern.

Scaling the DB Instance class of a Read Replica may reduce replication lag in some cases, particularly if your source DB Instance class is larger than your Read Replica DB Instance class. However, Read Replicas are not guaranteed to work in all cases. There may be scenarios and usage patterns where a Read Replica can never catch up with its source after initial creation, or otherwise remains too far behind its source to meet your use case requirements.

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: Can DB Snapshots or automated backups be taken of Read Replicas?

No. If you are looking to increase database write availability by taking backups from your Read Replica instead of its source DB Instance, you can accomplish the same objective by running your DB Instance as a Multi-AZ deployment. Backups will then be initiated from the Multi-AZ standby to minimize availability impact.

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 (MySQL) Read Replica will stay active and continue accepting read traffic even after its corresponding source DB Instance has been deleted. One of the Read 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: Can I directly access the event logs for my Database Instance?

With Amazon RDS for MySQL or Amazon RDS for MariaDB, you can use the mysqlbinlog utility to download or stream binary logs from your DB Instance. Amazon RDS for PostgreSQL does not currently provide access to the WAL files for your DB Instance.

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. Click here for more information on DB Instance billing visit this [FAQ](https://aws.amazon.com/rds/faqs/#15). 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 Amazon RDS [detail page](https://aws.amazon.com/rds/pricing/) 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.

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 RDS DB Instance hour rates until you issue a command to delete it.

Q: How does support for Read Replicas vary among the Amazon RDS engines that support this feature?

Read Replicas in both Amazon RDS for PostgreSQL, MySQL, and MariaDB allow you to elastically scale out beyond the capacity constraints of a single DB instance for read-heavy database workloads. There are similarities and differences in the implementations as they leverage native engine features. See the following table for details.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | PostgreSQL | MySQL | MariaDB |
| Maximum Read Replicas allowed per source DB Instance | 5 | 5 | 5 |
| Replication method | Asynchronous Physical | Asynchronous Logical | Asynchronous Logical |
| Must automatic backups be enabled for Read Replica support? | Yes | Yes | Yes |
| Engine versions for which Read Replicas are available | 9.3.5 or later | 5.5 or later | 10.0 or later |
| Promotion of Read Replica to a new standalone DB Instance | Supported | Supported | Supported |
| Creation of Indexes on Read Replica | Currently not supported | Supported | Supported |
| Creation of Backups of Read Replicas | Currently not supported | Supported | Supported |
| Chaining of Read Replicas (i.e., Read Replicas of Read Replicas) | Currently not supported | Supported | Supported |
| Cross-Region Read Replicas | Supported | Supported | Supported |

## **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.

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 [here](http://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_Monitoring.html).

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 for up to 1 hour ago, at a granularity of up to 1 second based on your setting.

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](https://aws.amazon.com/cloudwatch/pricing/) 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](http://docs.aws.amazon.com/AmazonCloudWatch/latest/DeveloperGuide/AlarmThatSendsEmail.html).

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](http://docs.aws.amazon.com/AmazonCloudWatch/latest/DeveloperGuide/Subscriptions.html) 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](http://docs.aws.amazon.com/AmazonCloudWatch/latest/DeveloperGuide/CWL_ES_Stream.html) 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](http://docs.aws.amazon.com/AmazonCloudWatch/latest/DeveloperGuide/WhatIsCloudWatch.html).

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](https://aws.amazon.com/cloudwatch/pricing/). 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 |

**Aws dynamoDB:**

## **What is DynamoDB**

Q: What is Amazon DynamoDB?

Amazon DynamoDB is a fully managed [NoSQL database](https://aws.amazon.com/nosql/) service that provides fast and predictable performance with seamless scalability. Amazon DynamoDB enables customers to offload the administrative burdens of operating and scaling distributed databases to AWS, so they don’t have to worry about hardware provisioning, setup and configuration, throughput capacity planning, replication, software patching, or cluster scaling.

Q: What does Amazon DynamoDB manage on my behalf?

Amazon DynamoDB takes away one of the main stumbling blocks of scaling databases, the management of the database software and the provisioning of hardware needed to run it. Customers can deploy a non-relational database in a matter of minutes. DynamoDB automatically scales throughput capacity to meet workload demands and partitions and re-partitions your data as your table size grows. In addition, Amazon DynamoDB synchronously replicates data across three facilities in an AWS Region, giving you high availability and data durability.

Q: What does read consistency mean? Why should I care?

Amazon DynamoDB stores three geographically distributed replicas of each table to enable high availability and data durability. Read consistency represents the manner and timing in which the successful write or update of a data item is reflected in a subsequent read operation of that same item. Amazon DynamoDB exposes logic that enables you to specify the consistency characteristics you desire for each read request within your application.

Q: What is the consistency model of Amazon DynamoDB?

When reading data from Amazon DynamoDB, users can specify whether they want the read to be eventually consistent or strongly consistent:

Eventually Consistent Reads (Default) – the eventual consistency option maximizes your read throughput. However, an eventually consistent read might not reflect the results of a recently completed write. Consistency across all copies of data is usually reached within a second. Repeating a read after a short time should return the updated data.

Strongly Consistent Reads — in addition to eventual consistency, Amazon DynamoDB also gives you the flexibility and control to request a strongly consistent read if your application, or an element of your application, requires it. A strongly consistent read returns a result that reflects all writes that received a successful response prior to the read.

Q: Does DynamoDB support in-place atomic updates?

Amazon DynamoDB supports fast in-place updates. You can increment or decrement a numeric attribute in a row using a single API call. Similarly, you can atomically add or remove to sets, lists, or maps. [View our documentation for more information on atomic updates.](http://docs.amazonwebservices.com/amazondynamodb/latest/developerguide/WorkingWithDDItems.html)

Q: Why is Amazon DynamoDB built on Solid State Drives?

Amazon DynamoDB runs exclusively on Solid State Drives (SSDs). SSDs help us achieve our design goals of predictable low-latency response times for storing and accessing data at any scale. The high I/O performance of SSDs also enables us to serve high-scale request workloads cost efficiently, and to pass this efficiency along in low request pricing.

Q: DynamoDB’s storage cost seems high. Is this a cost-effective service for my use case?

As with any product, we encourage potential customers of Amazon DynamoDB to consider the total cost of a solution, not just a single pricing dimension. The total cost of servicing a database workload is a function of the request traffic requirements and the amount of data stored. Most database workloads are characterized by a requirement for high I/O (high reads/sec and writes/sec) per GB stored. Amazon DynamoDB is built on SSD drives, which raises the cost per GB stored, relative to spinning media, but it also allows us to offer very low request costs. Based on what we see in typical database workloads, we believe that the total bill for using the SSD-based DynamoDB service will usually be lower than the cost of using a typical spinning media-based relational or non-relational database. If you have a use case that involves storing a large amount of data that you rarely access, then DynamoDB may not be right for you. We recommend that you use S3 for such use cases.

It should also be noted that the storage cost reflects the cost of storing multiple copies of each data item across multiple facilities within an AWS Region.

Q: Is DynamoDB only for high-scale applications?

No. DynamoDB offers seamless scaling so you can scale automatically as your application requirements increase. If you need fast, predictable performance at any scale then DynamoDB may be the right choice for you.

## **Getting Started**

Q: How do I get started with Amazon DynamoDB?

Click “Sign Up” to get started with Amazon DynamoDB today. From there, you can begin interacting with Amazon DynamoDB using either the [AWS Management Console](https://aws.amazon.com/console/) or Amazon DynamoDB APIs. If you are using the AWS Management Console, you can create a table with Amazon DynamoDB and begin exploring with just a few clicks.

Q: What kind of query functionality does DynamoDB support?

Amazon DynamoDB supports GET/PUT operations using a user-defined primary key. The primary key is the only required attribute for items in a table and it uniquely identifies each item. You specify the primary key when you create a table. In addition to that DynamoDB provides flexible querying by letting you query on non-primary key attributes using Global Secondary Indexes and Local Secondary Indexes.

A primary key can either be a single-attribute partition key or a composite partition-sort key. A single attribute partition primary key could be, for example, “UserID”. This would allow you to quickly read and write data for an item associated with a given user ID.

A composite partition-sort key is indexed as a partition key element and a sort key element. This multi-part key maintains a hierarchy between the first and second element values. For example, a composite partition-sort key could be a combination of “UserID” (partition) and “Timestamp” (sort). Holding the partition key element constant, you can search across the sort key element to retrieve items. This would allow you to use the Query API to, for example, retrieve all items for a single UserID across a range of timestamps.

For more information on Global Secondary Indexing and its query capabilities, see the [Secondary Indexes](https://aws.amazon.com/dynamodb/faqs/#gsi_anchor) section in FAQ.

Q: How do I update and query data items with Amazon DynamoDB?

After you have created a table using the AWS Management Console or CreateTable API, you can use the PutItem or BatchWriteItem APIs to insert items. Then you can use the GetItem, BatchGetItem, or, if composite primary keys are enabled and in use in your table, the Query API to retrieve the item(s) you added to the table.

Q: Does Amazon DynamoDB support conditional operations?

Yes, you can specify a condition that must be satisfied for a put, update, or delete operation to be completed on an item. To perform a conditional operation, you can define a ConditionExpression that is constructed from the following:

* Boolean functions: ATTRIBUTE\_EXIST, CONTAINS, and BEGINS\_WITH
* Comparison operators: =, <>, <, >, <=, >=, BETWEEN, and IN
* Logical operators: NOT, AND, and OR.

You can construct a free-form conditional expression that combines multiple conditional clauses, including nested clauses. Conditional operations allow users to implement optimistic concurrency control systems on DynamoDB. For more information on conditional operations, [please see our documentation](http://docs.amazonwebservices.com/amazondynamodb/latest/developerguide/WorkingWithDDItems.html).

Q: Are expressions supported for key conditions?

Yes, you can specify an expression as part of the Query API call to filter results based on values of primary keys on a table using the KeyConditionExpression parameter.

Q: Are expressions supported for partition and partition-sort keys?

Yes, you can use expressions for both partition and partition-sort keys. Refer to the [documentation page](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/QueryAndScan.html) for more information on which expressions work on partition and partition-sort keys.

Q: Does Amazon DynamoDB support increment or decrement operations?

Yes, Amazon DynamoDB allows atomic increment and decrement operations on scalar values.

Q: When should I use Amazon DynamoDB vs a relational database engine on Amazon RDS or Amazon EC2?

Today’s web-based applications generate and consume massive amounts of data. For example, an online game might start out with only a few thousand users and a light database workload consisting of 10 writes per second and 50 reads per second. However, if the game becomes successful, it may rapidly grow to millions of users and generate tens (or even hundreds) of thousands of writes and reads per second. It may also create terabytes or more of data per day. Developing your applications against Amazon DynamoDB enables you to start small and simply dial-up your request capacity for a table as your requirements scale, without incurring downtime. You pay highly cost-efficient rates for the request capacity you provision, and let Amazon DynamoDB do the work over partitioning your data and traffic over sufficient server capacity to meet your needs. Amazon DynamoDB does the database management and administration, and you simply store and request your data. Automatic replication and failover provides built-in fault tolerance, high availability, and data durability. Amazon DynamoDB gives you the peace of mind that your database is fully managed and can grow with your application requirements.

While Amazon DynamoDB tackles the core problems of database scalability, management, performance, and reliability, it does not have all the functionality of a relational database. It does not support complex relational queries (e.g. joins) or complex transactions. If your workload requires this functionality, or you are looking for compatibility with an existing relational engine, you may wish to run a relational engine on Amazon RDS or Amazon EC2. While relational database engines provide robust features and functionality, scaling a workload beyond a single relational database instance is highly complex and requires significant time and expertise. As such, if you anticipate scaling requirements for your new application and do not need relational features, Amazon DynamoDB may be the best choice for you.

Q: How does Amazon DynamoDB differ from Amazon SimpleDB?

Which should I use? Both services are non-relational databases that remove the work of database administration. Amazon DynamoDB focuses on providing seamless scalability and fast, predictable performance. It runs on solid state disks (SSDs) for low-latency response times, and there are no limits on the request capacity or storage size for a given table. This is because Amazon DynamoDB automatically partitions your data and workload over a sufficient number of servers to meet the scale requirements you provide. In contrast, a table in Amazon SimpleDB has a strict storage limitation of 10 GB and is limited in the request capacity it can achieve (typically under 25 writes/second); it is up to you to manage the partitioning and re-partitioning of your data over additional SimpleDB tables if you need additional scale. While SimpleDB has scaling limitations, it may be a good fit for smaller workloads that require query flexibility. Amazon SimpleDB automatically indexes all item attributes and thus supports query flexibility at the cost of performance and scale.

[Amazon CTO Werner Vogels' DynamoDB blog post](http://allthingsdistributed.com/2012/01/amazon-dynamodb.html) provides additional context on the evolution of non-relational database technology at Amazon.

Q: When should I use Amazon DynamoDB vs Amazon S3?

Amazon DynamoDB stores structured data, indexed by primary key, and allows low latency read and write access to items ranging from 1 byte up to 400KB. Amazon S3 stores unstructured blobs and suited for storing large objects up to 5 TB. In order to optimize your costs across AWS services, large objects or infrequently accessed data sets should be stored in Amazon S3, while smaller data elements or file pointers (possibly to Amazon S3 objects) are best saved in Amazon DynamoDB.

Q: Can DynamoDB be used by applications running on any operating system?

Yes. DynamoDB is a fully managed cloud service that you access via API. DynamoDB can be used by applications running on any operating system (e.g. Linux, Windows, iOS, Android, Solaris, AIX, HP-UX, etc.). We recommend using the AWS SDKs to get started with DynamoDB. You can find a list of the AWS SDKs on our [Developer Resources](http://aws.amazon.com/dynamodb/developer-resources/) page. If you have trouble installing or using one of our SDKs, please let us know by posting to the relevant [AWS Forum](https://forums.aws.amazon.com/index.jspa).

## **Data Models and APIs**

Q: What is the Data Model?

The data model for Amazon DynamoDB is as follows:

Table: A table is a collection of data items – just like a table in a relational database is a collection of rows. Each table can have an infinite number of data items. Amazon DynamoDB is schema-less, in that the data items in a table need not have the same attributes or even the same number of attributes. Each table must have a primary key. The primary key can be a single attribute key or a “composite” attribute key that combines two attributes. The attribute(s) you designate as a primary key must exist for every item as primary keys uniquely identify each item within the table.

Item: An Item is composed of a primary or composite key and a flexible number of attributes. There is no explicit limitation on the number of attributes associated with an individual item, but the aggregate size of an item, including all the attribute names and attribute values, cannot exceed 400KB.

Attribute: Each attribute associated with a data item is composed of an attribute name (e.g. “Color”) and a value or set of values (e.g. “Red” or “Red, Yellow, Green”). Individual attributes have no explicit size limit, but the total value of an item (including all attribute names and values) cannot exceed 400KB.

Q: Is there a limit on the size of an item?

The total size of an item, including attribute names and attribute values, cannot exceed 400KB.

Q: Is there a limit on the number of attributes an item can have?

There is no limit to the number of attributes that an item can have. However, the total size of an item, including attribute names and attribute values, cannot exceed 400KB.

Q: What are the APIs?

* CreateTable – Creates a table and specifies the primary index used for data access.
* UpdateTable – Updates the provisioned throughput values for the given table.
* DeleteTable – Deletes a table.
* DescribeTable – Returns table size, status, and index information.
* ListTables – Returns a list of all tables associated with the current account and endpoint.
* PutItem – Creates a new item, or replaces an old item with a new item (including all the attributes). If an item already exists in the specified table with the same primary key, the new item completely replaces the existing item. You can also use conditional operators to replace an item only if its attribute values match certain conditions, or to insert a new item only if that item doesn’t already exist.
* BatchWriteItem – Inserts, replaces, and deletes multiple items across multiple tables in a single request, but not as a single transaction. Supports batches of up to 25 items to Put or Delete, with a maximum total request size of 16 MB.
* UpdateItem – Edits an existing item's attributes. You can also use conditional operators to perform an update only if the item’s attribute values match certain conditions.
* DeleteItem – Deletes a single item in a table by primary key. You can also use conditional operators to perform a delete an item only if the item’s attribute values match certain conditions.
* GetItem – The GetItem operation returns a set of Attributes for an item that matches the primary key. The GetItem operation provides an eventually consistent read by default. If eventually consistent reads are not acceptable for your application, use ConsistentRead.
* BatchGetItem – The BatchGetItem operation returns the attributes for multiple items from multiple tables using their primary keys. A single response has a size limit of 16 MB and returns a maximum of 100 items. Supports both strong and eventual consistency.
* Query –  Gets one or more items using the table primary key, or from a secondary index using the index key. You can narrow the scope of the query on a table by using comparison operators or expressions. You can also filter the query results using filters on non-key attributes. Supports both strong and eventual consistency. A single response has a size limit of 1 MB.
* Scan – Gets all items and attributes by performing a full scan across the table or a secondary index. You can limit the return set by specifying filters against one or more attributes.

Q: What is the consistency model of the Scan operation?

The Scan operation supports eventually consistent and consistent reads. By default, the Scan operation is eventually consistent. However, you can modify the consistency model using the optional ConsistentRead parameter in the Scan API call. Setting the ConsistentRead parameter to true will enable you make consistent reads from the Scan operation. For more information, read the [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/API_Scan.html) for the Scan operation.

Q: How does the Scan operation work?

You can think of the Scan operation as an iterator. Once the aggregate size of items scanned for a given Scan API request exceeds a 1 MB limit, the given request will terminate and fetched results will be returned along with a LastEvaluatedKey (to continue the scan in a subsequent operation).

Q: Are there any limitations for a Scan operation?

A Scan operation on a table or secondary index has a limit of 1MB of data per operation. After the 1MB limit, it stops the operation and returns the matching values up to that point, and a LastEvaluatedKey to apply in a subsequent operation, so that you can pick up where you left off.

Q: How many read capacity units does a Scan operation consume?

The read units required is the number of bytes fetched by the scan operation, rounded to the nearest 4KB, divided by 4KB. Scanning a table with consistent reads consumes twice the read capacity as a scan with eventually consistent reads.

Q: What data types does DynamoDB support?

DynamoDB supports four scalar data types: Number, String, Binary, and Boolean. Additionally, DynamoDB supports collection data types: Number Set, String Set, Binary Set, heterogeneous List and heterogeneous Map. DynamoDB also supports NULL values.

Q: What types of data structures does DynamoDB support?

DynamoDB supports key-value and document data structures.

Q: What is a key-value store?

A key-value store is a database service that provides support for storing, querying and updating collections of objects that are identified using a key and values that contain the actual content being stored.

Q: What is a document store?

A document store provides support for storing, querying and updating items in a document format such as JSON, XML, and HTML.

Q: Does DynamoDB have a JSON data type?

No, but you can use the document SDK to pass JSON data directly to DynamoDB. DynamoDB’s data types are a superset of the data types supported by JSON. The document SDK will automatically map JSON documents onto native DynamoDB data types.

Q: Can I use the AWS Management Console to view and edit JSON documents?

Yes. The AWS Management Console provides a simple UI for exploring and editing the data stored in your DynamoDB tables, including JSON documents. To view or edit data in your table, please log in to the AWS Management Console, choose DynamoDB, select the table you want to view, then click on the “Explore Table” button.

Q: Is querying JSON data in DynamoDB any different?

No. You can create a Global Secondary Index or Local Secondary Index on any top-level JSON element. For example, suppose you stored a JSON document that contained the following information about a person: First Name, Last Name, Zip Code, and a list of all of their friends. First Name, Last Name and Zip code would be top-level JSON elements. You could create an index to let you query based on First Name, Last Name, or Zip Code. The list of friends is not a top-level element, therefore you cannot index the list of friends. For more information on Global Secondary Indexing and its query capabilities, see the [Secondary Indexes section](https://aws.amazon.com/dynamodb/faqs/#gsi_anchor) in this FAQ.

Q: If I have nested JSON data in DynamoDB, can I retrieve only a specific element of that data?

Yes. When using the GetItem, BatchGetItem, Query, or Scan APIs, you can define a ProjectionExpression to determine which attributes should be retrieved from the table. Those attributes can include scalars, sets, or elements of a JSON document.

Q. If I have nested JSON data in DynamoDB, can I update only a specific element of that data?

Yes. When updating a DynamoDB item, you can specify the sub-element of the JSON document that you want to update.

Q:What is the Document SDK?

The Document SDK is a datatypes wrapper for JavaScript that allows easy interoperability between JS and DynamoDB datatypes. With this SDK, wrapping for requests will be handled for you; similarly for responses, datatypes will be unwrapped. For more information and downloading the SDK see our GitHub respository [here](https://github.com/awslabs/dynamodb-document-js-sdk).

## **Scalability, Availability & Durability**

Q: Is there a limit to how much data I can store in Amazon DynamoDB?

No. There is no limit to the amount of data you can store in an Amazon DynamoDB table. As the size of your data set grows, Amazon DynamoDB will automatically spread your data over sufficient machine resources to meet your storage requirements.

Q: Is there a limit to how much throughput I can get out of a single table?

No, you can increase the maximum capacity limit setting for Auto Scaling or increase the throughput you have manually provisioned for your table using the API or the AWS Management Console. DynamoDB is able to operate at massive scale and there is no theoretical limit on the maximum throughput you can achieve. DynamoDB automatically divides your table across multiple partitions, where each partition is an independent parallel computation unit. DynamoDB can achieve increasingly high throughput rates by adding more partitions.

If you wish to exceed throughput rates of 10,000 writes/second or 10,000 reads/second, you must first [contact Amazon through this online form](http://portal.aws.amazon.com/gp/aws/html-forms-controller/DynamoDB_Limit_Increase_Form).

Q: Does Amazon DynamoDB remain available when Auto Scaling triggers scaling or when I ask it to scale up or down by changing the provisioned throughput?

Yes. Amazon DynamoDB is designed to scale its provisioned throughput up or down while still remaining available, whether managed by Auto Scaling or manually.

Q: Do I need to manage client-side partitioning on top of Amazon DynamoDB?

No. Amazon DynamoDB removes the need to partition across database tables for throughput scalability.

Q: How highly available is Amazon DynamoDB?

The service runs across Amazon’s proven, high-availability data centers. The service replicates data across three facilities in an AWS Region to provide fault tolerance in the event of a server failure or Availability Zone outage.

Q: How does Amazon DynamoDB achieve high uptime and durability?

To achieve high uptime and durability, Amazon DynamoDB synchronously replicates data across three facilities within an AWS Region.

## **Auto Scaling**

Q. What is DynamoDB Auto Scaling?  
  
DynamoDB Auto Scaling is a fully managed feature that automatically scales up or down provisioned read and write capacity of a DynamoDB table or a global secondary index, as application requests increase or decrease.  
  
Q. Why do I need to use Auto Scaling?

Auto Scaling eliminates the guesswork involved in provisioning adequate capacity when creating new tables and reduces the operational burden of continuously monitoring consumed throughput and adjusting provisioned capacity manually. Auto Scaling helps ensure application availability and reduces costs from unused provisioned capacity.  
  
Q. What application request patterns and workload are suited for Auto Scaling?  
  
Auto Scaling is ideally suited for request patterns that are uniform, predictable, with sustained high and low throughput usage that lasts for several minutes to hours.  
  
Q. How can I enable Auto Scaling for a DynamoDB table or global secondary index?  
  
From the DynamoDB console, when you create a new table, leave the 'Use default settings' option checked, to enable Auto Scaling and apply the same settings for global secondary indexes for the table. If you uncheck 'Use default settings', you can either set provisioned capacity manually or enable Auto Scaling with custom values for target utilization and minimum and maximum capacity. For existing tables, you can enable Auto Scaling or change existing Auto Scaling settings by navigating to the 'Capacity' tab and for indexes, you can enable Auto Scaling from under the 'Indexes' tab. Auto Scaling can also be programmatically managed using CLI or AWS SDK. Please refer to the [DynamoDB developer guide](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/AutoScaling.html) to learn more.  
  
Q. What are settings I can configure for Auto Scaling?  
  
There are three configurable settings for Auto Scaling: Target Utilization, the percentage of actual consumed throughput to total provisioned throughput, at a point in time, the Minimum capacity to which Auto Scaling can scale down to, and Maximum capacity, to which the Auto Scaling can scale up to. The default value for Target Utilization is 70% (allowed range is 20% - 80% in one percent increments), minimum capacity is 1 unit and maximum capacity is the table limit for your account in the region. Please refer to the [Limits in DynamoDB](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Limits.html#limits-capacity-units-provisioned-throughput) page for region-level default table limits.  
  
Q. Can I change the settings of an existing Auto Scaling policy?  
  
Yes, you can change the settings of an existing Auto Scaling policy at any time, by navigating to the 'Capacity' tab in the management console or programmatically from the CLI or SDK using the [AutoScaling APIs](http://docs.aws.amazon.com/ApplicationAutoScaling/latest/APIReference/Welcome.html).  
  
Q. How does Auto Scaling work?  
  
When you create a new Auto Scaling policy for your DynamoDB table, Amazon CloudWatch alarms are created with thresholds for target utilization you specify, calculated based on consumed and provisioned capacity metrics published to CloudWatch. If the table's actual utilization deviates from target for a specific length of time, the CloudWatch alarms activates Auto Scaling, which evaluates your policy and in turn makes an UpdateTable API request to DynamoDB to dynamically increase (or decrease) the table's provisioned throughput capacity to bring the actual utilization closer to the target.  
  
Q. Can I enable a single Auto Scaling policy across multiple tables in multiple regions?  
  
No, an Auto Scaling policy can only be set to a single table or a global secondary indexes within a single region.  
  
Q. Can I force an Auto Scaling policy to scale up to maximum capacity or scale down to minimum capacity instantly?  
  
No, scaling up instantly to maximum capacity or scaling down to minimum capacity is not supported. Instead, you can temporarily disable Auto Scaling, set desired capacity you need manually for required duration, and re-enable Auto Scaling later.

Q. Where can I monitor the scaling actions triggered by Auto Scaling?  
  
You can monitor status of scaling actions triggered by Auto Scaling under the 'Capacity' tab in the management console and from CloudWatch graphs under the 'Metrics' tab.  
  
Q. How can I tell if a table has an active Auto Scaling policy or not?  
  
From the DynamoDB console, click on Tables in the left menu, to bring up the list view of all DynamoDB tables in your account. For tables with an active Auto Scaling policy, the 'Auto Scaling' column shows either READ\_CAPACITY, WRITE\_CAPACITY or READ\_AND\_WRITE depending on whether Auto Scaling is enabled for read or write or both. Additionally, under the 'Table details' section of the 'Overview' tab of a table, the provisioned capacity label shows whether Auto Scaling is enabled for read, write or both.  
  
Q. What happens to the Auto Scaling policy when I delete a table or global secondary index with an active policy?  
  
When you delete a table or global secondary index from the console, its Auto Scaling policy and supporting Cloud Watch alarms are also deleted.  
  
Q. Are there any additional costs to use Auto Scaling?  
  
No, there are no additional cost to using Auto Scaling, beyond what you already pay for DynamoDB and CloudWatch alarms. To learn about DynamoDB pricing, please visit the [DynamoDB pricing page](https://aws.amazon.com/dynamodb/pricing/).  
  
Q. How does throughput capacity managed by Auto Scaling work with my Reserved Capacity?  
  
Auto Scaling works with reserved capacity in the same manner as manually provisioned throughput capacity does today. Reserved Capacity is applied to the total provisioned capacity for the region you purchased it in. Capacity provisioned by Auto Scaling will consume the reserved capacity first, billed at discounted prices, and any excess capacity will be charged at standard rates. To limit total consumption to the reserved capacity you purchased, distribute maximum capacity limit across all tables with Auto Scaling enabled, to be cumulatively less than total reserved capacity amount you have purchased.

## **Global Secondary Indexes**

Q: What are global secondary indexes?

Global secondary indexes are indexes that contain a partition or partition-and-sort keys that can be different from the table's primary key.

For efficient access to data in a table, Amazon DynamoDB creates and maintains indexes for the primary key attributes. This allows applications to quickly retrieve data by specifying primary key values. However, many applications might benefit from having one or more secondary (or alternate) keys available to allow efficient access to data with attributes other than the primary key. To address this, you can create one or more secondary indexes on a table, and issue Query requests against these indexes.

Amazon DynamoDB supports two types of secondary indexes:

* Local secondary index — an index that has the same partition key as the table, but a different sort key. A local secondary index is "local" in the sense that every partition of a local secondary index is scoped to a table partition that has the same partition key.
* Global secondary index — an index with a partition or a partition-and-sort key that can be different from those on the table. A global secondary index is considered "global" because queries on the index can span all items in a table, across all partitions.

Secondary indexes are automatically maintained by Amazon DynamoDB as sparse objects. Items will only appear in an index if they exist in the table on which the index is defined. This makes queries against an index very efficient, because the number of items in the index will often be significantly less than the number of items in the table.

Global secondary indexes support non-unique attributes, which increases query flexibility by enabling queries against any non-key attribute in the table.

Consider a gaming application that stores the information of its players in a DynamoDB table whose primary key consists of *UserId*(partition) and *GameTitle* (sort). Items have attributes named *TopScore*, *Timestamp*, *ZipCode*, and others. Upon table creation, DynamoDB provides an implicit index (primary index) on the primary key that can support efficient queries that return a specific user’s top scores for all games.

However, if the application requires top scores of users for a particular game, using this primary index would be inefficient, and would require scanning through the entire table. Instead, a global secondary index with GameTitle as the partition key element and TopScore as the sort key element would enable the application to rapidly retrieve top scores for a game.

A GSI does not need to have a sort key element. For instance, you could have a GSI with a key that only has a partition element *GameTitle*. In the example below, the GSI has no projected attributes, so it will just return all items (identified by primary key) that have an attribute matching the *GameTitle* you are querying on.

Q: When should I use global secondary indexes?

Global secondary indexes are particularly useful for tracking relationships between attributes that have a lot of different values. For example, you could create a DynamoDB table with *CustomerID* as the primary partition key for the table and *ZipCode* as the partition key for a global secondary index, since there are a lot of zip codes and since you will probably have a lot of customers. Using the primary key, you could quickly get the record for any customer. Using the global secondary index, you could efficiently query for all customers that live in a given zip code.

To ensure that you get the most out of your global secondary index's capacity, please review our [best practices documentation on uniform workloads](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/GuidelinesForTables.html#GuidelinesForTables.UniformWorkload">best practices documentation on uniform workloads).

Q: How do I create a global secondary index for a DynamoDB table?

GSIs associated with a table can be specified at any time. For detailed steps on creating a Table and its indexes, see [here](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/GettingStartedCreateTables.html). You can create a maximum of [5 global secondary indexes](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Limits.html) per table.

Q: Does the local version of DynamoDB support global secondary indexes?

Yes. The local version of DynamoDB is useful for developing and testing DynamoDB-backed applications. You can download the local version of DynamoDB [here](http://dynamodb-local.s3-website-us-west-2.amazonaws.com/dynamodb_local_latest.zip).

Q: What are projected attributes?

The data in a secondary index consists of attributes that are projected, or copied, from the table into the index. When you create a secondary index, you define the alternate key for the index, along with any other attributes that you want to be projected in the index. Amazon DynamoDB copies these attributes into the index, along with the primary key attributes from the table. You can then query the index just as you would query a table.

Q: Can a global secondary index key be defined on non-unique attributes?

Yes. Unlike the primary key on a table, a GSI index does not require the indexed attributes to be unique. For instance, a GSI on *GameTitle* could index all items that track scores of users for every game. In this example, this GSI can be queried to return all users that have played the game "TicTacToe."

Q: How do global secondary indexes differ from local secondary indexes?

Both global and local secondary indexes enhance query flexibility. An [LSI](https://aws.amazon.com/dynamodb/faqs/#lsi_anchor) is attached to a specific partition key value, whereas a GSI spans all partition key values. Since items having the same partition key value share the same partition in DynamoDB, the "Local" Secondary Index only covers items that are stored together (on the same partition). Thus, the purpose of the LSI is to query items that have the same partition key value but different sort key values. For example, consider a DynamoDB table that tracks Orders for customers, where *CustomerId* is the partition key.

An LSI on *OrderTime* allows for efficient queries to retrieve the most recently ordered items for a particular customer.

In contrast, a GSI is not restricted to items with a common partition key value. Instead, a GSI spans all items of the table just like the primary key. For the table above, a GSI on *ProductId* can be used to efficiently find all orders of a particular product. Note that in this case, no GSI sort key is specified, and even though there might be many orders with the same *ProductId*, they will be stored as separate items in the GSI.

In order to ensure that data in the table and the index are co-located on the same partition, LSIs limit the total size of all elements (tables and indexes) to 10 GB per partition key value. GSIs do not enforce data co-location, and have no such restriction.

When you write to a table, DynamoDB atomically updates all the LSIs affected. In contrast, updates to any GSIs defined on the table are eventually consistent.

LSIs allow the [Query API](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/API_Query.html) to retrieve attributes that are not part of the projection list. This is not supported behavior for GSIs.

Q: How do global secondary indexes work?

In many ways, GSI behavior is similar to that of a DynamoDB table. You can query a GSI using its partition key element, with conditional filters on the GSI sort key element. However, unlike a [primary key](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/DataModel.html#DataModelPrimaryKey) of a DynamoDB table, which must be unique, a GSI key can be the same for multiple items. If multiple items with the same GSI key exist, they are tracked as separate GSI items, and a GSI query will retrieve all of them as individual items. Internally, DynamoDB will ensure that the contents of the GSI are updated appropriately as items are added, removed or updated.

DynamoDB stores a GSI’s projected attributes in the GSI data structure, along with the GSI key and the matching items’ primary keys. GSI’s consume storage for projected items that exist in the source table. This enables queries to be issued against the GSI rather than the table, increasing query flexibility and improving workload distribution. Attributes that are part of an item in a table, but not part of the GSI key, primary key of the table, or projected attributes are thus not returned on querying the GSI index. Applications that need additional data from the table after querying the GSI, can retrieve the primary key from the GSI and then use either the [GetItem](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/API_GetItem.html) or [BatchGetItem](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/API_BatchGetItem.html) APIs to retrieve the desired attributes from the table. As GSI’s are eventually consistent, applications that use this pattern have to accommodate item deletion (from the table) in between the calls to the GSI and GetItem/BatchItem.

DynamoDB automatically handles item additions, updates and deletes in a GSI when corresponding changes are made to the table. When an item (with GSI key attributes) is added to the table, DynamoDB updates the GSI asynchronously to add the new item. Similarly, when an item is deleted from the table, DynamoDB removes the item from the impacted GSI.

Q: Can I create global secondary indexes for partition-based tables and partition-sort schema tables?

Yes, you can create a global secondary index regardless of the type of primary key the DynamoDB table has. The table's primary key can include just a partition key, or it may include both a partition key and a sort key.

Q: What is the consistency model for global secondary indexes?

GSIs support eventual consistency. When items are inserted or updated in a table, the GSIs are not updated synchronously. Under normal operating conditions, a write to a global secondary index will propagate in a fraction of a second. In unlikely failure scenarios, longer delays may occur. Because of this, your application logic should be capable of handling GSI query results that are potentially out-of-date. Note that this is the same behavior exhibited by other DynamoDB APIs that support eventually consistent reads.

Consider a table tracking top scores where each item has attributes *UserId*, *GameTitle* and *TopScore*. The partition key is *UserId*, and the primary sort key is *GameTitle*. If the application adds an item denoting a new top score for *GameTitle* "TicTacToe" and *UserId*"GAMER123," and then subsequently queries the GSI, it is possible that the new score will not be in the result of the query. However, once the GSI propagation has completed, the new item will start appearing in such queries on the GSI.

Q: Can I provision throughput separately for the table and for each global secondary index?

Yes. GSIs manage throughput independently of the table they are based on. When you enable Auto Scaling for a new or existing table from the console, you can optionally choose to apply the same settings to GSIs. You can also provision different throughput for tables and global secondary indexes manually.

Depending upon on your application, the request workload on a GSI can vary significantly from that of the table or other GSIs. Some scenarios that show this are given below:

* A GSI that contains a small fraction of the table items needs a much lower write throughput compared to the table.
* A GSI that is used for infrequent item lookups needs a much lower read throughput, compared to the table.
* A GSI used by a read-heavy background task may need high read throughput for a few hours per day.

As your needs evolve, you can change the provisioned throughput of the GSI, independently of the provisioned throughput of the table.

Consider a DynamoDB table with a GSI that projects all attributes, and has the GSI key present in 50% of the items. In this case, the GSI’s provisioned write capacity units should be set at 50% of the table’s provisioned write capacity units. Using a similar approach, the read throughput of the GSI can be estimated. Please see DynamoDB GSI Documentation for more details.

Q: How does adding a global secondary index impact provisioned throughput and storage for a table?

Similar to a DynamoDB table, a GSI consumes provisioned throughput when reads or writes are performed to it. A write that adds or updates a GSI item will consume write capacity units based on the size of the update. The capacity consumed by the GSI write is in addition to that needed for updating the item in the table.

Note that if you add, delete, or update an item in a DynamoDB table, and if this does not result in a change to a GSI, then the GSI will not consume any write capacity units. This happens when an item without any GSI key attributes is added to the DynamoDB table, or an item is updated without changing any GSI key or projected attributes.

A query to a GSI consumes read capacity units, based on the size of the items examined by the query.

Storage costs for a GSI are based on the total number of bytes stored in that GSI. This includes the GSI key and projected attributes and values, and an overhead of 100 bytes for indexing purposes.

Q: Can DynamoDB throttle my application writes to a table because of a GSI’s provisioned throughput?

Because some or all writes to a DynamoDB table result in writes to related GSIs, it is possible that a GSI’s provisioned throughput can be exhausted. In such a scenario, subsequent writes to the table will be throttled. This can occur even if the table has available write capacity units.

Q: How often can I change provisioned throughput at the index level?

Tables with GSIs have the same [daily limits](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Limits.html) on the number of throughput change operations as normal tables.

Q: How am I charged for DynamoDB global secondary index?

You are charged for the aggregate provisioned throughput for a table and its GSIs by the hour. When you provision manually, while not required, You are charged for the aggregate provisioned throughput for a table and its GSIs by the hour. In addition, you are charged for the data storage taken up by the GSI as well as standard data transfer (external) fees. If you would like to change your GSI’s provisioned throughput capacity, you can do so using the [DynamoDB Console](https://console.aws.amazon.com/dynamodb/) or the [UpdateTable API](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/API_UpdateTable.html) or the [PutScalingPolicy API](http://docs.aws.amazon.com/ApplicationAutoScaling/latest/APIReference/API_PutScalingPolicy.html) for updating Auto Scaling policy settings.

Q: Can I specify which global secondary index should be used for a query?

Yes. In addition to the common query parameters, a GSI [*Query command*](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/API_Query.html) explicitly includes the name of the GSI to operate against. Note that a query can use only one GSI.

Q: What API calls are supported by a global secondary index?

The API calls supported by a GSI are [*Query*](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/API_Query.html) and [*Scan*](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/API_Scan.html). A Query operation only searches index key attribute values and supports a subset of comparison operators. Because GSIs are updated asynchronously, you cannot use the [*ConsistentRead*](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/API_Query.html#DDB-Query-request-ConsistentRead) parameter with the query. Please see [here](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/GSI.html) for details on using GSIs with queries and scans.

Q: What is the order of the results in scan on a global secondary index?

For a global secondary index, with a partition-only key schema there is no ordering. For global secondary index with partition-sort key schema the ordering of the results for the same partition key is based on the sort key attribute.

Q. Can I change Global Secondary Indexes after a table has been created?

Yes, Global Secondary Indexes can be changed at any time, even after the table has been created.

Q. How can I add a Global Secondary Index to an existing table?

You can add a Global Secondary Indexes through the console or through an API call. On the DynamoDB console, first select the table for which you want to add a Global Secondary Index and click the “Create Index” button to add a new index. Follow the steps in the index creation wizard and select “Create” when done. You can also add or delete a Global Secondary Index using the UpdateTable API call with the GlobalSecondaryIndexes parameter.You can learn more by reading our [documentation page](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/GSI.OnlineOps.html).

Q. How can I delete a Global Secondary Index?

You can delete a Global Secondary Index from the console or through an API call. On the DynamoDB console, select the table for which you want to delete a Global Secondary Index. Then, select the “Indexes” tab under “Table Items” and click on the “Delete” button next to delete the index. You can also delete a Global Secondary Index using the UpdateTable API call.You can learn more by reading our[documentation page](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/GSI.OnlineOps.html).

Q. Can I add or delete more than one index in a single API call on the same table?  
  
You can only add or delete one index per API call.

Q. What happens if I submit multiple requests to add the same index?

Only the first add request is accepted and all subsequent add requests will fail till the first add request is finished.

Q. Can I concurrently add or delete several indexes on the same table?

No, at any time there can be only one active add or delete index operation on a table.

Q. Should I provision additional throughput to add a Global Secondary Index?

With Auto Scaling, it is recommended that you apply the same settings to Global Secondary Index as the table. When you provision manually, while not required, it is highly recommended that you provision additional write throughput that is separate from the throughput for the index. If you do not provision additional write throughput, the write throughput from the index will be consumed for adding the new index. This will affect the write performance of the index while the index is being created as well as increase  the time to create the new index.

Q. Do I have to reduce the additional throughput on a Global Secondary Index once the index has been created?

Yes, you would have to dial back the additional write throughput you provisioned for adding an index, once the process is complete.

Q. Can I modify the write throughput that is provisioned for adding a Global Secondary Index?

Yes, you can dial up or dial down the provisioned write throughput for index creation at any time during the creation process.

Q. When a Global Secondary Index is being added or deleted, is the table still available?

Yes, the table is available when the Global Secondary Index is being updated.

Q. When a Global Secondary Index is being added or deleted, are the existing indexes still available?

Yes, the existing indexes are available when the Global Secondary Index is being updated.

Q. When a Global Secondary Index is being created added, is the new index available?

No, the new index becomes available only after the index creation process is finished.

Q. How long does adding a Global Secondary Index take?

The length of time depends on the size of the table and the amount of additional provisioned write throughput for Global Secondary Index creation. The process of adding or deleting an index could vary from a few minutes to a few hours. For example, let's assume that you have a 1GB table that has 500 write capacity units provisioned and you have provisioned 1000 additional write capacity units for the index and new index creation. If the new index includes all the attributes in the table and the table is using all the write capacity units, we expect the index creation will take roughly 30 minutes.

Q. How long does deleting a Global Secondary Index take?

Deleting an index will typically finish in a few minutes. For example, deleting an index with 1GB of data will typically take less than 1 minute.

Q. How do I track the progress of add or delete operation for a Global Secondary Index?

You can use the DynamoDB console or DescribeTable API to check the status of all indexes associated with the table. For an add index operation, while the index is being created, the status of the index will be “CREATING”. Once the creation of the index is finished, the index state will change from “CREATING” to “ACTIVE”. For a delete index operation, when the request is complete, the deleted index will cease to exist.

Q. Can I get a notification when the index creation process for adding a Global Secondary Index is complete?

You can request a notification to be sent to your email address confirming that the index addition has been completed. When you add an index through the console, you can request a notification on the last step before creating the index. When the index creation is complete, DynamoDB will send an SNS notification to your email.

Q. What happens when I try to add more Global Secondary Indexes, when I already have 5?

You are currently limited to 5 GSIs. The “Add” operation will fail and you will get an error.

Q. Can I reuse a name for a Global Secondary Index after an index with the same name has been deleted?

Yes, once a Global Secondary Index has been deleted, that index name can be used again when a new index is added.

Q. Can I cancel an index add while it is being created?

No, once index creation starts, the index creation process cannot be canceled.

Q: Are GSI key attributes required in all items of a DynamoDB table?

No. GSIs are sparse indexes. Unlike the requirement of having a primary key, an item in a DynamoDB table does not have to contain any of the GSI keys. If a GSI key has both partition and sort elements, and a table item omits either of them, then that item will not be indexed by the corresponding GSI. In such cases, a GSI can be very useful in efficiently locating items that have an uncommon attribute.

Q: Can I retrieve all attributes of a DynamoDB table from a global secondary index?

A query on a GSI can only return attributes that were specified to be included in the GSI at creation time. The attributes included in the GSI are those that are projected by default such as the GSI’s key attribute(s) and table’s primary key attribute(s), and those that the user specified to be projected. For this reason, a GSI query will not return attributes of items that are part of the table, but not included in the GSI. A GSI that specifies all attributes as projected attributes can be used to retrieve any table attributes. See [here](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/GSI.html) for documentation on using GSIs for queries.

Q: How can I list GSIs associated with a table?

The [DescribeTable API](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/API_DescribeTable.html) will return detailed information about global secondary indexes on a table.

Q: What data types can be indexed?

All scalar data types (Number, String, Binary, and Boolean) can be used for the sort key element of the local secondary index key. Set, list, and map types cannot be indexed.

Q: Are composite attribute indexes possible?

No. But you can concatenate attributes into a string and use this as a key.

Q: What data types can be part of the projected attributes for a GSI?

You can specify attributes with any data types (including set types) to be projected into a GSI.

Q: What are some scalability considerations of GSIs?

[Performance considerations](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/GuidelinesForTables.html#GuidelinesForTables.UniformWorkload) of the primary key of a DynamoDB table also apply to GSI keys. A GSI assumes a relatively random access pattern across all its keys. To get the most out of secondary index provisioned throughput, you should select a GSI partition key attribute that has a large number of distinct values, and a GSI sort key attribute that is requested fairly uniformly, as randomly as possible.

Q: What new metrics will be available through CloudWatch for global secondary indexes?

Tables with GSI will provide aggregate metrics for the table and GSIs, as well as breakouts of metrics for the table and each GSI.

Reports for individual GSIs will support a subset of the CloudWatch metrics that are supported by a table. These include:

* Read Capacity (Provisioned Read Capacity, Consumed Read Capacity)
* Write Capacity (Provisioned Write Capacity, Consumed Write Capacity)
* Throttled read events
* Throttled write events

For more details on metrics supported by DynamoDB tables and indexes see [here](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/MonitoringDynamoDB.html).

Q: How can I scan a Global Secondary Index?

Global secondary indexes can be scanned via the Console or the Scan API.

To scan a global secondary index, explicitly reference the index in addition to the name of the table you’d like to scan. You must specify the index partition attribute name and value. You can optionally specify a condition against the index key sort attribute.

Q: Will a Scan on Global secondary index allow me to specify non-projected attributes to be returned in the result set?

Scan on global secondary indexes will not support fetching of non-projected attributes.

Q: Will there be parallel scan support for indexes?

Yes, parallel scan will be supported for indexes and the semantics are the same as that for the main table.

## **Local Secondary Indexes**

Q: What are local secondary indexes?

Local secondary indexes enable some common queries to run more quickly and cost-efficiently, that would otherwise require retrieving a large number of items and then filtering the results. It means your applications can rely on more flexible queries based on a wider range of attributes.

Before the launch of local secondary indexes, if you wanted to find specific items within a partition (items that share the same partition key), DynamoDB would have fetched all objects that share a single partition key, and filter the results accordingly. For instance, consider an e-commerce application that stores customer order data in a DynamoDB table with partition-sort schema of customer id-order timestamp. Without LSI, to find an answer to the question “Display all orders made by Customer X with shipping date in the past 30 days, sorted by shipping date”, you had to use the Query API to retrieve all the objects under the partition key “X”, sort the results by shipment date and then filter out older records.

With local secondary indexes, we are simplifying this experience. Now, you can create an index on “shipping date” attribute and execute this query efficiently and just retieve only the necessary items. This significantly reduces the latency and cost of your queries as you will retrieve only items that meet your specific criteria. Moreover, it also simplifies the programming model for your application as you no longer have to write customer logic to filter the results. We call this new secondary index a ‘local’ secondary index because it is used along with the partition key and hence allows you to search locally within a partition key bucket. So while previously you could only search using the partition key and the sort key, now you can also search using a secondary index in place of the sort key, thus expanding the number of attributes that can be used for queries which can be conducted efficiently.

Redundant copies of data attributes are copied into the local secondary indexes you define. These attributes include the table partition and sort key, plus the alternate sort key you define. You can also redundantly store other data attributes in the local secondary index, in order to access those other attributes without having to access the table itself.

Local secondary indexes are not appropriate for every application. They introduce some constraints on the volume of data you can store within a single partition key value. For more information, see the FAQ items below about item collections.

Q: What are Projections?

The set of attributes that is copied into a local secondary index is called a projection. The projection determines the attributes that you will be able to retrieve with the most efficiency. When you query a local secondary index, Amazon DynamoDB can access any of the projected attributes, with the same performance characteristics as if those attributes were in a table of their own. If you need to retrieve any attributes that are not projected, Amazon DynamoDB will automatically fetch those attributes from the table.

When you define a local secondary index, you need to specify the attributes that will be projected into the index. At a minimum, each index entry consists of: (1) the table partition key value, (2) an attribute to serve as the index sort key, and (3) the table sort key value.

Beyond the minimum, you can also choose a user-specified list of other non-key attributes to project into the index. You can even choose to project all attributes into the index, in which case the index replicates the same data as the table itself, but the data is organized by the alternate sort key you specify.

Q: How can I create a LSI?

You need to create a LSI at the time of table creation. It can’t currently be added later on. To create an LSI, specify the following two parameters:

Indexed Sort key – the attribute that will be indexed and queried on.

Projected Attributes – the list of attributes from the table that will be copied directly into the local secondary index, so they can be returned more quickly without fetching data from the primary index, which contains all the items of the table. Without projected attributes, local secondary index contains only primary and secondary index keys.

Q: What is the consistency model for LSI?

Local secondary indexes are updated automatically when the primary index is updated. Similar to reads from a primary index, LSI supports both strong and eventually consistent read options.

Q: Do local secondary indexes contain references to all items in the table?

No, not necessarily. Local secondary indexes only reference those items that contain the indexed sort key specified for that LSI. DynamoDB’s flexible schema means that not all items will necessarily contain all attributes.

This means local secondary index can be sparsely populated, compared with the primary index. Because local secondary indexes are sparse, they are efficient to support queries on attributes that are uncommon.

For example, in the Orders example described above, a customer may have some additional attributes in an item that are included only if the order is canceled (such as CanceledDateTime, CanceledReason). For queries related to canceled items, an local secondary index on either of these attributes would be efficient since the only items referenced in the index would be those that had these attributes present.

Q: How do I query local secondary indexes?

Local secondary indexes can only be queried via the Query API.

To query a local secondary index, explicitly reference the index in addition to the name of the table you’d like to query. You must specify the index partition attribute name and value. You can optionally specify a condition against the index key sort attribute.

Your query can retrieve non-projected attributes stored in the primary index by performing a table fetch operation, with a cost of additional read capacity units.

Both strongly consistent and eventually consistent reads are supported for query using local secondary index.

Q: How do I create local secondary indexes?

Local secondary indexes must be defined at time of table creation. The primary index of the table must use a partition-sort composite key.

Q: Can I add local secondary indexes to an existing table?

No, it’s not possible to add local secondary indexes to existing tables at this time. We are working on adding this capability and will be releasing it in the future. When you create a table with local secondary index, you may decide to create local secondary index for future use by defining a sort key element that is currently not used. Since local secondary index are sparse, this index costs nothing until you decide to use it.

Q: How many local secondary indexes can I create on one table?

Each table can have up to five local secondary indexes.

Q: How many projected non-key attributes can I create on one table?

Each table can have up to 20 projected non-key attributes, in total across all local secondary indexes within the table. Each index may also specifify that all non-key attributes from the primary index are projected.

Q: Can I modify the index once it is created?

No, an index cannot be modified once it is created. We are working to add this capability in the future.

Q: Can I delete local secondary indexes?

No, local secondary indexes cannot be removed from a table once they are created at this time. Of course, they are deleted if you also decide to delete the entire table. We are working on adding this capability and will be releasing it in the future.

Q: How do local secondary indexes consume provisioned capacity?

You don’t need to explicitly provision capacity for a local secondary index. It consumes provisioned capacity as part of the table with which it is associated.

Reads from LSIs and writes to tables with LSIs consume capacity by the standard formula of 1 unit per 1KB of data, with the following differences:

When writes contain data that are relevant to one or more local secondary indexes, those writes are mirrored to the appropriate local secondary indexes. In these cases, write capacity will be consumed for the table itself, and additional write capacity will be consumed for each relevant LSI.

Updates that overwrite an existing item can result in two operations– delete and insert – and thereby consume extra units of write capacity per 1KB of data.

When a read query requests attributes that are not projected into the LSI, DynamoDB will fetch those attributes from the primary index. This implicit GetItem request consumes one read capacity unit per 4KB of item data fetched.

Q: How much storage will local secondary indexes consume?

Local secondary indexes consume storage for the attribute name and value of each LSI’s primary and index keys, for all projected non-key attributes, plus 100 bytes per item reflected in the LSI.

Q: What data types can be indexed?

All scalar data types (Number, String, Binary) can be used for the sort key element of the local secondary index key. Set types cannot be used.

Q: What data types can be projected into a local secondary index?

All data types (including set types) can be projected into a local secondary index.

Q: What are item collections and how are they related to LSI?

In Amazon DynamoDB, an item collection is any group of items that have the same partition key, across a table and all of its local secondary indexes. Traditional partitioned (or sharded) relational database systems call these shards or partitions, referring to all database items or rows stored under a partition key.

Item collections are automatically created and maintained for every table that includes local secondary indexes. DynamoDB stores each item collection within a single disk partition.

Q: Are there limits on the size of an item collection?

Every item collection in Amazon DynamoDB is subject to a maximum size limit of 10 gigabytes. For any distinct partition key value, the sum of the item sizes in the table plus the sum of the item sizes across all of that table's local secondary indexes must not exceed 10 GB.

The 10 GB limit for item collections does not apply to tables without local secondary indexes; only tables that have one or more local secondary indexes are affected.

Although individual item collections are limited in size, the storage size of an overall table with local secondary indexes is not limited. The total size of an indexed table in Amazon DynamoDB is effectively unlimited, provided the total storage size (table and indexes) for any one partition key value does not exceed the 10 GB threshold.

Q: How can I track the size of an item collection?

DynamoDB’s write APIs (PutItem, UpdateItem, DeleteItem, and BatchWriteItem) include an option, which allows the API response to include an estimate of the relevant item collection’s size. This estimate includes lower and upper size estimate for the data in a particular item collection, measured in gigabytes.

We recommend that you instrument your application to monitor the sizes of your item collections. Your applications should examine the API responses regarding item collection size, and log an error message whenever an item collection exceeds a user-defined limit (8 GB, for example). This would provide an early warning system, letting you know that an item collection is growing larger, but giving you enough time to do something about it.

Q: What if I exceed the 10GB limit for an item collection?

If a particular item collection exceeds the 10GB limit, then you will not be able to write new items, or increase the size of existing items, for that particular partition key. Read and write operations that shrink the size of the item collection are still allowed. Other item collections in the table are not affected.

To address this problem , you can remove items or reduce item sizes in the collection that has exceeded 10GB. Alternatively, you can introduce new items under a new partition key value to work around this problem. If your table includes historical data that is infrequently accessed, consider archiving the historical data to Amazon S3, Amazon Glacier or another data store.

Q: How can I scan a local secondary index?

To scan a local secondary index, explicitly reference the index in addition to the name of the table you’d like to scan. You must specify the index partition attribute name and value. You can optionally specify a condition against the index key sort attribute.

Your scan can retrieve non-projected attributes stored in the primary index by performing a table fetch operation, with a cost of additional read capacity units.

Q: Will a Scan on a local secondary index allow me to specify non-projected attributes to be returned in the result set?

Scan on local secondary indexes will support fetching of non-projected attributes.

Q: What is the order of the results in scan on a local secondary index?

For local secondary index, the ordering within a collection will be the based on the order of the indexed attribute.

## **Security and Control**

Q: What is DynamoDB Fine-Grained Access Control?

Fine Grained Access Control (FGAC) gives a DynamoDB table owner a high degree of control over data in the table. Specifically, the table owner can indicate *who* (caller) can access *which* items or attributes of the table and perform *what* actions (read / write capability). FGAC is used in concert with [AWS Identity and Access Management (IAM)](https://aws.amazon.com/iam/), which manages the security credentials and the associated permissions.

Q: What are the common use cases for DynamoDB FGAC?

FGAC can benefit any application that tracks information in a DynamoDB table, where the end user (or application client acting on behalf of an end user) wants to read or modify the table directly, without a middle-tier service. For instance, a developer of a mobile app named *Acme* can use FGAC to track the top score of every *Acme* user in a DynamoDB table. FGAC allows the application client to modify only the top score for the user that is currently running the application.

Q: Can I use Fine Grain Access Control with JSON documents?

Yes. You can use Fine Grain Access Control (FGAC) to restrict access to your data based on top-level attributes in your document. You cannot use FGAC to restrict access based on nested attributes. For example, suppose you stored a JSON document that contained the following information about a person: ID, first name, last name, and a list of all of their friends. You could use FGAC to restrict access based on their ID, first name, or last name, but not based on the list of friends.

Q: Without FGAC, how can a developer achieve item level access control?

To achieve this level of control without FGAC, a developer would have to choose from a few potentially onerous approaches. Some of these are:

1. Proxy: The application client sends a request to a brokering proxy that performs the authentication and authorization. Such a solution increases the complexity of the system architecture and can result in a higher total cost of ownership (TCO).
2. Per Client Table: Every application client is assigned its own table. Since application clients access different tables, they would be protected from one another. This could potentially require a developer to create millions of tables, thereby making database management extremely painful.
3. Per-Client Embedded Token: A secret token is embedded in the application client. The shortcoming of this is the difficulty in changing the token and handling its impact on the stored data. Here, the key of the items accessible by this client would contain the secret token.

Q: How does DynamoDB FGAC work?

With FGAC, an application requests a security token that authorizes the application to access only specific items in a specific DynamoDB table. With this token, the end user application agent can make requests to DynamoDB directly. Upon receiving the request, the incoming request’s credentials are first evaluated by DynamoDB, which will use IAM to authenticate the request and determine the capabilities allowed for the user. If the user’s request is not permitted, FGAC will prevent the data from being accessed.

Q: How much does DynamoDB FGAC cost?

There is no additional charge for using FGAC. As always, you only pay for the provisioned throughput and storage associated with the DynamoDB table.

Q: How do I get started?

Refer to the [Fine-Grained Access Control](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/FGAC_DDB.html) section of the DynamoDB Developer Guide to learn how to create an access policy, create an IAM role for your app (e.g. a role named AcmeFacebookUsers for a Facebook app\_id of 34567), and assign your access policy to the role. The trust policy of the role determines which identity providers are accepted (e.g. Login with Amazon, Facebook, or Google), and the access policy describes which AWS resources can be accessed (e.g. a DynamoDB table). Using the role, your app can now to obtain temporary credentials for DynamoDB by calling the AssumeRoleWithIdentityRequest API of the AWS Security Token Service (STS).

Q: How do I allow users to Query a Local Secondary Index, but prevent them from causing a table fetch to retrieve non-projected attributes?

Some Query operations on a Local Secondary Index can be more expensive than others if they request attributes that are not projected into an index. You an restrict such potentially expensive “fetch” operations by limiting the permissions to only projected attributes, using the "dynamodb:Attributes" context key.

Q: How do I prevent users from accessing specific attributes?

The recommended approach to preventing access to specific attributes is to follow the principle of least privilege, and Allow access to only specific attributes.

Alternatively, you can use a *Deny* policy to specify attributes that are disallowed. However, this is not recommended for the following reasons:

1. With a *Deny*policy, it is possible for the user to discover the hidden attribute names by issuing repeated requests for every possible attribute name, until the user is ultimately denied access.
2. *Deny* policies are more fragile, since DynamoDB could introduce new API functionality in the future that might allow an access pattern that you had previously intended to block.

Q: How do I prevent users from adding invalid data to a table?

The available FGAC controls can determine which items changed or read, and which attributes can be changed or read. Users can add new items without those blocked attributes, and change any value of any attribute that is modifiable.

Q: Can I grant access to multiple attributes without listing all of them?

Yes, the IAM policy language supports a rich set of comparison operations, including StringLike, StringNotLike, and many others.  For additional details, please see the [IAM Policy Reference](http://docs.aws.amazon.com/IAM/latest/UserGuide/reference_policies.html).

Q: How do I create an appropriate policy?

We recommend that you use the DynamoDB Policy Generator from the DynamoDB console. You may also compare your policy to those listed in the Amazon DynamoDB Developer Guide to make sure you are following a recommended pattern. You can post policies to the AWS Forums to get thoughts from the DynamoDB community.

Q: Can I grant access based on a canonical user id instead of separate ids for the user based on the identity provider they logged in with?

Not without running a “token vending machine”. If a user retrieves federated access to your IAM role directly using Facebook credentials with STS, those temporary credentials only have information about that user’s Facebook login, and not their Amazon login, or Google login. If you want to internally store a mapping of each of these logins to your own stable identifier, you can run a service that the user contacts to log in, and then call STS and provide them with credentials scoped to whatever partition key value you come up with as their canonical user id.

Q: What information cannot be hidden from callers using FGAC?

Certain information cannot currently be blocked from the caller about the items in the table:

* Item collection metrics. The caller can ask for the estimated number of items and size in bytes of the item collection.
* Consumed throughput The caller can ask for the detailed breakdown or summary of the provisioned throughput consumed by operations.
* Validation cases. In certain cases, the caller can learn about the existence and primary key schema of a table when you did not intend to give them access. To prevent this, follow the principle of least privilege and only allow access to the tables and actions that you intended to allow access to.
* If you deny access to specific attributes instead of whitelisting access to specific attributes, the caller can theoretically determine the names of the hidden attributes if “allow all except for” logic. It is safer to whitelist specific attribute names instead.

Q: Does Amazon DynamoDB support IAM permissions?

Yes, DynamoDB supports API-level permissions through AWS Identity and Access Management (IAM) service integration.

For more information about IAM, go to:

* [AWS Identity and Access Management](https://aws.amazon.com/iam/)
* [AWS Identity and Access Management Getting Started Guide](http://docs.amazonwebservices.com/IAM/latest/GettingStartedGuide/)
* [Using AWS Identity and Access Management](http://docs.amazonwebservices.com/IAM/latest/UserGuide/)

Q: I wish to perform security analysis or operational troubleshooting on my DynamoDB tables. Can I get a history of all DynamoDB 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. Details about DynamoDB support for CloudTrail can be found [here](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/logging-using-cloudtrail.html). Learn more about CloudTrail at the [AWS CloudTrail detail page](http://aws.amazon.com/cloudtrail/), and turn it on via CloudTrail's [AWS Management Console](https://console.aws.amazon.com/cloudtrail/home) home page.

## **Pricing**

Q: How will I be charged for my use of Amazon DynamoDB?

Each DynamoDB table has provisioned read-throughput and write-throughput associated with it. You are billed by the hour for that throughput capacity if you exceed the free tier.

Please note that you are charged by the hour for the throughput capacity, whether or not you are sending requests to your table. If you would like to change your table’s provisioned throughput capacity, you can do so using the AWS Management Console, the UpdateTable API or the PutScalingPolicy API for Auto Scaling..

In addition, DynamoDB also charges for indexed data storage as well as the standard internet data transfer fees

To learn more about DynamoDB pricing, please visit the [DynamoDB pricing page](https://aws.amazon.com/dynamodb/pricing/).

Q: What are some pricing examples?

Here is an example of how to calculate your throughput costs using US East (Northern Virginia) Region pricing. To view prices for other regions, [visit our pricing page](https://aws.amazon.com/dynamodb/pricing/).

If you create a table and request 10 units of write capacity and 200 units of read capacity of provisioned throughput, you would be charged:

$0.01 + (4 x $0.01) = $0.05 per hour

If your throughput needs changed and you increased your reserved throughput requirement to 10,000 units of write capacity and 50,000 units of read capacity, your bill would then change to:

(1,000 x $0.01) + (1,000 x $0.01) = $20/hour

To learn more about DynamoDB pricing, please visit the [DynamoDB pricing page](https://aws.amazon.com/dynamodb/pricing/).

Q: Do your prices include taxes?

For details on taxes, see [Amazon Web Services Tax Help](https://aws.amazon.com/tax-help/).

Q: What is provisioned throughput?

Amazon DynamoDB Auto Scaling adjusts throughput capacity automatically as request volumes change, based on your desired target utilization and minimum and maximum capacity limits, or lets you specify the request throughput you want your table to be able to achieve manually. Behind the scenes, the service handles the provisioning of resources to achieve the requested throughput rate. Rather than asking you to think about instances, hardware, memory, and other factors that could affect your throughput rate, we simply ask you to provision the throughput level you want to achieve. This is the provisioned throughput model of service.

During creation of a new table or global secondary index, Auto Scaling is enabled by default with default settings for target utilization, minimum and maximum capacity; or you can specify your required read and write capacity needs manually; and Amazon DynamoDB automatically partitions and reserves the appropriate amount of resources to meet your throughput requirements.

Q: How does selection of primary key influence the scalability I can achieve?

When storing data, Amazon DynamoDB divides a table into multiple partitions and distributes the data based on the partition key element of the primary key. While allocating capacity resources, Amazon DynamoDB assumes a relatively random access pattern across all primary keys. You should set up your data model so that your requests result in a fairly even distribution of traffic across primary keys. If a table has a very small number of heavily-accessed partition key elements, possibly even a single very heavily-used partition key element, traffic is concentrated on a small number of partitions – potentially only one partition. If the workload is heavily unbalanced, meaning disproportionately focused on one or a few partitions, the operations will not achieve the overall provisioned throughput level. To get the most out of Amazon DynamoDB throughput, build tables where the partition key element has a large number of distinct values, and values are requested fairly uniformly, as randomly as possible. An example of a good primary key is CustomerID if the application has many customers and requests made to various customer records tend to be more or less uniform. An example of a heavily skewed primary key is “Product Category Name” where certain product categories are more popular than the rest.

Q: What is a read/write capacity unit?

How do I estimate how many read and write capacity units I need for my application? A unit of Write Capacity enables you to perform one write per second for items of up to 1KB in size. Similarly, a unit of Read Capacity enables you to perform one strongly consistent read per second (or two eventually consistent reads per second) of items of up to 4KB in size. Larger items will require more capacity. You can calculate the number of units of read and write capacity you need by estimating the number of reads or writes you need to do per second and multiplying by the size of your items (rounded up to the nearest KB).

Units of Capacity required for writes = Number of item writes per second x item size in 1KB blocks

Units of Capacity required for reads\* = Number of item reads per second x item size in 4KB blocks

\* If you use eventually consistent reads you’ll get twice the throughput in terms of reads per second.

If your items are less than 1KB in size, then each unit of Read Capacity will give you 1 strongly consistent read/second and each unit of Write Capacity will give you 1 write/second of capacity. For example, if your items are 512 bytes and you need to read 100 items per second from your table, then you need to provision 100 units of Read Capacity.

If your items are larger than 4KB in size, then you should calculate the number of units of Read Capacity and Write Capacity that you need. For example, if your items are 4.5KB and you want to do 100 strongly consistent reads/second, then you would need to provision 100 (read per second) x 2 (number of 4KB blocks required to store 4.5KB) = 200 units of Read Capacity.

Note that the required number of units of Read Capacity is determined by the number of items being read per second, not the number of API calls. For example, if you need to read 500 items per second from your table, and if your items are 4KB or less, then you need 500 units of Read Capacity. It doesn’t matter if you do 500 individual GetItem calls or 50 BatchGetItem calls that each return 10 items.

Q: Will I always be able to achieve my level of provisioned throughput?

Amazon DynamoDB assumes a relatively random access pattern across all primary keys. You should set up your data model so that your requests result in a fairly even distribution of traffic across primary keys. If you have a highly uneven or skewed access pattern, you may not be able to achieve your level of provisioned throughput.

When storing data, Amazon DynamoDB divides a table into multiple partitions and distributes the data based on the partition key element of the primary key. The provisioned throughput associated with a table is also divided among the partitions; each partition's throughput is managed independently based on the quota allotted to it. There is no sharing of provisioned throughput across partitions. Consequently, a table in Amazon DynamoDB is best able to meet the provisioned throughput levels if the workload is spread fairly uniformly across the partition key values. Distributing requests across partition key values distributes the requests across partitions, which helps achieve your full provisioned throughput level.

If you have an uneven workload pattern across primary keys and are unable to achieve your provisioned throughput level, you may be able to meet your throughput needs by increasing your provisioned throughput level further, which will give more throughput to each partition. However, it is recommended that you considering modifying your request pattern or your data model in order to achieve a relatively random access pattern across primary keys.

Q: If I retrieve only a single element of a JSON document, will I be charged for reading the whole item?

Yes. When reading data out of DynamoDB, you consume the throughput required to read the entire item.

Q: What is the maximum throughput I can provision for a single DynamoDB table?

DynamoDB is designed to scale without limits However, if you wish to exceed throughput rates of 10,000 write capacity units or 10,000 read capacity units for an individual table, you must first [contact Amazon through this online form](http://portal.aws.amazon.com/gp/aws/html-forms-controller/DynamoDB_Limit_Increase_Form). If you wish to provision more than 20,000 write capacity units or 20,000 read capacity units from a single subscriber account you must first [contact us](http://portal.aws.amazon.com/gp/aws/html-forms-controller/DynamoDB_Limit_Increase_Form) using the form described above.

Q: What is the minimum throughput I can provision for a single DynamoDB table?

The smallest provisioned throughput you can request is 1 write capacity unit and 1 read capacity unit for both Auto Scaling and manual throughput provisioning..

This falls within the free tier which allows for 25 units of write capacity and 25 units of read capacity. The free tier applies at the account level, not the table level. In other words, if you add up the provisioned capacity of all your tables, and if the total capacity is no more than 25 units of write capacity and 25 units of read capacity, your provisioned capacity would fall into the free tier.

Q: Is there any limit on how much I can change my provisioned throughput with a single request?

You can increase the provisioned throughput capacity of your table by any amount using the UpdateTable API. For example, you could increase your table’s provisioned write capacity from 1 write capacity unit to 10,000 write capacity units with a single API call. Your account is still subject to table-level and account-level limits on capacity, as described in our [documentation page](http://docs.amazonwebservices.com/amazondynamodb/latest/developerguide/WorkingWithDDItems.html). If you need to raise your provisioned capacity limits, you can visit our [Support Center](https://aws.amazon.com/support), click “Open a new case”, and file a service limit increase request.

Q: How am I charged for provisioned throughput?

Every Amazon DynamoDB table has pre-provisioned the resources it needs to achieve the throughput rate you asked for. You are billed at an hourly rate for as long as your table holds on to those resources. For a complete list of prices with examples, see the [DynamoDB pricing page](https://aws.amazon.com/dynamodb/pricing/).

Q: How do I change the provisioned throughput for an existing DynamoDB table?

There are two ways to update the provisioned throughput of an Amazon DynamoDB table. You can either make the change in the management console, or you can use the UpdateTable API call. In either case, Amazon DynamoDB will remain available while your provisioned throughput level increases or decreases.

Q: How often can I change my provisioned throughput?

You can increase your provisioned throughput as often as you want. You can decrease up to four times any time per day. A day is defined according to the GMT time zone. Additionally, if there was no decrease in the past four hours, an additional dial down is allowed, effectively bringing maximum number of decreases in a day to 9 (4 decreases in the first 4 hours, and 1 decrease for each of the subsequent 4 hour windows in a day).

Keep in mind that you can’t change your provisioned throughput if your Amazon DynamoDB table is still in the process of responding to your last request to change provisioned throughput. Use the management console or the DescribeTables API to check the status of your table. If the status is “CREATING”, “DELETING”, or “UPDATING”, you won’t be able to adjust the throughput of your table. Please wait until you have a table in “ACTIVE” status and try again.

Q: Does the consistency level affect the throughput rate?

Yes. For a given allocation of resources, the read-rate that a DynamoDB table can achieve is different for strongly consistent and eventually consistent reads. If you request “1,000 read capacity units”, DynamoDB will allocate sufficient resources to achieve 1,000 strongly consistent reads per second of items up to 4KB. If you want to achieve 1,000 eventually consistent reads of items up to 4KB, you will need half of that capacity, i.e., 500 read capacity units. For additional guidance on choosing the appropriate throughput rate for your table, see our provisioned throughput guide.

Q: Does the item size affect the throughput rate?

Yes. For a given allocation of resources, the read-rate that a DynamoDB table can achieve does depend on the size of an item. When you specify the provisioned read throughput you would like to achieve, DynamoDB provisions its resources on the assumption that items will be less than 4KB in size. Every increase of up to 4KB will linearly increase the resources you need to achieve the same throughput rate. For example, if you have provisioned a DynamoDB table with 100 units of read capacity, that means that it can handle 100 4KB reads per second, or 50 8KB reads per second, or 25 16KB reads per second, and so on.

Similarly the write-rate that a DynamoDB table can achieve does depend on the size of an item. When you specify the provisioned write throughput you would like to achieve, DynamoDB provisions its resources on the assumption that items will be less than 1KB in size. Every increase of up to 1KB will linearly increase the resources you need to achieve the same throughput rate. For example, if you have provisioned a DynamoDB table with 100 units of write capacity, that means that it can handle 100 1KB writes per second, or 50 2KB writes per second, or 25 4KB writes per second, and so on.

For additional guidance on choosing the appropriate throughput rate for your table, see our provisioned throughput guide.

Q: What happens if my application performs more reads or writes than my provisioned capacity?

If your application performs more reads/second or writes/second than your table’s provisioned throughput capacity allows, requests above your provisioned capacity will be throttled and you will receive 400 error codes. For instance, if you had asked for 1,000 write capacity units and try to do 1,500 writes/second of 1 KB items, DynamoDB will only allow 1,000 writes/second to go through and you will receive error code 400 on your extra requests. You should use CloudWatch to monitor your request rate to ensure that you always have enough provisioned throughput to achieve the request rate that you need.

Q: How do I know if I am exceeding my provisioned throughput capacity?

DynamoDB publishes your consumed throughput capacity as a CloudWatch metric. You can set an alarm on this metric so that you will be notified if you get close to your provisioned capacity.

Q: How long does it take to change the provisioned throughput level of a table?

In general, decreases in throughput will take anywhere from a few seconds to a few minutes, while increases in throughput will typically take anywhere from a few minutes to a few hours.

We strongly recommend that you do not try and schedule increases in throughput to occur at almost the same time when that extra throughput is needed. We recommend provisioning throughput capacity sufficiently far in advance to ensure that it is there when you need it.

## **Reserved Capacity**

Q: What is Reserved Capacity?

Reserved Capacity is a billing feature that allows you to obtain discounts on your provisioned throughput capacity in exchange for:

* A one-time up-front payment
* A commitment to a minimum monthly usage level for the duration of the term of the agreement.

Reserved Capacity applies within a single AWS Region and can be purchased with 1-year or 3-year terms. Every DynamoDB table has provisioned throughput capacity associated with it, whether managed by Auto Scaling or provisioned manually when you create or update a table. This capacity is what determines the read and write throughput rate that your DynamoDB table can achieve. Reserved Capacity is a billing arrangement and has no direct impact on the performance or capacity of your DynamoDB tables. For example, if you buy 100 write capacity units of Reserved Capacity, you have agreed to pay for that much capacity for the duration of the agreement (1 or 3 years) in exchange for discounted pricing.

Q: How do I buy Reserved Capacity?

Log into the [AWS Management Console](https://aws.amazon.com/console/), go to the DynamoDB console page, and then click on "Reserved Capacity”. This will take you to the "Reserved Capacity Usage" page. Click on "Purchase Reserved Capacity" and this will bring up a form you can fill out to purchase Reserved Capacity. Make sure you have selected the AWS Region in which your Reserved Capacity will be used. After you have finished purchasing Reserved Capacity, you will see purchase you made on the "Reserved Capacity Usage" page.

Q: Can I cancel a Reserved Capacity purchase?

No, you cannot cancel your Reserved Capacity and the one-time payment is not refundable. You will continue to pay for every hour during your Reserved Capacity term regardless of your usage.

Q: What is the smallest amount of Reserved Capacity that I can buy?

The smallest Reserved Capacity offering is 100 capacity units (reads or writes).

Q: Are there APIs that I can use to buy Reserved Capacity?

Not yet. We will provide APIs and add more Reserved Capacity options over time.

Q: Can I move Reserved Capacity from one Region to another?

No. Reserved Capacity is associated with a single Region.

Q: Can I provision more throughput capacity than my Reserved Capacity?

Yes. When you purchase Reserved Capacity, you are agreeing to a minimum usage level and you pay a discounted rate for that usage level. If you provision more capacity than that minimum level, you will be charged at standard rates for the additional capacity.

Q: How do I use my Reserved Capacity?

Reserved Capacity is automatically applied to your bill. For example, if you purchased 100 write capacity units of Reserved Capacity and you have provisioned 300, then your Reserved Capacity purchase will automatically cover the cost of 100 write capacity units and you will pay standard rates for the remaining 200 write capacity units.

Q: What happens if I provision less throughput capacity than my Reserved Capacity?

A Reserved Capacity purchase is an agreement to pay for a minimum amount of provisioned throughput capacity, for the duration of the term of the agreement, in exchange for discounted pricing. If you use less than your Reserved Capacity, you will still be charged each month for that minimum amount of provisioned throughput capacity.

Q: Can I use my Reserved Capacity for multiple DynamoDB tables?

Yes. Reserved Capacity is applied to the total provisioned capacity within the Region in which you purchased your Reserved Capacity. For example, if you purchased 5,000 write capacity units of Reserved Capacity, then you can apply that to one table with 5,000 write capacity units, or 100 tables with 50 write capacity units, or 1,000 tables with 5 write capacity units, etc.

Q: Does Reserved Capacity apply to DynamoDB usage in Consolidated Billing accounts?

Yes. If you have multiple accounts linked with Consolidated Billing, Reserved Capacity units purchased either at the Payer Account level or Linked Account level are shared with all accounts connected to the Payer Account. Reserved capacity will first be applied to the account which purchased it and then any unused capacity will be applied to other linked accounts.

## **DynamoDB Cross-region Replication**

Q: What is a DynamoDB cross-region replication?

DynamoDB cross-region replication allows you to maintain identical copies (called replicas) of a DynamoDB table (called master table) in one or more AWS regions. After you enable cross-region replication for a table, identical copies of the table are created in other AWS regions. Writes to the table will be automatically propagated to all replicas.

Q: When should I use cross-region replication?

You can use cross-region replication for the following scenarios.

* Efficient disaster recovery: By replicating tables in multiple data centers, you can switch over to using DynamoDB tables from another region in case a data center failure occurs.
* Faster reads: If you have customers in multiple regions, you can deliver data faster by reading a DynamoDB table from the closest AWS data center.
* Easier traffic management: You can use replicas to distribute the read workload across tables and thereby consume less read capacity in the master table.
* Easy regional migration: By creating a read replica in a new region and then promoting the replica to be a master, you migrate your application to that region more easily.
* Live data migration: To move a DynamoDB table from one region to another, you can create a replica of the table from the source region in the destination region. When the tables are in sync, you can switch your application to write to the destination region.

Q: What cross-region replication modes are supported?

Cross-region replication currently supports single master mode. A single master has one master table and one or more replica tables.

Q. How can I set up single master cross-region replication for a table?

You can create cross-region replicas using the [DynamoDB Cross-region Replication library](http://github.com/awslabs/dynamodb-cross-region-library).

Q: How do I know when the bootstrapping is complete?

On the replication management application, the state of the replication changes from Bootstrapping to Active.

Q: Can I have multiple replicas for a single master table?

Yes, there are no limits on the number of replicas tables from a single master table. A DynamoDB Streams reader is created for each replica table and copies data from the master table, keeping the replicas in sync.

Q: How much does it cost to set up cross-region replication for a table?

[DynamoDB cross-region replication](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Streams.CrossRegionRepl.html) is enabled using the [DynamoDB Cross-region Replication Library](https://github.com/awslabs/dynamodb-cross-region-library). While there is no additional charge for the cross-region replication library, you pay the usual prices for the following resources used by the process. You will be billed for:

* Provisioned throughput (Writes and Reads) and storage for the replica tables.
* Data Transfer across regions.
* Reading data from DynamoDB Streams to keep the tables in sync.
* The EC2 instances provisioned to host the replication process. The cost of the instances will depend on the instance type you choose and the region hosting the instances.

Q: In which region does the Amazon EC2 instance hosting the cross-region replication run?

The cross-region replication application is hosted in an Amazon EC2 instance in the same region where the cross-region replication application was originally launched. You will be charged the instance price in this region.

Q: Does the Amazon EC2 instance Auto Scale as the size and throughput of the master and replica tables change?

Currently, we will not auto scale the EC2 instance. You will need to pick the instance size when configuring DynamoDB Cross-region Replication.

Q: What happens if the Amazon EC2 instance managing the replication fails?

The Amazon EC2 instance runs behind an auto scaling group, which means the application will automatically fail over to another instance. The application underneath uses the Kinesis Client Library (KCL), which checkpoints the copy. In case of an instance failure, the application knows to find the checkpoint and resume from there.

Q: Can I keep using my DynamoDB table while a Read Replica is being created?

Yes, creating a replica is an online operation. Your table will remain available for reads and writes while the read replica is being created. The bootstrapping uses the Scan operation to copy from the source table. We recommend that the table is provisioned with sufficient read capacity units to support the Scan operation.

Q: How long does it take to create a replica?

The time to initially copy the master table to the replica table depends on the size of the master table, the provisioned capacity of the master table and replica table. The time to propagate an item-level change on the master table to the replica table depends on the provisioned capacity on the master and replica tables, and the size of the Amazon EC2 instance running the replication application.

Q: If I change provisioned capacity on my master table, does the provisioned capacity on my replica table also update?

After the replication has been created, any changes to the provisioned capacity on the master table will not result in an update in throughput capacity on the replica table.

Q: Will my replica tables have the same indexes as the master table?

If you choose to create the replica table from the replication application, the secondary indexes on the master table will NOT be automatically created on the replica table. The replication application will not propagate changes made on secondary indices on the master table to replica tables. You will have to add/update/delete indexes on each of the replica tables through the AWS Management Console as you would with regular DynamoDB tables.

Q: Will my replica have the same provisioned throughput capacity as the master table?

When creating the replica table, we recommend that you provision at least the same write capacity as the master table to ensure that it has enough capacity to handle all incoming writes. You can set the provisioned read capacity of your replica table at whatever level is appropriate for your application.

Q: What is the consistency model for replicated tables?

Replicas are updated asynchronously. DynamoDB will acknowledge a write operation as successful once it has been accepted by the master table. The write will then be propagated to each replica. This means that there will be a slight delay before a write has been propagated to all replica tables.

Q: Are there CloudWatch metrics for cross-region replication?

CloudWatch metrics are available for every replication configuration. You can see the metric by selecting the replication group and navigating to the Monitoring tab. Metrics on throughput and number of record processed are available, and you can monitor for any discrepancies in the throughput of the master and replica tables.

Q: Can I have a replica in the same region as the master table?

Yes, as long as the replica table and the master table have different names, both tables can exist in the same region.

Q: Can I add or delete a replica after creating a replication group?

Yes, you can add or delete a replica from that replication group at any time.

Q: Can I delete a replica group after it is created ?

Yes, deleting the replication group will delete the EC2 instance for the group. However, you will have to delete the DynamoDB metadata table.

## **DynamoDB Triggers**

Q. What is DynamoDB Triggers?

DynamoDB Triggers is a feature which allows you to execute custom actions based on item-level updates on a DynamoDB table. You can specify the custom action in code.

Q. What can I do with DynamoDB Triggers?

There are several application scenarios where DynamoDB Triggers can be useful. Some use cases include sending notifications, updating an aggregate table, and connecting DynamoDB tables to other data sources.

Q. How does DynamoDB Triggers work?

The custom logic for a DynamoDB trigger is stored in an AWS Lambda function as code. To create a trigger for a given table, you can associate an AWS Lambda function to the stream (via DynamoDB Streams) on a DynamoDB table. When the table is updated, the updates are published to DynamoDB Streams. In turn, AWS Lambda reads the updates from the associated stream and executes the code in the function.

Q: What does it cost to use DynamoDB Triggers?

With DynamoDB Triggers, you only pay for the number of requests for your AWS Lambda function and the amount of time it takes for your AWS Lambda function to execute. Learn more about AWS Lambda pricing [here](http://aws.amazon.com/lambda/pricing). You are not charged for the reads that your AWS Lambda function makes to the stream (via DynamoDB Streams) associated with the table.

Q. Is there a limit to the number of triggers for a table?

There is no limit on the number of triggers for a table.

Q. What languages does DynamoDB Triggers support?

Currently, DynamoDB Triggers supports Javascript, Java, and Python for trigger functions.

Q. Is there API support for creating, editing or deleting DynamoDB triggers?

No, currently there are no native APIs to create, edit, or delete DynamoDB triggers. You have to use the AWS Lambda console to create an AWS Lambda function and associate it with a stream in DynamoDB Streams. For more information, see the [AWS Lambda FAQ page](http://aws.amazon.com/lambda/faqs).

Q. How do I create a DynamoDB trigger?

You can create a trigger by creating an AWS Lambda function and associating the event-source for the function to a stream in DynamoDB Streams. For more information, see the AWS Lambda FAQ page.

Q. How do I delete a DynamoDB trigger?

You can delete a trigger by deleting the associated AWS Lambda function. You can delete an AWS Lambda function from the AWS Lambda console or throughput an AWS Lambda API call. For more information, see the [AWS Lambda FAQ](http://aws.amazon.com/lambda/faqs) and [documentation page](http://docs.aws.amazon.com/lambda/latest/dg/API_DeleteFunction.html).

Q. I have an existing AWS Lambda function, how do I create a DynamoDB trigger using this function?

You can change the event source for the AWS Lambda function to point to a stream in DynamoDB Streams. You can do this from the DynamoDB console. In the table for which the stream is enabled, choose the stream, choose the Associate Lambda Function button, and then choose the function that you want to use for the DynamoDB trigger from the list of Lambda functions.

Q. In what regions is DynamoDB Triggers available?

DynamoDB Triggers is available in all AWS regions where AWS Lambda and DynamoDB are available.

## **DynamoDB Streams**

Q: What is DynamoDB Streams?

DynamoDB Streams provides a time-ordered sequence of item-level changes made to data in a table in the last 24 hours. You can access a stream with a simple API call and use it to keep other data stores up-to-date with the latest changes to DynamoDB or to take actions based on the changes made to your table.

Q: What are the benefits of DynamoDB Streams?

Using the DynamoDB Streams APIs, developers can consume updates and receive the item-level data before and after items are changed. This can be used to build creative extensions to your applications built on top of DynamoDB. For example, a developer building a global multi-player game using DynamoDB can use the DynamoDB Streams APIs to build a multi-master topology and keep the masters in sync by consuming the DynamoDB Streams for each master and replaying the updates in the remote masters. As another example, developers can use the DynamoDB Streams APIs to build mobile applications that automatically notify the mobile devices of all friends in a circle as soon as a user uploads a new selfie. Developers could also use DynamoDB Streams to keep data warehousing tools, such as Amazon Redshift, in sync with all changes to their DynamoDB table to enable real-time analytics. DynamoDB also integrates with Elasticsearch using the Amazon DynamoDB Logstash Plugin, thus enabling developers to add free-text search for DynamoDB content.

You can read more about DynamoDB Streams in our [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Streams.html).

Q: How long are changes to my DynamoDB table available via DynamoDB Streams?

DynamoDB Streams keep records of all changes to a table for 24 hours. After that, they will be erased.

Q: How do I enable DynamoDB Streams?

DynamoDB Streams have to be enabled on a per-table basis. To enable DynamoDB Streams for an existing DynamoDB table, select the table through the AWS Management Console, choose the Overview tab, click the Manage Stream button, choose a view type, and then click Enable.

For more information, see our [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Streams.html).

Q: How do I verify that DynamoDB Streams has been enabled?

After enabling DynamoDB Streams, you can see the stream in the AWS Management Console. Select your table, and then choose the Overview tab. Under Stream details, verify Stream enabled is set to Yes.

Q: How can I access DynamoDB Streams?

You can access a stream available through DynamoDB Streams with a simple API call using the DynamoDB SDK or using the Kinesis Client Library (KCL). KCL helps you consume and process the data from a stream and also helps you manage tasks such as load balancing across multiple readers, responding to instance failures, and checkpointing processed records.

For more information about accessing DynamoDB Streams, see our [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Streams.html).

Q: Does DynamoDB Streams display all updates made to my DynamoDB table in order?

Changes made to any individual item will appear in the correct order. Changes made to different items may appear in DynamoDB Streams in a different order than they were received.

For example, suppose that you have a DynamoDB table tracking high scores for a game and that each item in the table represents an individual player. If you make the following three updates in this order:

* Update 1: Change Player 1’s high score to 100 points
* Update 2: Change Player 2’s high score to 50 points
* Update 3: Change Player 1’s high score to 125 points

Update 1 and Update 3 both changed the same item (Player 1), so DynamoDB Streams will show you that Update 3 came after Update 1. This allows you to retrieve the most up-to-date high score for each player. The stream might not show that all three updates were made in the same order (i.e., that Update 2 happened after Update 1 and before Update 3), but updates to each individual player’s record will be in the right order.

Q: Do I need to manage the capacity of a stream in DynamoDB Streams?

No, capacity for your stream is managed automatically in DynamoDB Streams. If you significantly increase the traffic to your DynamoDB table, DynamoDB will automatically adjust the capacity of the stream to allow it to continue to accept all updates.

Q: At what rate can I read from DynamoDB Streams?

You can read updates from your stream in DynamoDB Streams at up to twice the rate of the provisioned write capacity of your DynamoDB table. For example, if you have provisioned enough capacity to update 1,000 items per second in your DynamoDB table, you could read up to 2,000 updates per second from your stream.

Q: If I delete my DynamoDB table, does the stream also get deleted in DynamoDB Streams?

No, not immediately. The stream will persist in DynamoDB Streams for 24 hours to give you a chance to read the last updates that were made to your table. After 24 hours, the stream will be deleted automatically from DynamoDB Streams.

Q: What happens if I turn off DynamoDB Streams for my table?

If you turn off DynamoDB Streams, the stream will persist for 24 hours but will not be updated with any additional changes made to your DynamoDB table.

Q: What happens if I turn off DynamoDB Streams and then turn it back on?

When you turn off DynamoDB Streams, the stream will persist for 24 hours but will not be updated with any additional changes made to your DynamoDB table. If you turn DynamoDB Streams back on, this will create a new stream in DynamoDB Streams that contains the changes made to your DynamoDB table starting from the time that the new stream was created.

Q: Will there be duplicates or gaps in DynamoDB Streams?

No, DynamoDB Streams is designed so that every update made to your table will be represented exactly once in the stream.

Q: What information is included in DynamoDB Streams?

A DynamoDB stream contains information about both the previous value and the changed value of the item. The stream also includes the change type (INSERT, REMOVE, and MODIFY) and the primary key for the item that changed.

Q: How do I choose what information is included in DynamoDB Streams?

For new tables, use the CreateTable API call and specify the ViewType parameter to choose what information you want to include in the stream.  
For an existing table, use the UpdateTable API call and specify the ViewType parameter to choose what information to include in the stream.

The ViewType parameter takes the following values:

*ViewType: {  
                    { KEYS\_ONLY,   
                      NEW\_IMAGE,   
                      OLD\_IMAGE,   
                      NEW\_AND\_OLD\_IMAGES}  
                }*

The values have the following meaning: KEYS\_ONLY: Only the name of the key of items that changed are included in the stream.

* NEW\_IMAGE: The name of the key and the item after the update (new item) are included in the stream.
* OLD\_IMAGE: The name of the key and the item before the update (old item) are included in the stream.
* NEW\_AND\_OLD\_IMAGES: The name of the key, the item before (old item) and after (new item) the update are included in the stream.

Q: Can I use my Kinesis Client Library to access DynamoDB Streams?

Yes, developers who are familiar with Kinesis APIs will be able to consume DynamoDB Streams easily. You can use the DynamoDB Streams Adapter, which implements the Amazon Kinesis interface, to allow your application to use the Amazon Kinesis Client Libraries (KCL) to access DynamoDB Streams. For more information about using the KCL to access DynamoDB Streams, please see our [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Streams.html).

Q: Can I change what type of information is included in DynamoDB Streams?

If you want to change the type of information stored in a stream after it has been created, you must disable the stream and create a new one using the UpdateTable API.

Q: When I make a change to my DynamoDB table, how quickly will that change show up in a DynamoDB stream?

Changes are typically reflected in a DynamoDB stream in less than one second.

Q: If I delete an item, will that change be included in DynamoDB Streams?

Yes, each update in a DynamoDB stream will include a parameter that specifies whether the update was a deletion, insertion of a new item, or a modification to an existing item. For more information on the type of update, see our [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Streams.html).

Q: After I turn on DynamoDB Streams for my table, when can I start reading from the stream?

You can use the DescribeStream API to get the current status of the stream. Once the status changes to ENABLED, all updates to your table will be represented in the stream.

You can start reading from the stream as soon as you start creating it, but the stream may not include all updates to the table until the status changes to ENABLED.

Q: What is the Amazon DynamoDB Logstash Plugin for Elasticsearch?

Elasticsearch is a popular open source search and analytics engine designed to simplify real-time search and big data analytics. Logstash is an open source data pipeline that works together with Elasticsearch to help you process logs and other event data. The Amazon DynamoDB Logstash Plugin make is easy to integrate DynamoDB tables with Elasticsearch clusters.

Q: How much does the Amazon DynamoDB Logstash Plugin cost?

The Amazon DynamoDB Logstash Plugin is free to download and use.

Q: How do I download and install the Amazon DynamoDB Logstash Plugin?

The Amazon DynamoDB Logstash Plugin is available on [GitHub](https://github.com/awslabs/logstash-input-dynamodb). Read our [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Tools.DynamoDBLogstash.html) page to learn more about installing and running the plugin.

## **DynamoDB Storage Backend for Titan**

Q: What is the DynamoDB Storage Backend for Titan?

The DynamoDB Storage Backend for Titan is a plug-in that allows you to use DynamoDB as the underlying storage layer for Titan graph database. It is a client side solution that implements index free adjacency for fast graph traversals on top of DynamoDB.

Q: What is a graph database?

A graph database is a store of vertices and directed edges that connect those vertices. Both vertices and edges can have properties stored as key-value pairs.

A graph database uses adjacency lists for storing edges to allow simple traversal. A graph in a graph database can be traversed along specific edge types, or across the entire graph. Graph databases can represent how entities relate by using actions, ownership, parentage, and so on.

Q: What applications are well suited to graph databases?

Whenever connections or relationships between entities are at the core of the data you are trying to model, a graph database is a natural choice. Therefore, graph databases are useful for modeling and querying social networks, business relationships, dependencies, shipping movements, and more.

Q: How do I get started using the DynamoDB Storage Backend for Titan?

The easiest way to get started is to launch an EC2 instance running Gremlin Server with the DynamoDB Storage Backend for Titan, using the CloudFormation templates referred to in this [documentation page](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Tools.TitanDB.GremlinServerEC2.html). You can also clone the project from the GitHub repository and start by following the Marvel and Graph-Of-The-Gods tutorials on your own computer by following the instructions in the [documentation here](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Tools.TitanDB.DownloadingAndRunning.html). When you’re ready to expand your testing or run in production, you can switch the backend to use the DynamoDB service. Please see the [AWS documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Tools.TitanDB.html) for further guidance.

Q: How does the DynamoDB Storage Backend differ from other Titan storage backends?

DynamoDB is a managed service, thus using it as the storage backend for Titan enables you to run graph workloads without having to manage your own cluster for graph storage.

Q: Is the DynamoDB Storage Backend for Titan a fully managed service?

No. The DynamoDB storage backend for Titan manages the storage layer for your Titan workload. However, the plugin does not do provisioning and managing of the client side. For simple provisioning of Titan we have developed a CloudFormation template that sets up DynamoDB Storage Backend for Titan with Gremlin Server; see the instructions available [here](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Tools.TitanDB.GremlinServerEC2.html).

Q: How much does using the DynamoDB Storage Backend for Titan cost?

You are charged the regular DynamoDB throughput and storage costs. There is no additional cost for using DynamoDB as the storage backend for a Titan graph workload.

Q: Does DynamoDB backend provide full compatibility with the Titan feature set on other backends?

A table comparing feature sets of different Titan storage backends is available in the [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Tools.TitanDB.html#Table).

Q: Which versions of Titan does the plugin support?

We have released DynamoDB storage backend plugins for Titan versions 0.5.4 and 1.0.0.

Q: I use Titan with a different backend today. Can I migrate to DynamoDB?

Absolutely. The DynamoDB Storage Backend for Titan implements the Titan KCV Store interface so you can switch from a different storage backend to DynamoDB with minimal changes to your application. For full comparison of storage backends for Titan please see our [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Tools.TitanDB.html#Table).

Q: I use Titan with a different backend today. How do I migrate to DynamoDB?

You can use [bulk loading](http://s3.thinkaurelius.com/docs/titan/1.0.0/bulk-loading.html) to copy your graph from one storage backend to the DynamoDB Storage Backend for Titan.

Q: How do I connect my Titan instance to DynamoDB via the plugin?

If you create a graph and Gremlin server instance with the DynamoDB Storage Backend for Titan installed, all you need to do to connect to DynamoDB is provide a principal/credential set to the [default AWS credential provider chain](http://docs.aws.amazon.com/AWSSdkDocsJava/latest/DeveloperGuide/credentials.html#credentials-default). This can be done with an EC2 instance profile, environment variables, or the credentials file in your home folder. Finally, you need to choose a DynamoDB endpoint to connect to.

Q: How durable is my data when using the DynamoDB Storage Backend for Titan?

When using the DynamoDB Storage Backend for Titan, your data enjoys the strong protection of DynamoDB, which runs across Amazon’s proven, high-availability data centers. The service replicates data across three facilities in an AWS Region to provide fault tolerance in the event of a server failure or Availability Zone outage.

Q: How secure is the DynamoDB Storage Backend for Titan?

The DynamoDB Storage Backend for Titan stores graph data in multiple DynamoDB tables, thus is enjoys the same high security available on all DynamoDB workloads. Fine-Grained Access Control, IAM roles, and AWS principal/credential sets control access to DynamoDB tables and items in DynamoDB tables.

Q: How does the DynamoDB Storage Backend for Titan scale?

The DynamoDB Storage Backend for Titan scales just like any other workload of DynamoDB. You can choose to increase or decrease the required throughput at any time.

Q: How many vertices and edges can my graph contain?

You are limited by [Titan’s limits](http://s3.thinkaurelius.com/docs/titan/1.0.0/limitations.html) for (2^60) for the maximum number of edges and half as many vertices in a graph, as long as you use the multiple-item model for edgestore. If you use the single-item model, the number of edges that you can store at a particular out-vertex key is limited by DynamoDB’s maximum item size, currently 400kb.

Q: How large can my vertex and edge properties get?

The sum of all edge properties in the multiple-item model cannot exceed 400kb, the maximum item size. In the multiple item model, each vertex property can be up to 400kb. In the single-item model, the total item size (including vertex properties, edges and edge properties) can’t exceed 400kb.

Q: How many data models are there? What are the differences?

There are two different storage models for the DynamoDB Storage Backend for Titan – single item model and multiple item model. In the single item storage model, vertices, vertex properties, and edges are stored in one item. In the multiple item data model, vertices, vertex properties and edges are stored in different items. In both cases, edge properties are stored in the same items as the edges they correspond to.

Q: Which data model should I use?

In general, we recommend you use the multiple-item data model for the edgestore and graphindex tables. Otherwise, you either limit the number of edges/vertex-properties you can store for one out-vertex, or you limit the number of entities that can be indexed at a particular property name-value pair in graph index. In general, you can use the single-item data model for the other 4 KCV stores in Titan versions 0.5.4 and 1.0.0 because the items stored in them are usually less than 400KB each. For full list of tables that the Titan plugin creates on DynamoDB please see [here](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Tools.TitanDB.BestPractices.html).

Q: Do I have to create a schema for Titan graph databases?

Titan supports automatic type creation, so new edge/vertex properties and labels will get registered on the fly (see [here](http://s3.thinkaurelius.com/docs/titan/1.0.0/schema.html) for details) with the first use. The Gremlin Structure (Edge labels=MULTI, Vertex properties=SINGLE) is used [by default](http://s3.thinkaurelius.com/docs/titan/1.0.0/titan-config-ref.html).

Q: Can I change the schema of a Titan graph database?

Yes, however, you cannot change the schema of existing vertex/edge properties and labels. For details please see [here](http://s3.thinkaurelius.com/docs/titan/1.0.0/limitations.html).

Q: How does the DynamoDB Storage Backend for Titan deal with supernodes?

DynamoDB deals with supernodes via vertex label partitioning. If you define a vertex label as partitioned in the management system upon creation, you can key different subsets of the edges and vertex properties going out of a vertex at different partition keys of the partition-sort key space in the edgestore table. This usually results in the virtual vertex label partitions being stored in different physical DynamoDB partitions, as long as your edgestore has more than one physical partition. To estimate the number of physical partitions backing your edgestore table, please see guidance in the [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/GuidelinesForTables.html#GuidelinesForTables.Partitions).

Q: Does the DynamoDB Storage Backend for Titan support batch graph operations?

Yes, the DynamoDB Storage Backend for Titan supports batch graph with the Blueprints BatchGraph implementation and through Titan’s bulk loading configuration options.

Q: Does the DynamoDB Storage Backend for Titan support transactions?

The DynamoDB Storage Backend for Titan supports optimistic locking. That means that the DynamoDB Storage Backend for Titan can condition writes of individual Key-Column pairs (in the multiple item model) or individual Keys (in the single item model) on the existing value of said Key-Column pair or Key.

Q: Can I have a Titan instance in one region and access DynamoDB in another?

Accessing a DynamoDB endpoint in another region than the EC2 Titan instance is possible but not recommended. When running a Gremlin Server out of EC2, we recommend connecting to the DynamoDB endpoint in your EC2 instance’s region, to reduce the latency impact of cross-region requests. We also recommend running the EC2 instance in a VPC to improve network performance. The [CloudFormation template](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Tools.TitanDB.GremlinServerEC2.html) performs this entire configuration for you.

Q: Can I use this plugin with other DynamoDB features such as update streams and cross-region replication?

You can use Cross-Region Replication with the DynamoDB Streams feature to create read-only replicas of your graph tables in other regions.

## **DynamoDB CloudWatch Metrics**

Q: Does Amazon DynamoDB report CloudWatch metrics?  
  
Yes, Amazon DynamoDB reports several table-level metrics on CloudWatch. You can make operational decisions about your Amazon DynamoDB tables and take specific actions, like setting up alarms, based on these metrics. For a full list of reported metrics, see the [*Monitoring DynamoDB with CloudWatch*](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/MonitoringDynamoDB.html) section of our documentation.  
  
Q: How can I see CloudWatch metrics for an Amazon DynamoDB table?  
  
On the Amazon DynamoDB console, select the table for which you wish to see CloudWatch metrics and then select the Metrics tab.  
  
Q: How often are metrics reported?  
  
Most CloudWatch metrics for Amazon DynamoDB are reported in 1-minute intervals while the rest of the metrics are reported in 5-minute intervals. For more details, see the [*Monitoring DynamoDB with CloudWatch*](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/MonitoringDynamoDB.html) section of our documentation.

## **Tagging for DynamoDB**

Q: What is a tag?

A tag is a label you assign to an AWS resource. Each tag consists of a key and a value, both of which you can define. AWS uses tags as a mechanism to organize your resource costs on your cost allocation report. For more about tagging, see the [*AWS Billing and Cost Management User Guide*](http://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/allocation-what.html).  
  
Q: What DynamoDB resources can I tag?

You can tag DynamoDB tables. Local Secondary Indexes and Global Secondary Indexes associated with the tagged tables are automatically tagged with the same tags. Costs for Local Secondary Indexes and Global Secondary Indexes will show up under the tags used for the corresponding DynamoDB table.  
  
Q: Why should I use Tagging for DynamoDB?

You can use Tagging for DynamoDB for cost allocation. Using tags for cost allocation enables you to label your DynamoDB resources so that you can easily track their costs against projects or other criteria to reflect your own cost structure.

Q: How can I use tags for cost allocation?

You can use cost allocation tags to categorize and track your AWS costs. AWS Cost Explorer and detailed billing reports support the ability to break down AWS costs by tag. Typically, customers use business tags such as cost center/business unit, customer, or project to associate AWS costs with traditional cost-allocation dimensions. However, a cost allocation report can include any tag. This enables you to easily associate costs with technical or security dimensions, such as specific applications, environments, or compliance programs.

Q: How can I see costs allocated to my AWS tagged resources?

You can see costs allocated to your AWS tagged resources through either Cost Explorer or your cost allocation report.

[Cost Explorer](http://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/cost-explorer-what-is.html) is a free AWS tool that you can use to view your costs for up to the last 13 months, and forecast how much you are likely to spend for the next three months. You can see your costs for specific tags by filtering by “Tag” and then choose the tag key and value (choose “No tag” if no tag value is specified).

The cost allocation report includes all of your AWS costs for each billing period. The report includes both tagged and untagged resources, so you can clearly organize the charges for resources. For example, if you tag resources with an application name, you can track the total cost of a single application that runs on those resources. More information on cost allocation can be found in [*AWS Billing and Cost Management User Guide*](http://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/cost-alloc-tags.html).

Q: Can DynamoDB Streams usage be tagged?

No, DynamoDB Streams usage cannot be tagged at present.

Q: Will Reserved Capacity usage show up under my table tags in my bill?

Yes, DynamoDB Reserved Capacity charges per table will show up under relevant tags. Please note that Reserved Capacity is applied to DynamoDB usage on a first come, first serve basis, and across all linked AWS accounts. This means that even if your DynamoDB usage across tables and indexes is similar from month to month, you may see differences in your cost allocation reports per tag since Reserved Capacity will be distributed based on which DynamoDB resources are metered first.

Q: Will data usage charges show up under my table tags in my bill?

No, DynamoDB data usage charges are not tagged. This is because data usage is billed at an account level and not at table level.

Q: Do my tags require a value attribute?

No, tag values can be null.

Q: Are tags case sensitive?

Yes, tag keys and values are case sensitive.

Q: How many tags can I add to single DynamoDB table?

You can add up to 50 tags to a single DynamoDB table. Tags with the prefix “aws:” cannot be manually created and do not count against your tags per resource limit.

Q: Can I apply tags retroactively to my DynamoDB tables?

No, tags begin to organize and track data on the day you apply them. If you create a table on January 1st but don’t designate a tag for it until February 1st, then all of that table’s usage for January will remain untagged.

Q: If I remove a tag from my DynamoDB table before the end of the month, will that tag still show up in my bill?

Yes, if you build a report of your tracked spending for a specific time period, your cost reports will show the costs of the resources that were tagged during that timeframe.

Q. What happens to existing tags when a DynamoDB table is deleted?

When a DynamoDB table is deleted, its tags are automatically removed.

Q. What happens if I add a tag with a key that is same as one for an existing tag?

Each DynamoDB table can only have up to one tag with the same key. If you add a tag with the same key as an existing tag, the existing tag is updated with the new value.

## **DynamoDB Time-to-Live (TTL)**

Q: What is DynamoDB Time-to-Live (TTL)?

DynamoDB Time-to-Live (TTL) is a mechanism that lets you set a specific timestamp to delete expired items from your tables. Once the timestamp expires, the corresponding item is marked as expired and is subsequently deleted from the table. By using this functionality, you do not have to track expired data and delete it manually. TTL can help you reduce storage usage and reduce the cost of storing data that is no longer relevant.  
  
Q: Why do I need to use TTL?

There are two main scenarios where TTL can come in handy:

* Deleting old data that is no longer relevant – data like event logs, usage history, session data, etc. when collected can get bloated over time and the old data though stored in the system may not be relevant any more. In such situations, you are better off clearing these stale records from the system and saving the money used for storing it.
* Sometimes you may want data to be kept in DynamoDB for a specified time period in order to comply with your data retention and management policies. You might want to eventually delete this data once the obligated duration expires. Please do know however that TTL works on a best effort basis to ensure there is throughput available for other critical operations. DynamoDB will aim to delete expired items within a two-day period. The actual time taken may be longer based on the size of the data.

Q: How does DynamoDB TTL work?

To enable TTL for a table, first ensure that there is an attribute that can store the expiration timestamp for each item in the table. This timestamp needs to be in the [epoch time format](https://en.wikipedia.org/wiki/Unix_time). This helps avoid time zone discrepancies between clients and servers.

DynamoDB runs a background scanner that monitors all the items. If the timestamp has expired, the process will mark the item as expired and queue it for subsequent deletion.

Note: TTL requires a numeric DynamoDB table attribute populated with an epoch timestamp to specify the expiration criterion for the data. You should be careful when setting a value for the TTL attribute since a wrong value could cause premature item deletion.

Q: How do I specify TTL?

To specify TTL, first enable the TTL setting on the table and specify the attribute to be used as the TTL value. As you add items to the table, you can specify a TTL attribute if you would like DynamoDB to automatically delete it after its expiration. This value is the expiry time, specified in [epoch time format](https://en.wikipedia.org/wiki/Unix_time). DynamoDB takes care of the rest. TTL can be specified from the console from the overview tab for the table. Alternatively, developers can invoke the TTL API to configure TTL on the table. See our [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/TTL.html) and our [API guide](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/Welcome.html).

Q: Can I set TTL on existing tables?

Yes. If a table is already created and has an attribute that can be used as TTL for its items, then you only need to enable TTL for the table and designate the appropriate attribute for TTL. If the table does not have an attribute that can be used for TTL, you will have to create such an attribute and update the items with values for TTL.

Q: Can I delete an entire table by setting TTL on the whole table?

No. While you need to define an attribute to be used for TTL at the table level, the granularity for deleting data is at the item level. That is, each item in a table that needs to be deleted after expiry will need to have a value defined for the TTL attribute. There is no option to automatically delete the entire table.

Q: Can I set TTL only for a subset of items in the table?

Yes. TTL takes affect only for those items that have a defined value in the TTL attribute. Other items in the table remain unaffected.

Q: What is the format for specifying TTL?

The TTL value should use the [epoch time format](https://en.wikipedia.org/wiki/Unix_time), which is number of seconds since January 1, 1970 UTC. If the value specified in the TTL attribute for an item is not in the right format, the value is ignored and the item won’t be deleted.

Q: How can I read the TTL value for items in my table?

The TTL value is just like any attribute on an item. It can be read the same way as any other attribute. In order to make it easier to visually confirm TTL values, the DynamoDB Console allows you to hover over a TTL attribute to see its value in human-readable local and UTC time.

Q: Can I create an index based on the TTL values assigned to items in a table?

Yes. TTL behaves like any other item attribute. You can create indexes the same as with other item attributes.

Q: Can the TTL attribute be projected to an index?

Yes. TTL attribute can be projected onto an index just like any other attribute.

Q: Can I edit the TTL attribute value once it has been set for an item?

Yes. You can modify the TTL attribute value just as you modify any other attribute on an item.

Q: Can I change the TTL attribute for a table?

Yes. If a table already has TTL enabled and you want to specify a different TTL attribute, then you need to disable TTL for the table first, then you can re-enable TTL on the table with a new TTL attribute. Note that Disabling TTL can take up to one hour to apply across all partitions, and you will not be able to re-enable TTL until this action is complete.

Q: Can I use AWS Management Console to view and edit the TTL values?

Yes. The AWS Management Console allows you to easily view, set or update the TTL value.

Q: Can I set an attribute within a JSON document to be the TTL attribute?

No. We currently do not support specifying an attribute in a JSON document as the TTL attribute. To set TTL, you must explicitly add the TTL attribute to each item.

Q: Can I set TTL for a specific element in a JSON Document?

No. TTL values can only be set for the whole document. We do not support deleting a specific item in a JSON document once it expires.

Q: What if I need to remove the TTL on specific items?

Removing TTL is as simple as removing the value assigned to the TTL attribute or removing the attribute itself for an item.

Q: What if I set the TTL timestamp value to sometime in the past?

Updating items with an older TTL values is allowed. Whenever the background process checks for expired items, it will find, mark and subsequently delete the item. However, if the value in the TTL attribute contains an [epoch value](https://en.wikipedia.org/wiki/Epoch_(reference_date)) for a timestamp that is over 5 years in the past, DynamoDB will ignore the timestamp and not delete the item. This is done to mitigate accidental deletion of items when really low values are stored in the TTL attribute.

Q: What is the delay between the TTL expiry on an item and the actual deletion of that item?

TTL scans and deletes expired items using background throughput available in the system. As a result, the expired item may not be deleted from the table immediately. DynamoDB will aim to delete expired items within a two-day window on a best-effort basis, to ensure availability of system background throughput for other data operations. The exact duration within which an item truly gets deleted after expiration will be specific to the nature of the workload and the size of the table.

Q: What happens if I try to query or scan for items that have been expired by TTL?

Given that there might be a delay between when an item expires and when it actually gets deleted by the background process, if you try to read items that have expired but haven’t yet been deleted, the returned result will include the expired items. You can filter these items out based on the TTL value if the intent is to not show expired items.

Q: What happens to the data in my Local Secondary Index (LSI) if it has expired?

The impact is the same as any delete operation. The local secondary index is stored in the same partition as the item itself. Hence if an item is deleted it immediately gets removed from the Local Secondary Index.

Q: What happens to the data in my Global Secondary Index (GSI) if it has expired?

The impact is the same as any delete operation. A Global Secondary Index (GSI) is eventually consistent and so while the original item that expired will be deleted it may take some time for the GSI to get updated.

Q: How does TTL work with DynamoDB Streams?

The expiry of data in a table on account of the TTL value triggering a purge is recorded as a delete operation. Therefore, the Streams will also have the delete operation recorded in it. The delete record will have an additional qualifier so that you can distinguish between your deletes and deletes happening due to TTL. The stream entry will be written at the point of deletion, not the TTL expiration time, to reflect the actual time at which the record was deleted. See our [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/TTL.html) and our [API guide](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/Welcome.html).

Q: When should I use the delete operation vs TTL?

TTL is ideal for removing expired records from a table. However, this is intended as a best-effort operation to help you remove unwanted data and does not provide a guarantee on the deletion timeframe. As a result, if data in your table needs to be deleted within a specific time period (often immediately), we recommend using the delete command.

Q: Can I control who has access to set or update the TTL value?

Yes. The TTL attribute is just like any other attribute on a table. You have the ability to control access at an attribute level on a table. The TTL attribute will follow the regular access controls specified for the table.

Q: Is there a way to retrieve the data that has been deleted after TTL expiry?

No. Expired items are not backed up before deletion. You can leverage the DynamoDB Streams to keep track of the changes on a table and restore values if needed. The delete record is available in Streams for 24 hours since the time it is deleted.

Q: How can I know whether TTL is enabled on a table?

You can get the status of TTL at any time by invoking the DescribeTable API or viewing the table details in the DynamoDB console. See our [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/TTL.html) and our [API guide](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/Welcome.html).

Q: How do I track the items deleted by TTL?

If you have DynamoDB streams enabled, all TTL deletes will show up in the DynamoDB Streams and will be designated as a system delete in order to differentiate it from an explicit delete done by you. You can read the items from the streams and process them as needed. They can also write a Lambda function to archive the item separately. See our [documentation](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/TTL.html) and our [API guide](http://docs.aws.amazon.com/amazondynamodb/latest/APIReference/Welcome.html).

Q: Do I have to pay a specific fee to enable the TTL feature for my data?

No. Enabling TTL requires no additional fees.

Q: How will enabling TTL affect my overall provisioned throughput usage?

The scan and delete operations needed for TTL are carried out by the system and does not count toward your provisioned throughput or usage.

Q: Will I have to pay for the scan operations to monitor TTL?

No. You are not charged for the internal scan operations to monitor TTL expiry for items. Also these operations will not affect your throughput usage for the table.

Q: Do expired items accrue storage costs till they are deleted?

Yes. After an item has expired it is added to the delete queue for subsequent deletion. However, until it has been deleted, it is just like any regular item that can be read or updated and will incur storage costs.

Q: If I query for an expired item, does it use up my read capacity?

Yes. This behavior is the same as when you query for an item that does not exist in the table.

## **DynamoDB Accelerator (DAX)**

Q. What is Amazon DynamoDB Accelerator (DAX)?

Amazon DynamoDB Accelerator (DAX) is a fully managed, highly available, in-memory [cache](https://aws.amazon.com/caching/) for [DynamoDB](https://aws.amazon.com/dynamodb/) that enables you to benefit from fast in-memory performance for demanding applications. DAX improves the performance of read-intensive DynamoDB workloads so repeat reads of cached data can be served immediately with extremely low latency, without needing to be re-queried from DynamoDB. DAX will automatically retrieve data from DynamoDB tables upon a cache miss. Writes are designated as write-through (data is written to DynamoDB first and then updated in the DAX cache).

Just like DynamoDB, DAX is fault-tolerant and scalable. A DAX cluster has a primary node and zero or more read-replica nodes. Upon a failure for a primary node, DAX will automatically fail over and elect a new primary. For scaling, you may add or remove read replicas.

To get started, create a DAX cluster, download the DAX SDK for Java or Node.js (compatible with the DynamoDB APIs), re-build your application to use the DAX client as opposed to the DynamoDB client, and finally point the DAX client to the DAX cluster endpoint. You do not need to implement any additional caching logic into your application as DAX client implements the same API calls as DynamoDB.

Q. What does "DynamoDB-compatible" mean?

It means that most of the code, applications, and tools you already use today with DynamoDB can be used with DAX with little or no change. The DAX engine is designed to support the DynamoDB APIs for reading and modifying data in DynamoDB. Operations for table management such as CreateTable/DescribeTable/UpdateTable/DeleteTable are not supported.

Q. What is in-memory caching, and how does it help my application?

Caching improves application performance by storing critical pieces of data in memory for low-latency and high throughput access. In the case of DAX, the results of DynamoDB operations are cached. When an application requests data that is stored in the cache, DAX can serve that data immediately without needing to run a query against the regular DynamoDB tables. Data is aged or evicted from DAX by specifying a Time-to-Live (TTL) value for the data or, once all available memory is exhausted, items will be evicted based on the Least Recently Used (LRU) algorithm.

Q. What is the consistency model of DAX?

When reading data from DAX, users can specify whether they want the read to be eventually consistent or strongly consistent:

Eventually Consistent Reads (Default) – the eventual consistency option maximizes your read throughput and minimizes latency. On a cache hit, the DAX client will return the result directly from the cache. On a cache miss, DAX will query DynamoDB, update the cache, and return the result set. It should be noted that an eventually consistent read might not reflect the results of a recently completed write. If your application requires full consistency, then we suggest using strongly consistent reads.

Strongly Consistent Reads — in addition to eventual consistency, DAX also gives you the flexibility and control to request a strongly consistent read if your application, or an element of your application, requires it. A strongly consistent read is pass-through for DAX, does not cache the results in DAX, and returns a result that reflects all writes that received a successful response in DynamoDB prior to the read.

Q. What are the common use cases for DAX?  
  
DAX has a number of use cases that are not mutually exclusive:

Applications that require the fastest possible response times for reads. Some examples include real-time bidding, social gaming, and trading applications. DAX delivers fast, in-memory read performance for these use cases.

Applications that read a small number of items more frequently than others. For example, consider an e-commerce system that has a one-day sale on a popular product. During the sale, demand for that product (and its data in DynamoDB) would sharply increase, compared to all of the other products. To mitigate the impacts of a "hot" key and a non-uniform data distribution, you could offload the read activity to a DAX cache until the one-day sale is over.

Applications that are read-intensive, but are also cost-sensitive. With DynamoDB, you provision the number of reads per second that your application requires. If read activity increases, you can increase your table’s provisioned read throughput (at an additional cost). Alternatively, you can offload the activity from your application to a DAX cluster, and reduce the amount of read capacity units you'd need to purchase otherwise.

Applications that require repeated reads against a large set of data. Such an application could potentially divert database resources from other applications. For example, a long-running analysis of regional weather data could temporarily consume all of the read capacity in a DynamoDB table, which would negatively impact other applications that need to access the same data. With DAX, the weather analysis could be performed against cached data instead.

How It Works  
  
Q. What does DAX manage on my behalf?

DAX is a fully-managed cache for DynamoDB. It manages the work involved in setting up dedicated caching nodes, from provisioning the server resources to installing the DAX software. Once your DAX cache cluster is set up and running, the service automates common administrative tasks such as failure detection and recovery, and software patching. DAX provides detailed CloudWatch monitoring metrics associated with your cluster, enabling you to diagnose and react to issues quickly. Using these metrics, you can set up thresholds to receive CloudWatch alarms. DAX handles all of the data caching, retrieval, and eviction so your application does not have to. You can simply use the DynamoDB API to write and retrieve data, and DAX handles all of the caching logic behind the scenes to deliver improved performance.

Q. What kinds of data does DAX cache?

All read API calls will be cached by DAX, with strongly consistent requests being read directly from DynamoDB, while eventually consistent reads will be read from DAX if the item is available. Write API calls are write-through (synchronous write to DynamoDB which is updated in the cache upon a successful write).

The following API calls will result in examining the cache. Upon a hit, the item will be returned. Upon a miss, the request will pass through, and upon a successful retrieval the item will be cached and returned.  
  
• GetItem  
• BatchGetItem  
• Query  
• Scan

The following API calls are write-through operations.  
  
• BatchWriteItem  
• UpdateItem  
• DeleteItem  
• PutItem

Q. How does DAX handle data eviction?

DAX handles cache eviction in three different ways. First, it uses a Time-to-Live (TTL) value that denotes the absolute period of time that an item is available in the cache. Second, when the cache is full, a DAX cluster uses a Least Recently Used (LRU) algorithm to decide which items to evict. Third, with the write-through functionality, DAX evicts older values as new values are written through DAX. This helps keep the DAX item cache consistent with the underlying data store using a single API call.

Q. Does DAX work with DynamoDB GSIs and LSIs?

Just like DynamoDB tables, DAX will cache the result sets from both query and scan operations against both DynamoDB GSIs and LSIs.

Q. How does DAX handle Query and Scan result sets?

Within a DAX cluster, there are two different caches: 1) item cache and 2) query cache. The item cache manages GetItem, PutItem, and DeleteItem requests for individual key-value pairs. The query cache manages the result sets from Scan and Query requests. In this regard, the Scan/Query text is the “key” and the result set is the “value”. While both the item cache and the query cache are managed in the same cluster (and you can specify different TTL values for each cache), they do not overlap. For example, a scan of a table does not populate the item cache, but instead records an entry in the query cache that stores the result set of the scan.

Q. Does an update to the item cache either update or invalidate result sets in my query cache?

No. The best way to mitigate inconsistencies between result sets in the item cache and query cache is to set the TTL for the query cache to be of an acceptable period of time for which your application can handle such inconsistencies.

Q. Can I connect to my DAX cluster from outside of my VPC?

The only way to connect to your DAX cluster from outside of your VPC is through a VPN connection.

Q. When using DAX, what happens if my underlying DynamoDB tables are throttled?

If DAX is either reading or writing to a DynamoDB table and receives a throttling exception, DAX will return the exception back to the DAX client. Further, the DAX service does not attempt server-side retries.

Q. Does DAX support pre-warming of the cache?

DAX utilizes lazy-loading to populate the cache. What this means is that on the first read of an item, DAX will fetch the item from DynamoDB and then populate the cache. While DAX does not support cache pre-warming as a feature, the DAX cache can be pre-warmed for an application by running an external script/application that reads the desired data.

Q. How does DAX work with the DynamoDB TTL feature?

Both DynamoDB and DAX have the concept of a "TTL" (or Time to Live) feature. In the context of DynamoDB, TTL is a feature that enables customers to age out their data by tagging the data with a particular attribute and corresponding timestamp. For example, if customers wanted data to be deleted after the data has aged for one month, they would use the DynamoDB TTL feature to accomplish this task as opposed to managing the aging workflow themselves.

In the context of DAX, TTL specifies the duration of time in which an item in cache is valid. For instance, if a TTL is set for 5-minutes, once an item has been populated in cache it will continue to be valid and served from the cache until the 5-minute period has elapsed. Although not central to this conversation, TTL can be preempted by writes to the cache for the same item or if there is memory pressure on the DAX node and LRU evicts the items as it was the least recently used.

While TTL for DynamoDB and DAX will typically be operating in very different time scales (i.e., DAX TTL operating in the scope of minutes/hours and DynamoDB TTL operating in the scope of weeks/months/years), there is a potential when customers will need to be present of how these two features affect each other. For example, let's imagine a scenario in which the TTL value for DynamoDB is less than the TTL value for DAX. In this scenario, an item could conceivably be cached in DAX and subsequently deleted from DynamoDB via the DynamoDB TTL feature. The result would be an inconsistent cache. While we don’t expect this scenario to happen often as the time scales for the two features are typically order of magnitude apart, it is good to be aware of how the two features relate to each other.

Q. Does DAX support cross-region replication?

Currently DAX only supports DynamoDB tables in the same AWS region as the DAX cluster.

Q. Is DAX supported as a resource type in AWS CloudFormation?

Yes. You can create, update and delete DAX clusters, parameter groups, and subnet groups using AWS CloudFormation.

Getting Started  
  
Q. How do I get started with DAX?  
You can create a new DAX cluster through the AWS console or AWS SDK to obtain the DAX cluster endpoint. A DAX-compatible client will need to be downloaded and used in the application with the new DAX endpoint.

Q. How do I create a DAX Cluster?

You can create a DAX cluster using the AWS Console or the DAX CLI. DAX clusters range from a 13 GiB cache (dax.r3.large) to 216 GiB (dax.r3.8xlarge) in the R3 instance types and 15.25GiB cache (dax.r4.large) to 488 GiB (dax.r4.16xlarge) in the R4 instance types. With a few clicks in the AWS Console, or a single API call, you can add more replicas to your cluster (up to 10 replicas) for increased throughput.

The single node configuration enables you to get started with DAX quickly and cost-effectively and then scale out to a multi-node configuration as your needs grow. The multi-node configuration consists of a primary node that manages writes, and up to nine read replica nodes. The primary node is provisioned for you automatically.

Simply specify your preferred subnet groups/Availability Zones (optional), the number of nodes, node types, VPC subnet group, and other system settings. Once you've chosen your desired configuration, DAX will provision the required resources and set up your caching cluster specifically for DynamoDB.

Q. Does all my data need to fit in memory to use DAX?

No. DAX will utilize the available memory on the node. Using either TTL and/or LRU, items will be expunged to make space for new data when the memory space is exhausted.

Q. What languages does DAX support?

DAX provides DAX SDKs for Java and Node.js that you can download today. We are working on adding support for additional clients.

Q. Can I use DAX and DynamoDB at the same time?

Yes, you can access the DAX endpoint and DynamoDB at the same time through different clients. However, DAX will not be able to detect changes in data written directly to DynamoDB unless these changes are explicitly populated in to DAX through a read operation after the update was made directly to DynamoDB.

Q. Can I utilize multiple DAX clusters for the same DynamoDB table?

Yes, you can provision multiple DAX clusters for the same DynamoDB table. These clusters will provide different endpoints that can be used for different use cases, ensuring optimal caching for each scenario. Two DAX clusters will be independent of each other and will not share state or updates, so users are best served using these for completely different tables.

Q. How will I know what DAX node type I'll need for my workload?

Sizing of a DAX cluster is an iterative process. It is recommended to provision a three-node cluster (for high availability) with enough memory to fit the application's working set in memory. Based on the performance and throughput of the application, the utilization of the DAX cluster, and the cache hit/miss ratio you may need to scale your DAX cluster to achieve desired results.

Q. What kinds of EC2 instances can DAX run on?

Valid node types are as follows:

R3:

• dax.r3.large (13 GiB)  
• dax.r3.xlarge (26 GiB)  
• dax.r3.2xlarge (54 GiB)  
• dax.r3.4xlarge (108 GiB)  
• dax.r3.8xlarge (216 GiB)

R4:

• dax.r4.large (15.25 GiB)  
• dax.r4.xlarge (30.5 GiB)  
• dax.r4.2xlarge (61 GiB)  
• dax.r4.4xlarge (122 GiB)  
• dax.r4.8xlarge (244 GiB)  
• dax.r4.16xlarge (488 GiB)

Q. Does DAX support Reserved Instances or the AWS Free Usage Tier?

Currently DAX only supports on-demand instances.

Q. How is DAX priced?

DAX is priced per node-hour consumed, from the time a node is launched until it is terminated. Each partial node-hour consumed will be billed as a full hour. Pricing applies to all individual nodes in the DAX cluster. For example, if you have a three node DAX cluster, you will be billed for each of the separate nodes (three nodes in total) on an hourly basis.

Availability

Q. How can I achieve high availability with my DAX cluster?

DAX provides built-in multi-AZ support, letting you choose the preferred availability zones for the nodes in your DAX cluster. DAX uses asynchronous replication to provide consistency between the nodes, so that in the event of a failure, there will be additional nodes that can service requests. To achieve high availability for your DAX cluster, for both planned and unplanned outages, we recommend that you deploy at least three nodes in three separate availability zones. Each AZ runs on its own physically distinct, independent infrastructure, and is engineered to be highly reliable.

Q. What happens if a DAX node fails?

If the primary node fails, DAX automatically detects the failure, selects one of the available read replicas, and promotes it to become the new primary. In addition, DAX provisions a new node in the same availability zone of the failed primary; this new node replaces the newly-promoted read replica. If the primary fails due to a temporary availability zone disruption, the new replica will be launched as soon as the AZ has recovered. If a single-node cluster fails, DAX launches a new node in the same availability zone.

Scalability

Q. What type of scaling does DAX support?

DAX supports two scaling options today. The first option is read scaling to gain additional throughput by adding read replicas to a cluster. A single DAX cluster supports up to 10 nodes, offering millions of requests per second. Adding or removing additional replicas is an online operation. The second way to scale a cluster is to scale up or down by selecting larger or smaller r3 instance types. Larger nodes will enable the cluster to store more of the application's data set in memory and thus reduce cache misses and improve overall performance of the application. When creating a DAX cluster, all nodes in the cluster must be of the same instance type. Additionally, if you desire to change the instance type for your DAX cluster (i.e., scale up from r3.large to r3.2xlarge), you must create a new DAX cluster with the desired instance type. DAX does not currently support online scale-up or scale-down operations.

Q. How do I write-scale my application?

Within a DAX cluster, only the primary node handles write operations to DynamoDB. Thus, adding more nodes to the DAX cluster will increase the read throughput, but not the write throughput. To increase write throughput for your application, you will need to either scale-up to a larger instance size or provision multiple DAX clusters and shard your key-space in the application layer.

Monitoring

Q. How do I monitor the performance of my DAX cluster?  
Metrics for CPU utilization, cache hit/miss counts and read/write traffic to your DAX cluster are available via the AWS Management Console or Amazon CloudWatch APIs. You can also add additional, user-defined metrics via Amazon CloudWatch's custom metric functionality. In addition to CloudWatch metrics, DAX also provides information on cache hit, miss, query and cluster performance via the AWS Management Console.

Maintenance

Q. What is a maintenance window? Will my DAX cluster be available during software maintenance?

You can think of the DAX maintenance window as an opportunity to control when cluster modifications such as software patching occur. If a "maintenance" event is scheduled for a given week, it will be initiated and completed at some point during the maintenance window you identify.

Required patching is automatically scheduled only for patches that are security and reliability related. Such patching occurs infrequently (typically once every few months). If you do not specify a preferred weekly maintenance window when creating your cluster, a default value will be assigned. If you wish to modify when maintenance is performed on your behalf, you can do so by modifying your cluster in the AWS Management Console or by using the UpdateCluster API. Each of your clusters can have different preferred maintenance windows.

For multi-node clusters, updates in the cluster are performed serially, and one node will be updated at a time. After the node is updated, it will sync with one of the peers in the cluster so that the node has the current working set of data. For a single-node cluster, we will provision a replica (at no charge to you), sync the replica with the latest data, and then perform a failover to make the new replica the primary node. This way, you don’t lose any data during an upgrade for a one-node cluster.

## **VPC Endpoints for DynamoDB**

Q. What are VPC Endpoints for DynamoDB (VPCE for DynamoDB)?

Amazon Virtual Private Cloud (VPC) is an AWS service that provides users a virtual private cloud, by provisioning a logically isolated section of Amazon Web Services (AWS) Cloud. VPC Endpoint (VPCE) for DynamoDB is a logical entity within a VPC that creates a private connection between a VPC and DynamoDB without requiring access over the Internet, through a NAT device, or a VPN connection. More information on VPC endpoints, see the [Amazon VPC User Guide](http://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/vpc-endpoints.html).

Q. Why should I use VPCE for DynamoDB?

In the past, the main way of accessing DynamoDB from within a VPC was to traverse the Internet, which may have required complex configurations such as firewalls and VPNs. VPC Endpoints for DynamoDB improves privacy and security for customers, especially those dealing with sensitive workloads with compliance and audit requirements, by enabling private access to DynamoDB from within a VPC without the need for an Internet Gateway or NAT Gateway. In addition, VPC Endpoints for DynamoDB supports AWS Identity and Access Management (IAM) policies to simplify DynamoDB access control so you can now easily restrict access to your DynamoDB tables to a specific VPC endpoint.

Q. How do I get started using VPCE for DynamoDB?

You can create VPCE for DynamoDB by using the AWS Management Console, AWS SDK, or the AWS Command Line Interface (CLI). You need to specify the VPC and existing route tables in the VPC, and describe the IAM policy to attach to the endpoint. A route is automatically added to each of the specified VPC’s route tables.

Q. Does VPCE for DynamoDB ensure that traffic will not be routed outside of the Amazon Network?

Yes, when using VPCE for DynamoDB, data packets between DynamoDB and VPC will remain in the Amazon Network.

Q. Can I connect to a DynamoDB table in a region different from my VPC using VPCE for DynamoDB?

No, VPC endpoints can only be created for DynamoDB tables in the same region as the VPC.

Q. Does VPCE for DynamoDB limit throughput to DynamoDB?

No, you will continue to get the same throughput to DynamoDB as you do today from an instance with a public IP within your VPC.

Q. What is the price of using VPCE for DynamoDB?

There is no additional cost for using VPCE for DynamoDB.

Q. Can I access DynamoDB Streams using VPCE for DynamoDB?

At present, you cannot access DynamoDB Streams using VPCE for DynamoDB.

Q. I currently use an Internet Gateway and a NAT Gateway to send requests to DynamoDB. Do I need to change my application code when I use a VPCE?

Your application code does not need to change. Simply create a VPC endpoint, update your route table to point DynamoDB traffic at the DynamoDB VPCE, and access DynamoDB directly. You can continue using the same code and same DNS names to access DynamoDB.

Q. Can I use one VPCE for both DynamoDB and another AWS service?

No, each VPCE supports one service. But you can create one for DynamoDB and another for the other AWS service and use both of them in a route table.

Q. Can I have multiple VPC endpoints in a single VPC?

Yes, you can have multiple VPC endpoints in a single VPC. For example, you can have one VPCE for S3 and one VPCE for DynamoDB.

Q. Can I have multiple VPCEs for DynamoDB in a single VPC?

Yes, you can have multiple VPCEs for DynamoDB in a single VPC. Individual VPCEs can have different VPCE policies. For example, you could have a VPCE that is read only and one that is read/write. However, a single route table in a VPC can only be associated with a single VPCE for DynamoDB, since that route table will route all traffic to DynamoDB through the specified VPCE.

Q. What are the differences between VPCE for S3 and VPCE for DynamoDB?

The main difference is that these two VPCEs support different services – S3 and DynamoDB.

Q. What IP address will I see in AWS CloudTrail logs for traffic coming from the VPCE for DynamoDB?

AWS CloudTrail logs for DynamoDB will contain the private IP address of the EC2 instance in the VPC, and the VPCE identifier (e.g., sourceIpAddress=10.89.76.54, VpcEndpointId=vpce-12345678).

Q. How can I manage VPCEs using the AWS Command Line Interface (CLI)?

You can use the following CLI commands to manage VPCEs: create-vpc-endpoint, modify-vpc-endpoint, describe-vpc-endpoints, delete-vpc-endpoint and descrive-vpc-endpoint-services. You should specify the DynamoDB service name specific to your VPC and DynamoDB region, eg. ‘com.amazon.us.east-1.DynamoDB’. More information can be found [here](http://docs.aws.amazon.com/cli/latest/reference/ec2/create-vpc-endpoint.html).

Q. Does VPCE for DynamoDB require customers to know and manage the public IP addresses of DynamoDB?

No, customers don’t need to know or manage the public IP address ranges for DynamoDB in order to use this feature. A prefix list will be provided to use in route tables and security groups. AWS maintains the address ranges in the list. The prefix list name is: com.amazonaws.<region>.DynamoDB. For example: com.amazonaws.us-east-1.DynamoDB.

Q. Can I use IAM policies on a VPCE for DynamoDB?

Yes. You can attach an IAM policy to your VPCE and this policy will apply to all traffic through this endpoint. For example, a VPCE using this policy only allows describe\* API calls:  
{  
    "Statement":  [  
       {  
            "Sid": "Stmt1415116195105",  
            "Action": "dynamodb:describe\*",  
            "Effect": "Allow",  
            "Resource": "arn:aws:dynamodb:region:account-id:table/table-name",  
            "Principal": "\*"  
        }  
    ]  
}

Q. Can I limit access to my DynamoDB table from a VPC Endpoint?

Yes, you can create an IAM policy to restrict an IAM user, group, or role to a particular VPCE for DynamoDB tables.

This can be done by setting the IAM policy’s Resource element to a DynamoDB table and Condition element’s key to aws:sourceVpce. More details can be found in the [IAM User Guide.](http://docs.aws.amazon.com/IAM/latest/UserGuide/reference_policies_elements.html#Condition)

For example, the following IAM policy restricts access to DynamoDB tables unless sourceVpce matches “vpce-111bbb22”

{  
    "Statement":  [  
       {  
            "Sid": "Stmt1415116195105",  
            "Action": "dynamodb:\*",  
            "Effect": "Deny",  
            "Resource": "arn:aws:dynamodb:region:account-id:\*",  
            "Condition": { "StringNotEquals" : { "aws:sourceVpce": "vpce-111bbb22" } }  
        }  
    ]  
}

Q. Does VPCE for DynamoDB support IAM policy conditions for fine-grained access control (FGAC)?

Yes. VPCE for DynamoDB supports all FGAC access keys. You can use IAM policy conditions for FGAC to control access to individual data items and attributes. More information on FGAC can be found [here](http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/FGAC_DDB.html).

Q. Can I use the AWS Policy Generator to create VPC endpoint policies or DynamoDB?

You can use the [AWS Policy Generator](https://awspolicygen.s3.amazonaws.com/policygen.html) to create your VPC endpoint policies.  
  
Q. Does DynamoDB support resource-based policies similar to S3 bucket policies?

No, DynamoDB does not support resource based policies pertaining to individual tables, items, etc.

# Amazon Redshift

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 local disks, 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](https://aws.amazon.com/redshift/spectrum/), allowing you to directly run SQL queries against exabytes of unstructured data in Amazon S3. No loading or transformation is required, and you can use open data formats, including Avro, CSV, Grok, ORC, Parquet, RCFile, RegexSerDe, SequenceFile, TextFile, and TSV. 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.

Traditional 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, also makes it easy to analyze large amounts of data in its native format without requiring you to load the data.

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.](https://aws.amazon.com/redshift/partners/)

Q: What is Amazon Redshift Spectrum?

[Amazon Redshift Spectrum](https://aws.amazon.com/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 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.

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. In addition, Amazon Redshift doesn't require indexes or materialized views and so uses less space than traditional relational database systems. 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](https://aws.amazon.com/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 Redshift Spectrum, 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](https://aws.amazon.com/redshift/getting-started/) page to see how to try Amazon Redshift for free.

Q: In which AWS regions is Amazon Redshift available?

For information about Amazon Redshift regional availability, see the [Region Table](https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/) in the AWS Global Infrastructure page.

Q: In which AWS regions is Redshift Spectrum available?

Redshift Specturm is currently available in the US East (N. Virginia), US East (Ohio) and US West (Oregon) regions, and will expand to additional regions in the coming months.

Q: How do I create an Amazon Redshift data warehouse cluster?  
  
You can easily create an Amazon Redshift data warehouse cluster by using the [AWS Management Console](https://aws.amazon.com/console/) or the [Amazon Redshift APIs.](http://docs.aws.amazon.com/redshift/latest/APIReference/Welcome.html)You can start with a single node, 160GB data warehouse and scale all the way to a petabyte or more with a few clicks in the [AWS Console](https://aws.amazon.com/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. 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.

Q: What does a leader node do? What does a compute node do?   
  
A leader node receives queries from client applications, parses the queries and develops execution plans, which are an ordered set of steps to process these queries. The leader node then coordinates the parallel execution of these plans with the compute nodes, aggregates the intermediate results from these nodes and finally returns the results back to the client applications.

Compute nodes execute the steps specified in the execution plans and transmit data among themselves to serve these queries. The intermediate results are sent back to the leader node for aggregation before being sent back to the client applications.

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 Dense Storage (DS) nodes or Dense Compute nodes (DC). Dense Storage nodes allow you to create very large data warehouses using hard disk drives (HDDs) for a very low price point. Dense Compute nodes allow you to create very high performance data warehouses using fast CPUs, large amounts of RAM and solid-state disks (SSDs).

Dense Storage (DS) nodes 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. DS2.8XL 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. You can get started with a single Extra Large node, 2TB data warehouse for $0.85 per hour and scale up to a petabyte or more. You can pay by the hour or use reserved instance pricing to lower your price to under $1,000 per TB per year.

Dense Compute (DC) nodes are also 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 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.

Amazon Redshift's MPP architecture means you can increase your performance by increasing the number of nodes in your data warehouse cluster. The optimal amount of data per compute node depends on your application characteristics and your query performance needs.

Q: How many nodes can I specify per Amazon Redshift data warehouse cluster?  
  
An Amazon Redshift data warehouse cluster can contain from 1-128 compute nodes, depending on the node type. For details please see [our documentation](http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html).

Q: How do I access my running 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](http://docs.aws.amazon.com/redshift/latest/gsg/welcome.html).

Q: When would I use Amazon Redshift vs. Amazon RDS?  
  
Both Amazon Redshift and [Amazon RDS](https://aws.amazon.com/rds/) enable you to run traditional relational databases in the cloud while offloading database administration. Customers use Amazon RDS databases both for online-transaction processing (OLTP) and for reporting and analysis. 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 or if you want to prevent your reporting and analytic processing from interfering with the performance of your OLTP workload.

Q: When would I use Amazon Redshift 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 data and get superfast performance.

Q: Can Redshift Spectrum replace Amazon EMR?

No. While Redshift Spectrum 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 also use Redshift Spectrum together with EMR. Amazon Redshift Spectrum uses the same approach to store table definitions as Amazon EMR. So, 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 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 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: Can I use Redshift Spectrum to query data that I process using Amazon EMR?

Yes, Redshift Spectrum can support the same 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.

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.
* *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 S3.

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

## **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. 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. 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](https://aws.amazon.com/s3/pricing/).
* *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.

For Amazon Redshift pricing information, please visit the [Amazon Redshift pricing page.](https://aws.amazon.com/redshift/pricing/)

Q: When does billing of my Amazon Redshift data warehouse clusters begin and end?  
  
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.

Q: What defines billable Amazon Redshift instance hours?  
  
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.

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](https://aws.amazon.com/c-tax-faqs/).

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

## **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](https://aws.amazon.com/s3/), [Amazon DynamoDB](https://aws.amazon.com/dynamodb/), [Amazon EMR](https://aws.amazon.com/emr/), [AWS Data Pipeline](https://aws.amazon.com/datapipeline/) 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. For more details on loading data into Amazon Redshift please view our [Getting Started Guide](http://docs.aws.amazon.com/redshift/latest/gsg/welcome.html).

Q: Can I load data using SQL ‘INSERT' statements?  
  
Yes, 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.

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](http://docs.aws.amazon.com/redshift/latest/dg/r_COPY.html) 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.

In addition, many [ETL companies](https://aws.amazon.com/redshift/partners/) have certified Amazon Redshift for use with their tools, and a number are offering [free trials](https://aws.amazon.com/redshift/partners/) to help you get started loading your data. Finally, [AWS Data Pipeline](https://aws.amazon.com/datapipeline/) 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.

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](https://aws.amazon.com/snowball/) to transfer the data to Amazon S3 using portable storage devices. In addition, you can use [AWS Direct Connect](https://aws.amazon.com/directconnect/) 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.

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

## **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)](http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-HSM.html) or manage your keys through [AWS Key Management Service](https://aws.amazon.com/kms/).

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. Note that Redshift Spectrum does not currently support [Enhanced VPC Routing](https://docs.aws.amazon.com/redshift/latest/mgmt/enhanced-vpc-routing.html).

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.

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

## **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 in the event 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.

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

## **Backup and Restore**

Q: How does Amazon Redshift back up my data?  
  
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.

Q: How long does Amazon Redshift retain backups? Is it configurable?  
  
By default, Amazon Redshift retains backups for 1 day. You can configure this to be as long as 35 days.

Q: How do I restore my Amazon Redshift data warehouse cluster from 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: Do I need to enable backups for my data warehouse cluster or is it done automatically?  
  
By default, Amazon Redshift enables automated backups of your data warehouse cluster with a 1-day retention period. 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](https://aws.amazon.com/console/) or the [Amazon Redshift APIS](http://docs.aws.amazon.com/redshift/latest/APIReference/Welcome.html). For more information on automated snapshots, please refer to the Amazon Redshift [Management Guide](http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-snapshots.html). Amazon Redshift only backs up data that has changed so most snapshots only use up a small amount of your free backup storage.

Q: How do I manage the retention of my automated backups and snapshots?  
  
You can use the [AWS Management Console](https://aws.amazon.com/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 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, which 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](https://aws.amazon.com/s3/pricing/), unless you choose to delete them.

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

## **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 via the [AWS Management Console](https://aws.amazon.com/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](https://aws.amazon.com/console/) or Amazon CloudWatch APIs. You can also add additional, user-defined metrics via [Amazon Cloudwatch's](https://aws.amazon.com/cloudwatch/) custom metric functionality.

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?  
  
The existing data warehouse cluster remains available for read operations while a new data warehouse cluster gets created during scaling operations. When the new data warehouse cluster is ready, your existing data warehouse cluster will be temporarily unavailable while the canonical name record of the existing data warehouse cluster is flipped to point to the new data warehouse cluster. This period of unavailability typically lasts only a few minutes, and will occur during the maintenance window for your data warehouse cluster, unless you specify that the modification should be applied immediately. Amazon Redshift moves data in parallel from the compute nodes in your existing data warehouse cluster to the compute nodes in your new cluster. This enables your operation to complete as quickly as possible.

[Back to top »](https://aws.amazon.com/redshift/faqs/#redshift-faqs)

## **Querying and Analysis**

Q: Is Amazon Redshift 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 our [Console](https://console.aws.amazon.com/redshift/). We have validated integrations with [popular BI and ETL vendors](https://aws.amazon.com/redshift/partners/), a number of which are offering [free trials](https://aws.amazon.com/redshift/partners/) to help you get started loading and analyzing your data. You can also go to the [AWS Marketplace](https://aws.amazon.com/redshift/getting-started/) to deploy and configure solutions designed to work with Amazon Redshift in minutes.

Q: What kinds of queries does Redshift Spectrum support?

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 cluster. External tables are referenced using the schema name defined in the CREATE EXTERNAL SCHEMA command where they were registered.

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: What BI tools and SQL Clients does Redshift Spectrum support?

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.

Q: What data formats does Redshift Spectrum support?

Redshift Spectrum currently supports numerous open source data formats, including Avro, CSV, Grok, ORC, Parquet, RCFile, RegexSerDe, SequenceFile, TextFile, and TSV.

Q: What compression formats does Redshift Spectrum support?

Redshift Spectrum currently supports Gzip and Snappy compression.

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](https://aws.amazon.com/console/) or Amazon CloudWatch APIs. You can also add additional, user-defined metrics via [Amazon Cloudwatch’s](https://aws.amazon.com/cloudwatch/) custom metric functionality. In addition to CloudWatch metrics, 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. 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.

Q: I notice that some queries accessing data in my cluster are running slower than my Redshift Spectrum queries. Why is that?

Amazon Redshift queries are run on your cluster resources against local disk. Redshift Spectrum queries run using per-query scale-out resources against data in S3. For most queries, local disk will be faster, but for queries that scan a lot of data and do minimal compute processing, we can apply a lot of Redshift Spectrum workers and complete them quickly.

## **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 Amazon 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](https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html#rs-maintenance-windows) in the *Amazon Redshift Management Guide*.

# Amazon VPC FAQs

## **General Questions**

Q. What is Amazon Virtual Private Cloud (Amazon VPC)?

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.

#### **Manage Your AWS Resources**

[Sign in to the Console](https://console.aws.amazon.com/console/home)

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 (VPC): 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.
* Hardware VPN Connection: A hardware-based VPN connection between your Amazon VPC and your datacenter, home network, or co-location facility.
* Virtual Private Gateway: The Amazon VPC side of a VPN connection.
* Customer Gateway: Your side of a VPN connection.
* Router: Routers interconnect subnets and direct traffic between Internet gateways, virtual private gateways, NAT gateways, and subnets.
* Peering Connection: A peering connection enables you to route traffic via private IP addresses between two peered VPCs.
* VPC Endpoint: Enables private connectivity to Amazon services  from within your VPC without using an Internet gateway or NAT.
* 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, is exposed to the Internet. You can also leverage the greatly enhanced security options in Amazon VPC to provide more granular access both 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. VPC with a Single Public Subnet Only
2. VPC with Public and Private Subnets
3. VPC with Public and Private Subnets and Hardware VPN Access
4. VPC with a Private Subnet Only and Hardware VPN Access

## **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](http://aws.amazon.com/vpc).

Q. What defines billable VPN connection-hours?

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. 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](https://aws.amazon.com/c-tax-faqs/).

## **Connectivity**

Q. What are the connectivity options for my VPC?

You may connect your VPC to:

* The Internet (via an Internet gateway)
* Your corporate data center using a Hardware 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 VPCs (via VPC peering connections)

Q. What are the different types of VPC Endpoints available on Amazon VPC?

VPC endpoints enable you to privately connect your VPC to an AWS service without requiring an Internet gateway, a NAT device, 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 for S3 and DynamoDB. These endpoints will add an entry to you’re the route table you selected and route the traffic to the supported services through Amazon’s private network.

Interface type endpoints are available for Kinesis, EC2 Systems Manager (SSM), EC2 API and ELB API. These endpoints provide private connectivity to services powered by PrivateLink, and provide connectivity over Direct Connect. More services will be supported by these endpoints in the future. Please refer to [Amazon VPC pricing](https://aws.amazon.com/vpc/pricing/) for interface type endpoints' pricing.

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. How does a hardware VPN connection work with Amazon VPC?

A hardware 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 hardware VPN connection.

Q. What is IPsec?

[IPsec](http://en.wikipedia.org/wiki/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?

There are two types of VPN connections that you can create: 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 VPN connections must be able to:

* Establish Border Gateway Protocol (BGP) peerings
* Bind tunnels to logical interfaces (route-based VPN)
* Utilize IPsec Dead Peer Detection

Q. Which Diffie-Hellman Groups do you support?

We support the following Diffie-Hellman (DH) groups in Phase1 and Phase2.

* Phase1 DH groups 2, 14-18, 22, 23, 24
* Phase2 DH groups 2, 5, 14-18, 22, 23, 24

Q. What customer gateway devices are known to work with Amazon VPC?

The following devices meeting the aforementioned requirements are known to work with hardware VPN connections, and have support in the command line tools for automatic generation of configuration files appropriate for your device:

* Statically-routed VPN connections
  + [Cisco ASA 5500 Series](http://www.cisco.com/cisco/web/solutions/small_business/products/security/ASA_5500_series/index.html) version 8.2 (or later) software
  + [Cisco ISR](http://www.cisco.com/en/US/products/ps10906/Products_Sub_Category_Home.html) running Cisco IOS 12.4 (or later) software
  + [Dell SonicWALL Next Generation Firewalls (TZ, NSA, SuperMassive Series) running SonicOS5.8 (or later)](http://www.sonicwall.com/us/en/products/network-security.html)
  + [Juniper J-Series Service Router](http://www.juniper.net/us/en/products-services/routing/j-series/) running JunOS 9.5 (or later) software
  + [Juniper SRX-Series Services Gateway](http://www.juniper.net/us/en/products-services/security/srx-series/) running JunOS 9.5 (or later) software
  + [Juniper SSG](http://www.juniper.net/us/en/products-services/security/ssg-series/) running ScreenOS 6.1, or 6.2 (or later) software
  + [Juniper ISG](http://www.juniper.net/us/en/products-services/security/isg-series/) running ScreenOS 6.1, or 6.2 (or later) software
  + [Microsoft Windows Server 2008 R2](http://www.microsoft.com/en-us/server-cloud/windows-server/) (or later) software
  + [Yamaha RTX1200 router](http://www.rtpro.yamaha.co.jp/RT/docs/amazon-vpc/index.html)
* Dynamically-routed VPN connections (requires BGP)
  + [Astaro Security Gateway](http://www.astaro.com/products/hardware-appliances) running version 8.3 (or later)
  + [Astaro Security Gateway Essential Firewall Edition](http://www.astaro.com/landingpages/en-worldwide-essential-firewall) running version 8.3 (or later)
  + [Cisco ISR](http://www.cisco.com/en/US/products/ps10906/Products_Sub_Category_Home.html) running Cisco IOS 12.4 (or later) software
  + [Dell SonicWALL Next Generation Firewalls (TZ, NSA, SuperMassive Series) running SonicOS5.9 (or later)](http://www.sonicwall.com/us/en/products/network-security.html)
  + [Juniper J-Series Service Router](http://www.juniper.net/us/en/products-services/routing/j-series/) running JunOS 9.5 (or later) software
  + [Juniper SRX-Series Services Gateway](http://www.juniper.net/us/en/products-services/security/srx-series/) running JunOS 9.5 (or later) software
  + [Juniper SSG](http://www.juniper.net/us/en/products-services/security/ssg-series/) running ScreenOS 6.1, or 6.2 (or later) software
  + [Juniper ISG](http://www.juniper.net/us/en/products-services/security/isg-series/) running ScreenOS 6.1, or 6.2 (or later) software
  + [Palo Alto Networks PA Series](https://www.paloaltonetworks.com/products/platforms/firewalls.html) running PANOS 4.1.2 (or later) software
  + [Vyatta Network OS](http://www.vyatta.com/product/vyatta-network-os/specs) 6.5 (or later) software
  + [Yamaha RTX1200 router](http://www.rtpro.yamaha.co.jp/RT/docs/amazon-vpc/index.html)

Please note, these sample configurations are for the minimum requirement of AES128, SHA1, and DH Group 2. You will need to modify these sample configuration files to take advantage of AES256, SHA256, or other DH groups.

Q. If my device is not listed, where can I go for more information about using it with Amazon VPC?

We recommend checking the [Amazon VPC forum](http://developer.amazonwebservices.com/connect/forum.jspa?forumID=58) as other customers may be already using your device.

Q. Are there any VPN connection throughput limitations?

VPN connection throughput can depend on multiple factors, such as the capability of your Customer Gateway (CGW), the capacity of your connection, average packet size, the protocol being used (TCP vs. UDP), and the network latency between your CGW and the Virtual Private Gateway (VGW).

Q. What tools are available to me to help troubleshoot my Hardware VPN configuration?

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?

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)](http://en.wikipedia.org/wiki/Network_address_translation) on Amazon EC2 instances within a VPC accessed via a hardware VPN connection.

Q. Can I NAT my CGW behind a router or firewall?

Yes, you will need to enable NAT-T and open UDP port 4500 on your NAT device.

Q. What IP address do I use for my CGW address?

You will use the public IP address of your NAT device.

Q. How does my connection decide to use NAT-T?

If your device has NAT-T enabled on the tunnel, AWS will use it by default. You will need to open UDP port 4500 or else the tunnel will not establish.

Q. How do I disable NAT-T on my connection?

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 CGWs behind a NAT, what do I need to do to configure that?

You will use the public IP address of your NAT device for the CGW for each of your connections. You will also need to make sure UDP port 4500 is open.

Q. How many IPsec security associations can be established concurrently per tunnel?

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.

## **IP Addressing**

Q. What IP address ranges can I use within my VPC?

You can use any [IPv4](http://en.wikipedia.org/wiki/IPv4) address range, including [RFC 1918](https://tools.ietf.org/html/rfc1918) or publicly routable IP ranges for the primary CIDR block. For the secondary CIDR blocks certain [restrictions](http://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/VPC_Subnets.html#add-cidr-block-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 VPCs?

You assign a single [Classless Internet Domain Routing (CIDR)](http://en.wikipedia.org/wiki/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 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 hardware VPN, and to my VPC?

Yes, you can route traffic via the hardware VPN connection and advertise the address range from your home network.

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 a VPC's size?

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](https://aws.amazon.com/contact-us/vpc-request/).

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 the [EC2 User Guide](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html) 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.

## **Routing & Topology**

Q. What does an Amazon VPC router do?

An Amazon VPC router enables Amazon EC2 instances within subnets to communicate with Amazon EC2 instances in other subnets within the same VPC. The VPC router also enables subnets, Internet gateways, and virtual private gateways to communicate with each other. Network usage data is not available from the router; however, you can obtain network usage statistics from your instances using Amazon CloudWatch.

Q. Can I modify the VPC route tables?

Yes. You can create route rules to specify which subnets are routed to the Internet gateway, the virtual private gateway, or other instances.

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.

Q. Does Amazon VPC support [multicast](http://en.wikipedia.org/wiki/IP_multicast) or [broadcast](http://en.wikipedia.org/wiki/Broadcast_address#IP_network_broadcasting)?

No.

## **Security & 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, they can communicate using 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. Why can’t I ping the router, or my default gateway, that connects my subnets?

Ping (ICMP Echo Request and Echo Reply) requests to the router in your VPC is not supported. Ping between Amazon EC2 instances within VPC is supported as long as your operating system's firewalls, VPC security groups, and network ACLs permit such traffic.

Q. Can I monitor the network traffic in my VPC?

Yes. You can use the Amazon VPC Flow Logs feature to monitor the network traffic in your VPC.

## **Amazon VPC & EC2**

Q. Within which Amazon EC2 region(s) is Amazon VPC available?

Amazon VPC is currently available in multiple [Availability Zones](http://developer.amazonwebservices.com/connect/entry.jspa?externalID=1347) 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](http://aws.amazon.com/contact-us/vpc-request/).

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](http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/FAQ_Regions_Availability_Zones.html).

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](http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/FAQ_Regions_Availability_Zones.html).

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](http://aws.amazon.com/ec2/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](https://forums.aws.amazon.com/ann.jspa?annID=1875) 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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-vpc.html) 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. If you delete your default VPC, you can create a new default VPC later from the VPC Console or by using the CreateDefaultVPC API.This will create a new default VPC in the region, rather than restore the previous one.

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, contact [AWS Support](https://aws.amazon.com/contact-us/). We will review your request and your existing AWS services and EC2-Classic presence to determine if you are eligible for a default VPC.

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?

No. Peering connections are only available between VPCs in the same region.

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](http://aws.amazon.com/ec2/pricing/) 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](http://docs.aws.amazon.com/AmazonVPC/latest/PeeringGuide/) 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.

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](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/placement-groups.html).

## **PrivateLink**

Q. What is PrivateLink?

PrivateLink is a purpose-built technology designed for customers to access Amazon services in a highly available and scalable manner, while keeping all the network traffic within the AWS network. Customers can use this to privately access Amazon services from their Amazon Virtual Private Cloud (VPC) or their own data center, without using public IPs, and without requiring the traffic to traverse across the Internet.

Q. How can I use PrivateLink?

In order to use PrivateLink, you will need to create interface type VPC endpoints for Amazon services that are powered by PrivateLink. These service endpoints will appear as Elastic Network Interface (ENI) with private IPs in your VPCs. Once these endpoints are created, any traffic destined to these IPs will get privately routed to the corresponding Amazon services.

Q. Can I privately access services powered by PrivateLink over Direct Connect?

Yes. The application in your own data center can connect to the service endpoints in Amazon VPC over AWS Direct Connect.&nbsp; The service endpoints will automatically direct the traffic to Amazon Services powered by PrivateLink.

Q. Which services are currently available on PrivateLink?

Elastic Compute Cloud (EC2), Elastic Load Balancer (ELB), Kinesis, and EC2 Systems Manager are currently available on PrivateLink.

## **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](https://aws.amazon.com/ec2/pricing/).

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.

## **Virtual Private Gateway - Bring your own Autonomous System Number**

Q. What is this feature?

For any new VGWs, 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?   
There is no additional charge for this feature.

Q. How can I configure/assign my ASN to be advertised as Amazon side ASN?

You can configure/assign an ASN to be advertised as the Amazon side ASN during creation of the new Virtual Private Gateway (VGW). You can create a VGW using the VPC console or a EC2/CreateVpnGateway API call.

Q. What ASN did Amazon assign prior to this feature?

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?

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?

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.

Amazon will provide a default ASN for the VGW 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?

We will ask you to re-enter a private ASN once you attempt to create the VGW, 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?

Amazon will provide an ASN for the VGW 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?

You can view the Amazon side ASN in the VGW 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?

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 VGW with desired ASN, and create a new VIF with the newly created VGW. 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?

You will need to create a new VGW with the desired ASN, and recreate your VPN connections between your Customer Gateways and the newly created VGW.

Q. I already have a VGW 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 VGW, what Amazon side ASN will I be assigned?

Amazon will assign 64512 to the Amazon side ASN for the new VGW.

Q. I have a VGW 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?

You can configure/assign an ASN to be advertised as the Amazon side ASN during creation of the new Virtual Private Gateway (VGW). You can create VGW using console or EC2/CreateVpnGateway API call. As noted earlier, we will allow the use of the “legacy public ASN” for your newly created VGW.

Q. I have a VGW 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 VGW, what Amazon side ASN will I be assigned?

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 VGW and defaults to that ASN.

Q. I’m attaching multiple private VIFs to a single VGW. Can each VIF have a separate Amazon side ASN?

No, you can assign/configure separate Amazon side ASN for each VGW, not each VIF. Amazon side ASN for VIF is inherited from the Amazon side ASN of the attached VGW.

Q. I’m creating multiple VPN connections to a single VGW. Can each VPN connection have a separate Amazon side ASN?

No, you can assign/configure separate Amazon side ASN for each VGW, not each VPN connection. Amazon side ASN for VPN connection is inherited from the Amazon side ASN of the VGW.

Q. Where can I select my own ASN?

When creating a VGW 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 VGW is configured with Amazon side ASN, the private VIFs or VPN connections created using the VGW will use your Amazon side ASN.

Q. I use CloudHub today. Will I have to adjust my configurations 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 VGW 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 VGW and recreate a new VGW with the desired ASN.

Q. Is there a new API to configure/assign the Amazon side ASN?

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?

No. You can view the Amazon side ASN with the same EC2/DescribeVpnGateways API. We just added a new parameter (amazonSideAsn) to this API.

## **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, Internet gateways, customer gateways, virtual private gateways, and VPN connections 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 VPC
* Five virtual private gateways per AWS account per region
* Fifty customer gateways per AWS account per region
* Ten IPsec VPN Connections per virtual private gateway

See the [VPC User Guide](http://docs.aws.amazon.com/AmazonVPC/latest/UserGuide/VPC_Appendix_Limits.html) for more information on VPC limits.

Q. Does the Amazon VPC VPN Connection have a Service Level Agreement (SLA)?

Not currently.

Q. Can I obtain AWS Support with Amazon VPC?

Yes. [Click here](http://aws.amazon.com/premiumsupport/) for more information on AWS Support.

Q. Can I use [ElasticFox](http://sourceforge.net/projects/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.

# 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](https://portal.aws.amazon.com/billing/signup) 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 all files that can be served over HTTP. This includes dynamic web pages, such as HTML or PHP pages, any popular static files that are a part of your web application, such as website images, audio, video, media files or software downloads. For on-demand media files, you can also choose to stream your content using RTMP delivery. Amazon CloudFront also supports delivery of live media over HTTP.

Q. Does Amazon CloudFront support delivery of dynamic content?

Amazon CloudFront supports all files that can be served over HTTP. This includes dynamic web pages, such as HTML or PHP pages, any popular static files that are a part of your web application, such as website images, audio streams, video streams, media files or software downloads. For on-demand media files, you can also choose to stream your content using RTMP delivery. Amazon CloudFront also supports delivery of live media 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: 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](https://aws.amazon.com/cloudfront/sla/).

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](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.](https://aws.amazon.com/cloudfront/resources/)

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?

A default CloudFront feature that brings more of your content close to your viewers, even when the content is not popular to stay at a particular edge location. 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. These locations sit in between your origin webserver and the global edge locations that serve traffic directly to your viewers. As the popularity of your objects reduce, individual edge locations may evict those objects to make room for more popular content. Regional Edge Caches have larger cache width than any individual edge location, so objects remain in cache longer at these 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 instance, our 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 utilized 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](https://aws.amazon.com/cloudfront/details/).

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](https://aws.amazon.com/cloudfront/pricing/).

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](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/compliance.html) 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](https://aws.amazon.com/compliance/hipaa-compliance/) and our [developer's guide](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/compliance.html).

# 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](https://aws.amazon.com/caching/cdn/) 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](https://aws.amazon.com/cloudfront/custom-ssl-domains/).

# 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.](https://aws.amazon.com/cloudfront/custom-ssl-domains/)

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](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/SecureConnections.html#CNAMEsAndHTTPS) or the [SNI Wikipedia article.](http://en.wikipedia.org/wiki/Server_Name_Indication)

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](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/PrivateContent.html).

Q. How can I safeguard my web applications delivered via CloudFront from DDoS attacks?

As an AWS customer, you get [AWS Shield Standard](https://aws.amazon.com/shield/) 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](https://aws.amazon.com/shield/) 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](https://aws.amazon.com/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](http://docs.aws.amazon.com/console/waf) 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](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/forward-custom-headers.html).

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 to use in the cache key?

Yes, query string whitelisting feature allows you to easily configure Amazon CloudFront to only use certain parameters in the cache key, while still forwarding all 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 will automatically compress your text or binary data. To use the feature, simply specify within 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 browser do this by default)..For more information on this feature, please refer to [our developer guide.](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/ServingCompressedFiles.html)

# Streaming

Q. What is streaming? Why would I want to stream my content?

Generally, streaming refers to delivering audio and video to end users on the internet without having to download the media file prior to playback. The protocols used for streaming include proprietary ones such as Adobe’s Real Time Messaging Protocol (RTMP) and those that use HTTP for delivery such as Apple’s HTTP Live Streaming (HLS), Adobe’s HTTP Dynamic Streaming (HDS) and Microsoft’s Smooth Streaming. These protocols are different than the delivery of web pages and other content because streaming protocols deliver content in real time – the 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 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 computer when they finish watching a video.
* Streaming can help you reduce your costs, as it only delivers portions of a media file that the viewers actually watch. In contrast, with traditional downloads, frequently the whole media file will be downloaded by the viewers, even if they only watch a portion of the file.

Q. Does Amazon CloudFront support on-demand streaming protocols?

Yes, Amazon CloudFront provides you with multiple options to deliver on-demand content. If you have media files that have been converted to either HLS format or Microsoft Smooth Streaming format prior to storing 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. In addition 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. Another option, if you want to stream using RTMP, is to store your media files on Amazon S3 and use it as the origin for an Amazon CloudFront RTMP distribution.

Q. Does Amazon CloudFront support live streaming to multiple platforms?

Yes. Amazon CloudFront provides you three options to easily and cost-effectively deliver live events over HTTP to multiple platforms:

Live Streaming using Wowza Media Server 3.6: Using Amazon CloudFront with Wowza Media Server combines the benefits of Wowza Media Server with the reliability, scalability, low latency and cost-efficiency of Amazon CloudFront to stream live events to multiple streaming formats, including Apple HTTP Live Streaming (HLS), Adobe HTTP Dynamic Streaming (HDS) and Microsoft Smooth Streaming. We've made this simple for you by creating an AWS CloudFormation template that handles all of the provisioning and sequencing for all the AWS resources you need for this live streaming stack. Amazon CloudFront provides you the scale and flexible pay-as-you-go pricing model, while the use of HTTP protocols for streaming your live event offers your viewers easy access to your live content. Using Amazon CloudFront for live streaming also gives you full control of your Wowza origin server so you can configure it to best work with the specific nature of your event. In addition, you can choose the Amazon EC2 instance type and AWS region that best meet the needs of your live event. A detailed tutorial for setting-up live HTTP streaming using Amazon CloudFront is available [here](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/live-streaming-wowza.html).

Live Streaming using Adobe Media Server 5.0: Amazon CloudFront can be used with Amazon EC2 running Adobe Media Server (AMS 5.0) for live HTTP streaming to both Flash Player and Apple iOS devices. Amazon EC2 (running AMS 5.0) must be configured as the origin for a CloudFront web distribution. Similar to our other live streaming solutions we have setup an AWS CloudFormation template to make it easy for you to setup your pay-as-you-go streaming stack while providing you with full control of the AMS server running in the Amazon EC2 instances provisioned. A detailed tutorial (which also points to the AWS CloudFormation templates) for setting-up live HTTP streaming using CloudFront and AMS 5.0 is available [here](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/LiveStreamingAdobeMediaServer5.0.html).

Live Streaming using Windows Media Services: You can also use Amazon CloudFront and Amazon EC2 running Windows Media Services for live streaming. With this solution, you can deliver live media over HTTP to both Microsoft Silverlight clients and Apple iOS devices. We've made it simple to get started by creating a [tutorial](http://docs.amazonwebservices.com/AmazonCloudFront/latest/DeveloperGuide/IISLiveSmoothStreaming4.1.html) and an AWS CloudFormation template to automate the provisioning of AWS resources for your live streaming stack. You only pay for the AWS resources you consume, and have full control over the origin server (Amazon EC2 instance running Windows Media Services) so you can configure additional IIS Live Smooth Streaming functionality.

# 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](https://aws.amazon.com/cloudfront-request/), 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](http://docs.aws.amazon.com/general/latest/gr/aws_service_limits.html#limits_cloudfront) in the Amazon Web Services General Reference. To request a higher limit, please go to the [CloudFront Limit Increase Form](https://aws.amazon.com/support/createCase?type=service_limit_increase&serviceLimitIncreaseType=cloudfront-distributions).

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](http://aws.amazon.com/cloudfront/reporting/).

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](http://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/cost-alloc-tags.html). If you are ready to add tags to you CloudFront distributions, see [Amazon CloudFront Add Tags page](http://docs.aws.amazon.com/console/cloudfront/tagging).

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](https://console.aws.amazon.com/cloudtrail). For more information, visit [AWS CloudTrail home page](http://aws.amazon.com/cloudtrail/).

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](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/monitoring-using-cloudwatch.html) or simply navigate to the [Amazon CloudFront Management Console](https://console.aws.amazon.com/cloudfront/home) and select Monitoring & Alarming in the navigation pane.

# Lambda@Edge

Q: What is Lambda@Edge?

[Lambda@Edge](https://aws.amazon.com/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 code to [AWS Lambda](https://aws.amazon.com/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](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/lambda-at-the-edge.html).

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 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](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/lambda-at-the-edge.html).

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](https://cio.gov/worldclassdigitalservices/transition-to-ipv6/) – 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](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/AccessLogs.html#BasicDistributionFileFormat) 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’](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/RequestAndResponseBehaviorS3Origin.html#RequestS3IPAddresses) 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”](http://docs.aws.amazon.com/console/cloudfront/ipv6) 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](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/AccessLogs.html#BasicDistributionFileFormat) 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’](http://docs.aws.amazon.com/AmazonCloudFront/latest/DeveloperGuide/RequestAndResponseBehaviorS3Origin.html#RequestS3IPAddresses) 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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/resource-record-sets-choosing-alias-non-alias.html?shortFooter=true).

# 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](https://aws.amazon.com/free/), 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 will be charged for the volume of data transferred out of the Amazon CloudFront edge locations, measured in GB. If you are using other Amazon Web Services as the origins of your files, you will be charged separately for use of those services, including for storage, compute hours, GET requests and data transfer out of that service to Amazon CloudFront’s edge locations. Usage tiers for data transfer are measured separately for each geographic region. You can see the rates for Amazon CloudFront data transfer to the Internet [here](https://aws.amazon.com/cloudfront/pricing/).
* 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](https://aws.amazon.com/cloudfront/pricing/).
* 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](https://aws.amazon.com/cloudfront/pricing/).
* 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](https://aws.amazon.com/cloudfront/pricing/).
* 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](https://aws.amazon.com/support/createCase?type=service_limit_increase&serviceLimitIncreaseType=cloudfront-distributions).

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](https://aws.amazon.com/jp/c-tax-faqs/).

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](https://aws.amazon.com/cloudfront/pricing/).

## **Getting Started**

Q. [What is a Domain Name System (DNS) Service?](https://aws.amazon.com/route53/what-is-dns/)

[DNS](https://aws.amazon.com/route53/what-is-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](https://aws.amazon.com/route53/).
* 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](https://aws.amazon.com/route53/faqs/#which_dns_records_are_supported).
* 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 in 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](https://aws.amazon.com/route53/pricing/).

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](https://calculator.s3.amazonaws.com/index.html).

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](https://aws.amazon.com/iam/).

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](https://aws.amazon.com/contact-us/) 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](https://aws.amazon.com/route53/sla/).

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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/query-logs.html).

## **Domain Name System (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](https://aws.amazon.com/route53-request/) 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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/rrs-changes-import-console.html).

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](https://aws.amazon.com/websites/). 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](https://aws.amazon.com/websites/) solutions. Learn more about Amazon EC2 [here](https://aws.amazon.com/ec2/). 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](https://aws.amazon.com/cloudfront/).

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)
* Additionally, Amazon Route 53 offers ‘Alias’ records (an Amazon Route 53-specific virtual record). Alias records are used to map resource record sets in your hosted zone to Amazon Elastic Load Balancing load balancers, Amazon CloudFront distributions, AWS Elastic Beanstalk environments, or Amazon S3 buckets that are configured as websites. Alias records work like a CNAME record in that you can map one DNS name (example.com) to another ‘target’ DNS name (elb1234.elb.amazonaws.com). They differ from a CNAME record in that they are not visible to resolvers. Resolvers only see the A record and the resulting IP address of the target record.

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](https://aws.amazon.com/cloudtrail/) 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 your ELB DNS name (i.e. *elb1234.elb.amazonaws.com*). IP addresses associated with Amazon Elastic 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. Queries to Alias records that are mapped to ELB load balancers 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 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. 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](http://docs.amazonwebservices.com/Route53/latest/DeveloperGuide/CreatingLatencyRRSets.html).

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](https://aws.amazon.com/route53/pricing/) 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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/routing-policy.html).

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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/routing-policy.html).

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](https://aws.amazon.com/route53/pricing/) 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.

## **DNS 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?

No, it is not 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](https://aws.amazon.com/route53/pricing/).

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. Hoes 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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/routing-policy.html#routing-policy-geoproximity).

Q. How does geoproximity bias of an endpoint affect   dns traffic routing to other endpoints?

Changing geoproximity bias value on an endpoint either increases or decreases the value of the calculated distance relative to the other endpoints. However, the bias does not accurately predict the load factor but rather changes the sphere of influence. The amount of traffic shifting depends on how much queries are generated within the geographical sphere of influence of an endpoint. For more information please refer our [documentation](http://docs.aws.amazon.com/Route53/latest/APIReference/api-policies-traffic-policy-document-format.html#traffic-policy-document-format-rules-geoproximity).

 Q. Can I use bias for other traffic flow rules?

As of today, bias can only be applied to geoproximity rule.

## **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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/hosted-zones-private.html).

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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/hosted-zones-private.html) 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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/hosted-zone-private-associate-vpcs-different-accounts.html).

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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/health-checks.html) 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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/dns-failover.html).

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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/health-checks-creating-deleting.html).

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](http://aws.amazon.com/route53/pricing/) 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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/health-checks-creating-deleting.html).

Q. How much load should I expect a health check to generate on my endpoint (for example, a web server)?

Each heath 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:

1. Route 53 conducts a health check of your application. In this example, your application fails three consecutive health checks, triggering the following events.
2. 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 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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/health-checks-creating-deleting.html).

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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/monitoring-health-checks.html) 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](http://docs.aws.amazon.com/ElasticLoadBalancing/latest/DeveloperGuide/US_MonitoringLoadBalancerWithCW.html). 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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/dns-failover.html).

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](https://aws.amazon.com/forms/route53-unwanted-healthchecks), 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](https://ip-ranges.amazonaws.com/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](http://docs.aws.amazon.com/general/latest/gr/aws-ip-ranges.html) 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](https://aws.amazon.com/route53/domain-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](https://d32ze2gidvkk54.cloudfront.net/Amazon_Route_53_Domain_Registration_Pricing_20140731.pdf).

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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/domain-renew.html) 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](https://d32ze2gidvkk54.cloudfront.net/Amazon_Route_53_Domain_Registration_Pricing_20140731.pdf) and for the specific registration requirements for each, please see the [Amazon Route 53 Developer Guide](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/registrar-tld-list.html) and our [Domain Name Registration Agreement](https://aws.amazon.com/route53/domain-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](http://docs.aws.amazon.com/Route53/latest/APIReference/actions-on-reusable-delegation-sets.html).

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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/registrar-tld-list.html) 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](https://aws.amazon.com/route53-request/) 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](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/domain-configure-dnssec.html).

Q. How do I transfer a domain registration that has DNSSEC enabled to Amazon Route 53?

See our [documentation](http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/domain-configure-dnssec.html) for a step-by-step guide on transferring your DNSSEC-enabled domain to Amazon Route 53.

## **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.

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.

## **Pricing**

Q: How much does Amazon CloudWatch cost?

Please see our [pricing page](https://aws.amazon.com/cloudwatch/pricing/) for the latest information.

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:  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](https://aws.amazon.com/c-tax-faqs/).

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.

## **AWS Resource and 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.

## **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: Is CloudWatch Logs available in all regions?

Please refer to [Regional Products and Services](https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/) for details of CloudWatch Logs service availability by region.

Q: How much does CloudWatch Logs cost?

Please see our [pricing page](https://aws.amazon.com/cloudwatch/pricing/) 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](http://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/UsingConfig_WinAMI.html).

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](http://docs.aws.amazon.com/AmazonCloudWatch/latest/DeveloperGuide/WhatIsCloudWatchLogs.html).

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](https://aws.amazon.com/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.

## **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](https://aws.amazon.com/cloudwatch/pricing/) 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: What can I do with CloudWatch dashboards?

You can use CloudWatch Dashboards to monitor your applications and resources to quickly identify issues that might be impacting the health of your applications. You can save and revisit dashboards, add multiple graphs, or add text widgets into a dashboard to embed links and comments. For example, you can include graphs of your resource and application metrics to see when resource health problems might be impacting your applications. You can also view metrics from multiple regions on the same page.

Q: How do I get started with CloudWatch Dashboards?

To get started, visit the [Amazon CloudWatch Console](https://console.aws.amazon.com/cloudwatch) and select “Dashboards”. Click the “Create Dashboard” button.

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.

Amazon EC2 Systems Manager

Q: What is Amazon EC2 Systems Manager?

Amazon EC2 Systems Manager is a flexible and easy to use management service that enables enterprises to securely manage and administer their workloads, running on-premises or in AWS, using a single unified AWS experience. EC2 Systems Manager is designed to be highly automation-focused to enable configuration and management of instances at a large scale, while making it really simple to write and maintain automation artifacts.

Q: How do I get started with EC2 Systems Manager?

The best way to get started is to ensure your instance has met the necessary requirements in our Getting Started Guide. Once you've confirmed the requirements have been met, you can access the various EC2 Systems Manager capabilities from the left navigation bar in the EC2 Management Console or use the AWS SDKs and AWS Command Line Interface.

Q: Which operating systems does EC2 Systems Manager support?

EC2 Systems Manager is optimized to manage both [Windows](http://docs.aws.amazon.com/AWSEC2/latest/WindowsGuide/systems-manager-prereqs.html) and [Linux](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/systems-manager-prereqs.html) platforms from a single unified experience. Refer to Supported Operating Systems for more details on managing on-premises systems.

Q: Does EC2 Systems Manager manage instances running on-premises?

Yes, EC2 Systems Manager supports managing instances that are running on-premises data center. Refer to [EC2 Systems Manager Prerequisites](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/systems-manager-prereqs.html) for more details.

Q: In which AWS regions is EC2 Systems Manager available?

EC2 Systems Manager is available in multiple public regions. Refer to [AWS Regions and Endpoints](http://docs.aws.amazon.com/general/latest/gr/rande.html) for a complete list of supported regions.

Q: How much does EC2 Systems Manager cost?

There is no charge for EC2 Systems Manager.

Run Command

Q: What is Run Command?

Run Command is a feature of EC2 Systems Manager that provides a simple and secure way to remotely execute commands or run scripts against EC2 instances or on-premises servers, all from the EC2 API, CLI, or console. With Run Command, you can perform commands which make it easy to accomplish common administrative tasks like installing software, executing scripts, making configuration changes, and more.

Q: Who should use Run Command?

Run Command is designed for developers, system administrators, and other IT professionals who need to remotely manage their EC2 instances in a secure, reliable, and scalable way.

Q: Does AWS provide any predefined commands?

Yes. There are predefined commands available which are designed to help with commonly used administrative tasks. For Windows you can run a PowerShell command or script, configure Windows Update settings, deploy an MSI application and more. For Linux you run any Shell command or script, and remotely update the installed agent.

Q: Can I create my own commands?

Yes. Run Command allows you to easily create custom commands to perform common tasks required for your environment.

Q: What other types of commands or script can I run?

You can run any command or script that you can type into a command window on your EC2 instances.

Q: Can I send the same command to multiple EC2 instances at once?

Yes. You can easily issue a command to a fleet of instances by providing a list of instances when issuing a command.

Q: Can I retrieve the history of commands run against my instances?

Yes. Run Command keeps the output for each command for 30 days. In addition, you can have Run Command store a copy of all log files in [Amazon S3](https://aws.amazon.com/s3/) or capture the output of your commands using [AWS CloudTrail](https://aws.amazon.com/cloudtrail/).

Q: Can I control who can execute a command?

Yes. Using the published [AWS Identity and Access Management](https://aws.amazon.com/iam/) (IAM) permissions and policies, you can control who has access to execute commands or documents on specific instances. For example, you can specify an IAM user who can run PowerShell commands, but not join an instance to a domain. Another IAM user can only be given access to run a very specific command like restarting services, giving you the flexibility to specify how much access any given user can have.

Q: Can I check the status of a running command?

Run Command provides the status of a command for each instance it is running on. All of this can be retrieved from the AWS CLI, SDK, or the EC2 Management Console.

State Manager

Q: What is State Manager?

State Manager automates the process of defining and maintaining a consistent configuration of OS and applications across your entire fleet of systems. For example – configuring and enforcing firewall policies, keeping anti-malware definitions current. Through reapplication of your configuration policies, State Manager ensures that your systems are always compliant with your enterprise policies.

Q: Why should I use State Manager?

Businesses are moving towards automated IT with applications across environments and locations, including on AWS and on-premises data centers. However, ensuring that the infrastructure powering your applications is consistent is a challenge. State Manager allows you create policies, reapply these policies to prevent configuration drift, and monitor the status of your intended state.

Q: How do I create my policies?

Policies can be easily created through Systems Manager Documents. In addition, you also have predefined configurations that you can use for installing applications, joining instances to domain and so on.

Q: What are the targets that can be configured?

You have the flexibility to target instances or tags. This means you have the flexibility to have specific configurations for groups of instances such as webservers.

Patch Manager

Q: What is Patch Manager?

Patch Manager is a new automation-focused patching service which makes it easy for customers to keep their Windows instances up to date. Patch Manager helps you streamline your patching process through the implementation of built in best practices, such as maintenance windows and dynamic patch approval policies.

Q: How do I specify when I would like to patch an instance?

You use Maintenance Windows to define when patching occurs. Maintenance Windows are a new feature of EC2 which provide you the ability to define one or more recurring windows of time during which it is acceptable for your own maintenance to occur. By defining these windows and associating your instances with them, it is easier for you to ensure that any maintenance activities you perform on your instances which may impact the availability of a workload is done so during a well-defined window of time. Maintenance windows make it easy to schedule when you would like your own Run Command tasks to occur.

Q: How do I customize the patching process?

Patch Manager leverages Run Command to provide a fully automated patching process. While Patch Manager provides a pre-built Run Command document, you can easily customize the patching process by writing your own Run Command Document. For example, you can stop an NT service before rolling out the patches.

Q: What types of patches can I install with the Patch Manager?

Patch Manager supports the patching of Windows and Linux based instances. Please visit our [documentation](http://docs.aws.amazon.com/systems-manager/latest/userguide/systems-manager-patch.html) to see the versions currently supported.

Q: How do I pick the patches I want to install?

Patch Manager provides you with the ability to create Patch Baselines, which define the set of patches you have approved or blocked for deployment to your instances. In a Patch Baseline, you can select patches by the products (e.g. Windows Server 2008, Windows Server 2012, etc.), categories (e.g. Critical Updates, Security Updates, etc.) and severities for which you would like to review patches for deployment. For each category selected, you can then define a schedule on which the contained patches will be automatically approved for distribution. In addition to the rules, you can also specify a whitelist and blacklist of patches which indicate patches which are to be installed or blocked respectively. At the time of patching, Patch Manager will assess targeted instances for only the patches that have been approved prior to that point in time.

Q: How do I track the compliance levels of my instances?

With Patch Manager you can view patch compliance information which tells you the detailed results of the patching process. From the EC2 Management Console or API you can easily get aggregate compliance details per instance. In addition, you drill in further and for each instance you can determine which patches are installed, missing, not applicable, and which failed to install.

Inventory

Q: What is Inventory?

The Inventory capability in EC2 Systems Manager provides visibility into an instance's software catalog and configuration. You can set up Inventory to gather detail on a variety of instance attributes such as installed applications, AWS components and agents, network configuration, OS details, and more. Then use the powerful query feature to assess compliance and identify instances in need of remediation across your fleet.

Q: Who should use Inventory?

IT administrators and devops professionals will find this capability useful in understanding the configuration and composition of their fleets. Users can quickly determine which instances are missing a patch or are running an outdated application version. Similarly admins can run licensing audits to understand software usage. The net result is that systems administrators are better able to troubleshoot issues and assess security posture.

Q: Can I customize the information gathered by the Inventory?

Yes, you can create your own custom Inventory types and effectively extend Inventory's schema. For example, you can configure your instance to gather additional OS and CIM details, or record items like rack location and in-service date for on-premise servers.

Q: How can I track changes to my configuration over time?

Using [AWS Config](https://aws.amazon.com/config/), you can monitor an instance's compliance with a desired configuration through Config Rules. This capability allows security experts and compliance auditors to have a complete audit trail of instance configuration changes, as well as receive proactive notifications in the event of non-compliance.

Automation

Q: What is Automation?

The Automation capability in EC2 Systems Manager simplifies the process of building and maintaining Amazon Machine Images (AMIs). This provides you a repeatable process to apply patches, application updates, and other changes to your AMIs.

Q: What tasks can I automate?

AMI maintenance is greatly simplified by Automation feature of Systems Manager, allowing you to patch, update agents, or bake-in applications using a streamlined, repeatable, and auditable process. Alternately, you can use Run Command and AWS Lambda in your workflows orchestrate the configuration and management of instances and other AWS resources at scale.

Parameter Store

Q: What is Parameter Store?

Parameter Store makes it easy for you to store, reference and control access to your configuration parameters and sensitive information such as connection strings, and administrator passwords.

Q: Why should I use Parameter Store?

You can use Parameter Store to quickly store and reference configuration and sensitive information. Rather than storing data in config files or referencing them in plain text, you can leverage Parameter Store to obtain this information in your applications or scripts. Additionally, you control who has access to parameters so that only the right set of users have access to the appropriate information.

Q: How do you store sensitive data?

A secure string is any sensitive data that needs to be stored and referenced in a secure manner. If you have data that you do not want users to reference in clear text or have access to data that can be tampered or misused, you should use secure strings in Parameter Store. You can encrypt your sensitive data using your own [AWS Key Management Service](https://aws.amazon.com/kms/) (KMS) key or your user account default key provided by KMS.

Q: What services can I reference my parameters?

You can easily reference your parameters across EC2 Systems Manager services such as Run Command, State Manager and Automation.

Q: Can I track usage and provide access control to specific parameters?

Yes, you can provide granular access control through customized permissions to users and resources (such as instances) for parameters access using [AWS IAM](https://aws.amazon.com/iam/). This means you can control who can access which parameter on what resource. Additionally, you can also track and audit parameter API calls using [AWS CloudTrail](https://aws.amazon.com/cloudtrail/).

Maintenance Windows

Q: What is a maintenance window?

Maintenance windows is a feature of EC2 Systems Manager which provide you the ability to define one or more recurring windows of time during which it is acceptable for any disruptive operation to occur. By defining these windows and associating your instances with them, it is easier for you to ensure that any maintenance activities you perform on your instances which may impact the availability of a workload is done so during a well-defined window of time.

Q: Why should I use maintenance window?

Maintenance windows help improve availability and reliability of your workloads by automatically performing tasks in a well-defined window of time, significantly reducing the impact of any operational or infrastructure failures.

Q: What types of tasks can I schedule in a maintenance window?

You can create and schedule any Run Command, Amazon EC2 Systems Manager Automation, AWS Step Functions, or AWS Lambda functions as tasks.

Q: What are the types of schedules I can choose for my maintenance windows?

Maintenance windows can be scheduled for a recurring date (e.g. Weekly on Tuesdays at 22:00:00 or 1st Sunday of every month at 22:00:00). You can define your schedule using cron or rate expression.

## **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 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.

#### **Manage Your AWS Resources**

[Sign in to the Console](https://console.aws.amazon.com/console/home)

Q: How is AWS CloudFormation different from AWS Elastic Beanstalk?

These services are designed to complement each other. [AWS Elastic Beanstalk](https://aws.amazon.com/elasticbeanstalk/) 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 resources](http://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/cfn-supported-resources.html). 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: How do I get started with AWS CloudFormation?

You can easily access AWS CloudFormation through the [AWS Management Console](http://aws.amazon.com/console/), which gives you a point-and-click, web-based interface to deploy and manage stacks. You can create a new stack from inside the [AWS Management Console](http://aws.amazon.com/console/) in a few simple steps:

1. Give the stack a name: Provide a unique name for the stack.
2. Select a template: Select a template from your local file system or from a Amazon S3 URL. This may be one of the sample AWS CloudFormation templates, your own custom template, a template you are managing in a source control repository, or a template you got from a third party.
3. Specify any parameters: If the template allows you to configure the deployment, fill in any parameters or go with the specified defaults.
4. Click "Create": Start the deployment. You can see the current state of the deployment, with all the resource names and stack events in the [AWS Management Console](http://aws.amazon.com/console/).

Q: What resources does AWS CloudFormation support?

To see a complete list of supported AWS resources and their features, visit the [Supported AWS Services](http://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/cfn-supported-resources.html) page in the Release History of the documentation.

AWS CloudFormation [custom resources](http://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/crpg-walkthrough.html) enable management of additional AWS and non-AWS 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](https://s3.amazonaws.com/cloudformation-examples/BoostrappingApplicationsWithAWSCloudFormation.pdf).

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](https://s3.amazonaws.com/cloudformation-examples/IntegratingAWSCloudFormationWithOpscodeChef.pdf).

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](https://s3.amazonaws.com/cloudformation-examples/IntegratingAWSCloudFormationWithPuppet.pdf).

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 [detail page](https://aws.amazon.com/cloudformation/details/). After signing up, please refer to the AWS CloudFormation [documentation](https://aws.amazon.com/documentation/cloudformation/), 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](https://aws.amazon.com/cloudformation/aws-cloudformation-templates-old/) 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.

## **Billing**

Q: How much does AWS CloudFormation cost?

There is no additional charge for AWS CloudFormation. You only pay for the AWS resources that are created (e.g., Amazon EC2 instances, Elastic Load Balancing load balancers etc.)

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](https://aws.amazon.com/contact-us/cloudformation-request/), 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](http://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/using-cfn-endpoints.html).

Q: What are the AWS regions where AWS CloudFormation is currently available?

Please refer to [Regional Products and Services](https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/) for details of CloudFormation availability by region.

Q: What is AWS CloudTrail?  
AWS CloudTrail is a web service that records activity made on your account and delivers log files to your Amazon S3 bucket.

Q: What are the benefits of CloudTrail?   
CloudTrail provides visibility into user activity by recording actions taken on your account. CloudTrail records important information about each action, including who made the request, the services used, the actions performed, and parameters for the action, and the response elements returned by the AWS service. This information helps you to track changes made to your AWS resources and to troubleshoot operational issues. CloudTrail makes it easier to ensure compliance with internal policies and regulatory standards.For more details, refer to the AWS compliance white paper “[Security at scale: Logging in AWS](https://d1.awsstatic.com/whitepapers/aws-security-at-scale-logging-in-aws.ceacfb1e205cd35d0137b099ed6abff25ed4009e.pdf)”.

Q: Who should use CloudTrail?  
Customers who need to track changes to resources, answer simple questions about user activity, demonstrate compliance, troubleshoot, or perform security analysis 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?  
No, nothing is required to begin viewing your account activity. You can visit the [AWS CloudTrail console](https://console.aws.amazon.com/cloudtrail/) or AWS CLI and begin viewing up to the past 7 days of account activity.

Q: Does the CloudTrail Event History show all account activity within my account?  
AWS CloudTrail will only show the results of the CloudTrail Event History for the current region you are viewing for the last 7 days and support the AWS services found [here.](http://docs.aws.amazon.com/awscloudtrail/latest/userguide/view-cloudtrail-events-supported-services.html) These events are limited to Management Events with create, modify, and delete API calls and account activity. For a complete record of account activity, including all Management Events, Data Events, and read-only activity, you’ll need to configure a CloudTrail trail.

Q: What search filters can I use to view my account activity?  
You can specify Time range and one of the following attributes: Event name, User name, Resource name, Event source, Event ID, and Resource type.

Q: Can I use the lookup-events CLI command even if I don’t have a trail configured?  
Yes, you can visit the [CloudTrail console](https://console.aws.amazon.com/cloudtrail/) or use the CloudTrail API/CLI and begin viewing the past 7 days of account activity.

Q: What additional CloudTrail features are available by setting up CloudTrail and creating a trail?   
By setting up a CloudTrail trail you can deliver your CloudTrail events to Amazon S3, Amazon CloudWatch Logs, and Amazon CloudWatch Events. This enables you to leverage features to help you archive, analyze, and respond to changes in your AWS resources.

Q: Can I restrict access for users in my account from seeing the CloudTrail Event History?  
Yes, CloudTrail integrates with [AWS Identity and Access Management](https://aws.amazon.com/iam/) (IAM), which allows you to control access to CloudTrail and to other AWS resources that CloudTrail requires, including the ability to restrict permissions to view and search account activity. This is accomplished by removing the "cloudtrail:LookupEvents" from the Users IAM policy which will then prevent that IAM user from viewing account activity.

Q: Is there any cost associated with CloudTrail Event History being enabled on my account upon creation?   
There is no cost for viewing or searching account activity with CloudTrail Event History.

Q: Can I turn CloudTrail Event History off for my account?  
For any CloudTrail trails that you have created, you can stop logging or delete the trails which will also stop the delivery of account activity to the S3 bucket you had designated as part of your trail configuration as well as delivery to CloudWatch Logs if configured. Account activity for the past 7 days will still be collected and visible within the CloudTrail console and through the AWS CLI.

## **Services and Region Support**

Q: What services are supported by CloudTrail?  
AWS CloudTrail records account activity and service events from most AWS services. For the list of supported services, see [CloudTrail Supported Services](http://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-supported-services.html) in the *CloudTrail User Guide.*

Q: Are API calls made from the AWS Management Console recorded?   
Yes. CloudTrail records API calls made from any client. The AWS Management Console, AWS SDKs, command line tools, and higher level AWS services call AWS APIs, so these calls are recorded.

Q: Where are my log files stored and processed before they are delivered to my Amazon S3 bucket?  
Activity information for services with regional end points (EC2, RDS etc.) is captured and processed in the same region as to which the action is made and delivered to the region associated with your Amazon S3 bucket. Action information for services with single end points (IAM, STS, etc.) is captured in the region where the end point is located, processed in the region where the CloudTrail trail is configured and delivered to the region associated with your Amazon S3 bucket.

## **Applying a Trail to all Regions**

Q: What is applying a trail to all regions?  
Applying a trail to all regions refers to creating a trail that will record AWS account activity in all regions. This setting also applies to any new regions that are added. For more details on regions and partitions, refer to the [Amazon Resource Names and AWS Service Namespaces page.](http://docs.aws.amazon.com/general/latest/gr/aws-arns-and-namespaces.html)

Q: What are the benefits of applying a trail to all regions?  
You can create and manage a trail across all regions in the partition in one API call or few clicks. You will receive a record of account activity made in your AWS account across all regions to one S3 bucket or CloudWatch logs log group. When AWS launches a new region, you will receive the log files containing event history for the new region without taking any action.

Q: How do I apply a trail to all regions?   
In the CloudTrail console, you select yes to apply to all regions in the trail configuration page. If you are using the SDKs or AWS CLI, You set the IsMultiRegionTrail to true.

Q: What happens when I apply a trail to all regions?   
Once you apply a trail in all regions, CloudTrail will create a new trail in all regions by replicating the trail configuration. CloudTrail will record and process the log files in each region and will deliver log files containing account activity across all AWS regions to a single S3 bucket and a single CloudWatch Logs log group. If you specified an optional SNS topic, CloudTrail will deliver SNS notifications for all log files delivered to a single SNS topic.

Q: Can I apply an existing trail to all regions?  
Yes. You can apply an existing trail to all regions. When you apply an existing trail to all regions, CloudTrail will create a new trail for you in all regions. If you previously created trails in other regions, you can view, edit and delete those trails from the [CloudTrail console](https://console.aws.amazon.com/cloudtrail/home).

Q: How long will it take for CloudTrail to replicate the trail configuration to all regions?  
Typically, it will take less than 30 seconds to replicate the trail configuration to all regions.

## **Multiple Trails**

Q: How many trails can I create in an AWS region?   
You can create up to five trails in an AWS region. A trail that applies to all regions exists in each region and is counted as one trail in each region.

Q: What is the benefit of creating multiple trails in an AWS region?   
With multiple trails, different stakeholders such as security administrators, software developers and IT auditors can create and manage their own trails. For example, a security administrator can create a trail that applies to all regions and configure encryption using one KMS key. A developer can create a trail that applies to one region for troubleshooting operational issues.

Q: Does CloudTrail support resource level permissions?  
Yes. Using resource level permissions, you can write granular access control policies to allow or deny access to specific users for a particular trail. For more details, go to CloudTrail [documentation](http://docs.aws.amazon.com/awscloudtrail/latest/userguide/grant-custom-permissions-for-cloudtrail-users.html#grant-custom-permissions-for-cloudtrail-users-resource-level ).

## **Security and Expiration**

Q: How can I secure my CloudTrail log files?  
By default, CloudTrail log files are encrypted using S3 Server Side Encryption (SSE) and placed into your S3 bucket. You can control access to log files by applying IAM or S3 bucket policies. You can add an additional layer of security by enabling S3 [Multi Factor Authentication (MFA) Delete](http://docs.aws.amazon.com/AmazonS3/latest/dev/MultiFactorAuthenticationDelete.html) on your S3 bucket. For more details on creating and updating a trail, see the [CloudTrail documentation](http://docs.aws.amazon.com/awscloudtrail/latest/userguide/setupyourtrail.html).

Q: Where can I download a sample S3 bucket policy and an SNS topic policy?   
You can download a sample [S3 bucket policy](https://awscloudtrail.s3.amazonaws.com/policy/S3/AWSCloudTrail-S3BucketPolicy-2013-11-01.json) and an [SNS topic policy](https://awscloudtrail.s3.amazonaws.com/policy/SNS/AWSCloudTrail-SnsTopicPolicy-2013-11-01.json) from CloudTrail S3 bucket. You need to update the sample policies with your information before you apply them to your S3 bucket or SNS topic.

Q: How long can I store my activity log files?   
You control the retention policies for your CloudTrail log files. By default, log files are stored indefinitely. You can use [Amazon S3 object lifecycle management rules](http://docs.aws.amazon.com/AmazonS3/latest/dev/object-lifecycle-mgmt.html) to define your own retention policy. For example, you may want to delete old log files or archive them to Amazon Glacier.

## **Event Payload, Timeliness and Delivery Frequency**

Q: What information is available in an event?   
An event contains information about the associated activity: who made the request, the services used, the actions performed, and parameters for the action, and the response elements returned by the AWS service. For more details, see the [CloudTrail Event Reference](http://docs.aws.amazon.com/awscloudtrail/latest/userguide/eventreference.html)section of the user guide.

Q: How long does it take CloudTrail to deliver an event for an API call?   
Typically, CloudTrail delivers an event within 15 minutes of the API call.

Q: How often will CloudTrail deliver log files to my Amazon S3 bucket?   
CloudTrail delivers log files to your S3 bucket approximately every 5 minutes. CloudTrail does not deliver log files if no API calls are made on your account.

Q: Can I be notified when new log files are delivered to my Amazon S3 bucket?   
Yes. You can turn on Amazon SNS notifications so that you can take immediate action on delivery of new log files.

Q: What happens if CloudTrail is turned on for my account but my Amazon S3 bucket is not configured with the correct policy?   
CloudTrail log files are delivered in accordance with the S3 bucket policies that you have in place. If the bucket policies are misconfigured, CloudTrail will not be able to deliver log files.

## **Data Events**

Q: What are Data Events?   
Data Events represent API activity for data resources that you specify. If you enable Data Events for S3 Objects, you will capture all API activity, including calls such as Get Object and Get Object ACL, in CloudTrail. You can use these events to meet IT auditing or compliance requirements, perform security analysis, monitor for and alarm on specific patterns of user behavior in your AWS account or take immediate action on any object level API activity using [CloudWatch Events](http://docs.aws.amazon.com/AmazonCloudWatch/latest/events/WhatIsCloudWatchEvents.html).

Q: How can I consume Data Events?  
Data Events that are recorded by CloudTrail are delivered to S3, similar to Management Events. Once enabled, these events are also available in CloudWatch Events.

Q: What are S3 Data Events? How do I record them?  
S3 Data Events represent API activity on S3 Objects. To get CloudTrail to record these actions, you specify an S3 Bucket in the Event Selector. Any API actions on the objects within the specified S3 Bucket are recorded by CloudTrail.

## **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?   
Yes. You can configure one S3 bucket as the destination for multiple accounts. For detailed instructions, refer to [aggregating log files to a single Amazon S3 bucket section](http://docs.aws.amazon.com/awscloudtrail/latest/userguide/aggregatinglogs.html) of the AWS CloudTrail User Guide.

## **Integration with CloudWatch Logs**

Q: What is CloudTrail integration with CloudWatch Logs?  
CloudTrail integration with CloudWatch Logs delivers management and Data Events captured by CloudTrail to a CloudWatch Logs log stream in the CloudWatch Logs log group you specify.

Q: What are the benefits of CloudTrail integration with CloudWatch Logs?  
This integration enables you to receive SNS notifications of account activity captured by CloudTrail. For example, you can create CloudWatch alarms to monitor API calls that create, modify and delete Security Groups and Network ACL’s.

Q: How do I turn on CloudTrail integration with CloudWatch Logs?  
You can turn on CloudTrail integration with CloudWatch Logs from the CloudTrail console by specifying a CloudWatch Logs log group and an IAM role. You can also use the AWS SDKs or the AWS CLI to turn on this integration.

Q: What happens when I turn on CloudTrail integration with CloudWatch Logs?  
After you turn on the integration, CloudTrail continuously delivers account activity to a CloudWatch Logs log stream in the CloudWatch Logs log group you specified. CloudTrail also continues to deliver logs to your Amazon S3 bucket as before.

Q: In which AWS regions is CloudTrail integration with CloudWatch Logs supported?  
This integration is supported in the regions where CloudWatch Logs is supported. For more information, see [Regions and Endpoints](http://docs.aws.amazon.com/general/latest/gr/rande.html#cwl_region) in the *Amazon Web Services General Reference.*

Q: How does CloudTrail deliver events containing account activity to my CloudWatch Logs?  
CloudTrail assumes the IAM role you specify to deliver account activity to CloudWatch Logs. You limit the IAM role to only the permissions it requires to deliver events to your CloudWatch Logs log stream. To review IAM role policy, go to the [user guide](http://docs.aws.amazon.com/awscloudtrail/latest/userguide/cw_role_policy.html) of the CloudTrail documentation.

Q: What charges do I incur once I turn on CloudTrail integration with CloudWatch Logs?  
After you turn on CloudTrail integration with CloudWatch Logs, you incur standard CloudWatch Logs and CloudWatch charges. For details, go to CloudWatch [pricing](http://aws.amazon.com/cloudwatch/pricing/) page.

## **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?   
CloudTrail log file encryption using [SSE-KMS](http://docs.aws.amazon.com/AmazonS3/latest/dev/UsingKMSEncryption.html) allows you to add an additional layer of security to CloudTrail log files delivered to an Amazon S3 bucket by encrypting the log files with a KMS key. By default, CloudTrail will encrypt log files delivered to your Amazon S3 bucket using Amazon S3 server-side encryption.

Q: I have an application that ingests and processes CloudTrail log files. Do I need to make any changes to my application?   
With SSE-KMS, Amazon S3 will automatically decrypt the log files so that you do not need to make any changes your application. As always, you need to make sure that your application has appropriate permissions, i.e. Amazon S3 GetObject and KMS Decrypt permissions.

Q: How do I configure CloudTrail log file encryption?  
You can use the AWS Management Console, or AWS CLI or the AWS SDKs to configure log file encryption. For detailed instructions, refer to the [documentation](http://docs.aws.amazon.com/awscloudtrail/latest/userguide/encrypting-cloudtrail-log-files-with-aws-kms.html).

Q: What charges do I incur once I configure encryption using SSE-KMS?  
Once you configure encryption using SSE-KMS, you will incur standard AWS KMS charges. For details,go to [AWS KMS pricing](https://aws.amazon.com/kms/pricing/) page.

## **CloudTrail Log File Integrity Validation**

Q: What is CloudTrail log file integrity validation?  
CloudTrail log file integrity validation feature allows you to determine whether a CloudTrail log file was unchanged, deleted, or modified since CloudTrail delivered it to the specified Amazon S3 bucket.

Q: What is the benefit of CloudTrail log file integrity validation?   
You can use the log file integrity validation as an aid in your IT security and auditing processes.

Q: How do I enable CloudTrail log file integrity validation?   
You can enable the CloudTrail log file integrity validation feature from the AWS Management Console, AWS CLI or AWS SDKs.

Q: What happens once I turn on the log file integrity validation feature?  
Once you turn on the log file integrity validation feature, CloudTrail will deliver digest files on an hourly basis. The digest files contain information about the log files that were delivered to your Amazon S3 bucket, hash values for those log files, digital signatures for the previous digest file, and the digital signature for the current digest file in the Amazon S3 metadata section. For more information about digest files, digital signatures and hash values, go to [CloudTrail documentation.](http://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-log-file-validation-intro.html)

Q: Where are the digest files delivered to?   
The digest files are delivered to the same Amazon S3 bucket where your log files are delivered to. However, they are delivered to a different folder so that you can enforce granular access control policies. For details, refer to the digest file structure section of the [CloudTrail documentation.](http://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-log-file-validation-digest-file-structure.html)

Q: How can I validate the integrity of a log file or digest file delivered by CloudTrail?   
You can use the AWS CLI to validate that the integrity of log file or digest file. You can also build your own tools to do the validation. For more details on using the AWS CLI for validating the integrity of a log file, refer to the [CloudTrail documentation.](http://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-log-file-validation-cli.html)

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?   
Yes. CloudTrail will deliver the digest files across all regions and multiple accounts into the same Amazon S3 bucket.

## **AWS CloudTrail Processing Library**

Q: What is AWS CloudTrail Processing Library?  
AWS CloudTrail Processing Library is a Java library that makes it easy to build an application that reads and processes CloudTrail log files. You can download CloudTrail Processing Library from [GitHub](https://github.com/aws/aws-cloudtrail-processing-library).

Q: What functionality does CloudTrail Processing Library provide?  
CloudTrail Processing Library provides functionality to handle tasks such as continuously polling a SQS queue, reading and parsing SQS messages, downloading log files stored in S3, parsing and serializing events in the log file in a fault tolerant manner. For more information, go to the [user guide section](http://docs.aws.amazon.com/awscloudtrail/latest/userguide/using_processing_lib.html) of the CloudTrail documentation.

Q: What software do I need to start using the CloudTrail Processing Library?  
You need aws-java-sdk version 1.9.3 and Java 1.7 or higher.

## **Pricing**

Q: Does CloudTrail provide a free tier?   
AWS CloudTrail allows you to view and download the last 7 days of your account activity for create, modify, and delete operations of supported services free of charge.

There is no charge from AWS CloudTrail for creating a trail and the first copy of Management Events within each region is delivered to your trail free of charge. You will be charged for any Data Events or additional copies of Management Events recorded in that region, per the published [pricing plan](http://aws.amazon.com/cloudtrail/pricing/).

Q: If I have only one trail with Management Events, and apply it to all regions, will I incur charges?  
No. The first copy of Management Events is delivered free of charge in each region.

Q: If I enable Data Events on an existing trail with free Management Events, will I get charged?  
Yes. You will only be charged for the Data Events. The first copy of Management Events is delivered free of charge.

## **Partners**

Q: How do the AWS partner solutions help me analyze the events recorded by CloudTrail?  
Multiple partners offer integrated solutions to analyze CloudTrail log files. These solutions include features like change tracking, troubleshooting, and security analysis. For more information, see the [CloudTrail partners](http://aws.amazon.com/cloudtrail/partners/) section.

## **Other**

Q: Will turning on CloudTrail impact the performance of my AWS resources, or increase API call latency?  
No. Turning on CloudTrail has no impact on performance of your AWS resources or API call latency.

### **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 5.5, 5.3, 5.1, 2.3, and 1.5.

#### **Manage Your AWS Resources**

[Sign in to the Console](https://console.aws.amazon.com/console/home)

Q: What is an Amazon Elasticsearch domain?

Amazon Elasticsearch domains are Elasticsearch clusters created using the Amazon Elasticsearch Service console, API, or CLI. Each domain is an Elasticsearch cluster in the cloud with the compute and storage resources you specify. You can create and delete domains, define/refine infrastructure attributes of your domains, and control access and security. You can run one or more Amazon Elasticsearch 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 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 that powers your Amazon Elasticsearch instance. Amazon Elasticsearch Service is integrated 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 managing your domain based on your application needs.

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

Amazon Elasticsearch Service supports many of the commonly used Elasticsearch APIs, so code, applications, and popular tools that you're already using with your current Elasticsearch environments will work seamlessly. For a full list of supported Elasticsearch operations, see our [documentation](https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-gsg-supported-operations.html).

[Return to Top >>](https://aws.amazon.com/elasticsearch-service/faqs/#faqs)

### **Setup and Configuration**

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

Yes. You can create a new Amazon Elasticsearch domain with the Domain Creation Wizard in the console. 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 domains using the console.

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

Yes. [AWS CloudFormation](https://aws.amazon.com/cloudformation/) supports Amazon ES. For more information, see the [CloudFormation Template Reference](http://docs.aws.amazon.com/AWSCloudFormation/latest/UserGuide/aws-resource-elasticsearch-domain.html) documentation.

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

Yes. You can configure the dedicated master setting for an Amazon Elasticsearch domain. When choosing a dedicated master setup, you will be able to also set up the instance type and instance count for the dedicated master configuration.

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

Yes. You can create multiple Elasticsearch indices within the same Amazon Elasticsearch domain. Elasticsearch will manage storing of each index and any associated replica among 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, which is 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 set up your Amazon Elasticsearch domain as the backend store for all logs coming through your Logstash implementation.
* You can use native Elasticsearch APIs, such as PUT and \_bulk to load data into your domain.

If you use Amazon Kinesis Firehose, you can grant access using an IAM role that allows Firehose to perform the following actions against a particular domain: es:ESHttpPost, es:ESHttpPut, es:DescribeElasticsearchDomain, es:DescribeElasticsearchDomains, es:DescribeElasticsearchDomainConfig. If you use Logstash or Elasticsearch APIs, you can set up access control on your Amazon Elasticsearch domain to use request signing to authenticate API calls; or use resource based IAM policies to include IP addresses of instances that issue API calls.

Q: Does Amazon Elasticsearch Service support integration with Logstash?

Yes. Amazon Elasticsearch Service supports integration with Logstash. You can set up your Amazon Elasticsearch domain as the backend store for all logs coming through your Logstash implementation. You can set up access control on your Amazon Elasticsearch 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 built in support for Kibana that is deployed with your Amazon Elasticsearch 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](https://www.elastic.co/guide/en/kibana/4.0/index.html) from Elastic.co.

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

Customers will be able to choose between local on-instance storage available with the provisioned Amazon Elasticsearch instances or EBS volumes to store their Elasticsearch indices. During domain creation, if a customer selects the EBS storage option, the Domain Creation Wizard will prompt the customer to specify the type and size of EBS volume to be allocated to the domain. Customers will be able to modify the EBS settings after domain creation as well to increase or decrease the size and modify the volume type as needed.

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

Customers will be able to 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 domain?

Yes. Amazon Elasticsearch Service supports 1 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 100 instances per domain by creating a case with the [AWS Support Center](https://console.aws.amazon.com/support/home#/). With 100 instances, you can allocate about 150 TB of EBS storage to a single domain.

[Return to Top >>](https://aws.amazon.com/elasticsearch-service/faqs/#faqs)

### **Administration**

Q: How do I control access to Amazon Elasticsearch Service?

Customers will be able to set up IAM policies to control access to their Amazon Elasticsearch domains and sub resources like indices within the domains. IAM policies can also be set up to control access to the control plane API for operations like creating and scaling clusters and data plane API for operations like uploading documents and executing Elasticsearch requests.

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

Yes. You can set up IAM policies to allow programs running on servers outside of AWS to access your Amazon Elasticsearch domains. You can use either IP based access policies or resource based access policies with signed requests to access the Amazon Elasticsearch domains. [Click here](http://docs.aws.amazon.com/general/latest/gr/signature-version-4.html) for more information about signed requests.

Q: How can I migrate data from my existing Elasticsearch cluster to my new Amazon Elasticsearch 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 domain and load data from the snapshot into the newly created Amazon Elasticsearch domain using the Elasticsearch restore API.

Q: How can I scale an Amazon Elasticsearch domain?

Amazon Elasticsearch Service allows you to control the scaling of your Amazon Elasticsearch domains using the console, API, and CLI. You can scale your Amazon Elasticsearch 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 domains to enable you to make appropriate scaling decisions for your domains.

Q: Does scaling my Amazon Elasticsearch domain require downtime?

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

Q: What options does Amazon Elasticsearch Service provide for node failures?

Amazon Elasticsearch Service automatically detects node failures and replaces the node. The service will acquire new instances, and will then redirect Elasticsearch requests and document updates to the new instances. In the event that the node cannot be replaced, customers will be able to use any snapshots they have of their cluster to restart the domain with preloaded data.

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

Yes. Customers can enable Zone Awareness for their Amazon Elasticsearch domains either at domain creation time or by modifying a live domain. When Zone Awareness is enabled, Amazon Elasticsearch Service will distribute the instances supporting the domain across two different Availability Zones. Then, if replication is enabled in the Elasticsearch engine, Elasticsearch will allocate replicas of the domain across these different instances enabling cross-zone replication.

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](https://aws.amazon.com/cloudtrail/), and turn it on via [CloudTrail's AWS Management Console home page](https://console.aws.amazon.com/cloudtrail/home).

Q: What is a snapshot?

A snapshot is a copy of your Amazon Elasticsearch 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 domain with preloaded data or to create a new Amazon Elasticsearch 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 will automatically create daily snapshots of each Amazon Elasticsearch domain. The daily snapshots are setup to occur between midnight and 1AM UTC. Customers will also be able to modify the timing of the automated snapshot to better suit their needs.

Q: Can I change the default settings for the automated daily snapshot provided by Amazon Elasticsearch Service?

Yes. You will be able to change the timing of the automated daily snapshot to suit your application schedule.

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

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

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

There is no additional charge for the automated daily 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 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 domain and load data from the snapshot into the newly created Amazon Elasticsearch domain using the Elasticsearch restore API.

Q: What happens to my snapshots when I delete my Amazon Elasticsearch 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 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](https://www.elastic.co/guide/en/elasticsearch/reference/current/index-modules-slowlog.html).

Q: How can I enable slow logs on Amazon ES?

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](http://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-createupdatedomains.html#es-createdomain-configure-slow-logs).

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](http://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-createupdatedomains.html#es-createdomain-configure-slow-logs).

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

No. Turning on slow logs in Amazon ES 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](http://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-createupdatedomains.html#es-createdomain-configure-slow-logs).

Q: If I turn off the Slow Logs in Amazon ES, 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](http://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-createupdatedomains.html#es-createdomain-configure-slow-logs).

Q: Can I adjust the granularity of logging?

Yes. Elasticsearch exposes multiple levels of logging. You need to set the appropriate level in the configuration for your index. For more details on setting the index configuration for enabling slow logs, see our [documentation](http://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-createupdatedomains.html#es-createdomain-configure-slow-logs).

Q: Does enabling Slow Logs cost me anything?

When Slow Logs are enabled, Amazon ES starts publishing the generated logs to Amazon CloudWatch Logs. Amazon ES does not charge for enabling slow logs. However, [standard CloudWatch charges](https://aws.amazon.com/cloudwatch/pricing/) apply.

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](http://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-createupdatedomains.html#es-createdomain-configure-slow-logs).

Q: How can I consume logs from CloudWatch Logs?  
CloudWatch offers multiple ways to consume logs. You can [view log data](http://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/ViewingLogData.html), [export it to S3](http://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/S3Export.html), or [process it in real time](http://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/Subscriptions.html). To learn more, see the [CloudWatch Logs developer guide](http://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/WhatIsCloudWatchLogs.html).

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

Yes. slow logs can be enabled for all versions of Elasticsearch supported by Amazon ES. However, there are slight differences in the way log settings can be specified for each version of Elasticsearch. Please refer to our [documentation](http://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-createupdatedomains.html#es-createdomain-configure-slow-logs) 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.

[Return to Top >>](https://aws.amazon.com/elasticsearch-service/faqs/#faqs)

### **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 instance hours – Based on the class (e.g. Standard Small, Large, Extra Large) of the Amazon Elasticsearch instance consumed. Partial Amazon Elasticsearch instance hours consumed are billed as full hours.
* Storage (per GB per month) – EBS Storage capacity you have provisioned to your Amazon Elasticsearch instance. If you scale your provisioned storage capacity within the month, your bill will be pro-rated.
* Provisioned IOPS per month – 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](https://aws.amazon.com/elasticsearch-service/pricing/) for detailed pricing information.

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

Billing commences for an Amazon Elasticsearch instance as soon as the instance is available. Billing continues until the Amazon Elasticsearch 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 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 instance, you must delete the domain to avoid being billed for additional instance hours. Partial Amazon Elasticsearch instance hours consumed are billed as full hours.

## **General**

### **What is AWS Key Management Service (KMS)?**

AWS KMS is a managed encryption service that enables you to easily encrypt your data. AWS KMS provides a highly available key storage, management, and auditing solution for you to encrypt your data across AWS services and within your own applications.

### **Why should I use AWS KMS?**

If you are a developer who needs to encrypt data in your applications, you should use the AWS SDKs with AWS KMS support to easily use and protect encryption keys. If you’re an IT administrator looking for a scalable key management infrastructure to support your developers and their growing number of applications, you should use AWS KMS to reduce your licensing costs and operational burden. If you’re responsible for proving data security for regulatory or compliance purposes, you should use AWS KMS to verify that data is encrypted consistently across the applications where it is used and stored.

### **How do I get started with AWS KMS?**

The easiest way is to get started using AWS KMS is to check the box to encrypt your data within supported AWS services and use the default keys that are created in your account for each service. If you want further controls over the management of these keys, you can create keys in AWS KMS and assign them to be used in the supported AWS services when creating encrypted resources as well as use them directly within your own applications. AWS KMS can be accessed from the “Encryption Keys” section of the AWS Identity and Access Management (IAM) console for web-based access, and the AWS KMS Command Line Interface or AWS Software Development Kit for programmatic access. Visit the [Getting Started](https://aws.amazon.com/kms/getting-started/) page to learn more.

### **In what Regions is KMS available?**

Availability is listed on our global [Products and Services by Region](https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/) page.

### **What key management features are available in AWS KMS?**

You can perform the following key management functions in AWS KMS:

* Create keys with a unique alias and description
* Import your own keys
* 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 AWS KMS automatically rotate your keys on an annual basis
* Temporarily disable keys so they cannot be used by anyone
* Re-enable disabled keys
* Delete keys that you no longer use
* Audit use of keys by inspecting logs in AWS CloudTrail

### **How does AWS KMS work?**

AWS KMS allows you to centrally manage and securely store your keys. You can generate keys in KMS or import them from your key management infrastructure. These keys can be used from within your applications and supported AWS services to protect your data, but the key never leaves KMS AWS. You submit data to AWS KMS to be encrypted, or decrypted, under keys that you control. You set usage policies on these keys that determine which users can use them to encrypt and decrypt data. All requests to use these keys are logged in AWS CloudTrail so you can understand who used which key when.

### **Where is my data encrypted if I use AWS KMS?**

You can use AWS KMS to help encrypt data locally in your own applications or have it encrypted within a supported AWS service. You can use an AWS SDK with AWS KMS support to do the encryption wherever your applications run. You can also request a supported AWS service to encrypt your data as it is being stored. AWS CloudTrail provides access logs to allow you to audit how your keys were used in either situation.

### **Which AWS cloud services are integrated with AWS KMS?**

AWS Key Management Service is seamlessly integrated with several other AWS services to make encrypting data in those services as easy as checking a box and selecting the master key you want to use. See the [Product Details](https://aws.amazon.com/kms/details/#integration) page for the list of AWS services currently integrated with KMS. All use of your keys within integrated services appears in AWS CloudTrail logs. See the [AWS KMS Developer’s Guide](https://docs.aws.amazon.com/kms/latest/developerguide/service-integration.html) for more information on how integrated services use AWS KMS.

### **How do AWS cloud services use my keys to encrypt data?**

AWS cloud services integrated with AWS KMS use a method called envelope encryption to protect your data. Envelope encryption is an optimized method for encrypting data that uses two different keys. A data key is generated and used by the AWS service to encrypt each piece of data or resource. The data key is encrypted under a master key that you define in AWS KMS. The encrypted data key is then stored by the AWS service. When you need your data decrypted by the AWS service, the encrypted data key is passed to AWS KMS and decrypted under the master key that was originally encrypted under so the service can then decrypt your data.

### **Why use envelope encryption? Why not just send data to AWS KMS to encrypt directly?**

While AWS KMS does support sending data less than 4 KB to be encrypted, envelope encryption can offer significant performance benefits. When you encrypt data directly with KMS it must be transferred over the network. Envelope encryption reduces the network load for your application or AWS cloud service. Only the request and fulfillment of the data key through KMS must go over the network. Since the data key is always stored in encrypted form, it is easy and safe to distribute that key where you need it to go without worrying about it being exposed. Encrypted data keys are sent to AWS KMS and decrypted under master keys to ultimately allow you to decrypt your data. The data key is available directly in your application without having to send the entire block of data to AWS KMS and suffer network latency.

### **What’s the difference between a key I create vs. default master keys created for me for use within AWS cloud services?**

You have the option of selecting a specific master key to use when you want an AWS service to encrypt data on your behalf. A default master key specific to each service is created in your account as a convenience the first time you try to create an encrypted resource. This key is managed by AWS KMS but you can always audit its use in AWS CloudTrail. You can alternately create a customer master key in AWS KMS that you can then use in your own applications or from within a supported AWS service. AWS will update the policies on default master keys as needed to enable new features in supported services automatically. AWS does not modify policies on keys you create.

### **Why should I create a customer master key?**

Creating a key in AWS KMS gives you more control than you have with default service master keys. When you create a customer master key, you can choose to use key material generated by KMS on your behalf or import your own key material, define an alias, a description, and opt-in to have the key automatically rotated once per year if it backed by key material generated by KMS. You also can define permissions on the key to control who can use and manage the key. Management and usage activity related to the key is available for audit in AWS CloudTrail.

### **Can I import keys into KMS?**

Yes. You can import a copy of your key from your own key management infrastructure to KMS and use it with any integrated AWS service or from within your own applications.

### **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 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 delete the imported copy of the key on demand from AWS infrastructure once you no longer need the key.

### **What type of keys can I import?**

You can import 256-bit symmetric keys.

### **How is the key that I import into KMS protected in transit?**

During the import, your key must be wrapped by a KMS-provided public key using one of the two RSA PKCS#1 schemes. This ensures that your encrypted key can only be decrypted by KMS.

### **What’s the difference between a key I import vs. a key generated for me by KMS?**

There are two main differences between a key that you import vs. a key created for you by KMS:

1. You must securely maintain a copy of your imported keys in your key management infrastructure so that you can re-import them at any time. AWS ensures the availability, security, and durability of keys generated by KMS on your behalf until you schedule the keys for deletion.
2. You may set an expiration period for an imported key to automatically delete the key from KMS after the expiration period. You may also delete an imported key on demand without deleting the underlying customer master key. Further, you can manually disable or delete a customer master key with an imported key at any time. A key generated by KMS can only be disabled or scheduled for deletion, it cannot have an expiration time placed on it.

### **Can I rotate my keys?**

Yes. You can choose to have KMS automatically rotate keys generated by KMS on your behalf every year. Automatic key rotation is not supported for imported keys. If you choose to import keys to KMS, you can manually rotate them whenever you want.

### **Do I have to re-encrypt my data after keys in AWS KMS are rotated?**

If you choose to have KMS automatically rotate keys generated by KMS on your behalf, you don’t have to re-encrypt your data. AWS KMS 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 keys, you may have to re-encrypt your data depending on your application’s configuration.

### **Can I delete a key from AWS KMS?**

Yes. You can schedule a customer master key and associated metadata that you created in 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 the 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 gets 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.

### **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 KMS under the original customer master key so it can be used.

### **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.

### **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, and the Amazon S3 Encryption Client to facilitate encryption of data within your own applications wherever they run. AWS SDK in the Java, Ruby, .NET, and PHP platforms support AWS KMS APIs. Visit the [Developing on AWS](https://aws.amazon.com/developers/getting-started/) website for more information.

### **Is there a limit to the number of keys I can create in AWS KMS?**

You can create up to 1000 customer master keys per account per region. As both enabled and disabled customer master keys count towards the limit, we recommend deleting disabled keys that you no longer use. Default master keys 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 master key and used in your application or by AWS services to encrypt data on your behalf. You may request a limit increase for customer master keys by visiting the [AWS Support Center](https://aws.amazon.com/premiumsupport/).

## **Billing**

### **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 customer master keys 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](https://aws.amazon.com/kms/pricing/).

### **Is there a free tier?**

Yes. With the [AWS Free Usage Tier](https://aws.amazon.com/free/) you can get started with AWS KMS for free in all regions. Default master keys created on your behalf are free to store in your account. There is a free tier for usage as well that provides a free number of requests to AWS KMS each month. For current information on pricing, including the free tier, please visit the [AWS KMS pricing page](https://aws.amazon.com/kms/pricing/).

### **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](https://aws.amazon.com/c-tax-faqs/).

## **Security**

### **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.

### **Can AWS employees access my keys in AWS KMS?**

AWS KMS is designed so that no one has access to your master keys. The service is built on systems that are designed to protect your master keys with extensive hardening techniques such as never storing plaintext master keys on disk, not persisting them in memory, and limiting which systems can connect to the device. All access to update software on the service is controlled by a multi-level approval process that is audited and reviewed by an independent group within Amazon.

More details about these security controls can be found in the [AWS KMS Cryptographic Details whitepaper](https://d1.awsstatic.com/whitepapers/KMS-Cryptographic-Details.7d90f34ba02a50805cefbafad3d35edba3b4cb29.pdf). In addition, you can request a copy of the Service Organization Controls (SOC) report available from [AWS Compliance](https://aws.amazon.com/compliance/) to learn more about security controls AWS uses to protect your data and master keys.

### **Can I use KMS to help me comply with the encryption and key management requirements in the Payment Card Industry Data Security Standard (PCI DSS 3.1)?**

Yes. 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 3.1).

For more details on PCI DSS compliant services in AWS, you can read the [PCI DSS FAQs](https://aws.amazon.com/compliance/pci-dss-level-1-faqs/).

### **How does AWS KMS secure the data keys I export and use in my application?**

You can request that AWS KMS generate data keys that can be returned 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 used in encrypting the data key.

### **What length of keys does AWS KMS generate?**

Master keys in AWS KMS are 256-bits in length. Data keys can be generated at 128-bit or 256-bit lengths and encrypted under a master key you define. AWS KMS also provides the ability to generate random data of any length you define suitable for cryptographic use.

### **Can I export a master key from AWS KMS and use it in my own applications?**

No. Master keys are created and used only within AWS KMS to help ensure their security, enable your policies to be consistently enforced, and provide a centralized log of their use.

### **What geographic region are my keys stored in?**

Keys are only stored and used in the region in which they are 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.

### **How can I tell who used or changed the configuration of my keys in AWS KMS?**

Logs in AWS CloudTrail will show requests on your master keys, 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.

### **How does AWS KMS compare to AWS CloudHSM?**

AWS CloudHSM provides you with a FIPS 140-2 Level 3 validated single-tenant HSM in your Amazon Virtual Private Cloud (VPC) to store and use your keys. You have total control over your keys and the application software that uses them with AWS CloudHSM. In addition, you can use AWS CloudHSM to support a variety of use cases and applications, such as Digital Rights Management (DRM), Public Key Infrastructure (PKI), asymmetric cryptographic functions, document signing, and high-performance in-VPC cryptographic acceleration.

AWS KMS allows you to control the encryption keys used by your applications and supported AWS services in multiple regions around the world from a single console. Centralized management of all your keys in AWS KMS lets you enforce who can use your keys, 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.

Q: What is Amazon API Gateway?

Amazon API Gateway is a fully managed service that makes it easy for developers to publish, maintain, monitor, and secure APIs at any scale. With a few clicks in the AWS Management Console, you can create an API that acts as a “front door” for applications to access data, business logic, or functionality from your back-end services, such as applications running on Amazon Elastic Compute Cloud (Amazon EC2), code running on AWS Lambda, or any web application. Amazon API Gateway handles all of the tasks involved in accepting and processing up to hundreds of thousands of concurrent API calls, including traffic management, authorization and access control, monitoring, and API version management. Amazon API Gateway has no minimum fees or startup costs. You pay only for the API calls you receive and the amount of data transferred out.

Q: Why use Amazon API Gateway?

Amazon API Gateway provides developers with a simple, flexible, fully managed, pay-as-you-go service that handles all aspects of creating and operating robust APIs for application back ends.  With Amazon API Gateway, you can launch new services faster and with reduced investment so you can focus on building your core business services.  Amazon API Gateway was built to help you with several aspects of creating and managing APIs:

1) Metering. API Gateway helps you define plans that meter and restrict third-party developer access to your APIs. You can define a set of plans, configure throttling, and quota limits on a per API key basis. API Gateway automatically meters traffic to your APIs and lets you extract utilization data for each API key.

2) Security. API Gateway provides you with multiple tools to authorize access to your APIs and control service operation access. Amazon API Gateway allows you to leverage AWS administration and security tools, such as AWS Identity and Access Management (IAM) and Amazon Cognito, to authorize access to your APIs. Amazon API Gateway can verify signed API calls on your behalf using the same methodology AWS uses for its own APIs. Using custom authorizers written as AWS Lambda functions, API Gateway can also help you verify incoming bearer tokens, removing authorization concerns from your backend code.

3) Resiliency. Amazon API Gateway helps you manage traffic with throttling so that backend operations can withstand traffic spikes. Amazon API Gateway also helps you improve the performance of your APIs and the latency your end users experience by caching the output of API calls to avoid calling your backend every time.

4) Operations Monitoring. After an API is published and in use, API Gateway provides you with a metrics dashboard to monitor calls to your services. The Amazon API Gateway dashboard, through integration with Amazon CloudWatch, provides you with backend performance metrics covering API calls, latency data and error rates. You can enable detailed metrics for each method in your APIs and also receive error, access or debug logs in CloudWatch Logs.

5) Lifecycle Management. After an API has been published, you often need to build and test new versions that enhance or add new functionality. Amazon API Gateway lets you operate multiple API versions and multiple stages for each version simultaneously so that existing applications can continue to call previous versions after new API versions are published.

6) Designed for Developers. Amazon API Gateway allows you to quickly create APIs and assign static content for their responses to reduce cross-team development effort and time-to-market for your applications. Teams who depend on your APIs can begin development while you build your backend processes.

Q: How do I get started with Amazon API Gateway?

You can quickly and easily create a custom API using Amazon API Gateway. For a simple “Hello World” example, follow these steps:

1.  Go to the Amazon API Gateway console.

2.  Select an existing REST API or create a new one by entering a name for the API.

3.  On the REST API tree view, click “Create Resource”.

4.  Choose a name for your resource, such as “cars”.

5. With the new resource selected, click the button to create a new method and select the HTTP verb associated with the method (for example, GET).

6.  Select the integration type (for example, HTTP Proxy), and enter the URL the Amazon API Gateway should call.

7.  Define how requests and responses are transformed using a mapping template, or accept the default settings to pass all of the request and response data through without applying any transformation.

8.  Configure the method’s security settings.

9.  Deploy your new API to a stage.

10. From the Stage management page, set up caching and throttling.

11. On the Client Platforms tab in the Amazon API Gateway console, click the button to download the Android, iOS SDK, or JavaScript library that contains helper methods to call your sayHello operation. The SDK library makes calling your APIs similar to calling a local method. The client SDK automatically handles retries, informing the developer of network or other fault conditions. The SDK library includes the logic necessary to authenticate the client application to your APIs.

12. Integrate the downloaded SDK into your mobile application. Write the code to invoke your custom API. For example, to invoke the getCar(int carId) API in an iOS application:

–(void)getSampleCar

{

NSString \*response = [MyServiceClient getCar:1323];

NSLog( @”Response was [%@]”, response );

}

13. Run your application.

Q: Can I create HTTPS endpoints?

Yes, all of the APIs created with Amazon API Gateway expose HTTPS endpoints only. Amazon API Gateway does not support unencrypted (HTTP) endpoints. By default, Amazon API Gateway assigns an internal domain to the API that automatically uses the Amazon API Gateway certificate. When configuring your APIs to run under a custom domain name, you can provide your own certificate for the domain.

Q: What data types can I use with Amazon API Gateway?

APIs built on Amazon API Gateway can accept any payloads sent over HTTP. Typical data formats include JSON, XML, query string parameters, and request headers. You can declare any content type for your API’s responses, and then use the transform templates to change the back-end response into your desired format.

Q: With what backends can Amazon API Gateway communicate?

Amazon API Gateway can execute AWS Lambda functions in your account, start AWS Step Functions state machines, or call HTTP endpoints hosted on AWS Elastic Beanstalk, Amazon EC2, and also non-AWS hosted HTTP based operations that are accessible via the public Internet. API Gateway also allows you to specify a mapping template to generate static content to be returned, helping you mock your APIs before the backend is ready. You can also integrate API Gateway with other AWS services directly – for example, you could expose an API method in API Gateway that sends data directly to Amazon Kinesis.

Q: For which client platforms can Amazon API Gateway generate SDKs?

API Gateway generates custom SDKs for mobile app development with Android and iOS, and for web app development with JavaScript. Once an API and its models are defined in API Gateway, you can use the AWS console or the API Gateway APIs to generate and download a client SDK.

Q: In which AWS regions is Amazon API Gateway available?

Please refer to [Regional Products and Services](https://aws.amazon.com/about-aws/globalinfrastructure/regional-product-services/) for details of Amazon API Gateway service availability by region.

Q: What can I manage through the Amazon API Gateway console?

Through the Amazon API Gateway console, you can define the REST API and its associated resources and methods, manage the API lifecycle, generate client SDKs and view API metrics. You can also use the API Gateway console to define your APIs’ usage plans, manage developers’ API keys, and configure throttling and quota limits. All of the same actions are available through the API Gateway APIs.

Q: What is a REST API?

In Amazon API Gateway, a REST API is a group of resources and methods, or *endpoints*. REST APIs can be deployed to different stages and cloned to new versions.

Q: What is a resource?

A resource is a typed object that is part of your API’s domain. Each resource may have associated a data model, relationships to other resources, and can respond to different methods.You can also define resources as variables to intercept requests to multiple child resources.

Q: What is a method?

Each resource within a REST API can support one or more of the standard HTTP methods. You define which verbs should be supported for each resource (GET, POST, PUT, PATCH, DELETE, HEAD, OPTIONS) and their implementation. For example, a GET to the cars resource should return a list of cars. To connect all methods within a resource to a single backend endpoint, API Gateway also supports a special “ANY” method.

Q: What is an usage plan?  
Usage plans help you declare plans for third-party developers that restrict access only to certain APIs, define throttling and request quota limits, and associate them with API keys. You can also extract utilization data on an per-API key basis to analyze API usage and generate billing documents. For example, you can create a basic, professional, and enterprise plans – you can configure the basic usage plan to only allow 1,000 requests per day and a maximum of 5 requests per second (RPS).

Q: What is the Amazon API Gateway API lifecycle?

With Amazon API Gateway, each REST API can have multiple stages. Stages are meant to help with the development lifecycle of an API -- for example, after you’ve built your APIs and you deploy them to a development stage, or when you are ready for production, you can deploy them to a production stage.

Q: What is a stage?

In Amazon API Gateway, stages are similar to tags. They define the path through which the deployment is accessible. For example, you can define a development stage and deploy your cars API to it. The resource will be accessible at https://www.myapi.com/dev/cars. You can also set up custom domain names to point directly to a stage, so that you don’t have to use the additional path parameter. For example, if you pointed myapi.com directly to the development stage, you could access your cars resource at <https://www.myapi.com/cars>. Stages can be configured using variables that can be accessed from your API configuration or mapping templates.

Q: What are stage variables?  
Stage variables let you define key/value pairs of configuration values associated with a stage. These values, similarly to environment variables, can be used in your API configuration. For example, you could define the HTTP endpoint for your method integration as a stage variable, and use the variable in your API configuration instead of hardcoding the endpoint – this allows you to use a different endpoint for each stage (e.g. dev, beta, prod) with the same API configuration. Stage variables are also accessible in the mapping templates and can be used to pass configuration parameters to your Lambda or HTTP backend.

Q: What if I mistakenly deployed to a stage?

Amazon API Gateway saves the history of your deployments. At any point, using the Amazon API Gateway APIs or the console, you can roll back a stage to a previous deployment.

Q: Can I run multiple versions of the same REST API?

Yes. Amazon API Gateway gives you the ability to clone an existing API. When you are ready to start working on the next major version of your API, you will be able to keep working on your version 1 and version 2 APIs simultaneously.

Q: Can I use my Swagger API definitions?

Yes. You can use our open source [Swagger importer tool](https://github.com/awslabs/aws-apigateway-swagger-importer) to import your Swagger API definitions into Amazon API Gateway. With the Swagger importer tool you can create and deploy new APIs as well as update existing ones.

Q: How do I monetize my APIs on API Gateway?  
You can monetize your APIs on API Gateway by publishing them as products in AWS Marketplace. You will first need to [register as a seller](https://aws.amazon.com/marketplace/management/register/) in AWS Marketplace, and submit your usage plans on API Gateway as products. [Read here](https://aws.amazon.com/blogs/compute/) to learn more about API Monetization.

Q: How do I document my API on Amazon API Gateway?

API Gateway offers the ability to create, update, and delete documentation associated with each portion of your API, such as methods and resources. You can access documentation-related APIs through the AWS SDKs, CLI, via RESTful calls, or by editing the documentation strings directly in the API Gateway console. Documentation can also be imported as a Swagger file, either as part of the API or separately, allowing you to add or update the documentation without disturbing the API definition. API Gateway conforms to the [Open API specification](http://swagger.io/specification/) for documentation imported from, or exported to, Swagger files.

Q: How can I avoid creating redundant copies of error messages and other documentation that recurs frequently in my API?  
In addition to offering standards-conformant API documentation support, API Gateway additionally supports documentation inheritance, making it simple to define a documentation string once and then use it in multiple places. Inheritance simplifies the process of defining API documentation, and can be converted to the standard representation when exporting the API as a Swagger file.

[Back To Top >>](https://aws.amazon.com/api-gateway/faqs/#faqs)

## **Security and Authorization**

Q: How do I authorize access to my APIs?

With Amazon API Gateway, you can optionally set your API methods to require authorization. When setting up a method to require authorization you can leverage AWS Signature Version 4 or custom authorizers to support your own bearer token auth strategy.

Q: How does AWS Signature Version 4 work?

You can use AWS credentials -- access and secret keys – to sign requests to your service and authorize access like other AWS services. The signing of an Amazon API Gateway API request is managed by the custom API Gateway SDK generated for your service. You can retrieve temporary credentials associated with a role in your AWS account using Amazon Cognito.

Q: What is a custom authorizer?

Custom authorizers are AWS Lambda functions. With custom request authorizers, you will be able to authorize access to APIs using a bearer token auth strategy such as OAuth. When an API is called, API Gateway checks if a custom authorizer is configured, API Gateway then calls the Lambda function with the incoming authorization token. You can use Lambda to implement various authorization strategies (e.g. JWT verification, OAuth provider callout) that return IAM policies which are used to authorize the request. If the policy returned by the authorizer is valid, API Gateway will cache the policy associated with the incoming token for up to 1 hour.

Q: Can Amazon API Gateway generate API keys for distribution to third-party developers?

Yes. API Gateway can generate API keys and associate them with an usage plan. Calls received from each API key are monitored and included in the Amazon CloudWatch Logs you can enable for each stage. However, we do not recommend you use API keys for authorization. You should use API keys to monitor usage by third-party developers and leverage a stronger mechanism for authorization, such as signed API calls or OAuth.

Q: How can I address or prevent API threats or abuse?

Amazon API Gateway supports throttling settings for each method in your APIs. You can set a standard rate limit and a burst rate limit per second for each method in your REST APIs. Further, Amazon API Gateway automatically protects your backend systems from distributed denial-of-service (DDoS) attacks, whether attacked with counterfeit requests (Layer 7) or SYN floods (Layer 3).

Q: Can Amazon API Gateway work within an Amazon VPC?

No. Amazon API Gateway endpoints are always public to the Internet. Proxy requests to backend operations also need to be publicly accessible on the Internet. However, you can generate a client-side SSL certificate in Amazon API Gateway to verify that requests to your backend systems were sent by API Gateway using the public key of the certificate.

Q: Can I verify that it is API Gateway calling my backend?  
Yes. Amazon API Gateway can generate a client-side SSL certificate and make the public key of that certificate available to you. Calls to your backend can be made with the generated certificate, and you can verify calls originating from Amazon API Gateway using the public key of the certificate.

Q: Can I use AWS CloudTrail with Amazon API Gateway?

Yes. Amazon API Gateway is integrated with [AWS CloudTrail](https://aws.amazon.com/cloudtrail/) to give you a full auditable history of the changes to your REST APIs. All API calls made to the Amazon API Gateway APIs to create, modify, delete, or deploy REST APIs are logged to CloudTrail in your AWS account.

[Back To Top >>](https://aws.amazon.com/api-gateway/faqs/#faqs)

## **Management, Metrics and Logging**

Q: How can I monitor my Amazon API Gateway APIs?

Amazon API Gateway logs API calls, latency, and error rates to Amazon CloudWatch in your AWS account. The metrics are also available through the Amazon API Gateway console in a REST API dashboard. API Gateway also meters utilization by third-party developers, the data is available in the API Gateway console and through the APIs.

Q: Can I set up alarms on the Amazon API Gateway metrics?

Yes, Amazon API Gateway sends logging information and metrics to Amazon CloudWatch. You can utilize the Amazon CloudWatch console to set up custom alarms.

Q: How can I set up metrics for Amazon API Gateway?

By default, Amazon API Gateway monitors traffic at a REST API level. Optionally, you can enable detailed metrics for each method in your REST API from the deployment configuration APIs or console screen. Detailed metrics are also logged to Amazon CloudWatch and will be charged at the CloudWatch rates.

Q: Can I determine which version of the API my customers are using?

Yes. Metric details are specified by REST API and stage. Additionally, you can enable metrics for each method in your REST API.

Q: Does Amazon API Gateway provide logging support?

Yes. Amazon API Gateway integrates with Amazon CloudWatch Logs. You can optionally enable logging for each stage in your API. For each method in your REST APIs, you can set the verbosity of the logging, and if full request and response data should be logged.

Q: How quickly are logs available?

Logs, alarms, error rates and other metrics are stored in Amazon CloudWatch and are available near real time.

[Back To Top >>](https://aws.amazon.com/api-gateway/faqs/#faqs)

## **Throttling and Caching**

Q: How can I protect my backend systems and applications from traffic spikes?

Amazon API Gateway provides throttling at multiple levels including global and by service call. Throttling limits can be set for standard rates and bursts. For example, API owners can set a rate limit of 1,000 requests per second for a specific method in their REST APIs, and also configure Amazon API Gateway to handle a burst of 2,000 requests per second for a few seconds. Amazon API Gateway tracks the number of requests per second. Any requests over the limit will receive a 429 HTTP response. The client SDKs generated by Amazon API Gateway retry calls automatically when met with this response.

Q: Can I throttle individual developers calling my APIs?  
Yes. With usage plans you can set throttling limits for individual API keys.

Q: How does throttling help me?

Throttling ensures that API traffic is controlled to help your backend services maintain performance and availability.

Q: At which levels can Amazon API Gateway throttle inbound API traffic?

Throttling rate limits can be set at the method level. You can edit the throttling limits in your method settings through the Amazon API Gateway APIs or in the Amazon API Gateway console.

Q: How are throttling rules applied?  
First. API Gateway checks against your [AWS account limit](http://docs.aws.amazon.com/apigateway/latest/developerguide/limits.html). If the traffic is below the set account limit, API Gateway checks the limit you have set on a stage or method. If the traffic is below the stage limit, then API Gateway applies the usage plans limits you set on a per-API key basis.

Q: Does Amazon API Gateway provide API result caching?

Yes. You can add caching to API calls by provisioning an Amazon API Gateway cache and specifying its size in gigabytes. The cache is provisioned for a specific stage of your APIs. This improves performance and reduces the traffic sent to your back end. Cache settings allow you to control the way the cache key is built and the time-to-live (TTL) of the data stored for each method. Amazon API Gateway also exposes management APIs that help you invalidate the cache for each stage.

Q: What happens if a large number of end users try to invoke my API simultaneously?

If caching is not enabled and throttling limits have not been applied, then all requests will pass through to your backend service until the account level throttling limits are reached. If throttling limits are in place, then Amazon API Gateway will shed the necessary amount of requests and send only the defined limit to your back-end service. If a cache is configured, then Amazon API Gateway will return a cached response for duplicate requests for a customizable time, but only if under configured throttling limits. This balance between the backend and client ensures optimal performance of the APIs for the applications that it supports. Requests that are throttled will be automatically retried by the client-side SDKs generated by Amazon API Gateway. By default, Amazon API Gateway does not set any cache on your API methods.

Q: How do APIs scale?

Amazon API Gateway acts as a proxy to the backend operations that you have configured. Amazon API Gateway will automatically scale to handle the amount of traffic your API receives. Amazon API Gateway does not arbitrarily limit or throttle invocations to your backend operations and all requests that are not intercepted by throttling and caching settings in the Amazon API Gateway console are sent to your backend operations.

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](https://aws.amazon.com/what-is-cloud-computing/). 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: What are some example uses for Amazon SNS notifications?

The Amazon SNS service can support a wide variety of needs including 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. 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: 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: 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.

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, or SMS) and edit topic policies to control publisher and subscriber access. The AWS Management Console is provided free of charge at: [http://aws.amazon.com/console.](https://aws.amazon.com/console)

Q: What are the Amazon SNS service access points in each region?

Please refer to the [AWS Regions and Endpoints](http://docs.aws.amazon.com/general/latest/gr/rande.html#sns_region) 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](http://aws.amazon.com/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](http://docs.aws.amazon.com/sns/latest/dg/logging-using-cloudtrail.html).

To receive a history of SNS API calls made on your account, simply [turn on AWS CloudTrail in the AWS Management Console](https://console.aws.amazon.com/cloudtrail). To learn more about AWS CloudTrail, [click here](http://aws.amazon.com/cloudtrail/).

## **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 pay 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 Details 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](https://aws.amazon.com/c-tax-faqs/).

## **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 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](http://docs.aws.amazon.com/general/latest/gr/rande.html#sns_region) 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. The AWS Management Console is available at: [http://aws.amazon.com/console](https://aws.amazon.com/console)

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 end-point, 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](https://aws.amazon.com/documentation/sns/).

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](https://aws.amazon.com/documentation/sns/) 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](https://aws.amazon.com/health/providers-and-insurers/hipaa/inbound/) for more information.

To learn more, see the following resources:

* AWS [HIPAA Compliance](https://aws.amazon.com/compliance/hipaa-compliance/) page
* AWS [Cloud Computing in Healthcare](https://aws.amazon.com/health/) page

To see the current list of compliance programs that Amazon SNS is in scope for, see [AWS Services in Scope by Compliance Program](https://aws.amazon.com/compliance/services-in-scope/).

## **Reliability**

Q: How reliable is my data once published to Amazon SNS?

Amazon SNS stores all topic and message information within Amazon’s proven network infrastructure and datacenters. All messages are stored redundantly on multiple servers and in multiple data centers, which means that no single computer or network failure renders Amazon SNS inaccessible.

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?

When a message is published to a topic, Amazon SNS will attempt to deliver notifications to all subscribers registered for that topic. Due to potential Internet issues or Email delivery restrictions, sometimes the notification may not successfully reach an HTTP or Email end-point. In the case of HTTP, an SNS Delivery Policy can be used to control the retry pattern (linear, geometric, exponential backoff), maximum and minimum retry delays, and other parameters. If it is critical that all published messages be successfully processed, developers should have notifications delivered to an SQS queue (in addition to notifications over other transports).

## **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 limits 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 limits and how do they work?

Spend limits can be specified for an AWS account and for individual messages, and the limits apply only to the cost of sending SMS messages.

The default spend limit per account (if not specified) is 1.00 USD per month. If you want to raise the limit, submit an [SNS Limit Increase case](https://console.aws.amazon.com/support/home#/case/create?issueType=service-limit-increase&limitType=service-code-sns). For New limit value, enter your desired monthly spend limit. In the Use Case Description field, explain that you are requesting an SMS monthly spend limit increase.

Amazon SNS sends SMS messages that you publish while the total cost incurred for your SMS traffic is below your spend limit for that calendar month. Once the spend limit is exceeded, Amazon SNS stops delivering messages until you either increase the spend limit or a new calendar month begins. Similarly, you can also specify a spend limit for an individual message, and Amazon SNS will send the message only if the cost is below the limit. Amazon SNS will not send your SMS messages if the account-level spend limit is exceeded, regardless of whether the message-level spend limit 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.](https://console.aws.amazon.com/support/home#/) For more information, see [Reserving a Dedicated Short Code for SMS Messaging](http://docs.aws.amazon.com/sns/latest/dg/sms_shortcodes.html) in the *Amazon SNS Developer Guide*.

For pricing information, see [Worldwide SMS Pricing](https://aws.amazon.com/sns/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](http://docs.aws.amazon.com/sns/latest/dg/sms_shortcodes.html) 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](http://docs.aws.amazon.com/sns/latest/dg/sms_supported-countries.html) for a comprehensive list of supported calling countries.

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](http://en.wikipedia.org/wiki/E.164) for all phone numbers both in the ‘to’ and ‘from’ (when applicable) fields. Please refer to the [SMS Supported Country List](http://docs.aws.amazon.com/sns/latest/dg/sms_supported-countries.html) 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](http://docs.aws.amazon.com/sns/latest/dg/welcome.html).

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](http://docs.aws.amazon.com/sns/latest/dg/welcome.html).

Q: What is Amazon WorkSpaces?

Amazon WorkSpaces is a managed desktop computing service running on the [AWS cloud](https://aws.amazon.com/what-is-cloud-computing/). Amazon WorkSpaces allows customers to easily provision cloud-based desktops that allow end-users to access the documents, applications and resources they need on supported devices including Windows and Mac computers, Chromebooks, iPads, Fire tablets, Android tablets, and Chrome and Firefox web browsers. With a few clicks in the AWS Management Console, customers can provision a high-quality cloud desktop experience for any number of users at a cost that is competitive with traditional desktops and half the cost of most Virtual Desktop Infrastructure (VDI) solutions.

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 compute resources, storage space, and software applications that allow a user to perform day-to-day tasks just like using a traditional desktop. A user can connect to a WorkSpace from any supported device using the free Amazon WorkSpaces client application, 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, 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](https://aws.amazon.com/workspaces/details/#workspaces-bundles).

Q: Which versions of Windows are available for use with Amazon WorkSpaces?

Amazon WorkSpaces offers bundles that come with a Windows 7 or Windows 10 desktop experience, powered by Windows Server 2008 R2 and Windows Server 2016 respectively. If your organization is eligible to bring their own Windows Desktop licenses, you can run the Windows 7 or Windows 10 Enterprise operating system on your Amazon WorkSpaces. See [below](https://aws.amazon.com/workspaces/faqs/#BYOL) for more information.

Q: How do I select which desktop experience I want my Amazon WorkSpaces to run?

To select the desktop experience you need for your Amazon WorkSpaces, choose the bundle type that best describes your requirements when you launch new WorkSpaces. For the Windows 7 desktop experience, select bundles that include “Windows 7”. For the Windows 10 desktop experience, select bundles that include “Windows 10”.

Q: Can I migrate users from an Amazon WorkSpaces Windows 7 bundle to a Windows 10 bundle?

No. To offer existing users a Windows 10 desktop experience, you need to delete their existing Amazon WorkSpace and create a new one using a Windows 10 WorkSpaces bundle. To migrate data and documents, we recommend that you use the sync feature available with Amazon WorkDocs. Note that Amazon WorkDocs comes with 50GB of free storage for every Amazon WorkSpace.

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, 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, 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 (D:) on the WorkSpace is backed up every 12 hours. In the case of a WorkSpace failure, AWS can restore this volume from the 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](http://health.amazonworkspaces.com/). You can also refer to the [Regional Products and Services](https://aws.amazon.com/about-aws/global-infrastructure/regional-product-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](https://aws.amazon.com/tools/); you can learn more about the APIs in the [documentation](http://docs.aws.amazon.com/workspaces/latest/devguide/).

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?

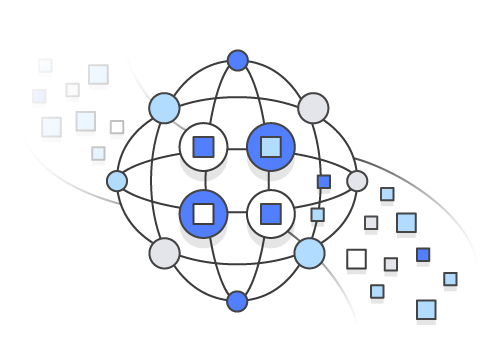
Yes. You can specify which Amazon WorkSpaces resources users can perform actions on. For details see the [documentation](http://docs.aws.amazon.com/workspaces/latest/adminguide/wsp_iam.html).

Q: Do I need to use the AWS Management Console to get started with Amazon WorkSpaces?

Yes. The first time set up for Amazon WorkSpaces relies on the AWS Management Console. Once you have created a directory and registered it with the Amazon WorkSpaces service, you can create and manage WorkSpaces using the Amazon WorkSpaces APIs.

[AMAZON EC2 CONTAINER SERVICE](https://aws.amazon.com/ecs/?sc_ichannel=ha&sc_icampaign=pac_2upweek0501_ecs&sc_isegment=en&sc_iplace=2up&sc_icontent=webp_ecs)

[Learn how to easily build and scale containerized microservices on AWS.](https://aws.amazon.com/ecs/?sc_ichannel=ha&sc_icampaign=pac_2upweek0501_ecs&sc_isegment=en&sc_iplace=2up&sc_icontent=webp_ecs)

[[](https://aws.amazon.com/ecs/?sc_ichannel=ha&sc_icampaign=pac_2upweek0501_ecs&sc_isegment=en&sc_iplace=2up&sc_icontent=webp_ecs)](https://aws.amazon.com/ecs/?sc_ichannel=ha&sc_icampaign=pac_2upweek0501_ecs&sc_isegment=en&sc_iplace=2up&sc_icontent=webp_ecs)

[AWS FREE TIER](https://aws.amazon.com/free/?sc_ichannel=ha&sc_icampaign=free-tier&sc_icontent=2234)

[Gain free, hands-on experience with AWS for 12 months](https://aws.amazon.com/free/?sc_ichannel=ha&sc_icampaign=free-tier&sc_icontent=2234)

[[](https://aws.amazon.com/free/?sc_ichannel=ha&sc_icampaign=free-tier&sc_icontent=2234)](https://aws.amazon.com/free/?sc_ichannel=ha&sc_icampaign=free-tier&sc_icontent=2234)

## **Bundles and Custom Images**

Q: What applications are available with Amazon WorkSpaces Windows 7 bundles?

Amazon WorkSpaces comes with a default set of applications at no additional cost that include Internet Explorer 11, Firefox, and 7-Zip. You can chose to add “Plus” application bundles to your Windows 7 Amazon WorkSpaces bundles which include Microsoft Office Professional 2010 or 2013, Trend Micro Worry-Free Business Security, and WinZip, for an additional monthly fee. Microsoft Office Professional 2016 is available with bundles that offer the Windows 10 desktop experience.

Q: What applications are available with Amazon WorkSpaces Windows 10 bundles?

Amazon WorkSpaces comes with a default set of applications at no additional cost that include Internet Explorer 11, Firefox, and 7-Zip. You can chose to add “Plus” application bundles to your Windows 10 Amazon WorkSpaces bundles which include Microsoft Office Professional 2016, Trend Micro Worry-Free Business Security, and WinZip, for an additional monthly fee. Microsoft Office Professional 2010 and 2013 continue to be available with bundles that offer the Windows 7 desktop experience.

Q: Can I create custom images for Amazon WorkSpaces?

Yes, as an administrator you can create a custom image from a running Amazon WorkSpace. Once you have customized an Amazon WorkSpace with your applications and settings, you can select the WorkSpace in the console and select “Create Image.” This will create an image with your applications and settings. Custom images created from Amazon WorkSpaces Graphics bundles can only be used with Graphics bundles, and custom images created from Value, Standard, Performance, or Power bundles can only be used with those bundles. Most Amazon WorkSpace images are available within 45 minutes. See the [custom image documentation](http://docs.aws.amazon.com/console/workspaces/images) 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](https://console.aws.amazon.com/support/home#/case/create?issueType=service-limit-increase&limitType=service-code-workspaces). To see the default limits for Amazon WorkSpaces, please visit our [documentation](http://docs.aws.amazon.com/workspaces/latest/adminguide/wsp_limits.html).

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: 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](http://docs.aws.amazon.com/workspaces/latest/adminguide/wsp_reset_workspace.html)). You can learn more about SSD EBS volumes [here](https://aws.amazon.com/ebs/details/), and magnetic EBS volumes [here](https://aws.amazon.com/ebs/previous-generation/).

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](http://docs.aws.amazon.com/console/workspaces/bundles).

Q: Which software can I install on an Amazon WorkSpace?

The Amazon WorkSpaces service does not have any technical restrictions on the kind of software that you can install, and any applications that are compatible with the Windows 7 experience provided by Windows Server 2008 R2 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.

## **Amazon WorkSpaces Graphics bundles**

Q: Does Amazon WorkSpaces offer GPU-enabled cloud desktops?

Yes. Amazon WorkSpaces offers Amazon WorkSpaces Graphics bundles, available in English and Japanese.

Q: What are Amazon WorkSpaces Graphics bundles?

Amazon WorkSpaces Graphics bundles are Windows Server 2008 R2 based desktops, for a Windows 7 Desktop Experience, or Windows Server 2016 based desktops for a Windows 10 Desktop Experience, and have a full NVIDIA GPU for graphics intensive applications. The Graphics bundle comes with 8vCPUs, 15GiB of RAM, 4GB of video memory, and 100GBs of storage on the user volume and 100 GBs of general purpose persistent root volume.

Q: What kind of GPU is included with Amazon WorkSpaces Graphics bundles?

The GPU used in Amazon WorkSpaces Graphics bundles is a high-performance NVIDIA GPU with 1,536 CUDA cores and 4GB of video memory. Each Amazon WorkSpaces Graphics bundle has its own GPU.

Q: When would I use Amazon WorkSpaces Graphics bundles?

Amazon WorkSpaces Graphics bundles are designed for engineers and 3D application developers to use as an alternative to expensive graphics-capable workstations. Graphics bundles can be used to run computer-aided design, manufacturing, and engineering software. Additionally, they provide support for OpenGL 4.x, DirectX 10, CUDA, OpenCL, and the GRID SDK for application developers who build 3D capable applications.

Q: Do Amazon WorkSpaces Graphics bundles support a 3D mouse?

The Amazon WorkSpaces Graphics bundles do not currently support a 3D mouse.

Q: What kinds of peripherals can I use with my Amazon WorkSpaces Graphics bundles?

You can use standard QWERTY and Japanese 106/109 keyboards, and most Bluetooth and USB pointing devices with your Amazon WorkSpaces Graphics bundles. You can expect any peripherals that work with Amazon WorkSpaces Value, Standard, Power, and Performance bundles to also work with Amazon WorkSpaces Graphics bundles.

Q: What is the maximum monitor resolution that I can use with my Amazon WorkSpaces Graphics bundles?

Amazon WorkSpaces Graphics bundles support a maximum resolution of 2560x1600, and all VESA compatible resolutions.

Q: How many monitors can I use with my Amazon WorkSpaces Graphics bundles?

Currently you can only use a single monitor with your Amazon WorkSpaces Graphics bundles.

Q: In which AWS Regions can I launch Amazon WorkSpaces Graphics bundles?

You can launch Graphics bundles in all AWS Regions where Amazon WorkSpaces is [available](https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/).

Q: Can I create a custom image for my Amazon WorkSpaces Graphics bundles?

Yes, you can create a custom image for your Amazon WorkSpaces Graphics bundles. Custom images created from Amazon WorkSpaces Graphics bundles can only be used with Graphics bundles, and custom images created from Value, Standard or Performance bundles can only be used with those bundles.

Q: How do I get started with Amazon WorkSpaces Graphics bundles?

You can launch Amazon WorkSpaces Graphics bundles using the Amazon WorkSpaces Management Console, or the Amazon WorkSpaces API. When launching a new Amazon WorkSpace, simply select the Graphics bundle.

Q: Can I enable hybrid IT scenarios with my Amazon WorkSpaces Graphics bundles?

Yes. You can integrate your Amazon WorkSpaces Graphics bundles with your on-premises environment just as you would with your existing Amazon WorkSpaces. You can connect your Amazon WorkSpaces Graphics bundles to your on-premises Active Directory using the AWS Directory Service. Once domain-joined to your on-premises Active Directory, you can access files from network file shares, print to network printers, and access intranet web sites and applications.

Q: Which operating systems can I use with Amazon WorkSpaces Graphics bundles?

Amazon WorkSpaces Graphics bundles provides users with the Windows 7 desktop experience, running Windows Server 2008 R2. In addition, you can run the Windows 7 Desktop operating system if your organization is eligible to bring their own Windows Desktop license.

Q: How much bandwidth does an Amazon WorkSpaces Graphics bundle consume?

Bandwidth used by an Amazon WorkSpaces Graphics bundle 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.

Q: Can I bring my own Windows Desktop licenses for Amazon WorkSpaces Graphics bundles?

Yes, you can. Please [contact us](https://aws.amazon.com/contact-us/aws-sales/) if this is something you’d like to do.

Q: Can I purchase Amazon WorkSpaces Graphics bundles using the monthly billing option?

Yes, you can. Please [contact us](https://aws.amazon.com/contact-us/aws-sales/) if this is something you’d like to do.

## **Amazon WorkSpaces Power bundles**

Q: What are Amazon WorkSpaces Power bundles?

Amazon WorkSpaces Power bundles are Windows Server 2008 R2 based desktops, for a Windows 7 Desktop Experience, or Windows Server 2016 based desktops for a Windows 10 Desktop Experience. The Power bundle comes with 4vCPUs, 16GiB of memory, 175GBs of SSD root volume, and 100GBs of SSD user volume.

Q: When would I use Amazon WorkSpaces Power bundles?

Amazon WorkSpaces Power bundles are designed for power users such as developers and analysts, who compile code and work with large datasets. If you’re already using the AWS cloud for software development or to store large datasets, Amazon WorkSpaces Power bundles provide a faster end-user experience because your apps stay close to your data.

Q: What is the maximum monitor resolution that I can use with my Amazon WorkSpaces Power bundles?

Amazon WorkSpaces Power bundles support a maximum resolution of 2560x1600, and all VESA compatible resolutions.

Q: How many monitors can I use with my Amazon WorkSpaces Power bundles?

Currently you can use up to two monitors with your Amazon WorkSpaces Power bundles.

Q: In which AWS Regions can I launch Amazon WorkSpaces Power bundles?

You can launch Power bundles in all [AWS Regions](https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/) where Amazon WorkSpaces is available.

Q: Can I create a custom image for my Amazon WorkSpaces Power bundles?

Yes, you can create a custom image for your Amazon WorkSpaces Power bundles. A Power bundles custom image can only be used with Value, Standard or Performance bundles if the image size is under 80 GB.

Q: How do I get started with Amazon WorkSpaces Power bundles?

You can launch Amazon WorkSpaces Power bundles using the Amazon WorkSpaces Management Console, or the Amazon WorkSpaces API. When launching a new Amazon WorkSpace, simply select the Power bundle.

Q: Which operating systems can I use with Amazon WorkSpaces Power bundles?

Amazon WorkSpaces Power bundles provide users with a Windows 7 Desktop Experience, running Windows Server 2008 R2, or a Windows 10 Desktop Experience running Windows Server 2016. In addition, you can run the Windows 7 or Windows 10 Desktop operating system if your organization is eligible to bring their own Windows Desktop license.

## **Bring Your Own Windows 7 And Windows 10 Desktop Licenses (BYOL)**

Q: Can I bring my Windows Desktop licenses to Amazon WorkSpaces?

Yes, you can bring your own Windows 7 and Windows 10 Desktop licenses to Amazon WorkSpaces. Amazon WorkSpaces provides you the ability to run on physically dedicated hardware, enabling you to run the Windows 7 or Windows 10 Desktop operating systems on your Amazon WorkSpaces when you are eligible to bring your own licenses.

Q: Can I bring my own Windows Desktop licenses for Amazon WorkSpaces Graphics bundles?

Yes, you can. Please [contact us](https://aws.amazon.com/contact-us/aws-sales/) 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 7 and 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](https://aws.amazon.com/workspaces/pricing/) 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 7 or 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 get started. Additionally, you can create a Technical Support case with Amazon WorkSpaces to get started with BYOL.

Q: How will I upload my Windows 7 or Windows 10 Desktop image to Amazon WorkSpaces?

Please use the VMImport ImportImage function to import your Windows desktop image and create an Amazon Machine Image (AMI). For additional details on importing your Windows desktop image, please consult our documentation [here](https://aws.amazon.com/ec2/vm-import/).

Q: How can I launch Amazon WorkSpaces using my Windows 7 or Windows 10 Desktop image?

In order for you to launch Amazon WorkSpaces using your Windows 7 or Windows 10 Desktop image, you first have to create a custom bundle with the image you imported. Once the new custom bundle has been created, you can launch WorkSpaces from that bundle through the AWS Management Console or using the WorkSpaces CLI or APIs. You can learn more about launching Amazon WorkSpaces using your own Windows 7 or Windows 10 images [here](http://docs.aws.amazon.com/workspaces/latest/adminguide/windows_images.html).

Q: How will I activate my Windows 7 or Windows 10 Desktop operating system on Amazon WorkSpaces?

You can activate your Windows 7 or 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 7 or Windows 10 Desktop image uploaded to Amazon WorkSpaces?

Yes. You can use the standard Amazon WorkSpaces image management functionality to further customize the Windows 7 or Windows 10 Desktop image and save it as a new Amazon WorkSpace image in your account.

Q: Can I launch Amazon WorkSpaces from a public bundle in the same directory as my Windows 7 or Windows 10 Desktop WorkSpaces?

No. Your Windows 7 and Windows 10 Desktop WorkSpaces are launched on physically dedicated hardware to enable you to bring your Windows Desktop licenses to Amazon WorkSpaces. Therefore, WorkSpaces launched in a directory marked for dedicated hardware can only be from a custom bundle that has your own Windows 7 or 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 7 and 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: Can I bring my Windows desktop licenses to all regions where the Amazon WorkSpaces service is available?

Yes. When you communicate with your sales representative or technical support, simply specify the region(s) in which you want to launch Amazon WorkSpaces using your own Windows desktop operating systems.

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.

Q: How long will it take before I can launch Amazon WorkSpaces using my own Windows desktop licenses and image?

You should expect to be able to launch Amazon WorkSpaces using your Windows 7 or Windows 10 Desktop operating systems within 4 weeks from when you begin the onboarding process.

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 or restarted on physically dedicated hardware?

Amazon WorkSpaces that are rebuilt or restarted can be placed on any available physical server allocated to your account. A re-start 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.

## **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](https://aws.amazon.com/health/providers-and-insurers/hipaa/inbound/) and we will put you in touch with a representative from our AWS sales team. For more information, see, [HIPAA Compliance](https://aws.amazon.com/compliance/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](https://aws.amazon.com/compliance/pci-dss-level-1-faqs/).

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, 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](http://docs.aws.amazon.com/workspaces/latest/adminguide/trusted-devices.html).

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. MFA solutions based on the TOTP (Time-based One-time Password) protocol are not currently supported.

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 OS X, Chromebooks, iOS, Fire, and Android. 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](https://aws.amazon.com/workspaces/details/#workspaces-bundles). 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](http://docs.aws.amazon.com/workspaces/latest/adminguide/create_iam_user.html).

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](http://docs.aws.amazon.com/workspaces/latest/adminguide/using_connect_directory.html#connect_update_details) 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](http://docs.aws.amazon.com/workspaces/latest/adminguide/using_connect_directory.html#connect_update_details) for more details.

## **Encryption**

Q: Does Amazon WorkSpaces support encryption?

Yes. Amazon WorkSpaces supports root volume (C: drive) and user volume (D: drive) 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 WorkSpace bundle types will support encryption?

Encryption is supported on all Amazon WorkSpaces hardware and software bundle types. This includes both Windows 7 and Windows 10 desktop experiences, and the Value, Standard, Performance, Power, and Graphics bundles. It also includes all Plus application bundles. Additionally, any custom bundles will 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 30 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](http://docs.aws.amazon.com/kms/latest/developerguide/limits.html)). 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 [here](http://docs.aws.amazon.com/kms/latest/developerguide/services-workspaces.html).

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 and rebuilds 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 and rebuilds 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 re-build 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 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.

## **Amazon WorkDocs Sync**

Q: What is the Amazon WorkDocs sync client?

The Amazon WorkDocs sync client is a client application that you can install on your Amazon WorkSpace, 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 PC 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](http://sync.amazonworkspaces.com/).

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 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](https://console.aws.amazon.com/), clicking the directory ID link for your directory and selecting the Apps & Services tab. For more information and detailed setup see our [documentation](http://docs.aws.amazon.com/console/directoryservice/enable_sso).

## **Amazon WorkSpaces Application Manager (Amazon 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](https://aws.amazon.com/workspaces/). 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 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 Amazon WorkSpaces users. 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](http://docs.aws.amazon.com/console/wam/packaging) and validating.

Q: What type of applications can be delivered using Amazon WAM?

Any application compatible with Microsoft Windows 7, Microsoft Windows 8, Microsoft Windows Server 2008 R2, and Microsoft Windows Server 2012 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?

Amazon WAM is currently available in the US East (N. Virginia), US West (Oregon), EU (Ireland), Asia Pacific (Sydney), and Asia Pacific (Singapore) AWS regions.

Q: Which Amazon WorkSpaces experiences work with Amazon WAM?

Today you can use Amazon WAM to deploy and manage applications for Amazon WorkSpaces running the Windows 7 desktop experience. Support for Amazon WorkSpaces running the Windows 10 desktop experience will be added at a later time.

Q. Which AWS Directory Service directories does Amazon WAM support?

Amazon WAM can be used with AWS Directory Services [AD Connector](https://aws.amazon.com/directoryservice/details/) and [Simple AD](https://aws.amazon.com/directoryservice/details/). Currently WAM cannot be used with AWS Directory Service [Microsoft Active Directory](https://aws.amazon.com/directoryservice/details/).

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](http://docs.aws.amazon.com/wam/latest/userguide/what_is.html).

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](http://docs.aws.amazon.com/workspaces/latest/adminguide/wsp_tag_workspace.html).

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](https://aws.amazon.com/s3/pricing/) 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 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.

## **Amazon WorkSpaces Client Applications, Web Access, and User Experience**

Q: Where can I download the Amazon WorkSpaces client application?

A: You can download the Amazon WorkSpaces client application for free on the [client download website](https://clients.amazonworkspaces.com/).

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, Mac OS X, 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 Mac OS X (10.8.1 and above)
* Google Chrome OS (45 and above)
* Apple iOS (8.0 and above)
* Google Android (4.2 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 4.2 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](https://forums.aws.amazon.com/forum.jspa?forumID=164).

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](https://forums.aws.amazon.com/forum.jspa?forumID=164).

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 visit the device finder [here](http://www.teradici.com/product-finder/aws) (site hosted by Teradici).

Q: Will my Amazon WorkSpace running in AutoStop running mode preserve the state of applications and data when it stops?

Yes. Because AutoStop causes your Amazon WorkSpace to stop, the state of your applications and data will be preserved. When you next connect, your Amazon WorkSpace will resume with all open documents and running programs intact.

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)
* 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. 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 launches. For existing WorkSpaces, Audio-in can be enabled with a reboot. 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.

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. If desired, it is possible to override the automatic settings by manually selecting a DPI configuration within Windows in an Amazon WorkSpace.

Q: Are dual monitors supported?

Dual monitor support is available for Value, Standard, Performance, and Power bundles, with a maximum resolution of 2560 X 1600. Dual monitor support is not available for Graphics bundles.

Q: Will the iPad and Android applications support Keyboard/Mouse input?

The iPad client supports keyboard input, and the Android client supports both keyboard and mouse input. 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 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. Web Access can be accessed [here](https://clients.amazonworkspaces.com/webclient).

Q: What is Amazon WorkSpaces Web Access?

Amazon WorkSpaces Web Access allows you to access your Amazon WorkSpace 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; users can choose between Web Access and native client applications. Web Access is available [here](https://clients.amazonworkspaces.com/webclient).

Q: Which web browsers can I use to access Amazon WorkSpaces Web Access?

Amazon WorkSpaces Web Access works with Google Chrome version 53 and higher, and Firefox version 49 and higher, 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?

No. Web Access support is only available on WorkSpaces with English based Windows. You can install language packs onto English based WorkSpaces and have Web Access support. However, WorkSpaces with Windows based on other languages will not have Web Access support through you can still use native clients to access those WorkSpaces.

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, which is available [here](https://clients.amazonworkspaces.com/webclient). 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 can be accessed using Web Access?

You can use Web Access to connect to the Value, Standard, Performance, and Power Amazon WorkSpaces bundles running Windows 7 or Windows Server 2008 R2 operating systems. Please note using Web Access to access Amazon WorkSpaces Graphics bundles, or bundles running Windows 10 or Windows Server 2016 operating systems is not currently supported.

Q: Which Amazon WorkSpaces operating systems can be accessed using Web Access?

Web Access can be used to connect to Amazon WorkSpaces running Windows 7 or Windows Server 2008 R2.

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 you can provision Amazon WorkSpaces. For the complete list, please visit [this](https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/) page.

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](http://health.amazonworkspaces.com/) compares your connection speed to each Amazon WorkSpaces region and recommends the fastest one.

Q: What languages are supported by Amazon WorkSpaces?

Amazon WorkSpaces bundles that provide the Windows 7 desktop experience currently support English (US) and Japanese. Amazon WorkSpaces bundles that provide the Windows 10 desktop experience currently support English (US) only. You can also download and install language packs for Windows directly from Microsoft. For more information, visit [this page](https://support.microsoft.com/en-us/help/14236/language-packs). Amazon WorkSpaces client applications currently support both English (US) and Japanese.

## **Setup and Maintenance**

Q: Does the Amazon WorkSpaces service have maintenance windows?

Yes. For AlwaysOn (monthly) WorkSpaces the current maintenance window is a four hour period from 00h00 – 04h00 (this time window will be based on the time zone of the AWS region where your Amazon WorkSpaces are located) each Sunday morning. During this time your WorkSpaces may not be available.

For AutoStop (hourly) WorkSpaces with Maintenance mode enabled, the maintenance window is typically from 00:00 to 05:00 on the 3rd Monday of the month.You can set the Maintenance mode for AutoStop WorkSpaces in the WorkSpaces management console. For more information see [Manage the WorkSpace Running Mode](http://docs.aws.amazon.com/workspaces/latest/adminguide/running-mode.html#set-maintenance-mode).  
  
The maintenance window for WorkSpaces is currently not configurable.

Q: Will my Amazon WorkSpaces require software updates?

Your Amazon WorkSpaces provide users with the Windows 7 experience, provided by Windows Server 2008 R2. The underlying OS, and any applications installed in the WorkSpace may need updates.

Q: How will my Amazon WorkSpaces be patched with software updates?

You have the ability to control how patching is configured your Amazon WorkSpaces. By default, Windows Update is turned on, but you have the ability to customize these settings, or use an alternative patch management approach if you prefer. 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](https://blogs.technet.microsoft.com/msrc/) for more information.

Q: How are updates for applications installed in my WorkSpaces provided?

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 can Amazon WorkSpaces be managed?

The WorkSpaces Management console lets you provision, reboot, rebuild, and delete WorkSpaces. To manage the underlying OS for the WorkSpaces, you can use standard Microsoft Active Directory tools such as Group Policy 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](http://docs.aws.amazon.com/workspaces/latest/adminguide/wsp_administration.html) for more information.

Q: Can I use tags to categorize my Amazon WorkSpaces?

Yes, you can assign tags to existing Amazon WorkSpaces, or during the launch of new Amazon WorkSpaces. You can assign up to 50 tags (key/value pairs) to each Amazon WorkSpace 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, follow the steps listed on this web page: [Tagging WorkSpaces](http://docs.aws.amazon.com/workspaces/latest/adminguide/wsp_tag_workspace.html).

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 rebooting and rebuilding an Amazon WorkSpace?

A reboot is just the same as a regular operating system (OS) reboot. A rebuild will retain the user volume on the WorkSpace (D:) but will return the WorkSpace to its original state (any changes made to the system drive (C:) will not be retained).

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](https://console.aws.amazon.com/support/home#/case/create?issueType=service-limit-increase&limitType=service-code-workspaces). To see the default limits for Amazon WorkSpaces, please visit our [documentation](http://docs.aws.amazon.com/workspaces/latest/adminguide/wsp_limits.html).

Q: Is there a minimum number of Amazon WorkSpaces or users I must provision?

No. There is no minimum requirement.

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: 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 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: Which AWS regions does Amazon WorkSpaces support?

Please refer to the [Regional Products and Services](https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/) page for details of Amazon WorkSpaces service availability by region.

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](http://docs.aws.amazon.com/workspaces/latest/adminguide/enable_zero_client.html) to enable the clients to successfully connect to Amazon WorkSpaces. Please consult the Amazon WorkSpaces documentation for a step by step [guide](http://docs.aws.amazon.com/workspaces/latest/adminguide/zero_client_help.html)on how to properly setup the PCoIP Connection Manager, and for help on how to [find and install the necessary firmware](http://docs.aws.amazon.com/workspaces/latest/adminguide/zero_client_help.html) required for your Zero Clients

Q: How do I get support with Amazon WorkSpaces?

You can get help from [AWS Support](https://aws.amazon.com/contact-us/), and you can also post in the [Amazon WorkSpaces Forum](https://forums.aws.amazon.com/forum.jspa?forumID=164).

## **Pricing and Billing**

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](https://aws.amazon.com/workspaces/pricing/) for the latest information.

Q: Is there a price difference between Amazon WorkSpaces Windows 7 and Windows 10 bundles?

No. There is no price difference between Amazon WorkSpaces Windows 7 and Windows 10 bundles. Note, however, that there is an additional charge for the Plus versions of both Windows 7 and Windows 10 bundles.

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.

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](https://aws.amazon.com/workspaces/pricing/).

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](https://aws.amazon.com/workspaces/pricing/) for the Plus software bundle even if you do not use those Amazon WorkSpaces in a particular month.

Q: Can I purchase Amazon WorkSpaces Graphics bundles using the monthly billing option?

Yes, you can. Please [contact us](https://aws.amazon.com/contact-us/aws-sales/) if this is something you’d like to do.

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 at current AWS bandwidth rates.

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](https://clients.amazonworkspaces.com/).

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](http://docs.aws.amazon.com/workspaces/latest/adminguide/wsp_tag_workspace.html). 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](http://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/configurecostallocreport.html).

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](https://aws.amazon.com/workspaces/applicationmanager/).

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.

## **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. The Standard bundle WorkSpace offers a cloud desktop with 2 vCPUs, 4 GB of memory, and 50 GB of SSD-based storage, and you can choose between a Windows 10 or Windows 7 desktop experience, both powered by Windows Server. As with all bundles, your WorkSpace comes with Internet Explorer 11, Mozilla Firefox, and 7-Zip pre-installed, and access to [Amazon WorkDocs](https://aws.amazon.com/workdocs/) with 50 GB included storage.

Q: What is included with the Amazon WorkSpaces Free Tier?

The WorkSpaces Free Tier includes two Standard bundle Amazon WorkSpaces, for 40 hours of combined use per month, for two calendar months. As with all bundles, your WorkSpace comes with Internet Explorer 11, Mozilla Firefox, and 7-Zip pre-installed, and access to [Amazon WorkDocs](https://aws.amazon.com/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](https://aws.amazon.com/workspaces/pricing/) 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 [Amazon WorkSpaces Pricing and Billing FAQ](https://aws.amazon.com/workspaces/faqs/#pricing).

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](https://aws.amazon.com/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](https://aws.amazon.com/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](http://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/billing-reports.html).

## **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](http://docs.aws.amazon.com/workspaces/latest/adminguide/windows_client_help.html) 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 in my Amazon WorkSpaces?

Yes. You can use IPv6 addresses for Amazon WorkSpaces Value, Standard, and Power bundles. At this time, IPv6 addresses are not supported in WorkSpaces Performance and Graphics 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](http://docs.aws.amazon.com/workspaces/latest/adminguide/trusted-devices.html).

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](http://docs.aws.amazon.com/workspaces/latest/adminguide/prep_connect.html#mfa_prereqs).

## **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](http://docs.aws.amazon.com/workspaces/latest/adminguide/managing_a_directory.html) 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](http://docs.aws.amazon.com/directoryservice/latest/admin-guide/setup_trust.html). To configure AD Connector, please see the documentation [here](http://docs.aws.amazon.com/directoryservice/latest/admin-guide/create_ad_connector.html).

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](https://docs.aws.amazon.com/directoryservice/latest/admin-guide/what_is.html) 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](http://docs.aws.amazon.com/workspaces/latest/adminguide/wsp_create_workspace.html#launch-workspace-trusted-domain).

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](http://docs.aws.amazon.com/workspaces/latest/adminguide/managing_a_directory.html) 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](https://aws.amazon.com/directoryservice/details/#Domain_Join) EC2 instances or [provide directory users access](https://aws.amazon.com/directoryservice/details/) 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](https://aws.amazon.com/directoryservice/pricing/). 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 Mirosoft Active Directory.

## **Amazon CloudWatch Monitoring**

Q: What does Amazon CloudWatch monitor for Amazon WorkSpaces?

You can use Amazon CloudWatch metrics for Amazon WorkSpaces 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.

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 for Amazon WorkSpaces is supported in all AWS regions in which Amazon WorkSpaces is available.

Q: What does it cost?

There is no additional cost for using Basic 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](https://aws.amazon.com/cloudwatch/pricing/) for more information.

Q: How do I get started?

Basic 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](http://docs.aws.amazon.com/AmazonCloudWatch/latest/DeveloperGuide/wsp-metricscollected.html) 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](http://docs.aws.amazon.com/AmazonCloudWatch/latest/DeveloperGuide/wsp-metricscollected.html) for more information on Amazon CloudWatch metrics with Amazon WorkSpaces.

## **Printing for Amazon WorkSpaces**

Q: Can I print from my Amazon WorkSpace?

Yes, Amazon WorkSpaces supports local printers, network printers, and cloud printing services.

Q: Which printer is set as the default in my Amazon WorkSpace?

If you have a local printer configured on the Windows or Mac computer you use to connect to your Amazon WorkSpace, then that printer will be set as the default printer when you connect to your WorkSpace.

Q: How do I print to my local printer?

If you have a local printer configured, it will be selected by default. 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 2008 R2 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,® and Google Cloud Print.

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® and Google Cloud Print. Local and network printing are not currently supported.