Global Infrastructure Study Notes

Andrew Brown

AWS Global Infrastructure

What is the AWS 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.

This is so you, the customer, can access their services no matter where you are located around the globe.

The AWS Global Infrastructure is made up of the following resources:

Currently there are:

- 24 Launched Regions
- 77 Availability Zones
- 97 Direct Connection Locations
- 216 Points of Presence
- 1 Local Zone

References

https://infrastructure.aws/

Regions

Region: a Geographically Distinct Location, which has multiple datacentres(AZs).

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

- Each region has at least two Availability Zones (AZs)
- AWS US-EAST-1 (North Virginia) is the largest region
- New services almost always become available first in US-EAST
- Not all AWS Services are available in all regions
- All your billing information appears in US-EAST-1 (North Virginia)
- AWS is always looking to expand their foot-print so new regions could be added in the upcoming years.

You can head over to **AWS's Global Infrastructure** page to learn more about Regions and see maps of their locations.

- In the AWS Console you can change AWS Regions. Whatever AWS resources you launch will be in that region.
- Some AWS Services such as CloudFront operate in multiple regions and you'll see the region change to Global

Availability Zones

Availability Zones are often referred to as `AZs`. An Availability Zone (AZ) is a physical location made up of one or more `datacenters`. A datacenter is a secured building that contains hundreds or thousands of computers. Datacenters within a region will be isolate from each other (different buildings). But they will be *close enough to provide low-latency (< 10ms)*.

A region will *generally contain 3 Availability Zones*.

The use of AZ's give customers the ability to operate production applications

and databases that are more:

- Highly available
- Fault tolerant
- Scalable

Its common practice to run workloads in at least 3 AZs to ensure services remain available in case one or two datacenters fail **(High Availability)**.

- **A subnet is associated with an Availability Zone**. So to launch an AWS resource into a specific AZ you need to launch your resource in a subnet that resides in that AZ.

Points of Presence

Points of Presence (PoP) are intermediate locations 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 data center owned by AWS or a a trusted partner that is utilized by AWS Services related for content delivery or expediated upload.

Edge Locations - are datacenters that hold cached (copy) on the most popular files (eg. 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.

he following AWS Services use PoPs for content delivery or expediated upload:

AWS CloudFront - is a 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 Locations around the world

AWS 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 to an 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 Accelerators are deployed within Edge Locations so you send user traffic to an Edge Location instead of directly to your web-application.

Direct Connect Locations

Direct Connect Locations - are *trusted third-party datacenters* that you can establish a dedicated high speed, low-latency connection from your on-premise to the AWS network.

- DirectConnect connections are always established through a *third-party provider*. These APN (Amazon Partner Network) Partners can help you establish network circuits between an AWS Direct Connect location and your datacenter, office, or colocation environment, or assist you in constructing a hybrid environment.

Local Zones

Local Zones - are `datacenters` located very close to a density 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

- It is a logical extension of the `US-West Region`
- The Identifier looks like the following: us-west-2-lax-1a
- Only specific AWS Services have been made available
  - EC2 Instance Types (T3, C5, R5, R5d, I3en, G4)
  - EBS (io1 and gp2)
  - Amazon FSx
  - Application Load Balancer
  - Amazon VPC

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

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

GovCloud

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

- GovCloud Regions are only operate by employees who are *U.S. citizens on U.S. soil*.
- Only accessible to U.S. entities and account state holders who pass a screening process.
- Customers can architect secure cloud solutions that comply with:
  - Fed**RAMP** High baseline
  - DOJ's Criminal Justice Information Systems (CJIS) Security Policy
  - U.S. International Traffic in Arms Regulations (TAR)
  - Export Administration Regulations (EAR)
  - Department of Defense (DoD) Cloud Computing Security Requirements

Guide

GovCloud Regions are currently only in the US.

- ◼ You can view their regions on <u>AWS Map of Regions and Edge Networks.</u>
- ◼ You can view more about GovCloud on the official <u>AWS GovCloud page.</u>