Cloud Computing Concepts

Cloud computing is the on-demand delivery of compute power, database storage, applications and other IT resources through a cloud services platform via the Internet with pay-as-you-go pricing.

Cloud computing provides a simple way to access servers, storage, databases and a broad set of application services over the Internet.

The following introductory-level articles cover some key concepts that relate to cloud computing:

- **Cloud Computing Basics – Compute**
- **Cloud Computing Basics – Storage**
- **Cloud Computing Basics – Network**

A cloud services platform such as Amazon Web Services owns and maintains the network-connected hardware required for these application services, while you provision and use what you need via a web application.

6 advantages of cloud:

1. Trade capital expense for variable expense.
2. Benefit from massive economies of scale.
3. Stop guessing about capacity.
4. Increase speed and agility.
5. Stop spending money running and maintaining data centres.
6. Go global in minutes.

Trade capital expense for variable expense

Instead of having to invest heavily in data centers and servers before you know how you're going to use them, you can pay only when you consume computing resources, and pay only for how much you consume.

Benefit from massive economies of scale

By using cloud computing, you can achieve a lower variable cost than you can get on your own. Because usage from hundreds of thousands of customers is aggregated in the cloud, providers such as AWS can achieve higher economies of scale, which translates into lower pay as-you-go price.

Stop guessing about capacity

Eliminate guessing on your infrastructure capacity needs. When you make a capacity decision prior to deploying an application, you often end up either sitting on expensive idle resources or dealing with limited capacity.

With cloud computing, these problems go away. You can access as much or as little capacity as you need, and scale up and down as required with only a few minutes' notice.

Increase speed and agility
In a cloud computing environment, new IT resources are only a click away, which means that you reduce the time to make those resources available to your developers from weeks to just minutes.

This results in a dramatic increase in agility for the organization, since the cost and time it takes to experiment and develop is significantly lower.

Stop spending money running and maintaining data centers
Focus on projects that differentiate your business, not the infrastructure. Cloud computing lets you focus on your own customers, rather than on the heavy lifting of racking, stacking, and powering servers.

Go global in minutes
Easily deploy your application in multiple regions around the world with just a few clicks. This means you can provide lower latency and a better experience for your customers at minimal cost.

Cloud Computing Models

3 types of cloud computing model:

1. Infrastructure as a service (IaaS).
2. Platform as a service (PaaS).
3. Software as a service (SaaS).

Infrastructure as a Service (IaaS)
Infrastructure as a Service (IaaS) contains the basic building blocks for cloud IT and typically provide access to networking features, computers (virtual or on dedicated hardware), and data storage space.

IaaS provides you with the highest level of flexibility and management control over your IT resources and is most similar to existing IT resources that many IT departments and developers are familiar with today.

Platform as a Service (PaaS)
Platform as a Service (PaaS) removes the need for your organization to manage the underlying infrastructure (usually hardware and operating

systems) and allows you to focus on the deployment and management of your applications.

This helps you be more efficient as you don't need to worry about resource procurement, capacity planning, software maintenance, patching, or any of the other undifferentiated heavy lifting involved in running your application.

Software as a Service (SaaS)
Software as a Service (SaaS) provides you with a completed product that is run and managed by the service provider. In most cases, people referring to Software as a Service are referring to end-user applications.

With a SaaS offering you do not have to think about how the service is maintained or how the underlying infrastructure is managed; you only need to think about how you will use that particular piece of software.

A common example of a SaaS application is web-based email which you can use to send and receive email without having to manage feature additions to the email product or maintain the servers and operating systems that the email program is running on.

Provides high availability, fault tolerance, scalability an elasticity.

Types of Cloud Deployment

There are 3 types of cloud deployment:

1. Public Cloud or simple "Cloud" – e.g. AWS, Azure, GCP.
2. Hybrid Cloud – mixture of public and private clouds.
3. Private Cloud (on-premise) – managed in your own data centre, e.g. Hyper-V, OpenStack, VMware.

Public Cloud
A cloud-based application is fully deployed in the cloud and all parts of the application run in the cloud. Applications in the cloud have either been created in the cloud or have been migrated from an existing infrastructure to take advantage of the benefits of cloud computing.

Cloud-based applications can be built on low-level infrastructure pieces or can use higher level services that provide abstraction from the management, architecting, and scaling requirements of core infrastructure.

Hybrid

A hybrid deployment is a way to connect infrastructure and applications between cloud-based resources and existing resources that are not located in the cloud.

The most common method of hybrid deployment is between the cloud and existing on-premises infrastructure to extend, and grow, an organization's infrastructure into the cloud while connecting cloud resources to the internal system.

On-premises
The deployment of resources on-premises, using virtualization and resource management tools, is sometimes called the "private cloud."

On-premises deployment doesn't provide many of the benefits of cloud computing but is sometimes sought for its ability to provide dedicated resources.

In most cases this deployment model is the same as legacy IT infrastructure while using application management and virtualization technologies to try and increase resource utilization.

AWS Compute

This article discusses AWS Compute in the context of the AWS Certified Cloud Practitioner Exam. This is one of the key technology areas covered in the exam blueprint.

Amazon EC2

Amazon Elastic Compute Cloud (Amazon EC2) is a web service in the AWS Compute suite of products that provides secure, resizable compute capacity in the cloud.

The EC2 simple web service interface allows you to obtain and configure capacity with minimal friction.

EC2 is designed to make web-scale cloud computing easier for developers.

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.

Amazon EC2 provides developers the tools to build failure resilient applications and isolate them from common failure scenarios.

Benefits of EC2 include:

- **Elastic Web-Scale computing** – you can increase or decrease capacity within minutes not hours and commission one to thousands of instances simultaneously.
- **Completely controlled** – You have complete control include root access to each instance and can stop and start instances without losing data and using web service APIs.
- **Flexible Cloud Hosting Services** – you can choose from multiple instance types, operating systems, and software packages as well as instances with varying memory, CPU and storage configurations.
- **Integrated** – EC2 is integrated with most AWS services such as S3, RDS, and VPC to provide a complete, secure solution.
- **Reliable** – EC2 offers a highly reliable environment where replacement instances can be rapidly and predictably commissioned with SLAs of 99.95% for each region.
- **Secure** – EC2 works in conjunction with VPC to provide a secure location with an IP address range you specify and offers Security Groups, Network ACLs, and IPSec VPN features.
- **Inexpensive** – Amazon passes on the financial benefits of scale by charging very low rates and on a capacity consumed basis.

An Amazon Machine Image (AMI) is a special type of virtual appliance that is used to create a virtual machine within the Amazon Elastic Compute Cloud (“EC2”).

An AMI includes the following:

- One or more EBS snapshots, or, for instance-store-backed AMIs, a template for the root volume of the instance (for example, 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

AMIs come in three main categories:

- **Community AMIs** – free to use, generally you just select the operating system you want.
- **AWS Marketplace AMIs** – pay to use, generally come packaged with additional, licensed software.
- **My AMIs** – AMIs that you create yourself.

Metadata and User Data:

- User data is data that is supplied by the user at instance launch in the form of a script.
- Instance metadata is data about your instance that you can use to configure or manage the running instance.
- User data is limited to 16KB.
- User data and metadata are not encrypted.
- Instance metadata is available at **http://169.254.169.254/latest/meta-data**.

The Instance Metadata Query tool allows you to query the instance metadata without having to type out the full URI or category names.

Pricing

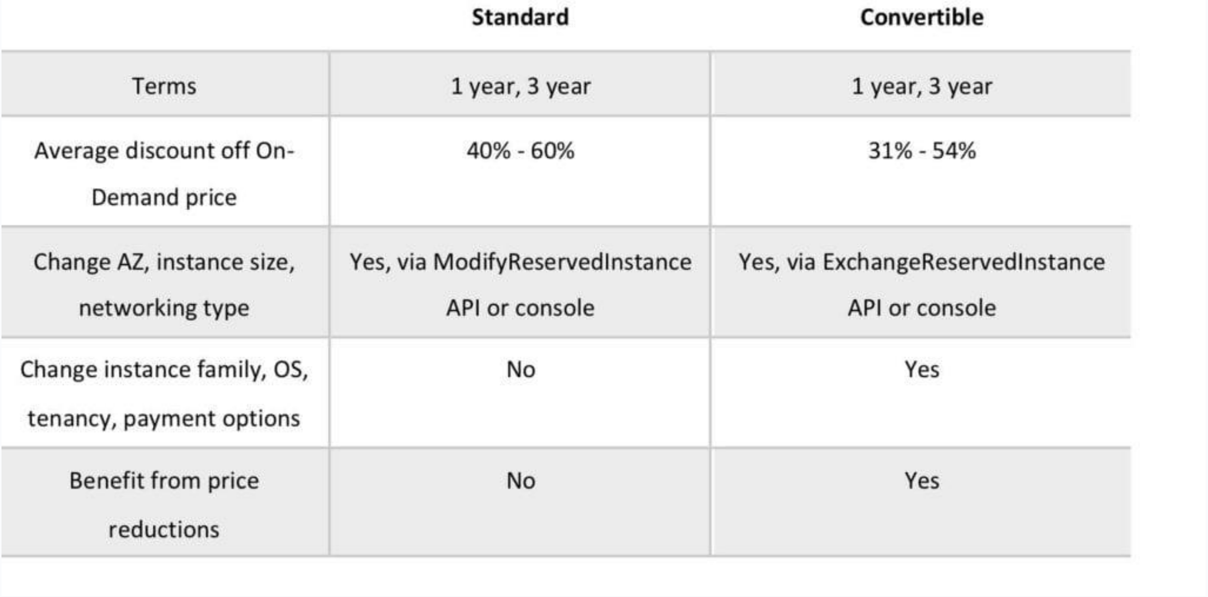
On-demand:

- Good for users that want the low cost and flexibility of EC2 without any up-front payment or long term commitment.
- Applications with short term, spiky, or unpredictable workloads that cannot be interrupted.
- Applications being developed or tested on EC2 for the first time.

Reserved:

- Applications with steady state or predictable usage.
- Applications that require reserved capacity.

- Users can make up-front payments to reduce their total computing costs even further.
- Standard Reserved Instances (RIs) provide up to 75% off on-demand price.
- Convertible RIs provide up to 54% off on-demand price – provides the capability to change the attributes of the RI as long as the exchange results in the creation of RIs of equal or greater value.
- Scheduled RIs are available to launch within the time window you reserve. This option allows you to match your capacity reservation to a predictable recurring schedule that only requires a fraction of a day, a week, or a month.



	Standard	Convertible
Terms	1 year, 3 year	1 year, 3 year
Average discount off On-Demand price	40% - 60%	31% - 54%
Change AZ, instance size, networking type	Yes, via ModifyReservedInstance API or console	Yes, via ExchangeReservedInstance API or console
Change instance family, OS, tenancy, payment options	No	Yes
Benefit from price reductions	No	Yes

Spot:

- Applications that have flexible start and end times.
- Applications that are only feasible at very low compute prices.
- Users with an urgent need for a large amount of additional compute capacity.
- If Amazon terminate your instances you do not pay, if you terminate you pay for the hour.

Dedicated hosts:

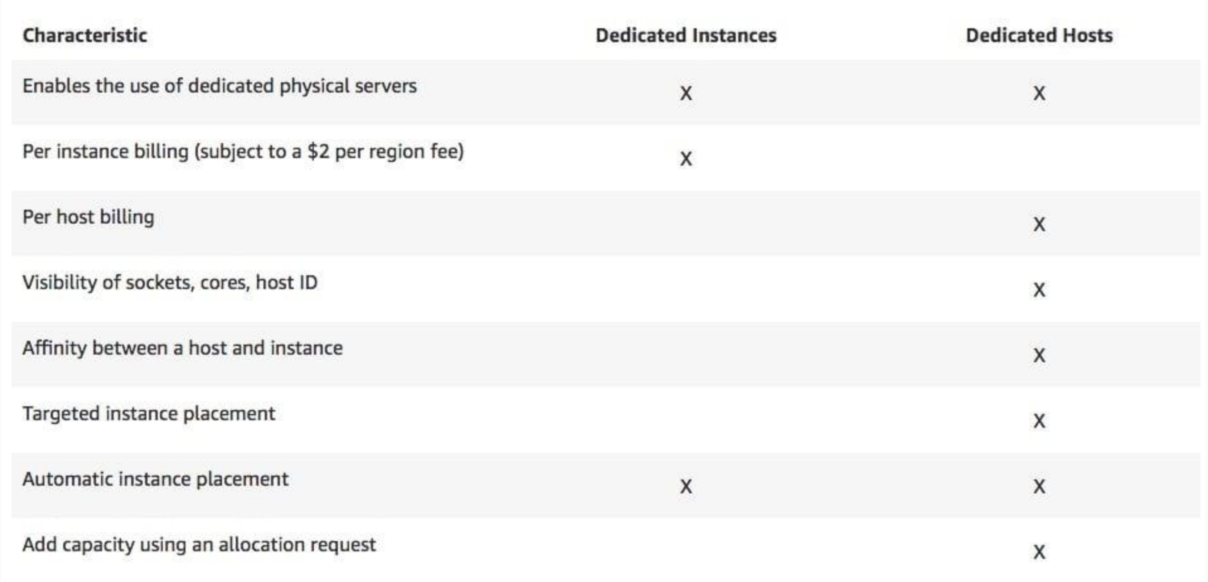
- Physical servers dedicated just for your use.
- You then have control over which instances are deployed on that host.
- Available as On-Demand or with Dedicated Host Reservation.
- Useful if you have server-bound software licences that use metrics like per-core, per-socket, or per-VM.

- Each dedicated host can only run one EC2 instance size and type.
- Good for regulatory compliance or licensing requirements.
- Predictable performance.
- Complete isolation.
- Most expensive option.
- Billing is per host.

Dedicated instances:

- Virtualized instances on hardware just for you.
- Also uses physically dedicated EC2 servers.
- Does not provide the additional visibility and controls of dedicated hosts (e.g. how instance are placed on a server).
- Billing is per instance.
- May share hardware with other non-dedicated instances in the same account.
- Available as On-Demand, Reserved Instances, and Spot Instances.
- Cost additional $2 per hour per region.

The following table describes some of the differences between dedicated instances and dedicated hosts:



Characteristic	Dedicated Instances	Dedicated Hosts
Enables the use of dedicated physical servers	X	X
Per instance billing (subject to a $2 per region fee)	X	
Per host billing		X
Visibility of sockets, cores, host ID		X
Affinity between a host and instance		X
Targeted instance placement		X
Automatic instance placement	X	X
Add capacity using an allocation request		X

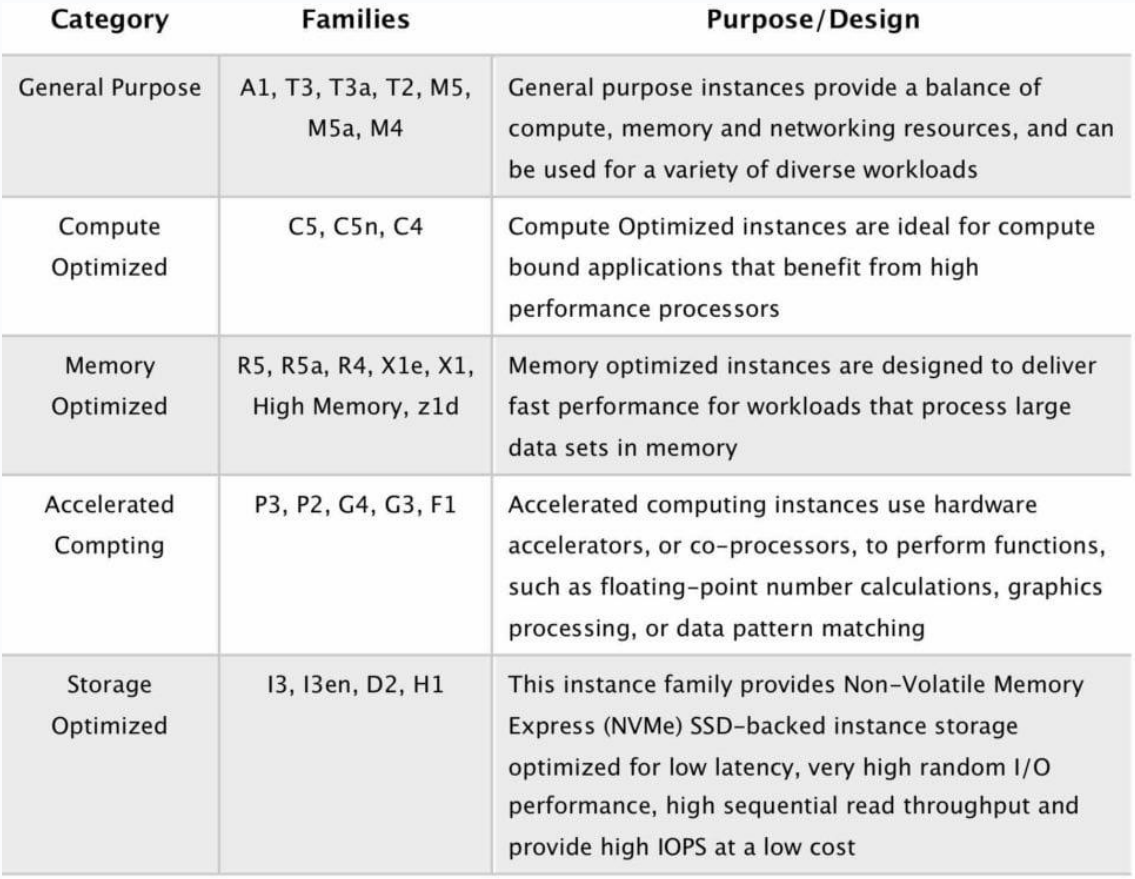
Instance Types;

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases.

Instance types comprise varying combinations of CPU, memory, storage, and networking capacity and give you the flexibility to choose the appropriate mix of resources for your applications.

Each instance type includes one or more instance sizes, allowing you to scale your resources to the requirements of your target workload.

The table below provides an overview of the different EC2 instance types:



Category	Families	Purpose/Design
General Purpose	A1, T3, T3a, T2, M5, M5a, M4	General purpose instances provide a balance of compute, memory and networking resources, and can be used for a variety of diverse workloads
Compute Optimized	C5, C5n, C4	Compute Optimized instances are ideal for compute bound applications that benefit from high performance processors
Memory Optimized	R5, R5a, R4, X1e, X1, High Memory, z1d	Memory optimized instances are designed to deliver fast performance for workloads that process large data sets in memory
Accelerated Compting	P3, P2, G4, G3, F1	Accelerated computing instances use hardware accelerators, or co-processors, to perform functions, such as floating-point number calculations, graphics processing, or data pattern matching
Storage Optimized	I3, I3en, D2, H1	This instance family provides Non-Volatile Memory Express (NVMe) SSD-backed instance storage optimized for low latency, very high random I/O performance, high sequential read throughput and provide high IOPS at a low cost

More information at: **https://digitalcloud.training/certification-training/aws-solutions-architect-associate/compute/amazon-ec2/**

Amazon EC2 Container Service (ECS)

Amazon Elastic Container Service (ECS) is another product in the AWS Compute category. It provides 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.

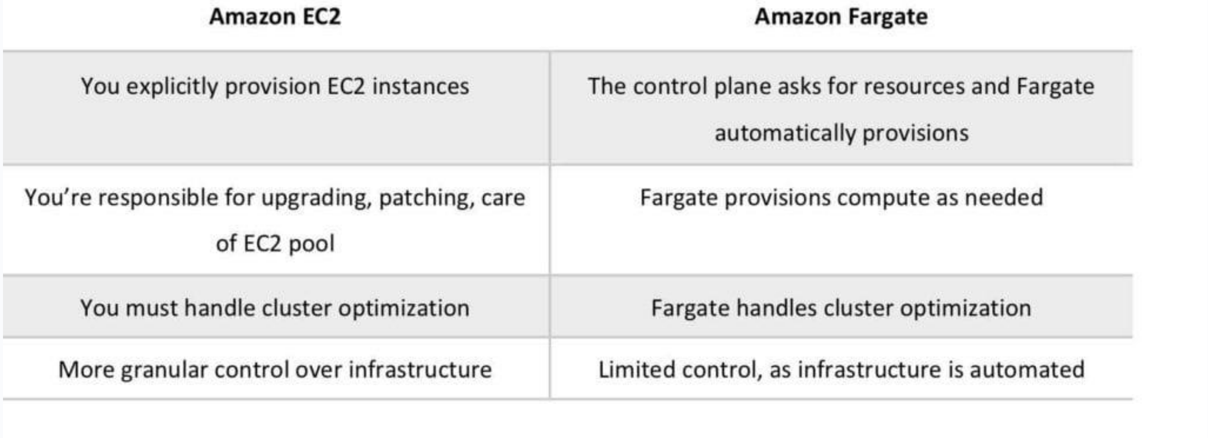
Amazon ECS eliminates the need for you to install, operate, and scale your own cluster management infrastructure.

Using API calls you can launch and stop container-enabled applications, query the complete state of clusters, and access many familiar features like security groups, Elastic Load Balancing, EBS volumes and IAM roles.

Amazon ECS can be used to schedule the placement of containers across clusters based on resource needs and availability requirements.

An Amazon ECS launch type determines the type of infrastructure on which your tasks and services are hosted.

There are two launch types and the table below describes some of the differences between the two launch types:



Amazon EC2	Amazon Fargate
You explicitly provision EC2 instances	The control plane asks for resources and Fargate automatically provisions
You're responsible for upgrading, patching, care of EC2 pool	Fargate provisions compute as needed
You must handle cluster optimization	Fargate handles cluster optimization
More granular control over infrastructure	Limited control, as infrastructure is automated

The Elastic container registry (ECR) is a managed AWS Docker registry service for storing, managing and deploying Docker images.

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.

Amazon ECR is integrated with Amazon EC2 Container Service (ECS).

With Amazon ECR, there are no upfront fees or commitments. You pay only for the amount of data you store in your repositories and data transferred to the Internet.

More information at: **https://digitalcloud.training/certification-training/aws-solutions-architect-associate/compute/amazon-ecs/**

AWS Lambda

AWS Lambda is a serverless computing technology that allows you to run code without provisioning or managing servers.

AWS Lambda executes code only when needed and scales automatically.

You pay only for the compute time you consume (you pay nothing when your code is not running).

Benefits of AWS Lambda:

- No servers to manage.
- Continuous scaling.
- Subsecond metering.
- Integrates with almost all other AWS services.

Primary use cases for AWS Lambda:

- Data processing.
- Real-time file processing.
- Real-time stream processing.
- Build serverless backends for web, mobile, IOT, and 3rd party API requests.

More information at: **https://digitalcloud.training/certification-training/aws-solutions-architect-associate/compute/aws-lambda/**

Amazon Lightsail

Amazon LightSail Instances

Amazon Lightsail is one of the newest services in the AWS Compute suite of products. Amazon Lightsail is great for users who do not have deep AWS technical expertise as it make it very easy to provision compute services.

Amazon Lightsail provides developers compute, storage, and networking capacity and capabilities to deploy and manage websites, web applications, and databases in the cloud.

Amazon Lightsail includes everything you need to launch your project quickly – a virtual machine, SSD-based storage, data transfer, DNS management, and a static IP.

Amazon Lightsail provides preconfigured virtual private servers (instances) that include everything required to deploy and application or create a database.

The underlying infrastructure and operating system is managed by Amazon Lightsail.

Best suited to projects that require a few dozen instances or fewer.

Provides a simple management interface.

Good for blogs, websites, web applications, e-commerce etc.

Can deploy load balancers and attach block storage.

Public API.

Limited to 20 Amazon Lightsail instances, 5 static IPs, 3 DNS zones, 20 TB block storage, 40 databases, and 5 load balancers per account.

Up to 20 certificates per calendar year.

Can connect to each other and other AWS resources through public Internet and private (VPC peering) networking.

Application templates include WordPress, WordPress Multisite, Drupal, Joomla!, Magento, Redmine, LAMP, Nginx (LEMP), MEAN, Node.js, and more.

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

Amazon LightSail Databases

Amazon Lightsail databases are instances that are dedicated to running databases.

An Amazon Lightsail database can contain multiple user-created databases, and you can access it by using the same tools and applications that you use with a stand-alone database.

Amazon Lightsail managed databases provide an easy, low maintenance way to store your data in the cloud.

Amazon Lightsail manages a range of maintenance activities and security for your database and its underlying infrastructure.

Amazon Lightsail automatically backs up your database and allows point in time restore from the past 7 days using the database restore tool.

Amazon Lightsail databases support the latest major versions of MySQL. Currently, these versions are 5.6, 5.7, and 8.0 for MySQL.

Amazon Lightsail databases are available in Standard and High Availability plans.

High Availability plans add redundancy and durability to your database, by automatically creating standby database in a separate Availability Zone.

Amazon Lightsail is very affordable.

Amazon Lightsail plans are billed on an on-demand hourly rate, so you pay only for what you use.

For every Amazon Lightsail plan you use, we charge you the fixed hourly price, up to the maximum monthly plan cost.

AWS Storage

Amazon Simple Storage Service (S3)

Amazon S3 is object storage built to store and retrieve any amount of data from anywhere – web sites and mobile apps, corporate applications, and data from IoT sensors or devices.

You can store any type of file in S3.

S3 is designed to deliver 99.999999999% durability, and stores data for millions of applications used by market leaders in every industry.

S3 provides comprehensive security and compliance capabilities that meet even the most stringent regulatory requirements.

S3 gives customers flexibility in the way they manage data for cost optimization, access control, and compliance.

Typical use cases include:

- **Backup and Storage** – Provide data backup and storage services for others.
- **Application Hosting** – Provide services that deploy, install, and manage web applications.
- **Media Hosting** – Build a redundant, scalable, and highly available infrastructure that hosts video, photo, or music uploads and downloads.
- **Software Delivery** – Host your software applications that customers can download.
- **Static Website** – you can configure a static website to run from an S3 bucket.

S3 provides query-in-place functionality, allowing you to run powerful analytics directly on your data at rest in S3. And Amazon S3 is the most supported cloud storage service available, with integration from the largest community of third-party solutions, systems integrator partners, and other AWS services.

Files can be anywhere from 0 bytes to 5 TB.

There is unlimited storage available.

Files are stored in buckets.

Buckets are root level folders.

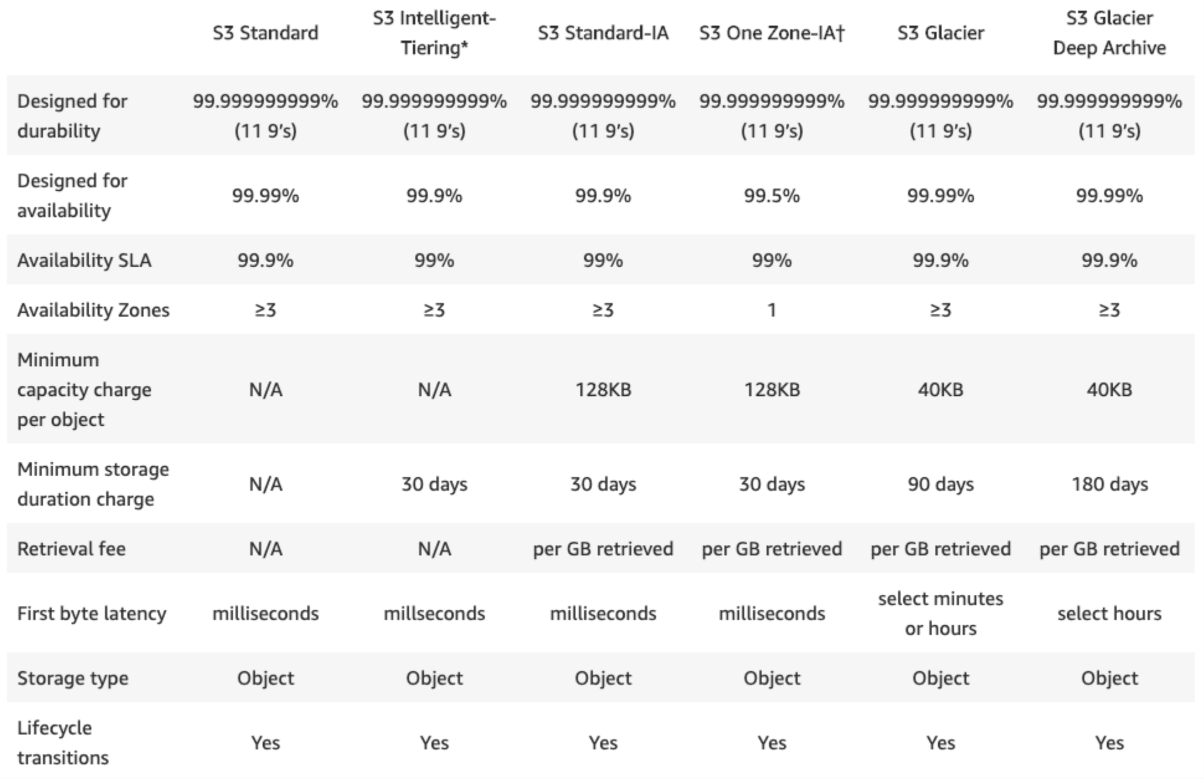
Any subfolder within a bucket is known as a “folder”.

S3 is a universal namespace so bucket names must be unique globally.

There are six S3 storage classes.

- S3 Standard (durable, immediately available, frequently accessed).
- S3 Intelligent-Tiering (automatically moves data to the most cost-effective tier).
- S3 Standard-IA (durable, immediately available, infrequently accessed).
- S3 One Zone-IA (lower cost for infrequently accessed data with less resilience).
- S3 Glacier (archived data, retrieval times in minutes or hours).
- S3 Glacier Deep Archive (lowest cost storage class for long term retention).

The table below provides the details of each Amazon S3 storage class:



	S3 Standard	S3 Intelligent-Tiering*	S3 Standard-IA	S3 One Zone-IA†	S3 Glacier	S3 Glacier Deep Archive
Designed for durability	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)
Designed for availability	99.99%	99.9%	99.9%	99.5%	99.99%	99.99%
Availability SLA	99.9%	99%	99%	99%	99.9%	99.9%
Availability Zones	≥3	≥3	≥3	1	≥3	≥3
Minimum capacity charge per object	N/A	N/A	128KB	128KB	40KB	40KB
Minimum storage duration charge	N/A	30 days	30 days	30 days	90 days	180 days
Retrieval fee	N/A	N/A	per GB retrieved	per GB retrieved	per GB retrieved	per GB retrieved
First byte latency	milliseconds	millseconds	milliseconds	milliseconds	select minutes or hours	select hours
Storage type	Object	Object	Object	Object	Object	Object
Lifecycle transitions	Yes	Yes	Yes	Yes	Yes	Yes

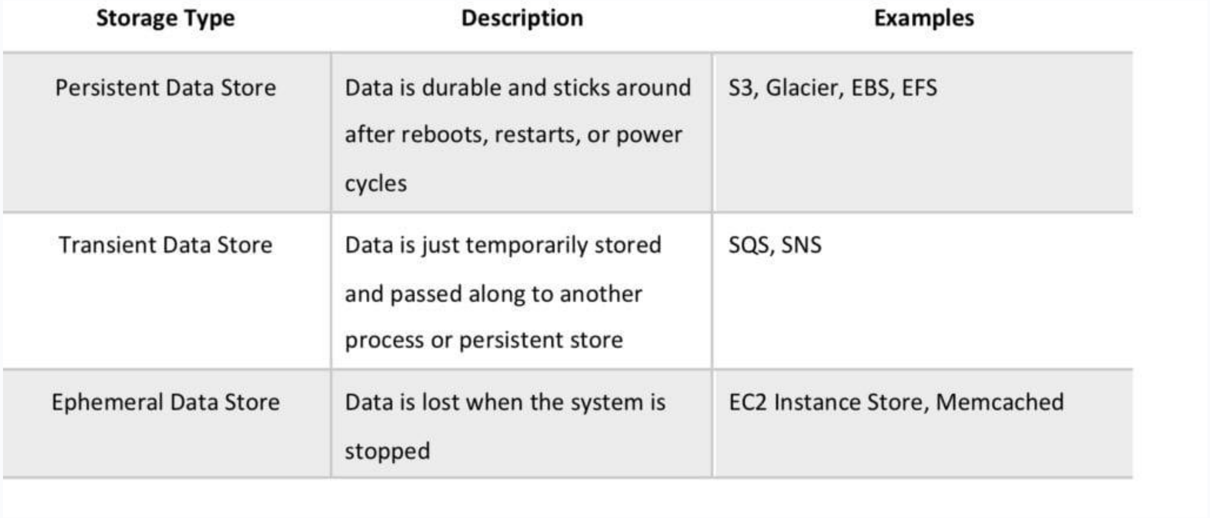
When you successfully upload a file to S3 you receive a HTTP 200 code.

S3 is a persistent, highly durable data store.

Persistent data stores are non-volatile storage systems that retain data when powered off.

This is in contrast to transient data stores and ephemeral data stores which lose the data when powered off.

The following table provides a description of persistent, transient and ephemeral data stores and which AWS service to use:



Storage Type	Description	Examples
Persistent Data Store	Data is durable and sticks around after reboots, restarts, or power cycles	S3, Glacier, EBS, EFS
Transient Data Store	Data is just temporarily stored and passed along to another process or persistent store	SQS, SNS
Ephemeral Data Store	Data is lost when the system is stopped	EC2 Instance Store, Memcached

Bucket names must follow a set of rules:

- Names must be unique across all of AWS.
- Names must be 3 to 63 characters in length.
- Names can only contain lowercase letters, numbers and hyphens.
- Names cannot be formatted as an IP address.

Data consistency:

- Read after write consistency for PUTS of new objects.
- Eventual consistency for overwrite PUTS and DELETES (takes time to propagate).

Objects consist of:

- Key (name of the object).
- Value (data made up of a sequence of bytes).
- Version ID (used for versioning).
- Metadata (data about the data that is stored).

Subresources:

- Access control lists.
- Torrent.

Built for 99.99 availability.

SLA is 99.9% availability.

Amazon guarantee 99.99999999% durability.

Object sharing – the ability to make any object publicly available via a URL.

Lifecycle management – set rules to transfer objects between storage classes at defined time intervals.

Versioning – automatically keep multiple versions of an object (when enabled).

Encryption.

Data secured using ACLs and bucket policies.

Tiers:

- S3 standard.
- S3-IA.
- S3 One Zone – IA.
- Glacier.

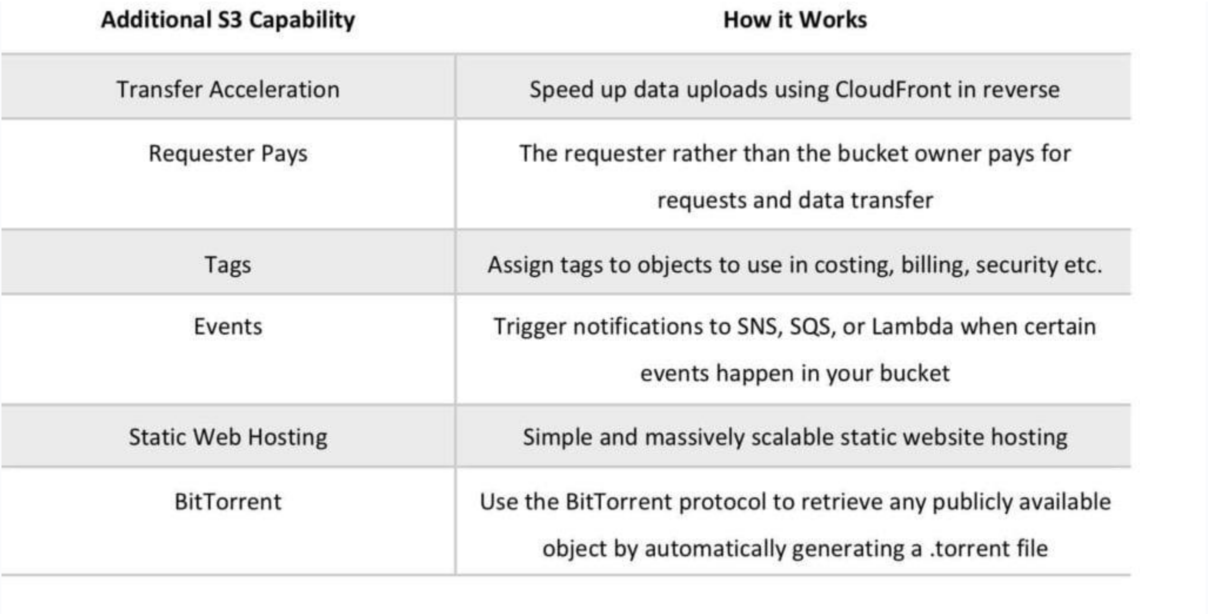
Charges:

- Storage.
- Requests.
- Storage management pricing.
- Data transfer pricing.
- Transfer acceleration.

When you create a bucket you need to select the region where it will be created.

It is a best practice to create buckets in regions that are physically closest to your users to reduce latency.

Additional capabilities offered by Amazon S3 include:



Additional S3 Capability	How it Works
Transfer Acceleration	Speed up data uploads using CloudFront in reverse
Requester Pays	The requester rather than the bucket owner pays for requests and data transfer
Tags	Assign tags to objects to use in costing, billing, security etc.
Events	Trigger notifications to SNS, SQS, or Lambda when certain events happen in your bucket
Static Web Hosting	Simple and massively scalable static website hosting
BitTorrent	Use the BitTorrent protocol to retrieve any publicly available object by automatically generating a .torrent file

The following link provides more information: **https://digitalcloud.training/certification-training/aws-solutions-architect-associate/storage/amazon-s3/**

AWS Snowball

With AWS Snowball (Snowball), you can transfer hundreds of terabytes or petabytes of data between your on-premises data centers and Amazon Simple Storage Service (Amazon S3).

Uses a secure storage device for physical transportation.

AWS Snowball Client is software that is installed on a local computer and is used to identify, compress, encrypt, and transfer data.

Uses 256-bit encryption (managed with the AWS KMS) and tamper-resistant enclosures with TPM.

Snowball (80TB) (50TB model available only in the USA).

Snowball Edge (100TB) comes with onboard storage and compute capabilities.

Snowmobile – exabyte scale with up to 100PB per Snowmobile.

Snowball can import to S3 or export from S3.

Import/export is when you send your own disks into AWS – this is being deprecated in favour of Snowball.

Snowball must be ordered from and returned to the same region.

To speed up data transfer it is recommended to run simultaneous instances of the AWS Snowball Client in multiple terminals and transfer small files as batches.

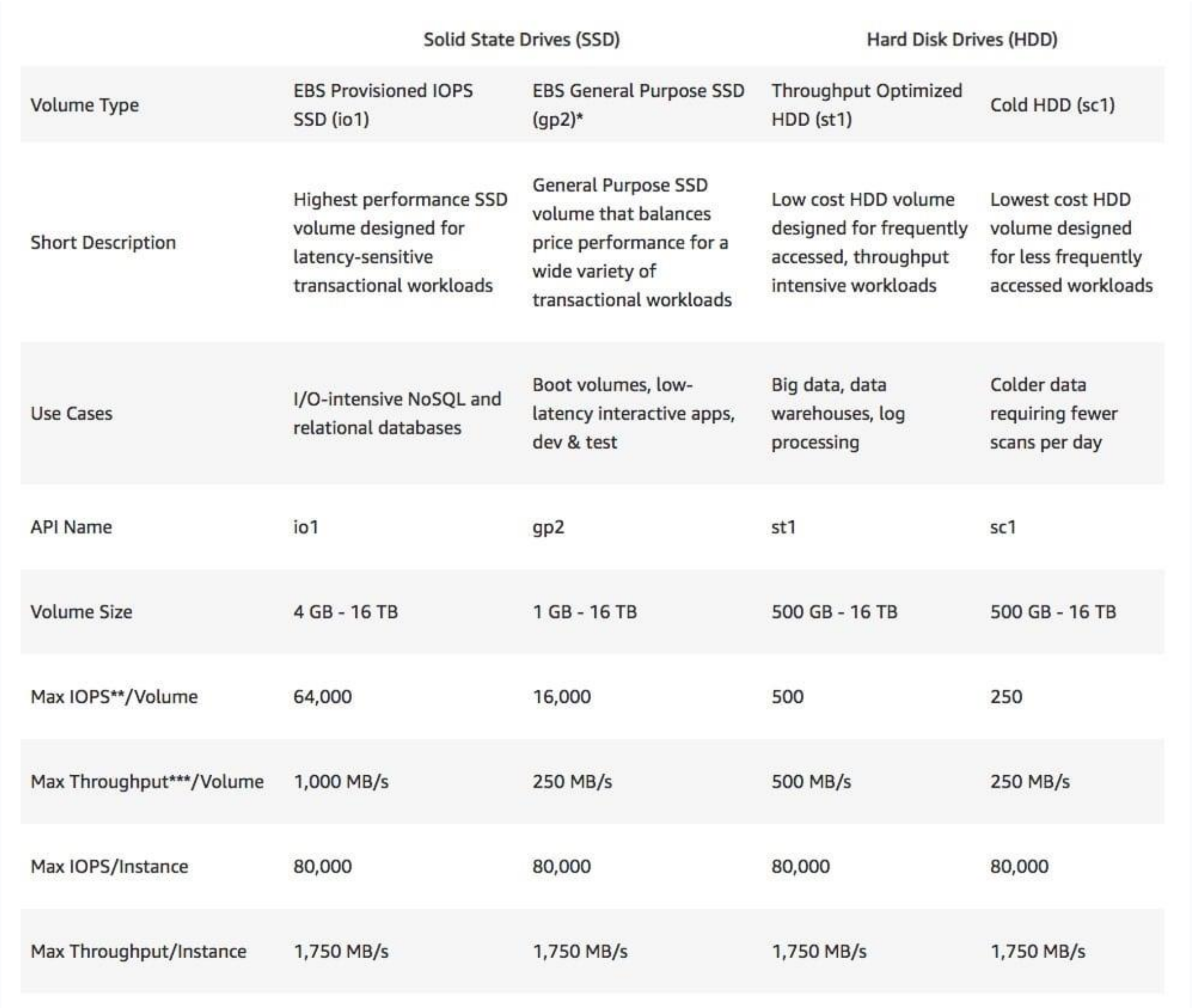
Amazon Elastic Block Store (EBS)

Amazon Elastic Block Store (Amazon EBS) provides persistent block storage volumes for use with Amazon EC2 instances in the AWS Cloud.

Each Amazon EBS volume is automatically replicated within its Availability Zone to protect you from component failure, offering high availability and durability.

Amazon EBS volumes offer the consistent and low-latency performance needed to run your workloads. With Amazon EBS, you can scale your usage up or down within minutes – all while paying a low price for only what you provision.

The following table shows a comparison of a few EBS volume types:



	Solid State Drives (SSD)		Hard Disk Drives (HDD)	
Volume Type	EBS Provisioned IOPS SSD (io1)	EBS General Purpose SSD (gp2)*	Throughput Optimized HDD (st1)	Cold HDD (sc1)
Short Description	Highest performance SSD volume designed for latency-sensitive transactional workloads	General Purpose SSD volume that balances price performance for a wide variety of transactional workloads	Low cost HDD volume designed for frequently accessed, throughput intensive workloads	Lowest cost HDD volume designed for less frequently accessed workloads
Use Cases	I/O-intensive NoSQL and relational databases	Boot volumes, low-latency interactive apps, dev & test	Big data, data warehouses, log processing	Colder data requiring fewer scans per day
API Name	io1	gp2	st1	sc1
Volume Size	4 GB - 16 TB	1 GB - 16 TB	500 GB - 16 TB	500 GB - 16 TB
Max IOPS**/Volume	64,000	16,000	500	250
Max Throughput***/Volume	1,000 MB/s	250 MB/s	500 MB/s	250 MB/s
Max IOPS/Instance	80,000	80,000	80,000	80,000
Max Throughput/Instance	1,750 MB/s	1,750 MB/s	1,750 MB/s	1,750 MB/s

EBS volume data persists independently of the life of the instance.

EBS volumes do not need to be attached to an instance.

You can attach multiple EBS volumes to an instance.

You cannot attach an EBS volume to multiple instances (use Elastic File Store instead).

EBS volumes must be in the same AZ as the instances they are attached to.

Termination protection is turned off by default and must be manually enabled (keeps the volume/data when the instance is terminated).

Root EBS volumes are deleted on termination by default.

Extra non-boot volumes are not deleted on termination by default.

The behaviour can be changed by altering the “DeleteOnTermination” attribute.

EBS Snapshots:

- Snapshots capture a point-in-time state of an instance.
- Snapshots are stored on S3.
- Does not provide granular backup (not a replacement for backup software).
- If you make periodic snapshots of a volume, the snapshots are incremental, which means that only the blocks on the device that have changed after your last snapshot are saved in the new snapshot.
- Even though snapshots are saved incrementally, the snapshot deletion process is designed so that you need to retain only the most recent snapshot in order to restore the volume.
- Snapshots can only be accessed through the EC2 APIs.
- EBS volumes are AZ specific but snapshots are region specific.

More information can be found here: **https://digitalcloud.training/certification-training/aws-solutions-architect-associate/compute/amazon-ebs/**

INSTANCE STORES

Instance store volumes are high performance local disks that are physically attached to the host computer on which an EC2 instance runs.

Instance stores are ephemeral which means the data is lost when powered off (non-persistent).

Instances stores are ideal for temporary storage of information that changes frequently, such as buffers, caches, or scratch data.

Instance store volume root devices are created from AMI templates stored on S3.

Instance store volumes cannot be detached/reattached.

Amazon Elastic File System (EFS)

EFS is a fully-managed service that makes it easy to set up and scale file storage in the Amazon Cloud.

Good for big data and analytics, media processing workflows, content management, web serving, home directories etc.

EFS uses the NFSv4.1 protocol.

Pay for what you use (no pre-provisioning required).

Can scale up to petabytes.

EFS is elastic and grows and shrinks as you add and remove data.

Can concurrently connect 1 to 1000s of EC2 instances, from multiple AZs.

A file system can be accessed concurrently from all AZs in the region where it is located.

By default you can create up to 10 file systems per account.

On-premises access can be enabled via Direct Connect or AWS VPN.

Can choose General Purpose or Max I/O (both SSD).

The VPC of the connecting instance must have DNS hostnames enabled.

EFS provides a file system interface, file system access semantics (such as strong consistency and file locking).

Data is stored across multiple AZ’s within a region.

Read after write consistency.

Need to create mount targets and choose AZ’s to include (recommended to include all AZ’s).

Instances can be behind an ELB.

There are two performance modes:

- “General Purpose” performance mode is appropriate for most file systems.
- “Max I/O” performance mode is optimized for applications where tens, hundreds, or thousands of EC2 instances are accessing the file system.

Amazon EFS is designed to burst to allow high throughput levels for periods of time.

AWS Networking

This article covers AWS Networking which is a key technology area in the Cloud Practitioner exam blueprint

Amazon Virtual Private Cloud (VPC)

A virtual private cloud (VPC) is a virtual network dedicated to your AWS account.

Analogous to having your own DC inside AWS.

It is logically isolated from other virtual networks in the AWS Cloud.

Provides complete control over the virtual networking environment including selection of IP ranges, creation of subnets, and configuration of route tables and gateways.

You can launch your AWS resources, such as Amazon EC2 instances, into your VPC.

When you create a VPC, you must specify a range of IPv4 addresses for the VPC in the form of a Classless Inter-Domain Routing (CIDR) block; for example, 10.0.0.0/16.

This is the primary CIDR block for your VPC.

A VPC spans all the Availability Zones in the region.

You have full control over who has access to the AWS resources inside your VPC.

You can create your own IP address ranges, and create subnets, route tables and network gateways.

When you first create your AWS account a default VPC is created for you in each AWS region.

A default VPC is created in each region with a subnet in each AZ.

By default you can create up to 5 VPCs per region.

You can define dedicated tenancy for a VPC to ensure instances are launched on dedicated hardware (overrides the configuration specified at launch).

A default VPC is automatically created for each AWS account the first time Amazon EC2 resources are provisioned.

The default VPC has all-public subnets.

Public subnets are subnets that have:

- "Auto-assign public IPv4 address" set to "Yes".
- The subnet route table has an attached Internet Gateway.

Instances in the default VPC always have both a public and private IP address.

AZs names are mapped to different zones for different users (i.e. the AZ "ap-southeast-2a" may map to a different physical zone for a different user).

Components of a VPC:

- **A Virtual Private Cloud**: A logically isolated virtual network in the AWS cloud. You define a VPC's IP address space from ranges you select.
- **Subnet**: A segment of a VPC's IP address range where you can place groups of isolated resources (maps to an AZ, 1:1).
- **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 Endpoints**: Enables private connectivity to services hosted in AWS, from within your VPC without using an Internet Gateway, VPN, Network Address Translation (NAT) devices, or firewall proxies.
- **Egress-only Internet Gateway**: A stateful gateway to provide egress only access for IPv6 traffic from the VPC to the Internet.

Options for securely connecting to a VPC are:

- AWS managed VPN – fast to setup.
- Direct Connect – high bandwidth, low-latency but takes weeks to months to setup.
- VPN CloudHub – used for connecting multiple sites to AWS.
- Software VPN – use 3rd party software.

An Elastic Network Interface (ENI) is a logical networking component that represents a NIC.

ENIs can be attached and detached from EC2 instances and the configuration of the ENI will be maintained.

Flow Logs capture information about the IP traffic going to and from network interfaces in a VPC.

Flow log data is stored using Amazon CloudWatch Logs.

Flow logs can be created at the following levels:

- VPC.
- Subnet.
- Network interface.

Peering connections can be created with VPCs in different regions (available in most regions now).

Data sent between VPCs in different regions is encrypted (traffic charges apply).

Subnets

After creating a VPC, you can add one or more subnets in each Availability Zone.

When you create a subnet, you specify the CIDR block for the subnet, which is a subset of the VPC CIDR block.

Each subnet must reside entirely within one Availability Zone and cannot span zones.

Types of subnet:

- If a subnet’s traffic is routed to an internet gateway, the subnet is known as a public subnet.
- If a subnet doesn’t have a route to the internet gateway, the subnet is known as a private subnet.
- If a subnet doesn’t have a route to the internet gateway, but has its traffic routed to a virtual private gateway for a VPN connection, the subnet is known as a VPN-only subnet.

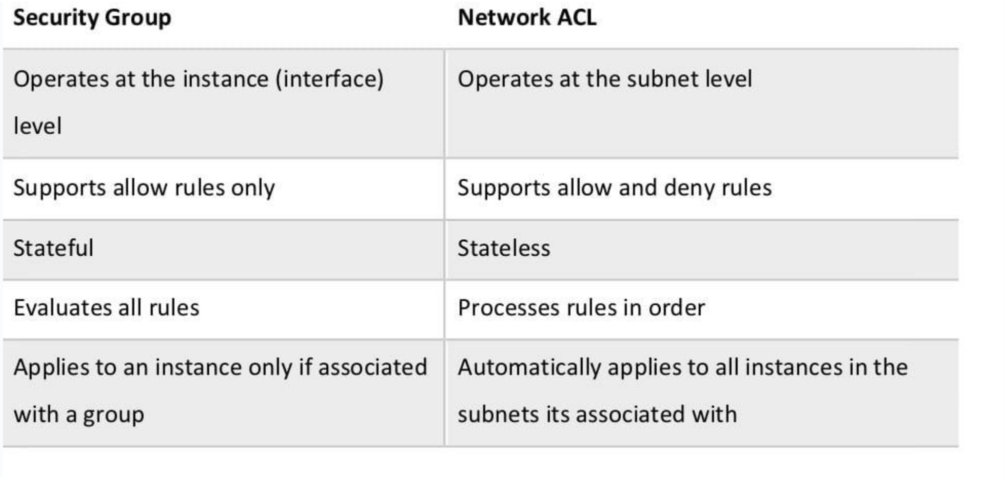
An Internet Gateway is a horizontally scaled, redundant, and highly available VPC component that allows communication between instances in your VPC and the internet.

Firewalls

Network Access Control Lists (ACLs) provide a firewall/security layer at the subnet level.

Security Groups provide a firewall/security layer at the instance level.

The table below describes some differences between Security Groups and Network ACLs:



Security Group	Network ACL
Operates at the instance (interface) level	Operates at the subnet level
Supports allow rules only	Supports allow and deny rules
Stateful	Stateless
Evaluates all rules	Processes rules in order
Applies to an instance only if associated with a group	Automatically applies to all instances in the subnets its associated with

VPC Wizard

The VPC Wizard can be used to create the following four configurations:

VPC with a Single Public Subnet:

- Your instances run in a private, isolated section of the AWS cloud with direct access to the Internet.
- Network access control lists and security groups can be used to provide strict control over inbound and outbound network traffic to your instances.
- Creates a /16 network with a /24 subnet. Public subnet instances use Elastic IPs or Public IPs to access the Internet.

VPC with Public and Private Subnets:

- In addition to containing a public subnet, this configuration adds a private subnet whose instances are not addressable from the Internet.
- Instances in the private subnet can establish outbound connections to the Internet via the public subnet using Network Address Translation (NAT).
- Creates a /16 network with two /24 subnets.
- Public subnet instances use Elastic IPs to access the Internet.
- Private subnet instances access the Internet via Network Address Translation (NAT).

VPC with Public and Private Subnets and Hardware VPN Access:

- This configuration adds an IPsec Virtual Private Network (VPN) connection between your Amazon VPC and your data center – effectively extending your data center to the cloud while also

providing direct access to the Internet for public subnet instances in your Amazon VPC.
- Creates a /16 network with two /24 subnets.
- One subnet is directly connected to the Internet while the other subnet is connected to your corporate network via an IPsec VPN tunnel.

VPC with a Private Subnet Only and Hardware VPN Access:

- Your instances run in a private, isolated section of the AWS cloud with a private subnet whose instances are not addressable from the Internet.
- You can connect this private subnet to your corporate data center via an IPsec Virtual Private Network (VPN) tunnel.
- Creates a /16 network with a /24 subnet and provisions an IPsec VPN tunnel between your Amazon VPC and your corporate network.

NAT Instances

NAT instances are managed **by** you.
Used to enable private subnet instances to access the Internet.

When creating NAT instances always disable the source/destination check on the instance.

NAT instances must be in a single public subnet.

NAT instances need to be assigned to security groups.

NAT Gateways

NAT gateways are managed **for** you by AWS.
NAT gateways are highly available in each AZ into which they are deployed.

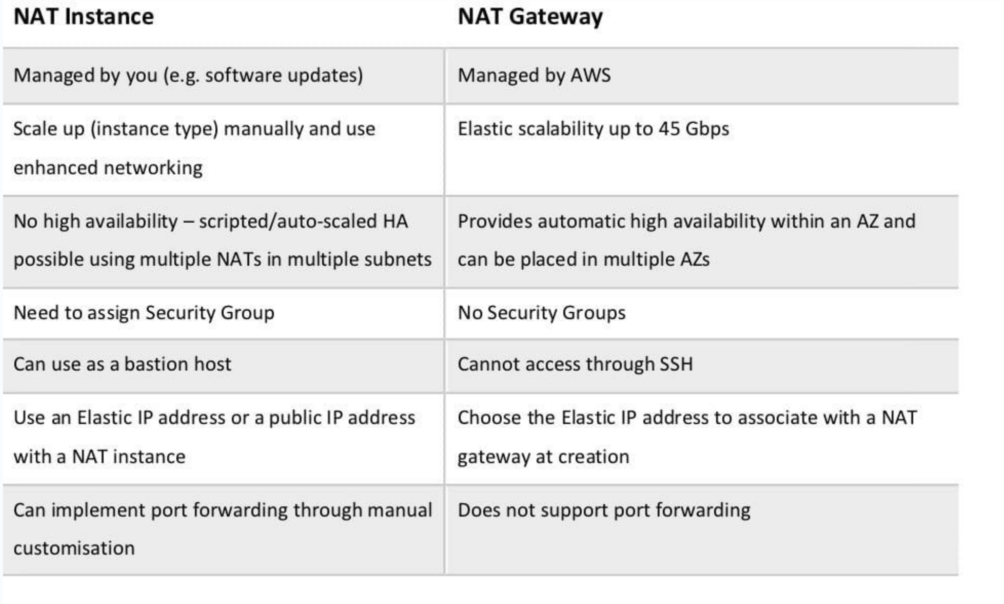
They are preferred by enterprises.

Can scale automatically up to 45Gbps.

No need to patch.

Not associated with any security groups.

The table below describes some differences between NAT instances and NAT gateways:



NAT Instance	NAT Gateway
Managed by you (e.g. software updates)	Managed by AWS
Scale up (instance type) manually and use enhanced networking	Elastic scalability up to 45 Gbps
No high availability – scripted/auto-scaled HA possible using multiple NATs in multiple subnets	Provides automatic high availability within an AZ and can be placed in multiple AZs
Need to assign Security Group	No Security Groups
Can use as a bastion host	Cannot access through SSH
Use an Elastic IP address or a public IP address with a NAT instance	Choose the Elastic IP address to associate with a NAT gateway at creation
Can implement port forwarding through manual customisation	Does not support port forwarding

Direct Connect

AWS Direct Connect is a network service that provides an alternative to using the Internet to connect a customer’s on premise sites to AWS.

Data is transmitted through a private network connection between AWS and a customer’s datacenter or corporate network.

Benefits:

- Reduce cost when using large volumes of traffic.
- Increase reliability (predictable performance).
- Increase bandwidth (predictable bandwidth).
- Decrease latency.

Each AWS Direct Connect connection can be configured with one or more virtual interfaces (VIFs).

Public VIFs allow access to public services such as S3, EC2, and DynamoDB.

Private VIFs allow access to your VPC.

From Direct Connect you can connect to all AZs **within the region.**
You can establish IPSec connections over public VIFs to remote regions.

Direct Connect is charged by port hours and data transfer.

Available in 1Gbps and 10Gbps.

Speeds of 50Mbps, 100Mbps, 200Mbps, 300Mbps, 400Mbps, and 500Mbps can be purchased through AWS Direct Connect Partners.

Uses Ethernet trunking (802.1q).

Each connection consists of a single dedicated connection between ports on the customer router and an Amazon router.

for HA you must have 2 DX connections – can be active/active or active/standby.

Route tables need to be updated to point to a Direct Connect connection.

VPN can be maintained as a backup with a higher BGP priority.

You cannot extend your on-premise VLANs into the AWS cloud using Direct Connect.

AWS Global Accelerator

AWS Global Accelerator is a service that improves the availability and performance of applications with local or global users.

It provides static IP addresses that act as a fixed entry point to application endpoints in a single or multiple AWS Regions, such as Application Load Balancers, Network Load Balancers or EC2 instances.

Uses the AWS global network to optimize the path from users to applications, improving the performance of TCP and UDP traffic.

AWS Global Accelerator continually monitors the health of application endpoints and will detect an unhealthy endpoint and redirect traffic to healthy endpoints in less than 1 minute.

Here’s a video providing an overview of how AWS Global Accelerator works:

Details and Benefits

Uses redundant (two) static anycast IP addresses in different network zones (A and B).

The redundant pair are globally advertized.

Uses AWS Edge Locations – addresses are announced from multiple edge locations at the same time.

Addresses are associated to regional AWS resources or endpoints.

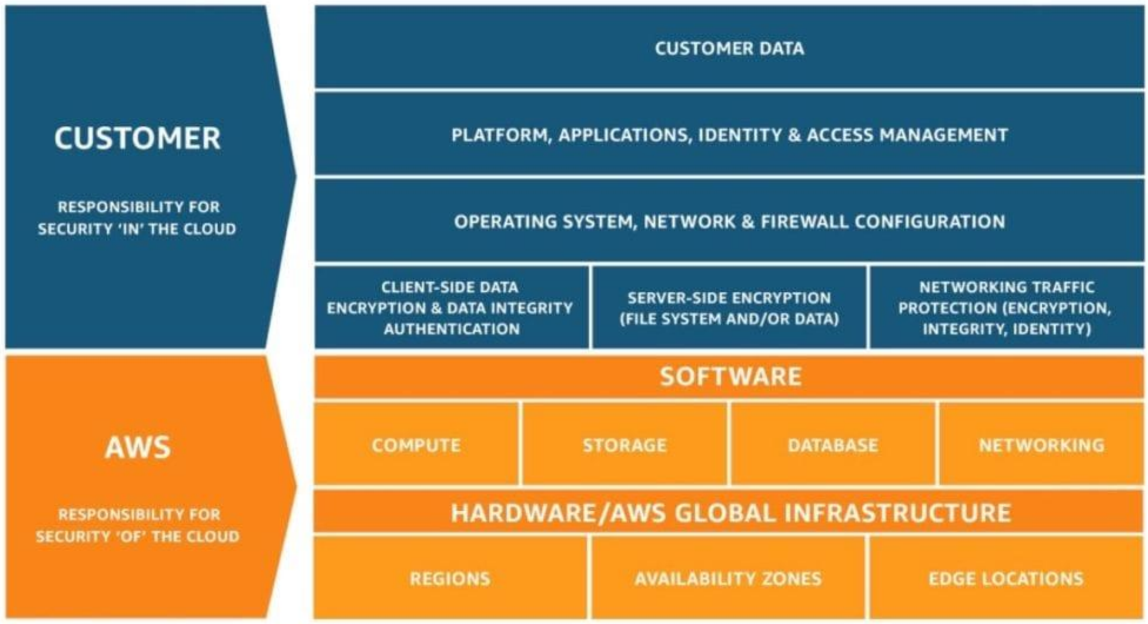
AWS Global Accelerator’s IP addresses serve as the frontend interface of applications.

Intelligent traffic distribution: Routes connections to the closest point of presence for applications.

Targets can be Amazon EC2 instances or Elastic Load Balancers (ALB and NLB).

By using the static IP addresses, you don’t need to make any client-facing changes or update DNS records as you modify or replace endpoints.

The addresses are assigned to your accelerator for as long as it exists, even if you disable the accelerator and it no longer accepts or routes traffic.


CUSTOMER
RESPONSIBILITY FOR SECURITY 'IN' THE CLOUD
CUSTOMER DATA
PLATFORM, APPLICATIONS, IDENTITY & ACCESS MANAGEMENT
OPERATING SYSTEM, NETWORK & FIREWALL CONFIGURATION
CLIENT-SIDE DATA ENCRYPTION & DATA INTEGRITY AUTHENTICATION
SERVER-SIDE ENCRYPTION (FILE SYSTEM AND/OR DATA)
NETWORKING TRAFFIC PROTECTION (ENCRYPTION, INTEGRITY, IDENTITY)
AWS
RESPONSIBILITY FOR SECURITY 'OF' THE CLOUD
SOFTWARE
COMPUTE
STORAGE
DATABASE
NETWORKING
HARDWARE/AWS GLOBAL INFRASTRUCTURE
REGIONS
AVAILABILITY ZONES
EDGE LOCATIONS

AWS Security

As an AWS customer you inherit all the best practices of AWS policies, architecture, and operational processes.

The AWS Cloud enables a shared responsibility model.

AWS manages security OF the cloud, you are responsible for security IN the cloud .

You retain control of the security you choose to implement to protect your own content, platform, applications, systems, and networks no differently than you would in an on-site data center.

Benefits of AWS Security

- **Keep Your Data Safe** – the AWS infrastructure puts strong safeguards in place to help.
- **Protect your privacy** – All data is stored in highly secure AWS data centers.
- **Meet Compliance Requirements** – AWS manages dozens of compliance programs in its infrastructure. This means that segments of your compliance have already been completed.
- **Save Money** – cut costs by using AWS data centers. Maintain the highest standard of s security without having to manage your own facility.

- **Scale Quickly** – security scales with your AWS Cloud usage. No matter the size of your business, the AWS infrastructure is designed to keep your data safe.

Compliance

AWS Cloud Compliance enables you to understand the robust controls in place at AWS to maintain security and data protection in the cloud.

As systems are built on top of AWS Cloud infrastructure, compliance responsibilities will be shared.

Compliance programs include:

-
  -
  - Certifications / attestations.
  - Laws, regulations, and privacy.
  - Alignments / frameworks.

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 regulatory compliance.

With AWS Config, you can discover existing and deleted AWS resources, determine your overall compliance against rules, and dive into configuration details of a resource at any point in time. AWS Config enables compliance auditing, security analysis, resource change tracking, and troubleshooting.

AWS Service Catalog

You can use AWS Service Catalog to create and manage catalogs of IT services that you have approved for use on AWS, including virtual machine images, servers, software, and databases to complete multi-tier application architectures.

AWS Service Catalog allows you to centrally manage commonly deployed IT services, and helps you achieve consistent governance to meet your compliance requirements, while enabling users to quickly deploy the approved IT services they need.

Amazon GuardDuty

Amazon GuardDuty offers threat detection and continuous security monitoring for malicious or unauthorized behavior to help you protect your AWS accounts and workloads.

The service monitors for activity that indicate a possible account compromise, potentially compromised instance, or reconnaissance by attackers or intellectual property, and continuously monitors data access activity for anomalies that might single unauthorized access or inadvertent data leaks.

AWS WAF & AWS Shield

WAF:

- AWS WAF is a web application firewall.
- Protects against common exploits that could compromise application availability, compromise security or consume excessive resources.

Shield:

- AWS Shield is a managed Distributed Denial of Service (DDoS) protection service.
- Safeguards web application running on AWS with always-on detection and automatic inline mitigations.
- Helps to minimize application downtime and latency.
- Two tiers – Standard and Advanced.

AWS Key Management Service

AWS Key Management Service gives you centralized control over the encryption keys used to protect your data.

You can create, import, rotate, disable, delete, define usage policies for, and audit the use of encryption keys used to encrypt your data.

AWS Key Management Service is integrated with most other AWS services making it easy to encrypt the data you store in these services with encryption keys you control.

AWS KMS is integrated with AWS CloudTrail which provides you the ability to audit who used which keys, on which resources, and when.

AWS KMS enables developers to easily encrypt data, whether through 1-click encryption in the AWS Management Console, or using the AWS SDK to easily add encryption in their application code.

https://aws.amazon.com/kms/features/

AWS CloudHSM

AWS CloudHSM is a cloud-based hardware security module (HSM) that enables you to easily generate and use your own encryption keys on the AWS Cloud.

With CloudHSM, you can manage your own encryption keys using FIPS 140-2 Level 3 validated HSMs.

CloudHSM offers you the flexibility to integrate with your applications using industry-standard APIs, such as PKCS#11, Java Cryptography Extensions (JCE), and Microsoft CryptoNG (CNG) libraries.

https://aws.amazon.com/cloudhsm/features/

AWS Artifact

AWS Artifact is your go-to, central resource for compliance-related information that matters to you.

It provides on-demand access to AWS’ security and compliance reports and select online agreements.

Reports available in AWS Artifact include our Service Organization Control (SOC) reports, Payment Card Industry (PCI) reports, and certifications from accreditation bodies across geographies and compliance verticals that validate the implementation and operating effectiveness of AWS security controls.

Agreements available in AWS Artifact include the Business Associate Addendum (BAA) and the Nondisclosure Agreement (NDA).

https://aws.amazon.com/artifact/

AWS Inspector and AWS Trusted Advisor

AWS Inspector:

- Inspector is an automated security assessment service that helps improve the security and compliance of applications deployed on AWS.
- Inspector automatically assesses applications for vulnerabilities or deviations from best practices.
- Uses an agent installed on EC2 instances.
- Instances must be tagged.

AWS Trusted Advisor:

- Trusted Advisor is an online resource that helps to reduce cost, increase performance and improve security by optimizing your AWS environment.
- Trusted Advisor provides real time guidance to help you provision your resources following best practices.
- Advisor will advise you on Cost Optimization, Performance, Security, and Fault Tolerance.

Trusted Advisor scans your AWS infrastructure and compares is to AWS best practices in five categories:

- Cost Optimization.
- Performance.
- Security.
- Fault Tolerance.
- Service Limits.

Trusted Advisor comes in two versions.

Core Checks and Recommendations (free):

-
    -
    - Access to the 7 core checks to help increase security and performance.
    - Checks include S3 bucket permissions, Security Groups, IAM use, MFA on root account, EBS public snapshots, RDS public snapshots.

Full Trusted Advisor Benefits (business and enterprise support plans):

-
    -
    - Full set of checks to help optimize your entire AWS infrastructure.
    - Advises on security, performance, cost, fault tolerance and service limits.
    - Additional benefits include weekly update notifications, alerts, automated actions with CloudWatch and programmatic access using the AWS Support API.

AWS Personal Health Dashboard

AWS Personal Health Dashboard provides alerts and remediation guidance when AWS is experiencing events that may impact you.

Personal Health Dashboard gives you a personalized view into the performance and availability of the AWS services underlying your AWS resources.

The dashboard displays relevant and timely information to help you manage events in progress.

Also provides proactive notification to help you plan for scheduled activities.

Alerts are triggered by changes in the health of AWS resources, giving you event visibility, and guidance to help quickly diagnose and resolve issues.

You get a personalized view of the status of the AWS services that power your applications, enabling you to quickly see when AWS is experiencing issues that may impact you.

Also provides forward looking notifications, and you can set up alerts across multiple channels, including email and mobile notifications, so you receive timely and relevant information to help plan for scheduled changes that may affect you.

Alerts include remediation details and specific guidance to enable you to take immediate action to address AWS events impacting your resources.

Can integrate with Amazon CloudWatch Events, enabling you to build custom rules and select targets such as AWS Lambda functions to define automated remediation actions.

The AWS Health API allows you to integrate health data and notifications with your existing in-house or third-party IT Management tools.

Penetration Testing

Penetration testing is the practice of testing one's own application's security for vulnerabilities by simulating an attack.

AWS allows penetration testing. There is a limited set of resources on which penetration testing can be performed.

You do not need permission to perform penetration testing against the following services:

- Amazon EC2 instances, NAT Gateways, and Elastic Load Balancers.
- Amazon RDS.
- Amazon CloudFront.
- Amazon Aurora.
- Amazon API Gateways.
- AWS Lambda and Lambda Edge functions.
- Amazon Lightsail resources.
- Amazon Elastic Beanstalk environments.

You can read the full vulnerability and penetration testing support policy **here**.
In case an account is or may be compromised, AWS recommend that the following steps are taken:

-
-

1. Change your AWS root account password.
2. Change all IAM user’s passwords.
3. Delete or rotate all programmatic (API) access keys.
4. Delete any resources in your account that you did not create.
5. Respond to any notifications you received from AWS through the AWS Support Center and/or contact AWS Support to open a support case.