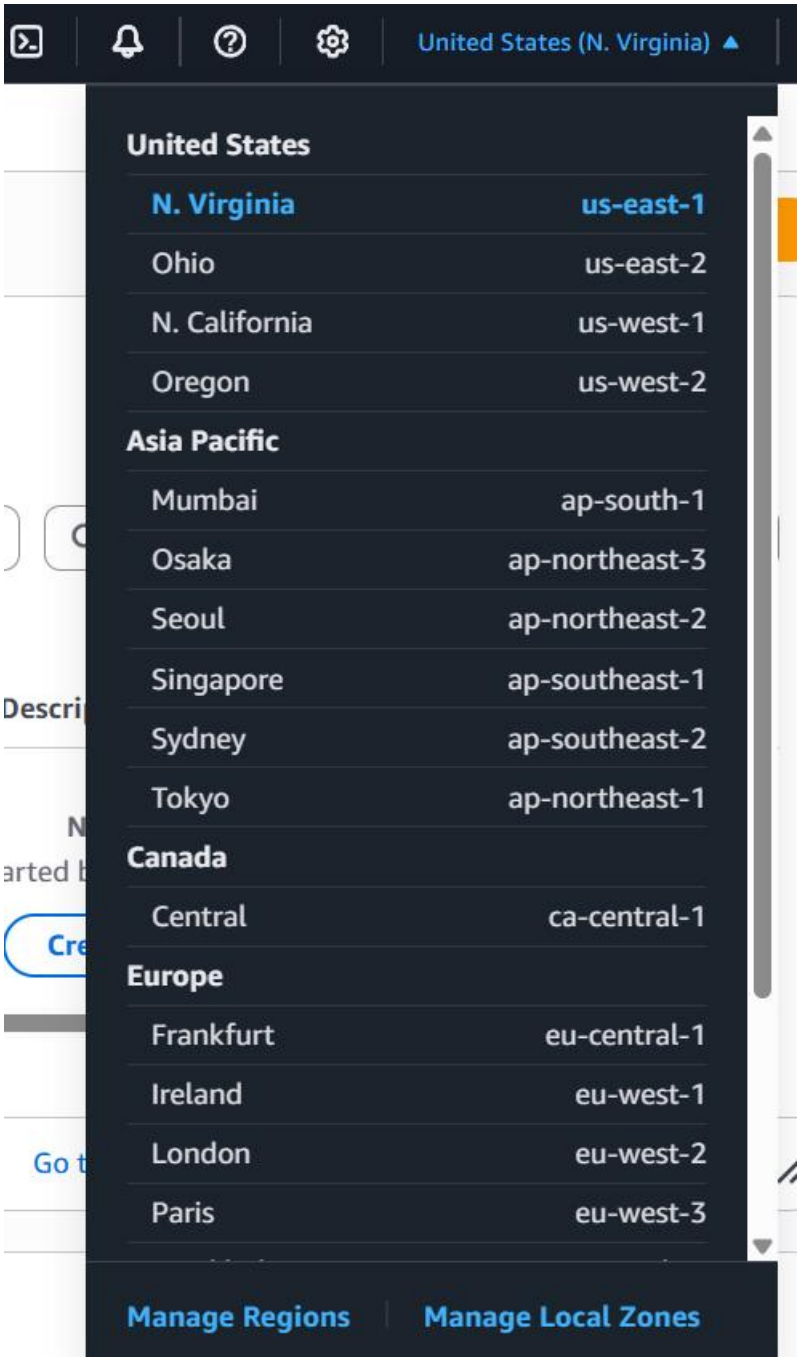
Regions is a geographical area that contains one or more data centers, while an availability zone is a physically separate data center within a region designed for high availability and fault tolerance. Each availability zone operates independently ensuring that if one zone fails the others continue to function

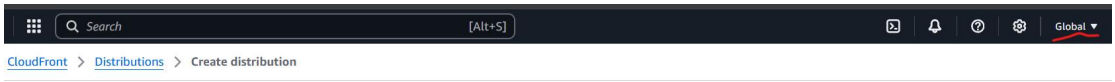
For AWS specific-
New services almost always become available first in US-EAST
Not all AWS services are available in all regions
The cost of AWS services vary per region
All billing information appears in US-EAST-1(North Virginia)

Factors to consider when choosing a region
1. What regulatory compliance does this region meet?
2. What is the cost of AWS services in this region?
3. What AWS services are available in this region?
4. What is the distance or latency to my-end-users?

Regional Services
AWS scopes their AWS Management console on a selected region. This will determine where an AWS service will be launched and what will be seen within an AWS service console. You generally do not explicitly set the region for a service at the time of creation



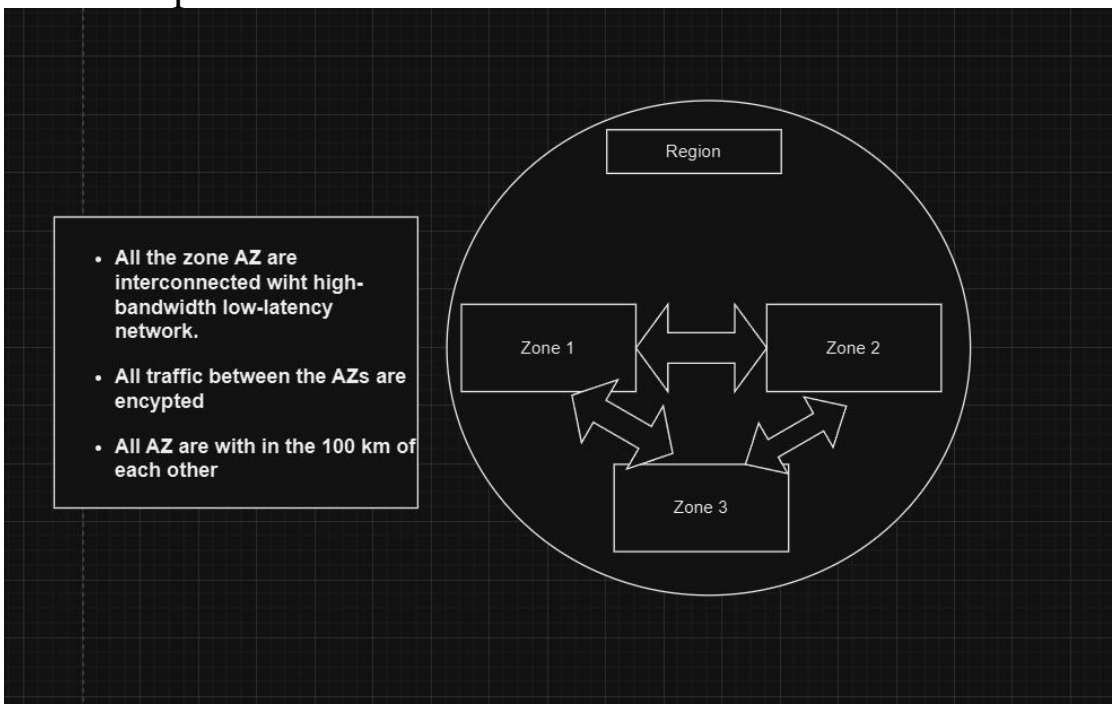
Global services
Some AWS services operate across multiple regions and the region will be fixex to “Global” Eg. S3,CloudFront, IAM
CloudFront-



Availability Zone(AZ)- A region will generally contain 3 AZs. AZs are represented by a Region Code followed by a letter eg- us-east-1a

A sub net is associated with an Availabilty Zone. We never choose AZ when lauching a resources we choose subnet associated with it.

Visual Represntation-



Global Infrastrucutre- Falut Tolerance
In short the data centers can be damaged so we setup a falut domain in which the damage does not propogate further.
A falut level is a collection of falut domain

AWS Global network
Edge Locations in AWS are data centers located around the world that deliver content to users with low latency. They are part of Amazon CloudFront and cache content like images, videos, or API responses closer to users for faster access.

How Edge Locations Work:

- When a user requests content (like an image), the request is routed to the nearest Edge Location.
- If the content is already cached there, it is delivered instantly.
- If not cached, the Edge Location fetches the content from the original source (like an S3 bucket).

- The content is then cached at the Edge Location for faster access on future requests.