AWS Global Infrastructure   Ankit Mohari

This article covers AWS Global Infrastructure training which is a key technology area covered in the Cloud Practitioner exam blueprint. The AWS infrastructure is built around Regions and Availability Zones (AZs).

An AWS Region is a physical location in the world where AWS have multiple AZs.

AZs consist of one or more discrete data centers, each with redundant power, networking, and connectivity, housed in separate facilities.

Each region is completely independent. Each Availability Zone is isolated, but the Availability Zones in a region are connected through low-latency links.

AWS are constantly expanding around the world and currently there are:

- 21 regions.

- 66 availability zones.

The following diagram shows the AWS global infrastructure with regions (orange circles, green are new regions), and availability zones (the number of AZs is specified within each region):

Regions

A region is a geographical area.

Each region consists of 2 or more availability zones.

Each Amazon Region is designed to be completely isolated from the other Amazon Regions.

Each AWS Region has multiple Availability Zones and data centers.

You can replicate data within a region and between regions using private or public Internet connections.

You retain complete control and ownership over the region in which your data is physically located, making it easy to meet regional compliance and data residency requirements.

Note that there is a charge for data transfer between regions.

When you launch an EC2 instance, you must select an AMI that's in the same region. If the AMI is in another region, you can copy the AMI to the region you're using.

Regions and Endpoints:

- When you work with an instance using the command line interface or API actions, you must specify its regional endpoint.

- To reduce data latency in your applications, most Amazon Web Services offer a regional endpoint to make your requests.

- An endpoint is a URL that is the entry point for a web service.

- For example, https://dynamodb.us-west-2.amazonaws.com is an entry point for the Amazon DynamoDB service.

Availability Zones

Availability Zones are physically separate and isolated from each other.

AZs span one or more data centers and have direct, low-latency, high throughput and redundant network connections between each other.

Each AZ is designed as an independent failure zone.

When you launch an instance, you can select an Availability Zone or let AWS choose one for you.

If you distribute your EC2 instances across multiple Availability Zones and one instance fails, you can design your application so that an instance in another Availability Zone can handle requests.

You can also use Elastic IP addresses to mask the failure of an instance in one Availability Zone by rapidly remapping the address to an instance in another Availability Zone.

An Availability Zone is represented by a region code followed by a letter identifier; for example, **us-east-1a.**

To ensure that resources are distributed across the Availability Zones for a region, AWS independently map Availability Zones to names for each AWS account.

For example, the Availability Zone **us-east-1a** for your AWS account might not be the same location as us-east-1a for another AWS account.

To coordinate Availability Zones across accounts, you must use the *AZ ID*, which is a unique and consistent identifier for an Availability Zone.

AZs are physically separated within a typical metropolitan region and are located in lower risk flood plains.

AZs use discrete UPS and onsite backup generation facilities and are fed via different grids from independent facilities.

AZs are all redundantly connected to multiple tier-1 transit providers.
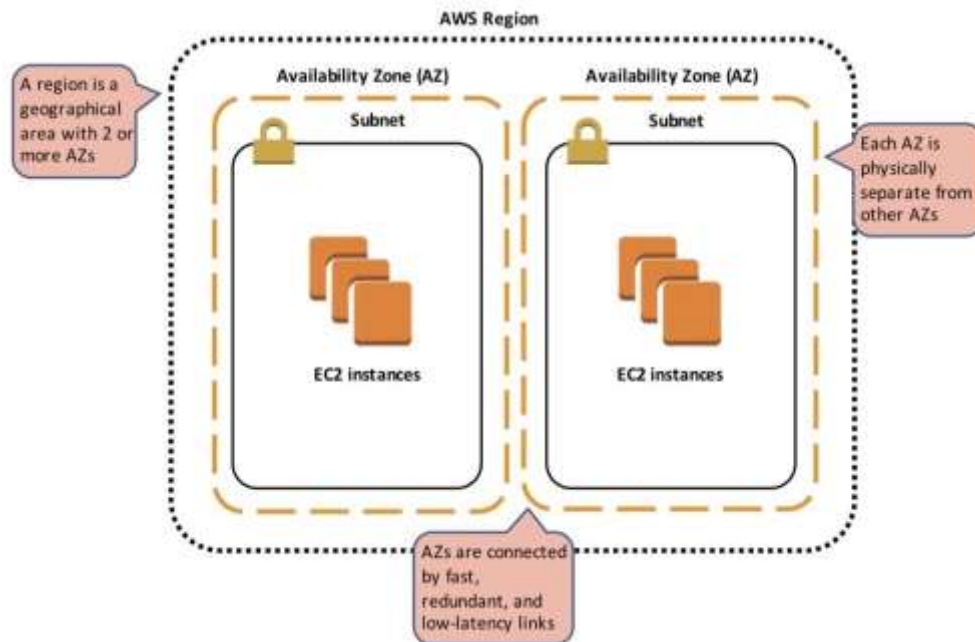
Edge Locations and Regional Edge Caches

Edge locations are Content Delivery Network (CDN) endpoints for CloudFront.

There are many more edge locations than regions.

Currently there are over 100 edge locations.

Regional Edge Caches sit between your CloudFront Origin servers and the Edge Locations.

A Regional Edge Cache has a larger cache-width than each of the individual Edge Locations.