# AWS

## AWS Intro:

* Aws was launched in in the leading cloud service provider in the world.
* Cloud Service Provider can be initialized as a CSPs.
* **Simple Queue Service (SQS)** was the first AWS service launched for public in 2004.
* **Simple Storage Service (S3)** was launched in March of 2006.
* **Elastic Compute Cloud (EC2)** was launched in August of 2006.
* **In November 2010,** It was reported that all of **Amazon.com’s retail sites had migrated to AWS.**
* To support industry-wide training and skills standardization, AWS began offering a certification program for computer engineers April 2013.

## What is a Cloud Service Provider (CSP)

A Cloud Service Provider(CSP) is a company which

* provides multiple Cloud Services e.g. tens to hundreds of services.
* those Cloud Services can be chained together to create cloud architectures.
* those Cloud Services are accessible via Single Unified API eg. AWS API.
* Those Cloud Services utilized metered billing based on usage e.g. per second, per hour.
* Those cloud services have rich monitoring built in eg. AWS CloudTrail.
* Those cloud services have an infrastructure as a Service (IaaS) Offering.
* Those cloud services offers automation via Infrastructure as Code (IaC).

A diagram of a software flow

Description automatically generated with medium confidence

## Common Cloud Services

A cloud service provider can have hundreds of cloud services that are grouped into various types of services. The four most common types of cloud services (4 core) for infrastructure as a Service (IaaS) would be:

* **Compute** : Imagine having a virtual computer that can run application, programs and code.
* **Networking**: Imagine having virtual network defining internet connections or networks isolations between services or outbound to the internet.
* **Storage**: imagine having a virtual hard-drive that can store files.
* **Database**: Imagine a virtual database for storing, reporting data or a database for general purpose web-application

Aws has over 200+ cloud services. The term ‘Cloud Computing’ can be used to refer to all categories even though compute in the name.

## Technology Overview

A screenshot of a computer

Description automatically generated

## Types of Cloud Computing

SaaS (Software as a Service) – **For Customer**: A product that is run and managed by the service provider. Don’t worry about how the service is maintained. It just works and remain available.

Pass (Platform as a Service) – **For Developers**: Focus on the deployment and management of your apps. Don’t worry about, provisioning, configuring or understanding the hardware or OS.

IaaS (Infrastructure as a Service) – **For Admine** : The basic building blocks for cloud IT. Provide access to networking feature, computers and data storage space. Don’t worry about IT staff, data centers and hardware.

## Cloud Computing Deployment Models

**Public Cloud** : Everything (The workload or project) is built on the CSP Also know as : \*Cloud-Native or Cloud First

A screenshot of a computer

Description automatically generated

Companies that are starting out today or are small enough to make the leap from VPS to a CSP.

* Startups
* SaaS offerings
* new projects and companies
* DropBox

**Private Cloud** : Everything built on company’s datacenters Also known as **On-Premise**. The cloud could be **OpenStack.**

A computer screen shot of a computer

Description automatically generated

Organization that started with their own datacenter, cant fully move to cloud due to effort of migration or security.

For Ex:

* Banks
* FinTech
* Investment Management
* Large Professional Service providers
* Legacy on-premise

Like CIBC, Deloitte and CPP INVESTMENT BOARD

**Hybrid** : Using both **On-Premise** and **Cloud Service Provider**

A screenshot of a computer

Description automatically generated

Organizations that cannot run on cloud due to strict regulatory compliance or the sheer size of their organization

For EX:

* Public sectors like government
* Super Sensitive Dara like Hospital
* Large Enterprises with heavy regulation like Insurance Companies

Like Canada Government, AIG

**Cross-Cloud** : Using Multiple Cloud Providers Aka **multi-cloud**

A diagram of a diagram

Description automatically generated with medium confidence

## The Benefits of Clouds

The benefits of the cloud is a summary of reasons why an organization would consider adopting or migrating to utilizing public cloud.

* **Agility** : Increase speed and agility
* **Pay-as-you go pricing** : Trade capital expense for variable expense
* **Economy of scale**: Benefit from massive economies of scale
* **Global Reach** : Go global in minutes
* **Security**
* **Reliability** : Stop spending money on running and maintaining data centers
* **High** **Availability**
* **Scalability** : Benefit from massive economies of scale
* **Elasticity**

Two Extra : Fault Tolerance and Disaster Recovery

The benefits of cloud are reworking and expansion of the **Six Advantage** of cloud

## Six Advantage to Cloud

1. **Trade capital expense for variable expense** : You can Pay On-Demand meaning there is no upfront-cost and you pay for only what you consume or pay by the hour, minutes or seconds. Instead of paying for upfronts costs of data center and server.
2. **Benefits from massive economies of** scale : You are sharing the cost with other customers to get unbeatable savings.

Hundres of thousands of customers utilizing a fraction of a server.

1. **Stop guessing capacity** : scale up or down to meet the current need. Launch and destroy services whenever instead of waiting days or weeks of your IT to implement the solution on-premise.
2. **Increase speed and agility** : Launch resources within a few clicks in minutes Instead of waiting days or weeks of your IT to implement the solution on-premise
3. **Stop spending money on running and maintaining data center** : Focus on your own customers, developing and configuring your applications Instead of operations such as racking, stacking, and powering servers
4. **Go global in minutes** : Deploy your app in multiple regions around the world with few clicks. Provides lower latency and a better experience for your customers at minimal cost.

## AWS Global Infrastructure

What is the Global Infrastructure ?

The AWS Global Infrastructure is globally distributed hardware and datacenters that are physically networked together to act as one large resource for the end customer

The AWS Global Infracture is made up of following resources:

* 32 Launched Regions
* 102 Availability Zones
* 115 Direct Connections Locations
* 550+ Points of Presence
* 35 Local Zone
* 29 Wavelength Zones

Note : The above number can be different from the new one so double check

## The difference between Reginal and Global Services :

**Regional services** : Aws scopes their AWS Management Console on a selected Region. This will determine where an AWS service will be launched and what will be seen within an AWS Service’s console.

You generally don’t explicitly set the region for a service at the time of creation.

**Global Services**: Some AWS Services operate across multiple regions and the region will be fixed to “Global” EX: Amazone S3, CloudFront, Route53, IAM

## Global Infrastructure – Availability Zone

A Availability Zone (AZ) is physical location made up oof one or more datacenters.

A datacenter is a secure building that contains hundreds of thousands of computers.

A region will \*generally contain 3 Availability zones.

Datacenter within a region will be isolate from each other (different building). But they will be close enough to provide low-latency (<10ms).

Its common practise to run workloads in at least 3 AZs to ensure services remain available in case one or two datacenters fail. (high availability).

AZs are represented by a Region Code, followed by a letter identifier eg. Us-east-1a (a represent the availability zone)

## Subnet

A subnet is associated with an Availability Zone.

You never choose the AZ when launching resources. You choose the Subnet which is associated to the AZ.

Example of an architectural diagram, representing two AZs, and EC2 instances (virtual Machines) launched in those subnets.

A screenshot of a computer

Description automatically generated

The US-EAST-1 region has 6 AZs (the most Availability zones of any regions)

A regions has multiple Availability zones.

An availability Zone is made up of one or more datacenters.

All AZs in an AWS Region are interconnected with high-bandwidth, low-latency networking, over fully redundant, dedicated metro fiber providing high-throughput, low-latency networking between

All traffic between AZs is encrypted

AZs are within 100 kms (60 miles) of each other

## Fault Tolerance

**What is a fault domain ?**

A fault domain is a section of a network that is vulnerable to a damage if a critical device or system fails. The purpose of a fault domain is that if a failure occurs it will not cascade outside that domain, limiting the damage possible.

You can have fault domains nested fault domains.

**What is a fault level ?**

A fault level is a collection of fault domains.

**The scope of domain could be :**

* Specific servers in a rack
* An entire rack in a datacenter
* An entire room in a datacenter
* The entire data center building

Its up to the Cloud Service Provider (CSPs) to define the boundaries of a domain.

An AWS Region would be a Fault Level EX: Fault level => us-east-1 (Region )

A Availability Zone would be a Fault Domain EX: Fault Domain => us-east-1a us-east-1b (AZ)

Each Amazone Region is designed to be completely isolated form the other Amazone Regions.

* This achieves the greatest possible fault tolerance and stability

Each Availability Zone is isolated, but the Availability Zone in a Region are connected to through low-latency links

Each Availability Zone is designed as an independent failure zone

* A “Failure Zone” is AWS describing a Fault Domain.

**Failure Zone**

* Availability Zones are physically separated within a typical metropolitan region and are located in lower risk flood plains
* Discrete Uninterruptible Power Supply (UPS) and onsite backup generation facilities independent substations to reduce the risk of an event on the power grid impacting more than one Availability Zone.
* Availability Zones are all redundantly connected to multiple tier-1 transit providers
* **Multi-AZ for High Availability**

If an application is partitioned across AZs, companies are better isolated and protected from such as power outages, lightning strikes, tornadoes, earthquakes and more.

## AWS Global Network

The AWS Global Network represent the interconnections between AWS Global Infrastructure. Commonly referred to as the “The Backbone of AWS”.

Think of it as private expressway, where things can move very fast between datacenters.

A map of the world

Description automatically generated

## Point of Presence (PoP)

Point of Presence (PoP) is an intermediate location between an AWS Region and the end user, and this location could be a datacenter or collection of hardware.

For AWS a Point of Presence is a datacenter owned by AWS or a trusted partner that is utilized by AWS Services related for content delivery or expediated upload

PoP resources are :

* Edge Locations
* Regional Edge catches

Edge Locations are datacenters that hold cached (Copy) on the most popular files (e.g. web pages, images and videos) so that the delivery of distance to the end users are reduce

Regional Edge Locations are datacenters that hold much larger caches of less-popular files to reduce a full round trip and also to reduce the cost of transfer fees.

A diagram of network connections

Description automatically generated

The following AWS Services use PoP for content delivery or expediated upload.

**Amazon** **CloudFront is Content Delivery Network (CDN)** service that:

* You point your website to CloudFront so that it will route requests to nearest Edge Location cache
* Allows you to choose an origin (such as a web-server or storage) that will be source of cached
* Caches the contents of what origin would returned to various Edge Location around the world

**Amazone S3 Transfer Acceleration** allows you to generate a special URL that can be used by end users to upload files to a nearby Edge Location. Once a file is uploaded toa n Edge Location, it can move much faster within the AWS Network to reach S3.

**Aws Global Accelerator** can find the optimal path from the end user to your web-servers. Global Accelerator are deployed within Edge Location so you send user traffic to an Edge Location instead of directly to your web-application.

## AWS Direct Connect

AWS Direct Connect is a **private/dedicated connection between your datacenter, office, co-location and AWS.**

**Direct Connect** has two very-fast network connections options:

1. Lower Bandwidth : 50MBps – 500MBps
2. Higher Bandwidth : 1GBps or 10GBps

Helps **reduce network costs** and **increase bandwidth throughput** (great for high traffic networks)

Provides a **more consistent** **network** **experience** than a typical internet-based connection. (reliable and secure)

A co-location (aka carrier-hotel) is a data center where equipment, space and bandwidth are available for rental to retail customers.

Direct Connect Location are trusted partners datacenters that you can establish a dedicated high speed, low-latency connection from your on-premise to AWS.

## Local Zones

**Local Zones** : Local Zones are datacenters located vary close to a densely populated area to provide single-digit millisecond low latency performance (eg 7ms) for that area.

* **Los Angeles, California** was the first Local Zone to be deployed

1. It is a logical extension of the US-West Region
2. The identifier looks like the following : us-west-2-lax-1a

* Only specific AWS Services have been made available

1. EC2 instance Types (T3, C5, R5d, l3en, G4)
2. EBS (io1 and gp2)
3. Amazone FSx
4. Application Load Balancer
5. Amazone VPC

The purpose of Local Zone is the support highly-demanding applications sensitive to latencies:

* Media & Entertainment
* Electronic Design Automation
* Ad-Tech
* Machine Learning

## Wavelength Zones

AWS Wavelength Zones allows for edge-computing on 5G Networks.

Applications will have ultra-low latency being as close as possible to the users.

AWS partnered with various Telecom companies to utilize their 5G networks ex: Verizon, Vodafone

You can create a Subnet tied to a Wavelength Zone and then you can launch VMs to the edge of the targeted 5G Networks.

## Data Residency

**What is Data Residency ?**

The physical or geographic location of where an organization or cloud resource reside.

**What is Compliance Boundaries ?**

A regulatory compliance (legal requirement) by a government or organisation that describes where data and cloud resources are allowed to reside.

**What is Data Sovereignty ?**

Data Sovereignty is the jurisdictional control or legal authority that can be asserted over data because location is within jurisdictional boundaries.

For workloads that need to meet compliance boundaries strictly defining the data residency of data and cloud resources in AWS you can use:

* **AWS Outposts** is **physical rack of servers** that you can puy in your data center. Your data will reside whenever the outpost Physically resides
* **AWS Config** is a Policy as Code service. You can create rules to continuous check AWS resources configuration. If they deviate from your expectations you are alerted or AWS Config can in some cases auto-remediate.
* **IAM Policies** can be written explicitly deny access to specific AWS Regions. A **Service Control Policy** (SCP) are permissions applied organization wide

## AWS for government

**What is Public Sector** ?

Public sectors include public goods and governmental services such as: military, law enforcement, infrastructure, public transit, public education , health care, the government itself

Aws can be utilized by public sector or organizations developing cloud workloads for the public sector

AWS achieve this by meeting regulatory compliance programs along with specific governance and security controls

Ex: HIPAA , FedRAMP, FIPS

AWS has special regions for US regulation called GovCloud

## GovCloud

Federal Risk and Authorization Management Program (FedRAMP) a US government-wide program that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services.

What is GovCloud ?

A Cloud Service Provider (CSP) generally will offer an isolated region to run FedRAMP workloads.

AWS GovCloud Regions allow customers to host sensitive Controlled Unclassified Information and other types of regulated workloads.

* GovCloud Regions are only operated by employees who are U.S. citizens on U.S. soil.
* They are only accessible to U.S. entities and root account holders who pass a screening process Customers can architect secure cloud solutions that comply with:
* FedRAMP high baseline
* DOJ’s Criminal Justice information System (CJIS) Security Policy
* U.S. International Traffic in Arms Regulations (ITAR)
* Exports Administration Regulations (EAR)
* Department of Defense(DOD) Cloud Computing Security Requirements Guide

## AWS in China

AWS China is completely isolate intentionally from AWS Global to meet regulatory compliance for Mainland China

AWS China is on its own domain at: amazonaws.cn

* In order to operate in AWS China Region you need to have a Chinese Business License (ICP license)
* Not all service are available in China eg. Route53
* Running in Mainland China (instead of Singapore) means you would not need to traverse the The Great Firewall

## Sustainability

Amazone co-founded the Climate Pledge to achieve Net-Zero Carbon Emissions by 2040 across all of Amazon’s business (this includes AWS)

AWS Cloud’s Sustainability goals are composed of three parts:

1. Renewable Energy : AWS is working towards having their AWS Global infrastructure powered by 100% renewable energy by 2025.

Aws purchases and retires environmental attributes to cover the non-renewable energy for AWS Global infrastructure:

Renewable Energy Credits (RECs), Guarantees of Origin (GOs)

1. Cloud Efficiency : AWS’s infrastructure is 3.6 times more energy efficient than the median of U.S. enterprises data centers surveyed.
2. Water Stewardship : Direct evaporative technology to cool our data center Use of non-potable water for cooling purposes (recycle water)

On-site water treatment allows us to remove scale-forming minerals and reuse water for more cycles

Water efficiency metrics to determine and monitor optimal water use for each AWS region

## AWS Ground Station

AWS Ground Station is a fully managed service that lets you control satellite communications, process data, and scale your operations without having to worry about building or managing your own ground station infrastructure.

|  |  |
| --- | --- |
| Use case for Ground station : | To use Ground Station: |
| Weather forecasting | You schedule a Contact (select satellite, start and end time, and the ground location) |
| Surface imaging | Use the AWS Ground Station EC2 AMI to launch EC2 instances that will uplink and downlink data during the contact or receive downlinked data in an Amazone S3 bucket. |
| Communication |  |
| Video broadcasts |  |

Use Case:

A company reaches an agreement with a Satellite Imagery Provider to take satellite photos of a specific region. They use AWS Ground Station to communicate that company’s Satellite and download S3 image data.

## AWS Outposts

AWS Outposts is a fully managed service that offers the same AWS infrastructure, AWS services, APIs, and tools to virtually any datacenter, co-location, or on-premises facility for a truly consistent hybrid experience.

AWS Outposts is rack of servers running AWS infrastructure on your physical location

**What is a Server Rack ?**

A frame design to hold and organize IT equipment. U stand for “rack units” or “U spaces” with is equal to 1.75 inches. The industry standard rack size is 48U (7 foot Rack)

Full-size rack cage is 42U high and the equipment size is typically 1U, 2U, 3U or 4U high

1. The full rank of servers provided by AWS is 42U. AWS delivers it to your preferred physical site fully assembled and ready to be rolled into final position. It is installed by AWS and the rack need to be simple plugged into power and network.
2. The 1U is suitable for 19-iches wide 24 -inches deep cabinets. AWS Graviton2 (up to 64 vCPUs) 128 GiB memory 4 TB of local NVME storage.
3. The 2U is suitable for 19-inches wide and 36-inches deep cabinets, intel processor (up to 128 vCPUs) 256 GiB memory 8TB of local NVMe storage

## AWS Cloud Architecture Terminologies

**What is a solution Architect ?**

A role in a technical organization that architects a technical solution using multiple systems via researching, documentation, experimentation.

**What is a cloud Architect ?**

A solution architect that is focused solely on architecting technical solution using cloud services.

A cloud architect need to understand the following terms and factor them into their designed architecture based on the business requirements.

* Availability – Your ability to ensure a service remains available eg. Highly Available (HA)
* Scalability – Your ability to grow rapidly or unimpeded
* Elasticity - Your ability to shrink and grow to meet the demand
* Fault Tolerance – Your ability to prevent a failure
* Disaster Recovery – Your ability to recover from a failure eg. Highly Durable (DR)

A solution Architect needs to always consider the following business factors:

* (Security) How secure is this solution ?
* (Cost) How much it going to be cost ?

### High Availability

High Availability : Your ability for your service to remain available by ensuring there is no single point of failure and/or ensure a certain level of performance

A diagram of a computer

Description automatically generated

Running your workload across multiple Availability Zones ensure that if 1 or 2 AZs becomes unavailable your service/ application remain available.

A purple square with white circles

Description automatically generated Elastic Load Balancer : A load balancer allows you to evenly distribute traffic to multiple servers in one or more datacenters. If a datacenters or server becomes unavailable the load balancer will route the traffic to only available datacentres with servers.

### High Scalability

Your ability to increase your capacity based on the increasing demand of traffic, memory and computing power.

A computer screen shot of a computer

Description automatically generated

Vertical Scaling Horizonal Scaling

Scaling Up Scaling Out

Upgrade to a bigger server Add more server of the same size

### High Elasticity

Your ability to automatically increase or decrease your capacity based on the current demand of traffic, memory and computing power

A close-up of a sign

Description automatically generated

Horizonal Scaling

Scaling Out – Add more servers of the same size

Scaling In – Removing underutilized servers of the same size

Vertical Scaling is generally hard for traditional architecture so you will usually only see horizontal scaling described with Elasticity.

### Highly Fault Tolerant

Your ability for your service to ensure there is no single point of failure. Preventing the chance of failure.

Fails-overs is when you have a plan to shift traffic to a redundant systems in case the primary fails.

A common example is having a copy (secondary) of your database where all ongoing changes are synced. The secondary systems is not in-use a fail over occurs and it becomes the primary database.

A diagram of a server

Description automatically generated

**A blue square with arrows and a logo

Description automatically generated RDS Multi-AZ** is when you run a duplicate standby database in another Availability zone in case your primary database fails.

### High Durability

Your ability to recover from a disaster and to prevent the loss of data solution that recover from a disaster is known as **Disaster Recovery (DR).**

* Do you have a backup ?
* How fast can you restore that backup ?
* Does your backup still work ?
* How do you ensure that current live data is not enough ?

A green square with a white line and a lightning bolt in the middle

Description automatically generated ColudEndure Disaster Recovery continuously replicates your machines into a low-cost staging area in your target AWS account and preferred Region enabling fast and reliable recovery in case of IT data center failures.

## Business Continuity Plan (BCP)

A Business Continuity Plan (BCP) is a document that outlines how a business will continue operating during an unplanned disruption in services.

Recover Point Objective (RPO) : The maximum acceptable amount of data loss after an unplanned data-loss incident, expressed as an amount of time

Recovery Time Objective (RTO) : The maximum amount of downtime your business can tolerate without incurring a significant financial loss.

A diagram of a fire

Description automatically generated

## Disaster Recovery Options

There are multiple options for recovery that trade cost vs time to recover.

A screen shot of a screen

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| You backup your data and restore it to new infrastructure | Data is replicate to another region with the minimal services running | Scaled down copy of your infrastructure running ready to scale up | Scaled up copy of your infrastructure in another region |
| Low priority use cases | Less stringent RTO & RPO | Business Critical Services | Zero downtime |
| Restore data after event | Core Services | Scale resources after event | Nero Zero loss |
| Deploy resources after event | Start & and scale resources after event |  | Mission Critical Services |
| Cost $ | Cost $$ | Cost $$$ | Cost $$$$$ |

## RTO

Recovery Time Objective (RTO) is the maximum acceptable delay between the interruption of service and restoration of service. This objective determines what is considered an acceptable time window when service is unavailable and is define by the organization.

A diagram of cost and cost of service

Description automatically generated

## RPO

Recovery Point Objective (RPO) is the maximum acceptable amount of time since the last data recovery point. This objective determines what is considered an acceptable loss of data between the last recovery point and the interruption of service and is defined by the organization.

A diagram of cost and cost reduction

Description automatically generated

## AWS Application Programming Interface (API)

What is an Application Programming Interface (API)

An API is software that allows two applications/services to talk to each other. The most common type of API is via HTTP/S requests.

AWS API is an HTTP API and you can interact by sending HTTPS requests, using an application interaction with APIs like Postman.

A screenshot of a computer

Description automatically generated

Each AWS Services has its own Service Endpoint which you send requests.

To authorities you will need generate a signed request.

You make a separate request with your AWS credentials and get back a token. Your need to provide an ACTION and accompanying parameters as the payload.

Rarely do users directly send HTTP requests directly to the AWS API. Its much easier to interact with the API via a variety of Developer Tools.A diagram of a software development process

Description automatically generated

## AWS PowerShell

A screenshot of a computer program

Description automatically generated

## Amazon Resource Name (ARNs)

Amazone Resources Names (ARNs) uniquely identify AWS resources. ARNs are required to specify a resources unambiguously across all of AES.

The ARN has following format variations

Arn:partition:service:region:account-id:resource-id

Arn:partition:service:region:account-id:resource-type/resource-id

Arn:partition:service:region:account-id:resource-type:resource-id

**Partition**

* aws - AWS Regions
* aws-cn – China Regions
* aws-us-gov-AWS GovCloud (US) Regions

**Service – Identifies the service**

* ec2
* s3
* iam

**Region – which AWS resource**

* us-east-1
* ca-central-1

**Account ID**

121212112112

123456789012

**Resource ID : could be number, name or path**

* user/Bob
* instance/i-1234567890abcdef0

In the AWS Management Console its common to be able to copy the AEN to your clipboard

## Paths in ARNs

Resources ARNs can include a path

Path can include a wildcard character, namely an asterisk (\*)

IAM Policy ARN Path: arn:aws:iam::123456789012:user/Development/product\_1234/\*

S3 ARN Path : arn:aws:s3:::my\_corporate\_bucket/Development/\*

## Aws Command Line Interface (CLI)

**What is CLI ?**

A Command Line Interface (CLI) processes command to a computer program in the form of lines of text. Operating systems implement a command-line interface in a shell.

**What is Terminal ?**

A terminal is a text only interface (input/output environment)

**What is console ?**

A console is a physical computer to physically input information into a terminal

**What is a shell ?**

A shell is the command line program that users interact with to input commands. Ex : Bash, PowerShell