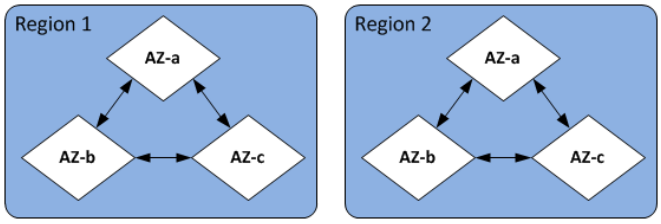
AWS Global Infrastructure

10 July 2023 19:12

The following are the components for AWS infrastructure:

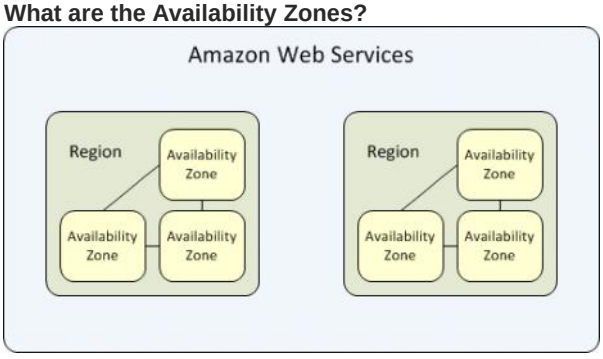
- Region
- Availability Zones
- Edge locations
- Regional Edge Caches

What are the AWS regions?

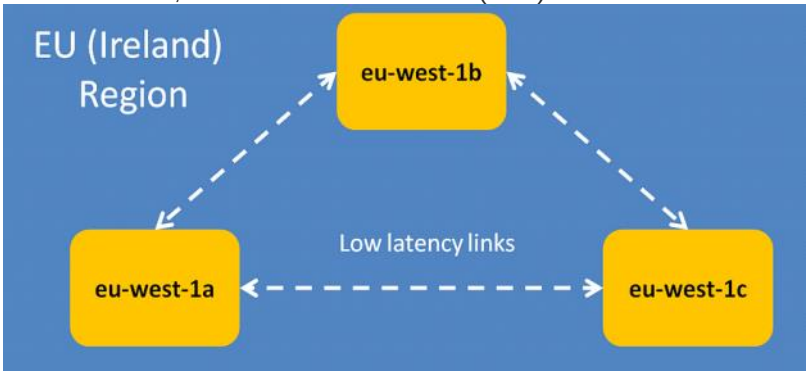


- A region represents a separate geographic area.
- The region is the collection of availability zones mapped to physical data centers in that region.
- Every region is physically isolated from and independent of each other in terms of location, power, water supply, etc but the availability zones within each region are connected via low-latency links to provide replication and fault tolerance.
- Whenever we deploy any application we need to choose regions closer to end-users to reduce the latency.

What are the Availability Zones?



- An Availability Zone (AZ) consists of one or more data centers at a location within an AWS Region.
- Each AZ has an independent cooling system, power supply, and physical security.
- Every AZ in an AWS Region is interconnected with low-latency networking and high-bandwidth to provide high-throughput, low-latency networking between AZ's.
- All AZ's are physically separated within 60 miles of each other.
- The AWS Cloud spans 81 Availability Zones within 25 geographic regions around the world.
- AWS also announced plans for 21 more Availability Zones and 7 more AWS Regions in Australia, India, Indonesia, Israel, Spain, Switzerland, and United Arab Emirates (UAE).



For example, Europe (Ireland) is of the region in AWS with **region code eu-west-1** and there are three data centers available within the region ie. **eu-west-1a, eu-west-1b, eu-west-1c** that are connected to each other.

What is the AWS Edge Locations?

- Edge locations are AWS data centers designed for caching content to deliver services with the lowest latency possible.
- Edge locations are mainly located in most of the major cities to distribute the content to end-users with reduced latency.
- Currently, there are over 218+ Edge Locations
- A site that CloudFront uses to cache copies of your content for faster delivery to users at any location.

Let's think a user is trying to access a site from India but the application is hosted in the North Virginia region in AWS.
When a user requests some image content from the website, the first request goes to the edge location that is nearest to the user.
In the edge location, CloudFront checks the cache for the requested content. If the content is available then returns to the user if not CloudFront sends the request to the origin server. The origin server retrieves the image and sends it back to the CloudFront edge location.
CloudFront adds the files to the cache in the edge location for the next time someone requests those files.
Amazon Web Services (AWS) is adding two new edge locations in India at Hyderabad and New Delhi.

What is Regional Edge Cache?

- AWS announced a new type of edge location in November 2016, known as a Regional Edge Cache.
- Regional Edge cache lies between CloudFront Origin servers and the edge locations.
- A regional edge cache has a large cache than an individual edge location.
- Data is removed from the cache at the edge location while the data is retained at the Regional Edge Caches.
- Currently, there are over 12 regional Edge Locations.
- When the user requests the data, then data is no longer available at the edge location. Therefore, the edge location retrieves the cached data from the Regional edge cache instead of the Origin servers that have high latency.

Scope of AWS Services — Global, Regional, Availability Zone resources

AWS provides a lot of services and these services are either Global, Regional or specific to the Availability Zone and cannot be accessed outside.

- IAM: Users, Groups, Roles, Accounts — **Global**
  Same AWS accounts, users, groups, and roles can be used in all regions.

- Key Pairs: Amazon EC2 created key pairs are specific to the region — **Regional**
  RSA key pair can be created and uploaded that can be used in all regions

- Virtual Private Cloud
- VPC — **Regional**
  VPC is created within a region

- Subnet — **Availability Zone**
  A subnet can span only a single Availability Zone

- Security groups — **Regional**
  A security group is tied to a region and can be assigned only to instances in the same region.

- VPC Endpoints — **Regional**
  You cannot create an endpoint between a VPC and an AWS service in a different region.

- VPC Peering — **Regional**
  VPC Peering can be performed across VPC in the same account of different AWS accounts but only within the same region. They cannot span across regions

- Elastic IP Address — **Regional**
  Elastic IP addresses created within the region can be assigned to instances within the region only

- EC2 Instance — **Regional**
  An instance is tied to the Availability Zones in which you launched it. However, note that its instance ID is tied to the region. Each resource identifier, such as an AMI ID, instance ID, EBS volume ID, or EBS snapshot ID, is tied to its region and can be used only in the region where you created the resource.

- EBS Volumes — **Availability Zone**
  Amazon EBS volume is tied to its **Availability Zone** and can be attached only to instances in the same Availability Zone.

- EBS Snapshot — **Regional**
  An EBS snapshot is tied to its region and can only be used to create volumes in the same region and has to be copied from one region to another if needed

- AMIs — **Regional**
  AMI provides templates to launch EC2 instances
  AMI is tied to the Region where its files are located with Amazon S3. For using AMI in different regions, the AMI can be copied to other regions

- Auto Scaling — **Regional**
  Auto Scaling spans across multiple Availability Zones within the same region but cannot span across regions

- Elastic Load Balancer — **Regional**
  Elastic Load Balancer distributes traffic across instances in multiple Availability Zones in the same region

- Cluster Placement Groups — **Availability Zone**
  Cluster Placement groups can be span across Instances within the same Availability Zones

- S3 — **Global but Data is Regional**
  S3 buckets are created within the selected region
  Objects stored are replicated across Availability Zones to provide high durability but are not cross-region replicated unless done explicitly

- Route53 — **Global**
  Route53 services are offered at AWS edge locations and are global

- DynamoDb — **Regional**
  All data objects are stored within the same region and replicated across multiple Availability Zones in the same region
  Data objects can be explicitly replicated across regions using cross-region replication

- WAF — **Global**
  Web Application Firewall (WAF) services protect web applications from common web exploits are offered at AWS edge locations and are global

- CloudFront — **Global**
  CloudFront is the global content delivery network (CDN) service are offered at AWS edge locations

- Storage Gateway — **Regional**
  AWS Storage Gateway stores volume, snapshot, and tape data in the AWS region in which the gateway is activated.