AWS Global Infrastructure

1. Overview
2. Regions
3. Regions vs Global Services
4. AZ
5. Regions vs AZs
6. Fault tolerance
7. AWS Global Network
8. PoP
9. AWS Direct connect

(1) Overview

What is AWS Global Infrastructures?

→ It is globally distributed hardware & datacenters that are physically networked together to act as one large resource for the end customer

It has

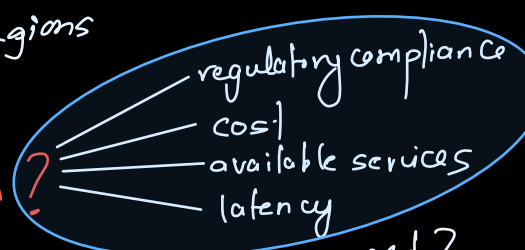
1. 32 Launched Regions
2. 102 AZs
3. 115 Direct connection locations
4. 550+ Point of Presence
5. 35 Local zones
6. 29 Wavelength zones

② AWS Regions

AWS Regions are geographically distinct locations consisting of one or more availability zones.

Every location is physically isolated from and independent of every other region in terms of location, power & water supply.

- One Region = 3 availability Zones
- Not all AWS services are available in all regions
- All billing info appears in US-East-1 (N. Virginia)
- All new services are first available in US-East-1
- Cost of regions vary upon regions



How to choose a Region?

(A) What regulatory compliance do this region meet?
(B) What is the cost of AWS services in that region?
(C) What services are available?
(D) What is the distance / latency to my end users?

Regional vs Global Services

Regional Services

AWS scopes their AWS management console on selected region. They will determine where an AWS service will be launched and what will be seen within the service console.

Global Services

Some of AWS services are not regional they operate across multiple regions and the region will be fixed to Global. Ex: Amazon S3, Cloud Front, Route 53, IAM etc. But why?

- Global Resource management
- Consistency & Efficiency
- Global communication & Availability
- Unified control and Monitoring
- Data Replication & Availability

③ Availability Zones

An AZ is physical location made up of one or more datacenters.

A datacenter is a secured building that contains hundreds of thousands of computers

A region will *generally* contain 3 availibility zones.

Datacenters within a region are isolated enough for disaster recovery but close enough to provide low latency

It is important to run work loads in at least 3 AZs to ensure services remain available in case of one or two data centers fail.

AZs are represented by region code followed by a letter 'a' or 'b' or 'c'.

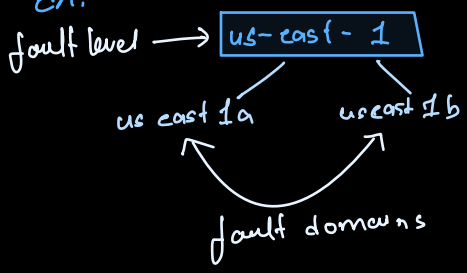
Ex: us-east1-a.

Fault tolerance

- Fault domain
  section of network vulnerable to damage if a critical device or system fails.

- Fault Level
  collection of fault domains

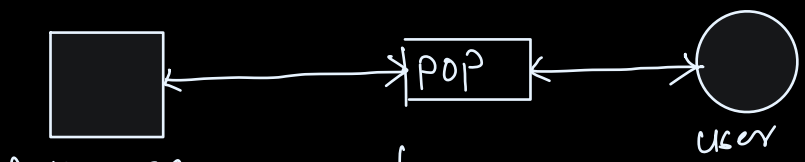
Ex:



- Each amazon region is designed to be isolated from others
  - ↳ greatest fault tolerance
  - → each AZ is isolated within Regions
  - → each AZ is independent failure zone

AWS Point of Presence (PoP)

It is an intermediate location between AWS region and end user.



↳ for content delivery & expediated upload

Ex:
- Edge locations
- Regional Edge caches

* Some AWS services that use PoPs are:

- **Amazon Cloud Front**
  ↳ Content delivery Network (CDN) that
  ① allows you to point your website to CF so that it will route requests to nearest edge location cache
  ② allows you to choose an origin (such as web-server) that will be source of cached

- **Amazon S3 Transfer Acceleration**

  - is a simple feature designed to speed up the upload & download of data to and from Amazon S3 by using AWS CF's globally distributed edge locations

- **Amazon Global Accelerator**

  - can find optimal path between end user to your web-servers. Global Accelerators are deployed within edge locations so you send user traffic to an edge location instead of your web-app directly.

AWS direct Connect

↳ a private/dedicated connection between your datacenter, office, co-location & AWS

Direct Connect has two very fast network options

① Lower Bandwidth (50 - 500 Mbps)

② Higher Bandwidth (1 GBPs or 10 GBPs)

AWS local Zones

These are data centers located in densely populated area to provide single digit low latency performance (ex: 7ms) for that area

→ LA was first place to have LZ
(us-west-2-lax-1a)

→ It has only some of AWS services

- EC2
- EBS
- Amazon FSx
- Application load balancer
- VPC

AWS Wavelength Zones

↳ allows for edge computing on 5G nets.

↳ Applications will have ultra low latency being as close as to the users

AWS Data Residency

What is Data Residency?

The physical or geographic location where cloud resources reside.

- **AWS Config** is a service that allows you to set policies to continuous check if your resources deviate from your expectations

· AWS Ground Station

It is a fully managed AWS 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.

AWS Outposts

Fully managed service that offers the same AWS infrastructure AWS services, APIs and tools to virtually and datacenters to locate space, or on-premises facility for truly consistent hybrid experience.