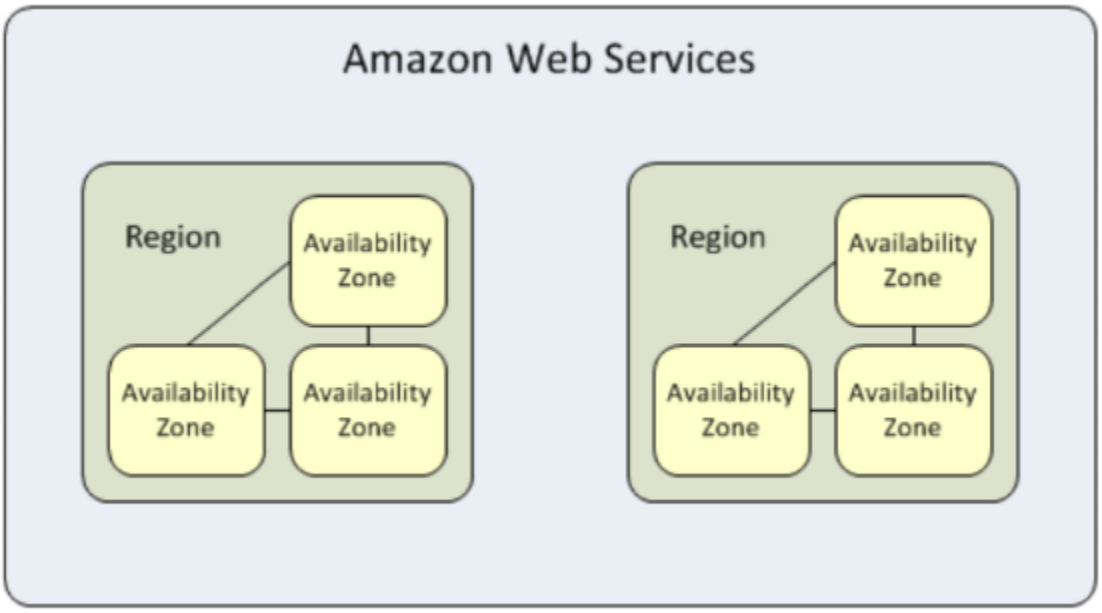
Regions, Availability Zones & Edge Locations

Amazon EC2 is hosted in multiple locations world-wide. These locations are composed of regions and Availability Zones. Each **region** is a separate geographic area.

Each region has multiple, isolated locations known as **Availability Zones.** 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 specifically.

Each region is completely independent. Each Availability Zone is isolated, but the Availability Zones in a region are connected through low-latency links. The below diagram illustrates the relationship between regions and Availability Zones.



So, let's see in detail **"What is An AWS Region"**

An amazon aws region is a physical location spread across globe to host your data to reduce latency. In each region there will be atleast two availability zones for fault tolerance. Below table illustrates some regions in AWS. Currently we have approx. 22 regions in AWS.

Region CodeName	Location
Ap-northeast-1	Asia Pacific (Tokyo)
Ap-southeast-1	Asia Pacific (Singapore)
Eu-central-1	EU(Frankfrut)
Us-east-1	US East (North Virgina)
Us-west-2	US West (Oregon)

What is an Amazon AWS Availability Zone?

An AZ is a combination of **one or more** data centers in a given region. These data centers need not be separated by multiple kilometers physically but by meters within a physical compound which are completely isolated from each other failure such as power, network in given AZ. Basically, it can be seen as a logical grouping of data centers in a given region for serving high availability. These AZ's in a region are connected with direct fiber optic links which have capacity of around 25Tbps bandwidth and a latency of 2ms to 1ms. Please see the below table for list of AZ's in a region:

Region CodeName	Location	List of AZ's
Ap-northeast-1	Asia Pacific (Tokyo)	Ap-northeast-1a, ap-northeast-1b, ap-northeast-1c
Ap-southeast-1	Asia Pacific (Singapore)	Ap-southeast-1a, ap-southeast-1b
Eu-central-1	EU(Frankfrut)	Eu-central-1a, eu-central-1b
Us-east-1	US East(North Virginia)	Us-east-1a, us-east-1b, us-east-1c, us-east-1d, us-east-1e
Us-west-2	US West(Oregon)	Us-west-2a,us-west-2b,us-west-2c

What is an Amazon AWS Edge Location?

An edge location is where end users access services located at AWS. They are located in most of the cities around the world and are specifically used by CloudFront (CDN) to distribute content to end users to reduce latency. It is like frontend for the service we access which are located in AWS cloud. Below table shows some of the Edge locations in AWS:

Region CodeName	Location	List of AZ's	Edge Locations
Ap-northeast-1	Asia Pacific (Tokyo)	Ap-northeast-1a, ap-northeast-1b, ap-northeast-1c	Chennai, India, Hong Kong, China (2), Melbourne, Australia, Mumbai, India, Osaka, Japan
Ap-southeast-1	Asia Pacific (Singapore)	Ap-southeast-1a, ap-southeast-1b	Chennai, India, Hong Kong, China (2), Melbourne, Australia, Mumbai, India, Osaka, Japan
Eu-central-1	EU(Frankfrut)	Eu-central-1a, eu-central-1b	Frankfrut(Germany)(3), Seoul, Korea, London Manila, The Philippines, etc..
Us-east-1	US East(North Virginia)	Us-east-1a, us-east-1b, us-east-1c, us-east-1d, us-east-1e	Paris, France (2), Stockholm, Sweden, and Warsaw, Poland
Us-west-2	US West(Oregon)	Us-west-2a,us-west-2b,us-west-2c	Ashburn, VA (3), Atlanta, GA, Dallas/Fort Worth, TX (2), Hayward, CA, Jacksonville, FL

Below is the summary of AWS Global Infrastructure:

22 Launched Regions Each with multiple Availability Zones (AZ's)	**5 Announced Regions**	**70 Availability Zones**	**1 Local Zone** For ultralow latency applications
2x More Regions With multiple AZ's than the next largest cloud provider	**245 Countries and Territories Served**	**97 Direct Connect Locations**	**216 Points of Presence** 205 Edge Locations and 11 Regional Edge Caches

Let's see the AWS Global Infrastructure Map:

Here if talk about Asia Pacific Region alone. It has 20 availability zones with six geographic regions and, with 34 Edge Network Locations.

Regions (Availability Zones):
Hong Kong SAR (3), Mumbai (3), Seoul (3), Singapore (3), Sydney (3), Tokyo (4), Osaka (1)

Edge Locations:

Bangalore, India (3); Chennai, India (2); Hong Kong SAR, China (3); Hyderabad, India (4); Kuala Lumpur, Malaysia; Mumbai, India (2); Manila, Philippines; New Delhi, India (5); Osaka, Japan; Seoul, South Korea (4); Singapore (3); Taipei, Taiwan(3); Tokyo, Japan (16); Melbourne; Perth; Sydney (2); Beijing, China (1); Shanghai, China (1); Zhongwei, China (1); Shenzhen, China (1)

Let's see what services are global and region specific:

- S3: S3 Namespace is *global* but buckets are created within the selected *region*. Cross Region Replication (CRR) is an Amazon S3 feature that automatically replicates data across AWS regions. With CRR, every object uploaded to an S3 bucket is automatically replicated to a destination bucket in a different AWS region that you choose.
- EC2: Resource Identifiers (e.g. AMI ID, instance ID) are *regional*. An instance is hosted on selected Availability Zone. There is a soft limit of 20 EC2 instances per region. EBS volume is attached only to the instance in the *same Availability Zone*. But EBS snapshot is used to create volumes in the *same region*. To make EBS snapshot available in the different region, copy from one region to other then share it. The same as AMI (Amazon Machine Images) is also a *regional* resource to provide templates to launch EC2 instances in the region.
- Identity & Access Management (IAM): IAM is a *global* service. Same AWS accounts, users, groups and roles can be used in all regions. However, EC2 Key Pairs are *region* specific. You can create an RSA key pair using a third-party tool and then import the public key to Amazon EC2 in all regions.
- Route 53: Route 53 is a globally distributed service and offered at AWS edge locations.
- Auto Scaling: Auto Scaling ensures that you are running your desired number of Amazon EC2 instances. It spans across *multiple Availability Zones* within the *same* region. It's easy to confuse it with Placement group. A placement group is a logical grouping of EC2 instances with in a *single* Availability Zone.
- Elastic Load Balancer (ELB): ELB distributes traffic across instances in *multiple Availability Zones* in the same region.

- Virtual Private Cloud (VPC): VPC is a logical data center within a *Region*. Subnet can span only a *single Availability Zone.* Security groups, Network ACLs, and Route Tables in VPC can span subnets/*AZs* in the same region. VPC Peering can be performed across VPCs in the same account or between another account within the *same region*.
- DynamoDB: DynamoDB is a regional service. So there is no need to explicitly create a multi-AZ deployment.
- Elastic IP Address(EIP): By default, all AWS accounts are limited to five (5) Elastic IP addresses per *region*, because public (IPv4) internet addresses are a scarce public resource.

For more info on AWS Global Infrastructure, visit:
https://aws.amazon.com/about-aws/global-infrastructure/