Cloud Computing

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

Advantages of Cloud Computing-

- ➢Go global in minutes
- ➢Trade capital expense
- ➢Benefit from massive economies of scale
- ➢No guessing about capacity
- ➢Increase speed and agility

Types of Cloud Computing

As cloud computing has grown in popularity, several different models and deployment strategies have emerged to help meet specific needs of different users**.** Each type of cloud service and deployment method provides you with different levels of control, flexibility, and management.

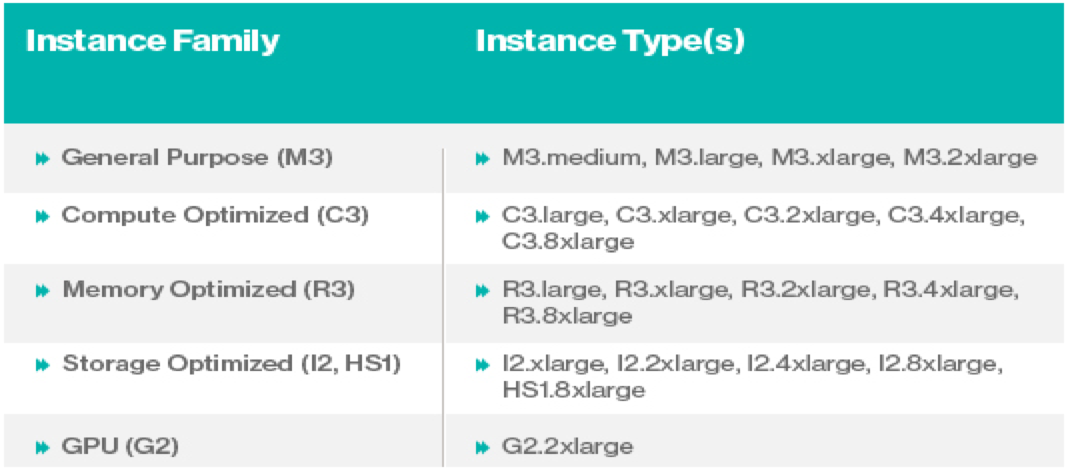
Infrastructure as a Service (IaaS) : Contains the basic building blocks for cloud IT and typically provide access to networking features, computers (virtual or on dedicated hardware), and data storage space. IaaS provides you with the **highest level of flexibility and management control** over your IT resources and is most similar to existing IT resources.

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**.

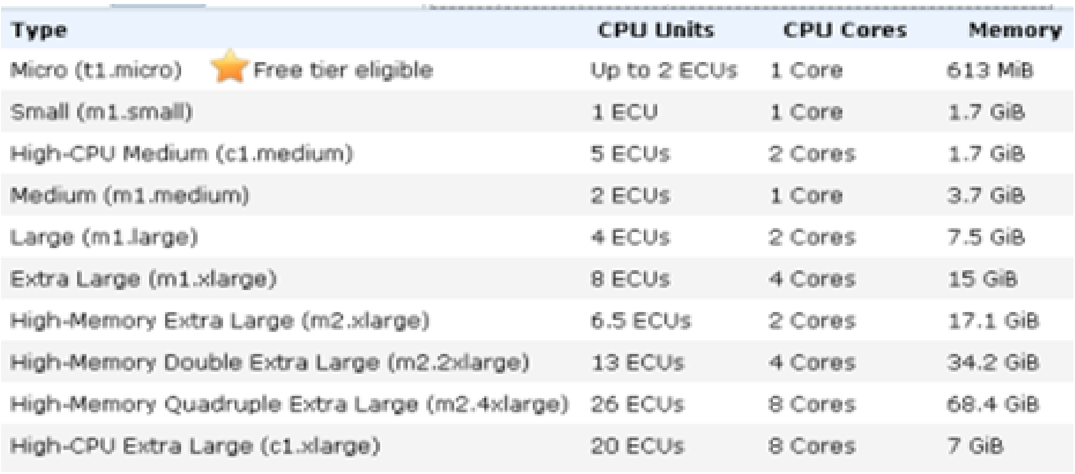
Software as a Service (SaaS): With a SaaS offering you do not have to think about how the service is maintained or how the underlying infrastructure is managed; you only need to think about how you will use that particular piece of software

AWS EC2:

- ➢Provides scalable virtual servers in the cloud
- ➢These virtual servers are known as instances & are made up of different types.
- ➢Various configurations of CPU, memory, storage, and networking capacity for your instances, known as *instance types*



Instance Family	Instance Type(s)
General Purpose (M3)	M3.medium, M3.large, M3.xlarge, M3.2xlarge
Compute Optimized (C3)	C3.large, C3.xlarge, C3.2xlarge, C3.4xlarge, C3.8xlarge
Memory Optimized (R3)	R3.large, R3.xlarge, R3.2xlarge, R3.4xlarge, R3.8xlarge
Storage Optimized (I2, HS1)	I2.xlarge, I2.2xlarge, I2.4xlarge, I2.8xlarge, HS1.8xlarge
GPU (G2)	G2.2xlarge



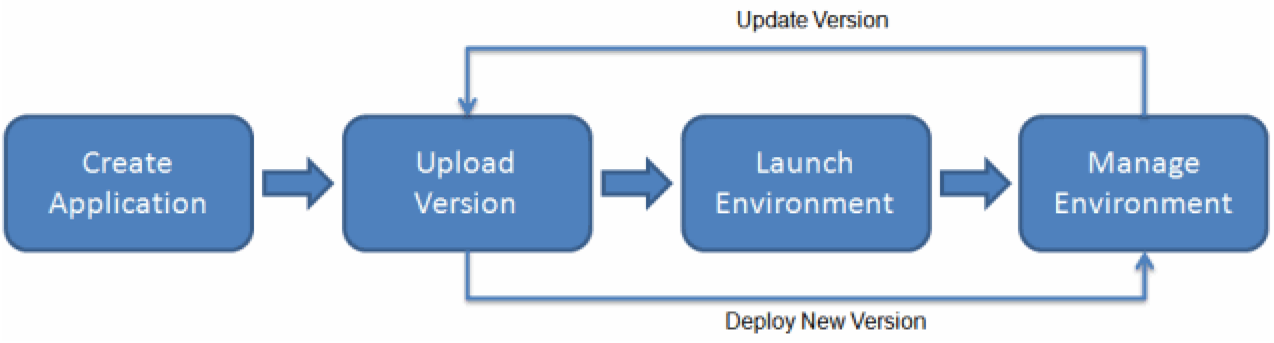
Type	CPU Units	CPU Cores	Memory
Micro (t1.micro) ★ Free tier eligible	Up to 2 ECUs	1 Core	613 MiB
Small (m1.small)	1 ECU	1 Core	1.7 GiB
High-CPU Medium (c1.medium)	5 ECUs	2 Cores	1.7 GiB
Medium (m1.medium)	2 ECUs	1 Core	3.7 GiB
Large (m1.large)	4 ECUs	2 Cores	7.5 GiB
Extra Large (m1.xlarge)	8 ECUs	4 Cores	15 GiB
High-Memory Extra Large (m2.xlarge)	6.5 ECUs	2 Cores	17.1 GiB
High-Memory Double Extra Large (m2.2xlarge)	13 ECUs	4 Cores	34.2 GiB
High-Memory Quadruple Extra Large (m2.4xlarge)	26 ECUs	8 Cores	68.4 GiB
High-CPU Extra Large (c1.xlarge)	20 ECUs	8 Cores	7 GiB

Elastic Beanstalk:

Amazon Web Services (AWS) comprises dozens of services, each of which exposes an area of functionality. While the variety of services offers flexibility for how you want to manage your AWS infrastructure, it can be challenging to figure out which services to use and how to provision them.

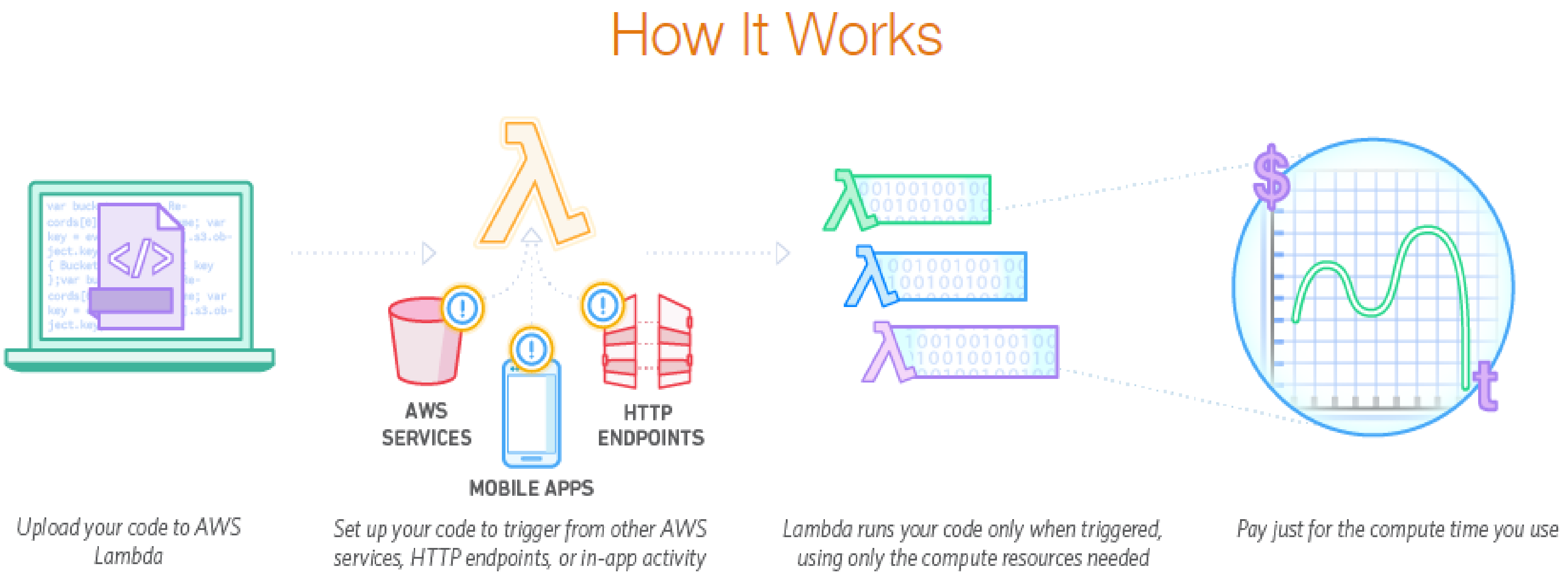
You can quickly deploy and manage applications in the AWS Cloud without worrying about the infrastructure that runs those applications.

Reduces management complexity without restricting choice or control. You simply upload your application, and Elastic Beanstalk automatically handles the details of capacity provisioning, load balancing, scaling, and application health monitoring.



Lambda:

- ➢Server-less computing platform.
- ➢Server-less means you can run code without provisioning & managing servers.
- ➢You pay only for the compute time you consume - there is no charge when your code is not running.



How It Works

Upload your code to AWS Lambda

Set up your code to trigger from other AWS services, HTTP endpoints, or in-app activity

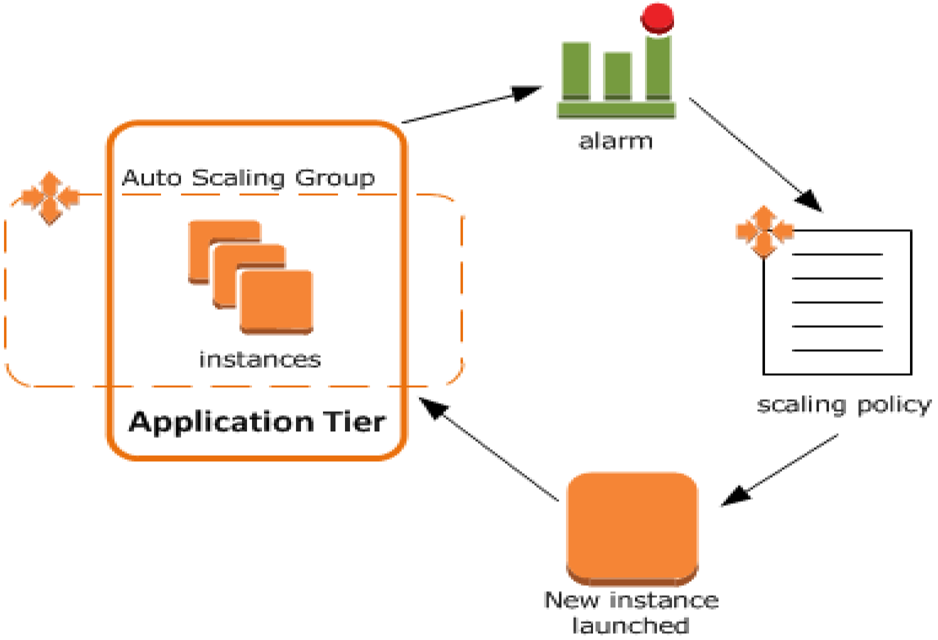
Lambda runs your code only when triggered, using only the compute resources needed

Pay just for the compute time you use

Auto Scaling:

Service that helps you ensure that you have the correct number of Amazon EC2 instances available to handle the load for your application.

- Better fault tolerance
- Better availability
- Better cost management

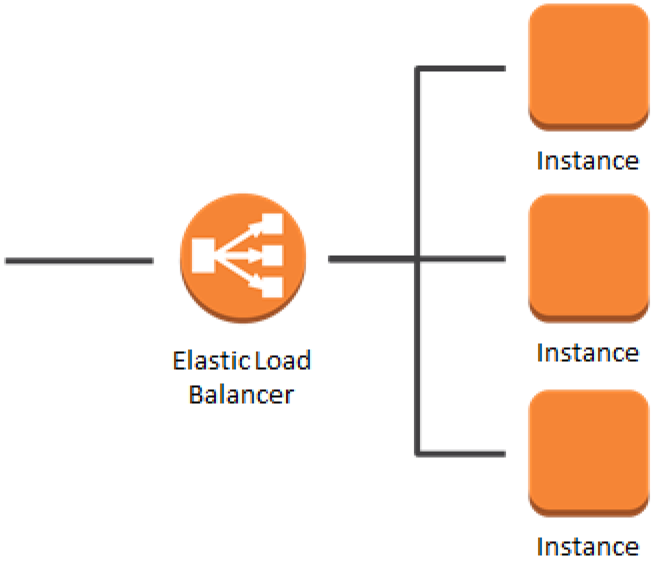


Elastic Load Balancing

Service that automatically distributes your incoming application traffic across to multiple EC2 instances that are associated with it.

It monitors the health of registered targets and routes traffic only to the healthy targets.

Commonly used with Auto-scaling to provide better availability & fault tolerance.

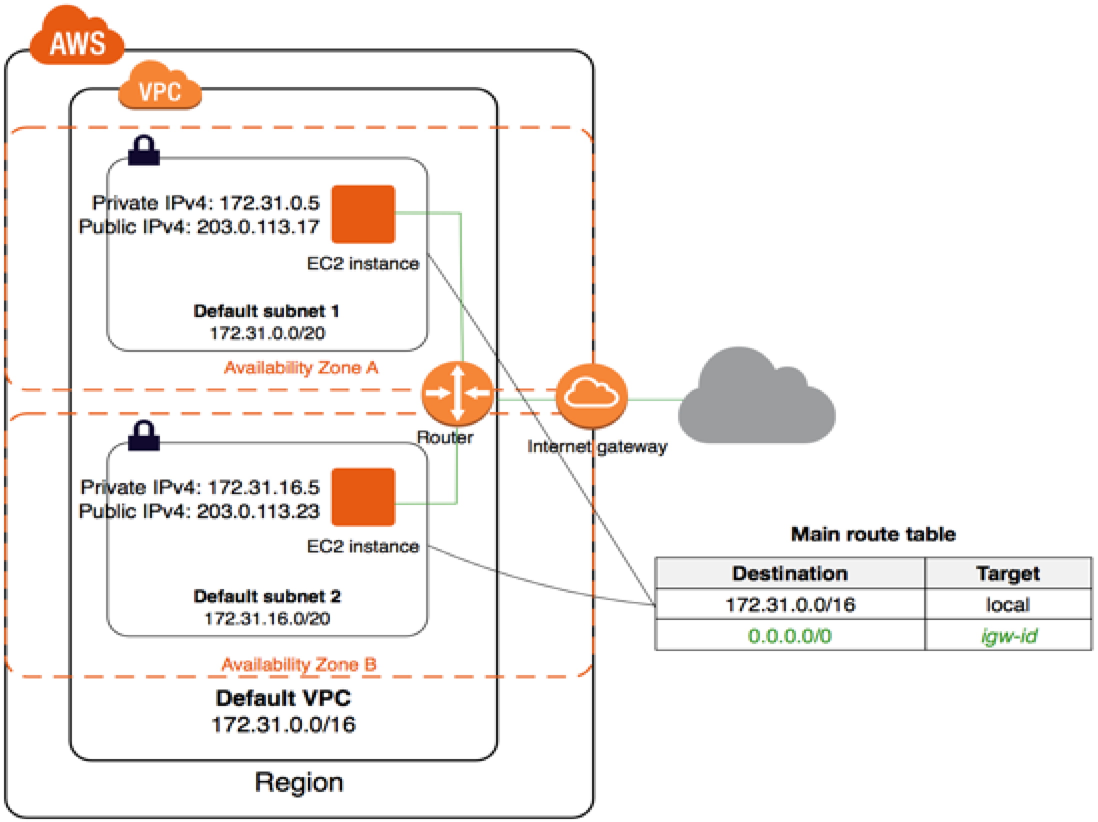


Amazon Virtual Private Cloud (Amazon VPC):

➢Lets you provision a logically isolated section of the Amazon Web Services (AWS) cloud where you can launch AWS resources in a virtual network that you define.

➢You have complete control over your virtual networking environment, including selection of your own IP address range, creation of subnets, and configuration of route tables and network gateways.

➢You can use both IPv4 and IPv6 in your VPC for secure and easy access to resources and applications.



Main route table

Destination	Target
172.31.0.0/16	local
0.0.0.0/0	*igw-id*

VPC Multiple Connectivity options:

➢Connect directly to the Internet
➢Connect to the Internet using Network Address Translation
➢Connect securely to your corporate datacenter using standard, encrypted IPsec hardware VPN connection or using **Direct Connect**
➢Connect privately to other VPCs- Peer VPCs together to share resources across multiple virtual networks owned by your or other AWS accounts.
➢Combine connectivity methods to match the needs of your application– You can connect your VPC to both the Internet and your corporate datacenter

Storage:

Elastic Block Storage:

Network attached storage.

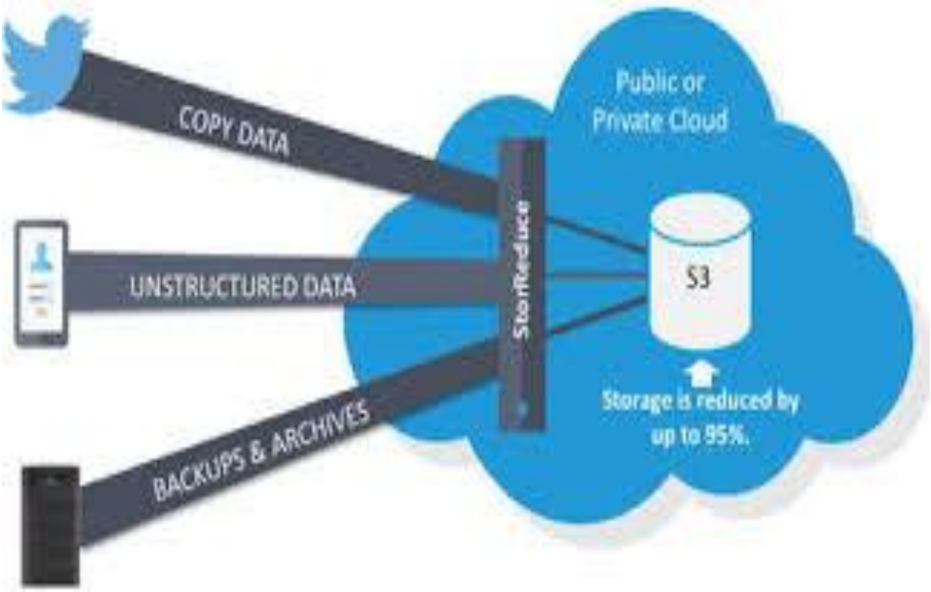
Provides persistent block storage volumes for use with EC2 instances in the AWS Cloud.

- ➢Each **EBS** volume is automatically replicated within its Availability Zone to protect you from component failure, offering high availability and durability.
- ➢Reliable, Secure Storage
- ➢Consistent, Low-latency Performance
- ➢Backup, Restore
- ➢Quickly Scale Up, Easily Scale Down
- ➢Geographic Flexibility
- ➢Optimized Performance

Simple storage Service (S3)

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. It is ideal for capturing data like mobile device photos and videos, mobile and other device backups, machine backups, machine-generated log files

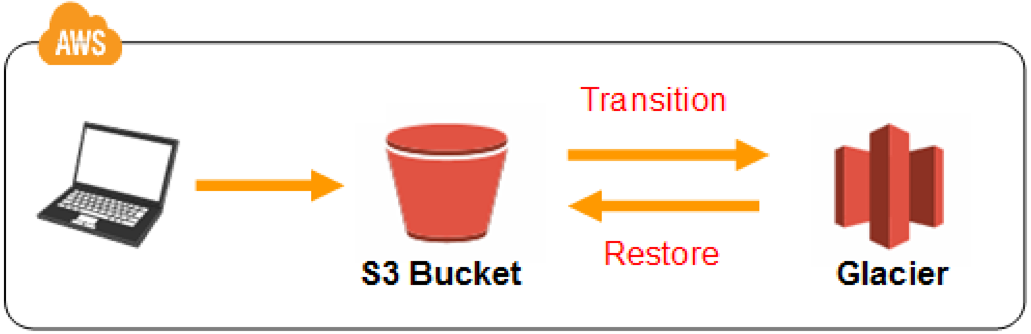
Secure
Highly Available & Durable: It is designed to deliver 99.999999999% durability.
Storage: Infinite: storing billions of objects and exabytes of data.
High Performance



Glacier:

It is an extremely low-cost and highly durable object storage service for long-term backup (months, years, or even decades) and archive of any type of data (infrequently used data, or "cold data.")

Secure
Highly Available & Durable
Storage: Infinite
Low cost
Integrated with S3

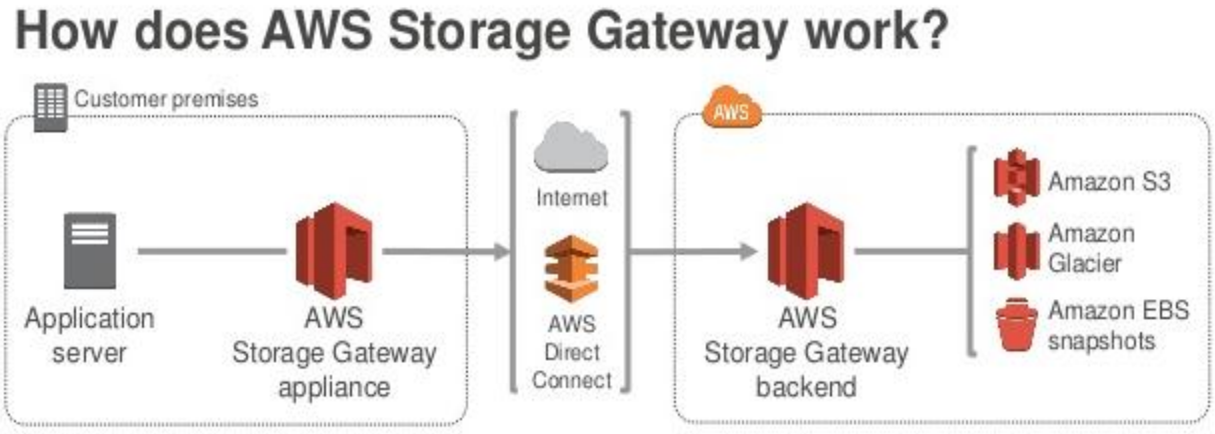


AWS Storage Gateway:

It is a hybrid storage service that enables your on-premises applications to seamlessly use storage in the AWS Cloud.
You can use the service for backup and archiving, disaster recovery,storage tiering, and migration. Your applications connect to the service through a gateway appliance using standard storage protocols, such as NFS and iSCSI.

The gateway connects to AWS storage services, such as Amazon S3, Amazon Glacier, and Amazon EBS, providing storage for files, volumes, and virtual tapes in AWS.

The service includes a highly-optimized data transfer mechanism, with bandwidth management, automated network resilience, and efficient data transfer, along with a local cache for low-latency on-premises access to your most active data.



Snowball:

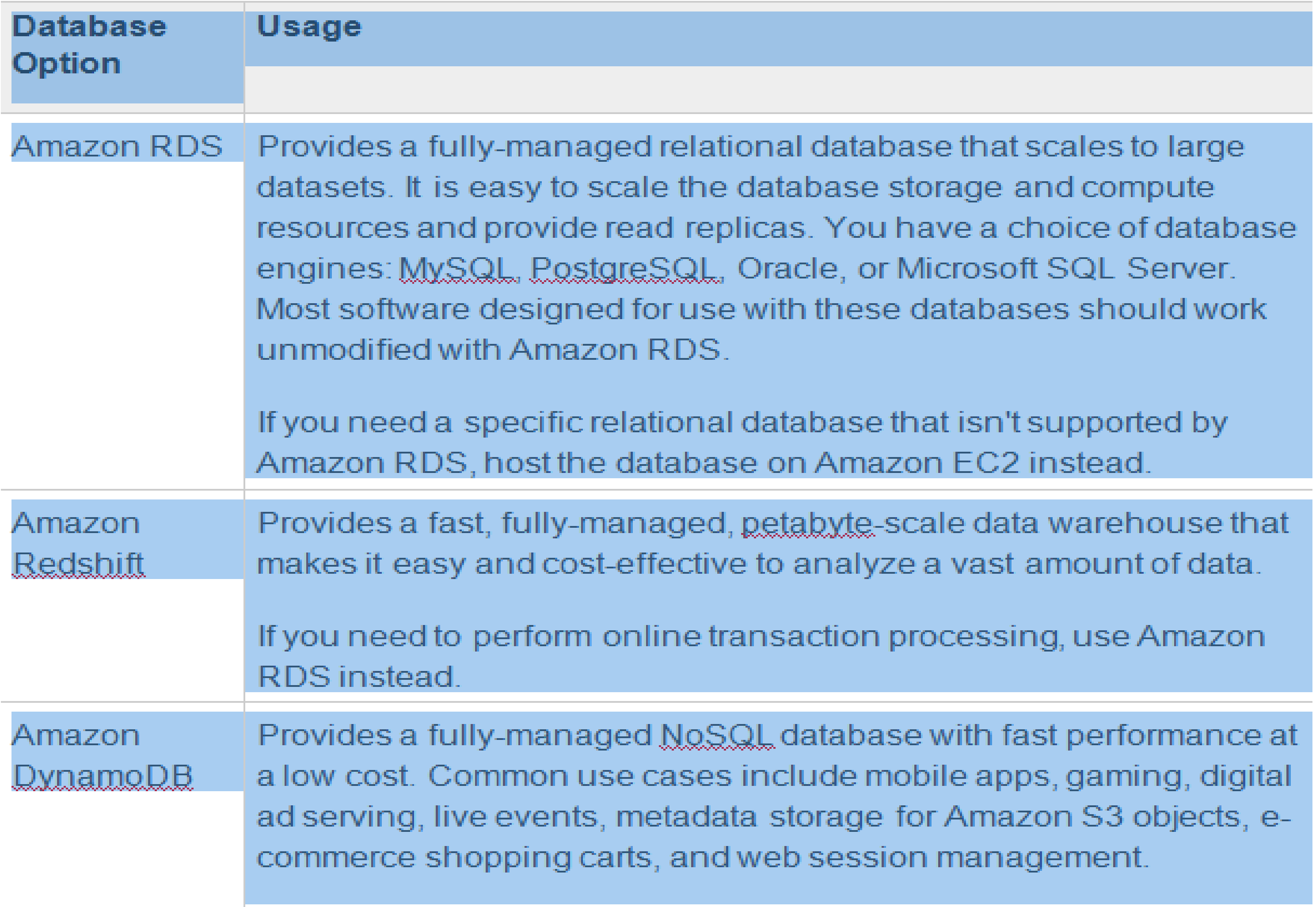
➢It is a petabyte-scale data transport solution that uses secure appliances to transfer large amounts of data into and out of the AWS cloud.

➢Using Snowball addresses common challenges with large-scale data transfers including high network costs, long transfer times, and security concerns.

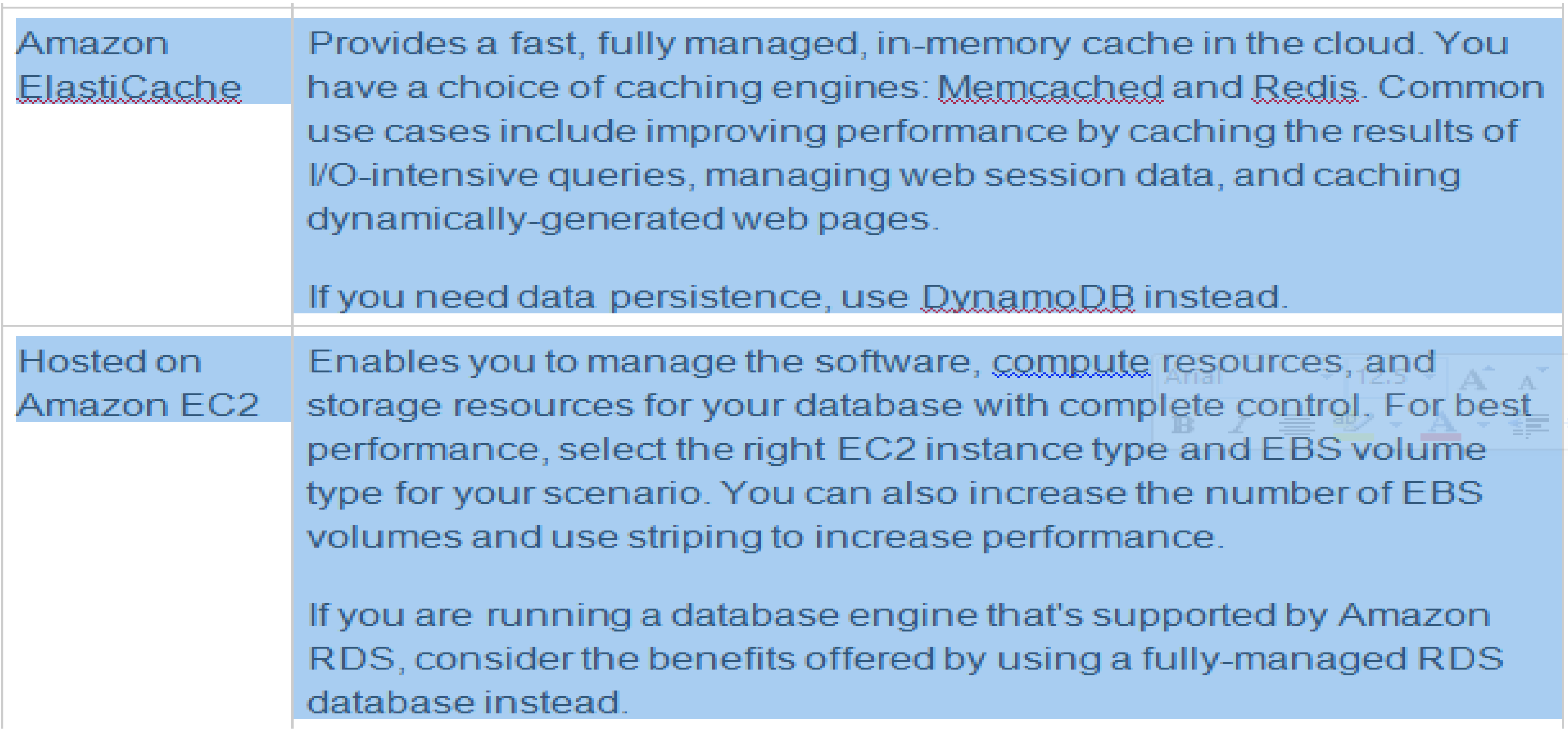
➢Transferring data with Snowball is simple, fast, secure, and can be as little as one-fifth the cost of high-speed Internet.

AWS: Database Services

AWS offers a wide range of database services to fit your application requirements. These database services are fully managed and can be launched in minutes with just a few clicks.



Database Option	Usage
Amazon RDS	Provides a fully-managed relational database that scales to large datasets. It is easy to scale the database storage and compute resources and provide read replicas. You have a choice of database engines: MySQL, PostgreSQL, Oracle, or Microsoft SQL Server. Most software designed for use with these databases should work unmodified with Amazon RDS. If you need a specific relational database that isn't supported by Amazon RDS, host the database on Amazon EC2 instead.
Amazon Redshift	Provides a fast, fully-managed, petabyte-scale data warehouse that makes it easy and cost-effective to analyze a vast amount of data. If you need to perform online transaction processing, use Amazon RDS instead.
Amazon DynamoDB	Provides a fully-managed NoSQL database with fast performance at a low cost. Common use cases include mobile apps, gaming, digital ad serving, live events, metadata storage for Amazon S3 objects, e-commerce shopping carts, and web session management.



Amazon ElastiCache	Provides a fast, fully managed, in-memory cache in the cloud. You have a choice of caching engines: Memcached and Redis. Common use cases include improving performance by caching the results of I/O-intensive queries, managing web session data, and caching dynamically-generated web pages. If you need data persistence, use DynamoDB instead.
Hosted on Amazon EC2	Enables you to manage the software, compute resources, and storage resources for your database with complete control. For best performance, select the right EC2 instance type and EBS volume type for your scenario. You can also increase the number of EBS volumes and use striping to increase performance. If you are running a database engine that's supported by Amazon RDS, consider the benefits offered by using a fully-managed RDS database instead.

Route 53:

It is Domain Name System service provided by AWS

- Domain Registration
- DNS services
- Health check
- Multiple routing algorithms
    - Simple
    - Latency
    - Fail-over
    - Weighted
    - Geo

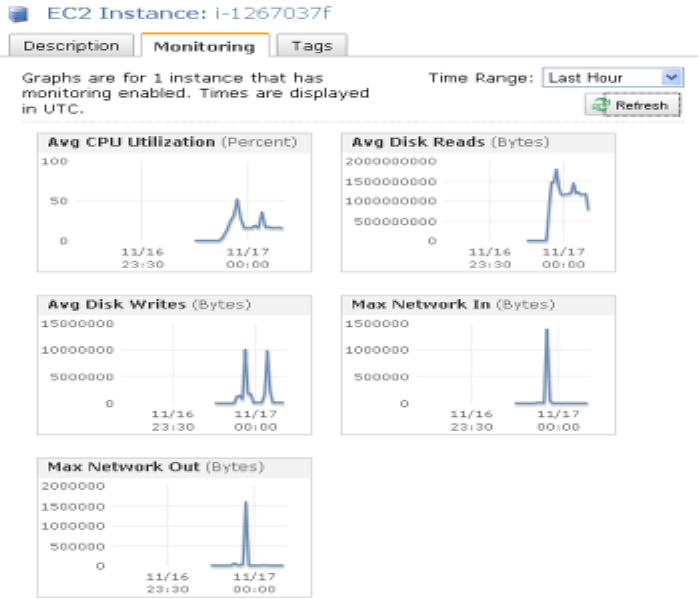
Management Tools

CloudWatch

Amazon CloudWatch is a monitoring service for AWS cloud resources and the applications you run on AWS.

To collect and track metrics, collect and monitor log files, set alarms, and automatically react to changes in your AWS resources. You can use these insights to react and keep your application running smoothly.

It 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.



EC2 Instance: i-1267037f

Graphs are for 1 instance that has monitoring enabled. Times are displayed in UTC.

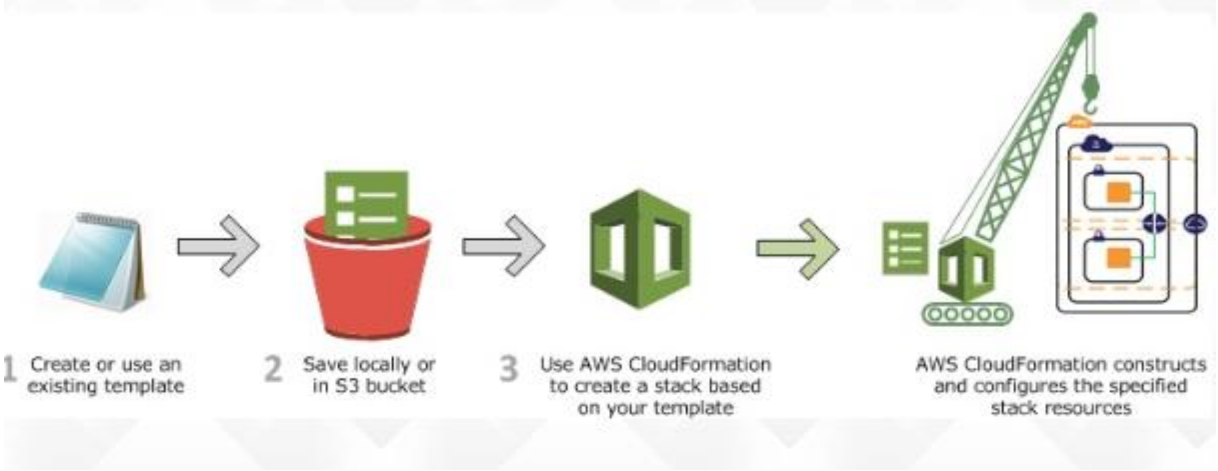
Time Range: Last Hour

AWS CloudFormation :

It gives developers and systems administrators an easy way to create and manage a collection of related AWS resources, provisioning and updating them in an orderly and predictable fashion.

You can use AWS CloudFormation's sample templates or create your own templates to describe the AWS resources, and any associated dependencies or runtime parameters, required to run your application.

Version control: We can modify and update resources in a controlled and predictable way, in effect applying version control to your AWS infrastructure.



1 Create or use an existing template
2 Save locally or in S3 bucket
3 Use AWS CloudFormation to create a stack based on your template
AWS CloudFormation constructs and configures the specified stack resources

AWS CloudTrail:

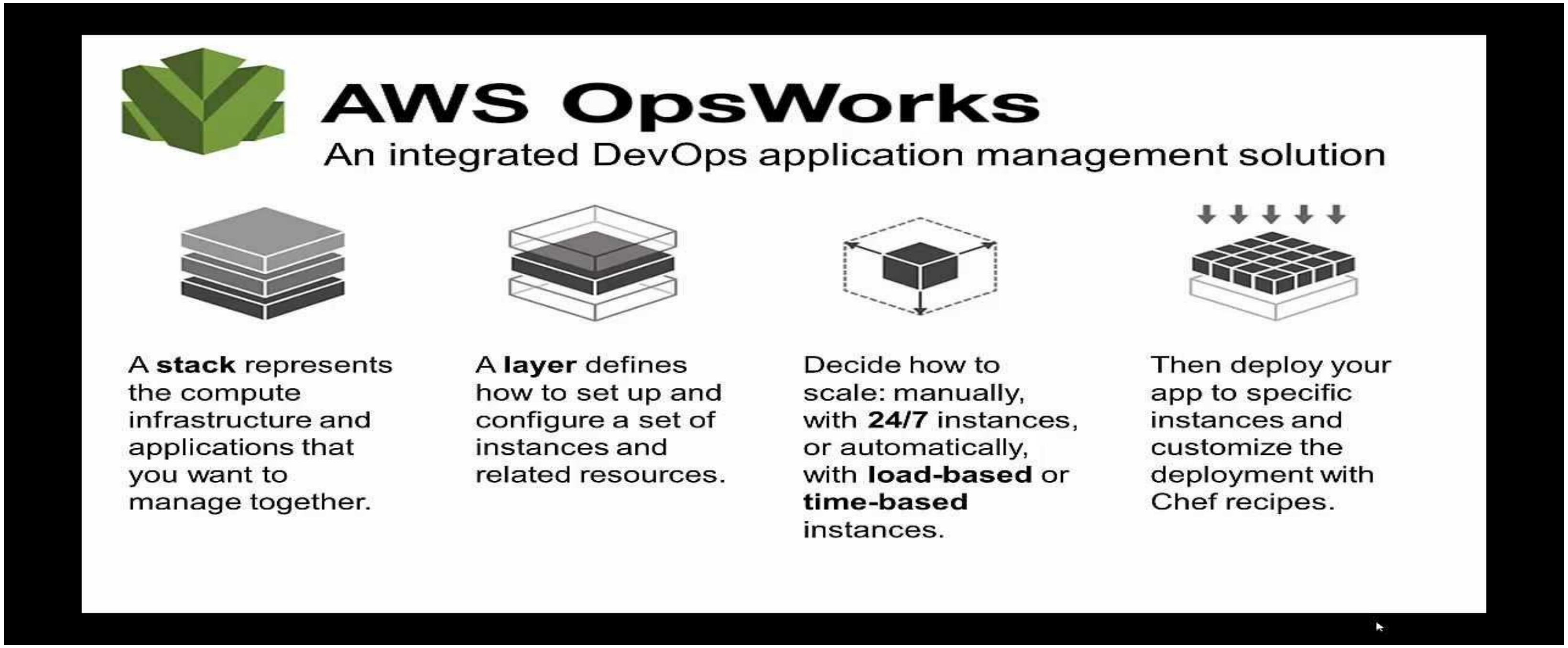
It is a service that enables governance, compliance, operational auditing, and risk auditing of your AWS account. With CloudTrail, you can log, continuously monitor, and retain events related to API calls across your AWS infrastructure.

CloudTrail provides a history of AWS API calls for your account, including API calls made through the AWS Management Console, AWS SDKs, command line tools, and other AWS services. This history simplifies security analysis, resource change tracking, and troubleshooting.

AWS OpsWorks:

It is a configuration management service that uses Chef, an automation platform that treats server configurations as code.

OpsWorks uses Chef to automate how servers are configured, deployed, and managed across your Amazon Elastic Compute Cloud (Amazon EC2) instances or on-premises compute environments



AWS OpsWorks

An integrated DevOps application management solution

A **stack** represents the compute infrastructure and applications that you want to manage together.

A **layer** defines how to set up and configure a set of instances and related resources.

Decide how to scale: manually, with **24/7** instances, or automatically, with **load-based** or **time-based** instances.

Then deploy your app to specific instances and customize the deployment with Chef recipes.

Trusted Advisor:

An online resource to help you reduce cost, increase performance, and improve security by optimizing your AWS environment, Provides real time guidance to help you provision your resources following AWS best practices.

Security & Identity

AWS Identity and Access Management (IAM)

It 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*).

AWS Key Management Service (KMS) :

It is a managed service that makes it easy for you to create and control the encryption keys used to encrypt your data, and uses Hardware Security Modules (HSMs) to protect the security of your keys.

AWS Key Management Service is integrated with several other AWS services to help you protect the data you store with these services.

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.

Application Services:

Amazon Simple Email Service (Amazon SES)

It is a cost-effective email service built on the reliable and scalable infrastructure that Amazon.com developed to serve its own customer base.

With Amazon SES, you can send and receive email with no required minimum commitments – you pay as you go, and you only pay for what you use.

Amazon Simple Queue Service (SQS)

It is a fully managed message queuing service that makes it easy to decouple and scale microservices, distributed systems, and serverless applications.

Building applications from individual components that each perform a discrete function improves scalability and reliability, and is best practice design for modern applications.

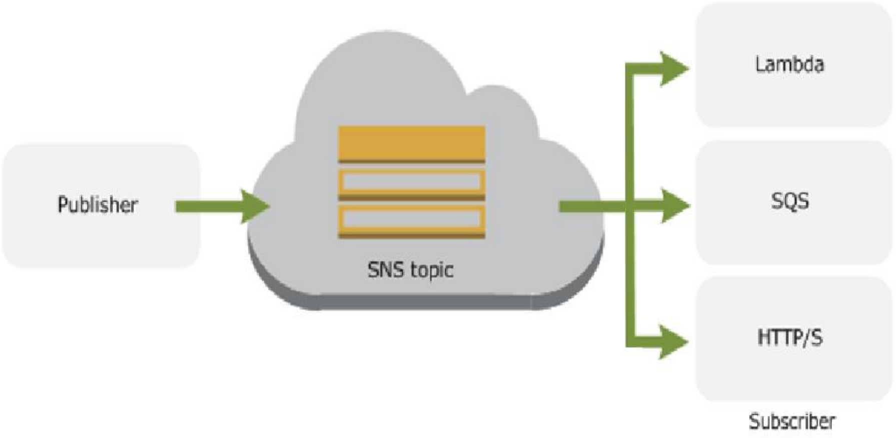
SQS makes it simple and cost-effective to decouple and coordinate the components of a cloud application.

Amazon Simple Notification Service (SNS)

It is a flexible, fully managed pub/sub messaging and mobile notifications service for coordinating the delivery of messages to subscribing endpoints and clients.

With SNS you can fan-out messages to a large number of subscribers, including distributed systems and services, and mobile devices.

It is easy to set up, operate, and reliably send notifications to all your endpoints – at any scale.

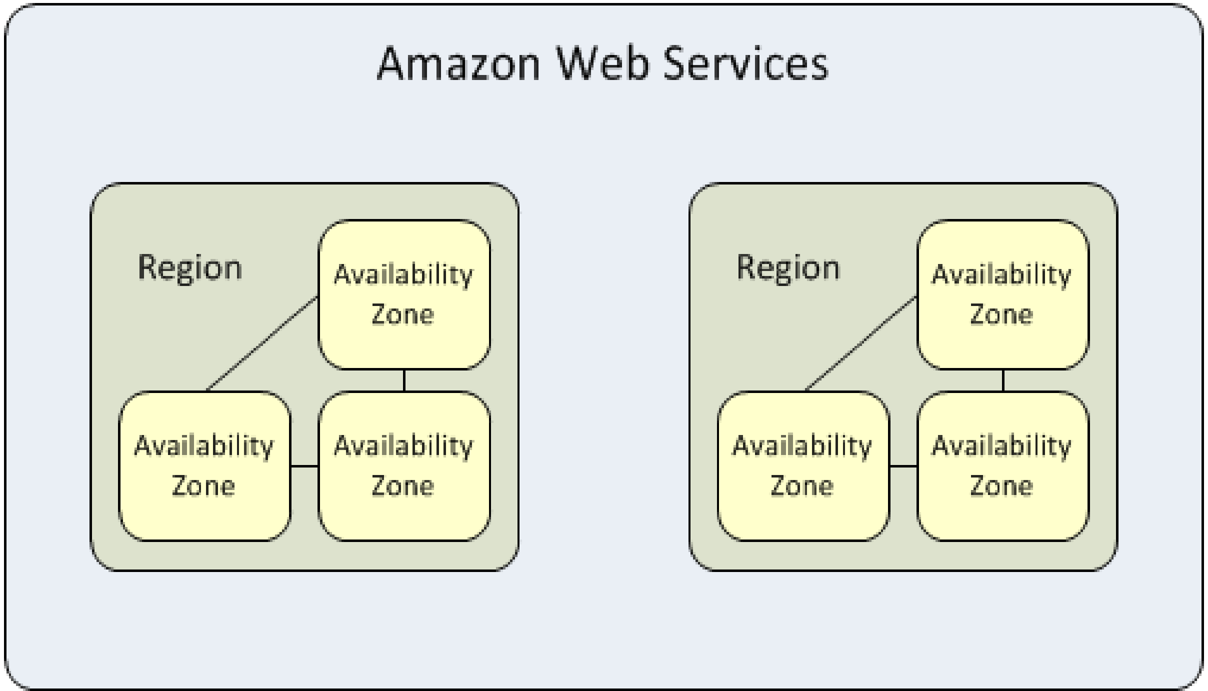


AWS infrastructure

Regions & Availability zones:

Each *region* is a separate geographic area. Each Amazon region is designed to be completely isolated from the other Amazon EC2 regions. This achieves the greatest possible fault tolerance and stability.

Each region has multiple, isolated locations known as *Availability Zones* but the Availability Zones in a region are connected through low-latency links. Amazon operates state-of-the-art, highly-available data centers. Although rare, failures can occur that affect the availability of instances that are in the same location. If you host all your instances in a single location that is affected by such a failure, none of your instances would be available.



- Amazon EC2 provides you the ability to place resources, such as instances, and data in multiple locations.
- Resources aren't replicated across regions unless you do so specifically.

Edge location:

It is where end users access services located at **AWS**. They are located in most of the major cities around the world and are specifically used by CloudFront (CDN) to distribute content to end user to reduce latency. It is like frontend for the service we access which are located in **AWS** .

Amazon CloudFront:

It is a web service that speeds up distribution of your to your users. CloudFront delivers your content through a worldwide network of data centers called edge locations.

The request from user is routed to the edge location that provides the lowest latency (time delay), so that content is delivered with the best possible performance.
If the content is already in the edge location with the lowest latency, CloudFront delivers it immediately.
If the content is not in that edge location, CloudFront retrieves it from an Amazon S3 bucket or an HTTP server (for example, a web server) that you have identified as the source for the definitive version of your content.

EC2 Purchasing Options:

On-Demand

With On-Demand instances, you pay for compute capacity by the hour with no long-term commitments or upfront payments.

You can increase or decrease your compute capacity depending on the demands of your application and only pay the specified hourly rate for the instances you use.

On-Demand instances are recommended for:

- Users that prefer the low cost and flexibility of Amazon 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 Amazon EC2 for the first time

Spot Instances

Amazon EC2 Spot instances allow you to bid on spare Amazon EC2 computing capacity for up to 90% off the On-Demand price.

Spot instances are recommended for:

- Applications that have flexible start and end times
- Applications that are only feasible at very low compute prices
- Users with urgent computing needs for large amounts of additional capacity

Reserved Instances

Reserved Instances provide you with a significant discount (up to 75%) compared to On-Demand instance pricing.
In addition, when Reserved Instances are assigned to a specific Availability Zone, they provide a capacity reservation, giving you additional confidence in your ability to launch instances when you need them.

For applications that have steady state or predictable usage, Reserved Instances can provide significant savings compared to using On-Demand instances.

Reserved Instances are recommended for:

➢Applications that may require reserved capacity
➢Customers that can commit to using EC2 over a 1 or 3 year term to reduce their total computing costs

Placement group:

It 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