Global Infrastructure Section

Why make a global application?

- A **global application** is an application deployed in **multiple geographies**
- On AWS: this could be **Regions** and / or **Edge Locations**
- **Decreased Latency**
  - Latency is the time it takes for a network packet to reach a server
  - It takes time for a packet from Asia to reach the US
  - Deploy your applications closer to your users to decrease latency, better experience
- **Disaster Recovery (DR)**
  - If an AWS region goes down (earthquake, storms, power shutdown, politics)…
  - You can fail-over to another region and have your application still working
  - A DR plan is important to increase the availability of your application
- **Attack protection:** distributed global infrastructure is harder to attack

Global AWS Infrastructure

- **Regions:** For deploying applications and infrastructure

- **Availability Zones:** Made of multiple data centers

- **Edge Locations (Points of Presence):** for content delivery as close as possible to users

- More at: https://infrastructure.aws/

Global Applications in AWS

- **Global DNS: Route 53**
  - Great to route users to the closest deployment with least latency
  - Great for disaster recovery strategies
- **Global Content Delivery Network (CDN): CloudFront**
  - Replicate part of your application to AWS Edge Locations – decrease latency
  - Cache common requests – improved user experience and decreased latency
- **S3 Transfer Acceleration**
  - Accelerate global uploads & downloads into Amazon S3
- **AWS Global Accelerator:**
  - Improve global application availability and performance using the AWS global network
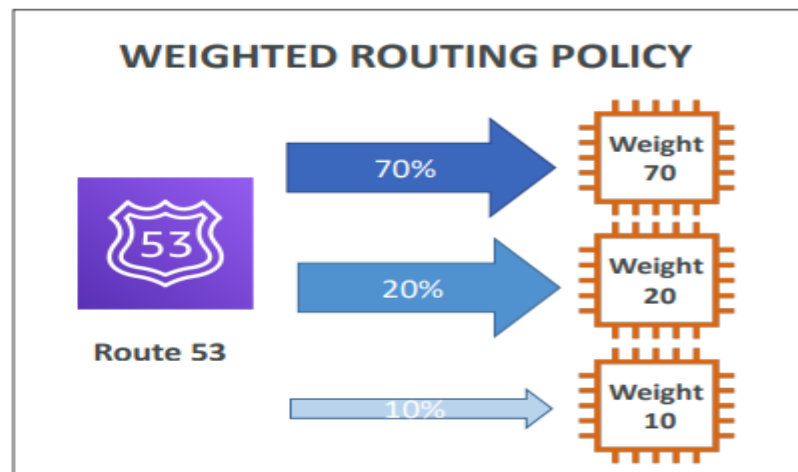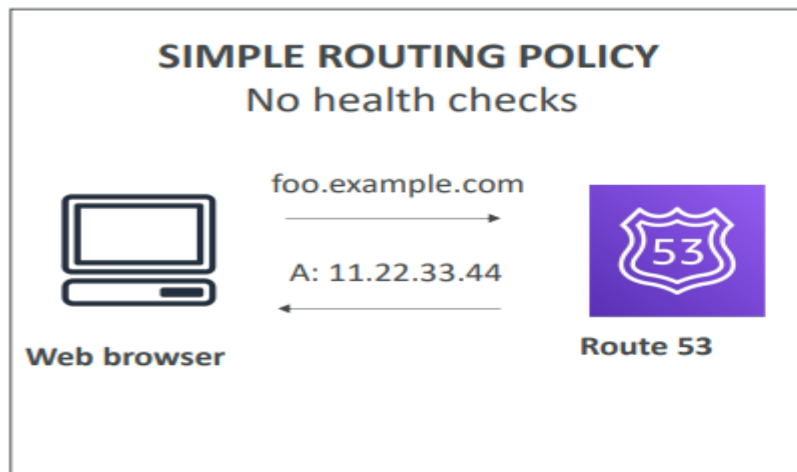
Amazon Route 53 Overview

- Route53 is a Managed DNS (Domain Name System)
- DNS is a collection of rules and records which helps clients understand how to reach a server through URLs.

- In AWS, the most common records are:
    - www.google.com => 12.34.56.78 == **A record (IPv4)**
    - www.google.com => 2001:0db8:85a3:0000:0000:8a2e:0370:7334 == **AAAA IPv6**
    - search.google.com => www.google.com == **CNAME: hostname to hostname**
    - example.com => AWS resource == **Alias (ex: ELB, CloudFront, S3, RDS, etc…)**

Route 53 – Diagram for A Record

Route 53 Routing Policies

- Need to know them at a high-level for the Cloud Practitioner Exam

SIMPLE ROUTING POLICY
No health checks

foo.example.com

A: 11.22.33.44

Web browser

Route 53

WEIGHTED ROUTING POLICY

Route 53

70% → Weight 70

20% → Weight 20

10% → Weight 10

Route 53 Routing Policies
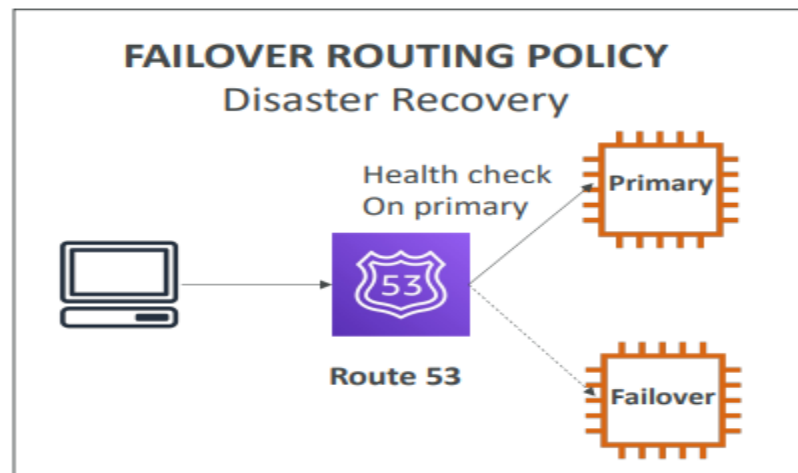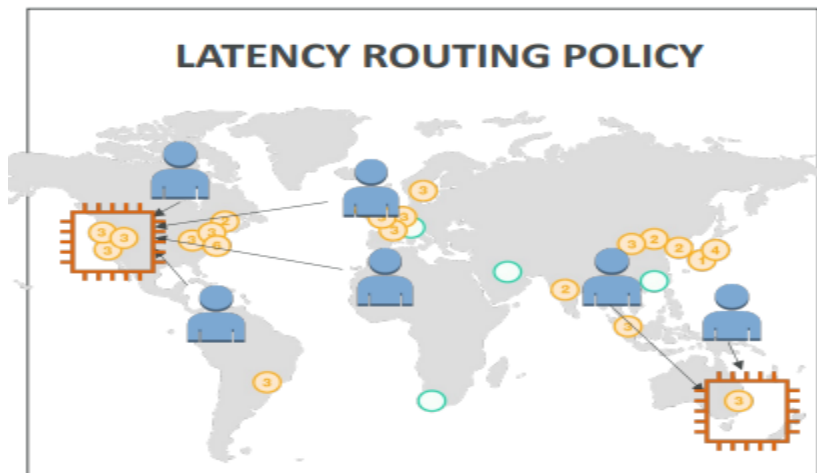
Amazon CloudFront

- Content Delivery Network (CDN)
- **Improves read performance, content is cached at the edge**
- Improves users experience
- 216 Point of Presence globally (edge locations)
- DDoS protection (because worldwide), integration with Shield, AWS Web Application Firewall
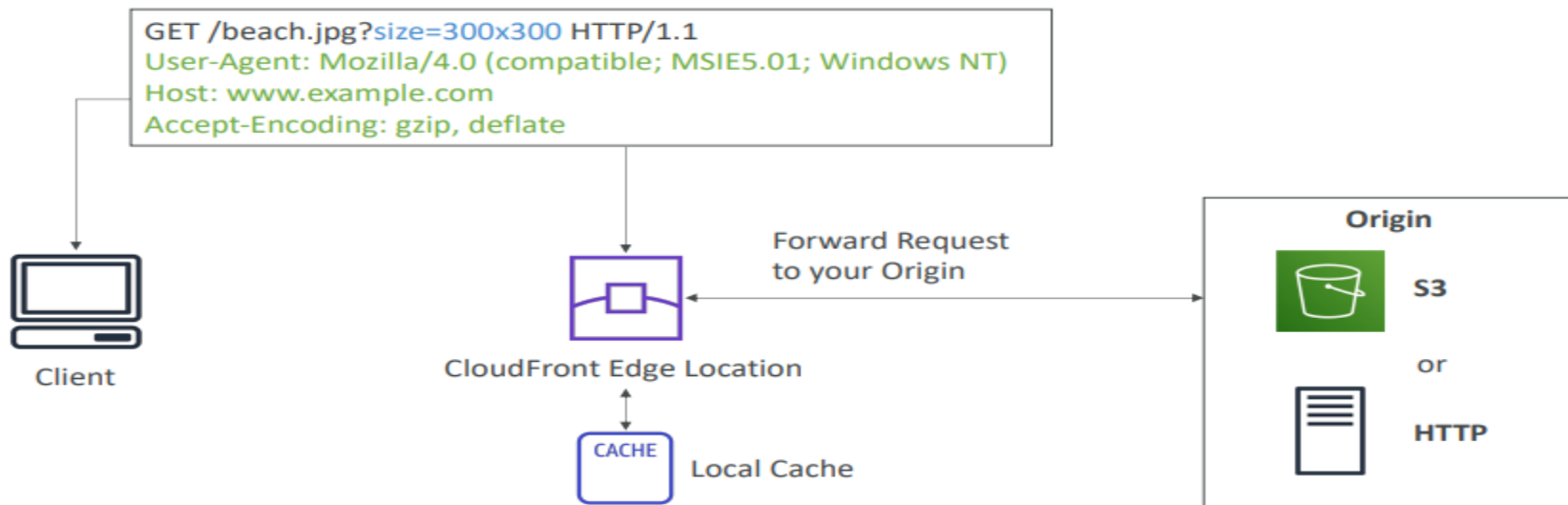

Source: https://aws.amazon.com/cloudfront/features/?nc=sn&loc=2

CloudFront – Origins

- **S3 bucket**
  - For distributing files and caching them at the edge
  - Enhanced security with CloudFront **Origin Access Control (OAC)**
  - OAC is replacing Origin Access Identity (OAI)
  - CloudFront can be used as an ingress (to upload files to S3)

- **Custom Origin (HTTP)**
  - Application Load Balancer
  - EC2 instance
  - S3 website (must first enable the bucket as a static S3 website)
  - Any HTTP backend you want

CloudFront at a high level

GET /beach.jpg?size=300x300 HTTP/1.1
User-Agent: Mozilla/4.0 (compatible; MSIE5.01; Windows NT)
Host: www.example.com
Accept-Encoding: gzip, deflate

Client

CloudFront Edge Location

CACHE
Local Cache

Forward Request
to your Origin

Origin

S3

or

HTTP

CloudFront – S3 as an Origin
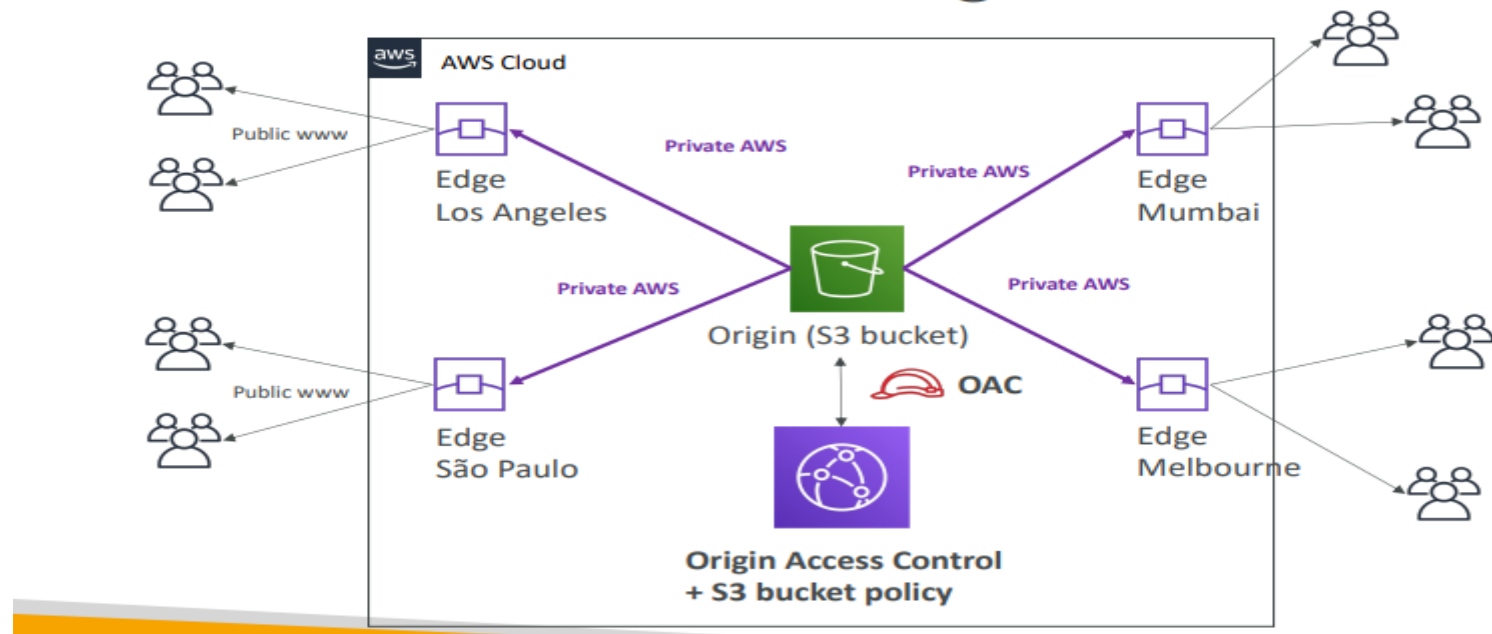
CloudFront vs S3 Cross Region Replication

- CloudFront:
  - Global Edge network
  - Files are cached for a TTL (maybe a day)
  - **Great for static content that must be available everywhere**

- S3 Cross Region Replication:
  - Must be setup for each region you want replication to happen
  - Files are updated in near real-time
  - Read only
  - **Great for dynamic content that needs to be available at low-latency in few regions**

S3 Transfer Acceleration

- Increase transfer speed by transferring file to an AWS edge location which will forward the data to the S3 bucket in the target region

Test the tool at: https://s3-accelerate-speedtest.s3-accelerate.amazonaws.com/en/accelerate-speed-comparsion.html

AWS Global Accelerator

- Improve global application <u>availability</u> and <u>performance</u> using the AWS global network

- Leverage the AWS internal network to optimize the route to your application (60% improvement)

- 2 Anycast IP are created for your application and traffic is sent through **Edge Locations**

- The Edge locations send the traffic to your application
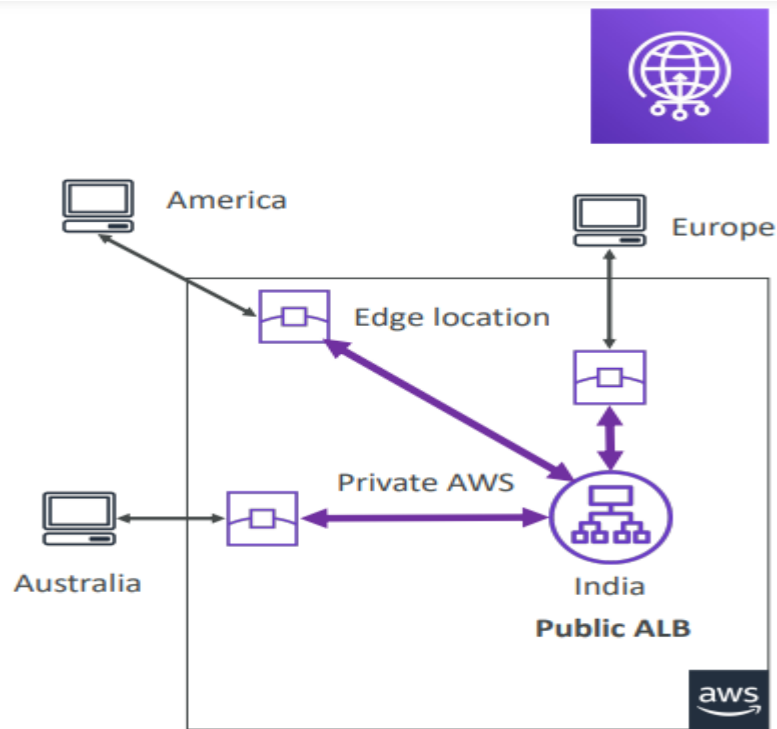
AWS Global Accelerator

Without Global Accelerator

With Global Accelerator

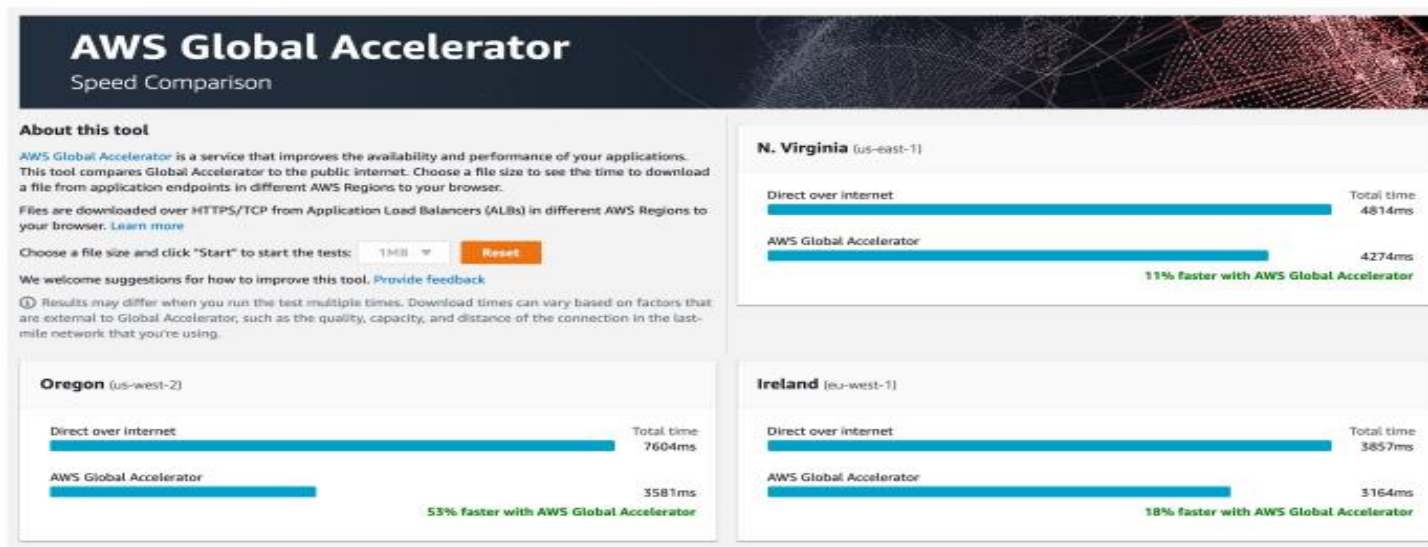https://aws.amazon.com/global-accelerator/

AWS Global Accelerator vs CloudFront

- They both use the AWS global network and its edge locations around the world
- Both services integrate with AWS Shield for DDoS protection.
- **CloudFront – Content Delivery Network**
  - Improves performance for your cacheable content (such as images and videos)
  - Content is <u>served at the edge</u>

- **Global Accelerator**
  - No caching, <u>proxying packets</u> at the edge to applications running in one or more AWS Regions.
  - Improves performance for a wide range of applications over TCP or UDP
  - Good for HTTP use cases that require static IP addresses
  - Good for HTTP use cases that required deterministic, fast regional failover

https://speedtest.globalaccelerator.aws/#/

AWS Global Accelerator
Speed Comparison

About this tool

AWS Global Accelerator is a service that improves the availability and performance of your applications. This tool compares Global Accelerator to the public internet. Choose a file size to see the time to download a file from application endpoints in different AWS Regions to your browser.

Files are downloaded over HTTPS/TCP from Application Load Balancers (ALBs) in different AWS Regions to your browser. Learn more

Choose a file size and click "Start" to start the tests: [1MB ▾]   [Reset]

We welcome suggestions for how to improve this tool. Provide feedback

ⓘ Results may differ when you run the test multiple times. Download times can vary based on factors that are external to Global Accelerator, such as the quality, capacity, and distance of the connection in the last-mile network that you're using.

N. Virginia (us-east-1)

Direct over internet   Total time
                                        4814ms

AWS Global Accelerator
                                        4274ms
              11% faster with AWS Global Accelerator

Oregon (us-west-2)

Direct over internet   Total time
                                        7604ms

AWS Global Accelerator
                                        3581ms
        53% faster with AWS Global Accelerator

Ireland (eu-west-1)

Direct over internet   Total time
                                        3857ms

AWS Global Accelerator
                                        3164ms
        18% faster with AWS Global Accelerator

AWS Outposts

- **Hybrid Cloud**: businesses that keep an on-premises infrastructure alongside a cloud infrastructure

- Therefore, two ways of dealing with IT systems:
  - One for the AWS cloud (using the AWS console, CLI, and AWS APIs)
  - One for their on-premises infrastructure

- **AWS Outposts are "server racks"** that offers the same AWS infrastructure, services, APIs & tools to build your own applications on-premises just as in the cloud

- **AWS will setup and manage "Outposts Racks"** within your on-premises infrastructure and you can start leveraging AWS services on-premises

- You are responsible for the Outposts Rack physical security

aws **AWS Cloud**

Corporate data center

On-prem servers

Extension of AWS services

Outposts Racks
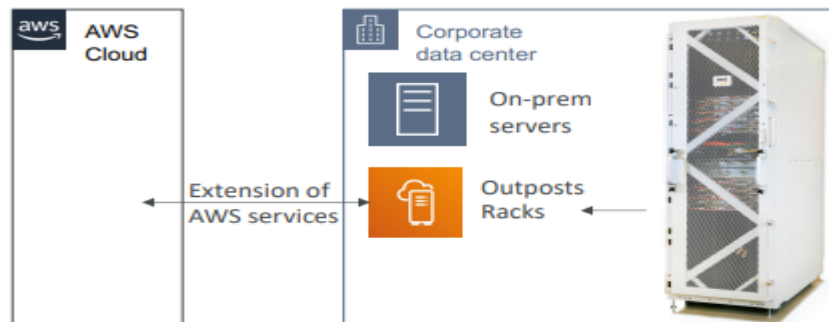
AWS Outposts

- Benefits:
  - Low-latency access to on-premises systems
  - Local data processing
  - Data residency
  - Easier migration from on-premises to the cloud
  - Fully managed service
- Some services that work on Outposts:

Amazon EC2   Amazon EBS   Amazon S3   Amazon EKS   Amazon ECS   Amazon RDS   Amazon EMR

AWS WaveLength

- **WaveLength Zones** are infrastructure deployments embedded within the telecommunications providers' datacenters at the edge of the 5G networks
- Brings AWS services to the edge of the 5G networks
- Example: EC2, EBS, VPC…
- Ultra-low latency applications through 5G networks
- Traffic doesn't leave the Communication Service Provider's (CSP) network
- High-bandwidth and secure connection to the parent AWS Region
- No additional charges or service agreements
- Use cases: Smart Cities, ML-assisted diagnostics, Connected Vehicles, Interactive Live Video Streams, AR/VR, Real-time Gaming, …
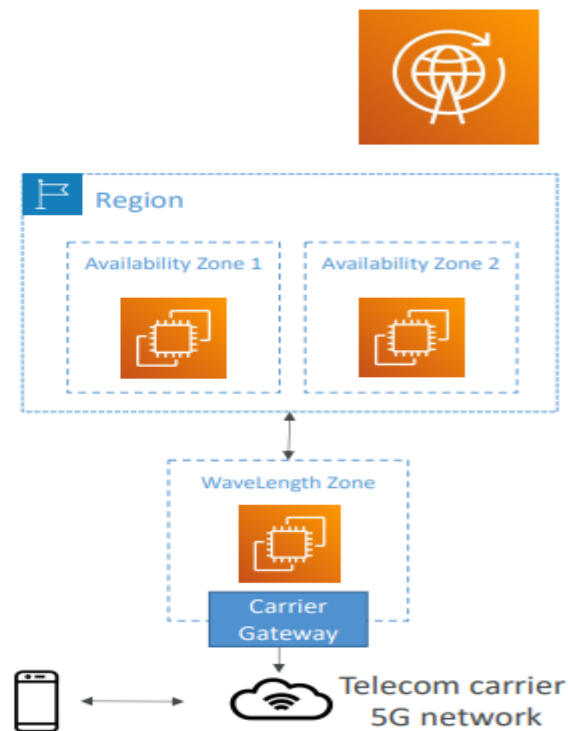
AWS Local Zones

- Places AWS compute, storage, database, and other selected AWS services **closer to end users to run latency-sensitive applications**

- Extend your VPC to more locations – "**Extension of an AWS Region**"

- Compatible with EC2, RDS, ECS, EBS, ElastiCache, Direct Connect ...

- Example:
  - **AWS Region:** N. Virginia (us-east-1)
  - **AWS Local Zones:** Boston, Chicago, Dallas, Houston, Miami, ...
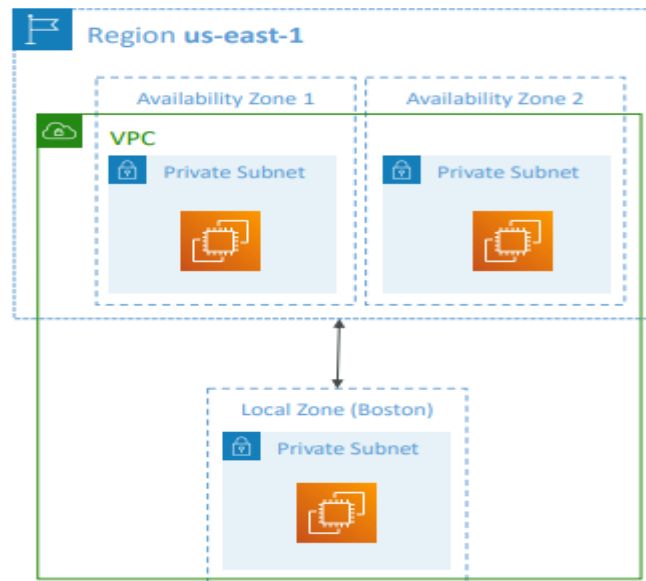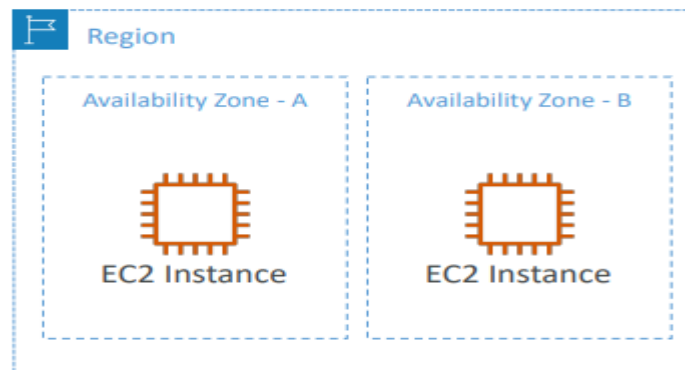
Global Applications Architecture

Single Region, Single AZ

- ✗ High Availability
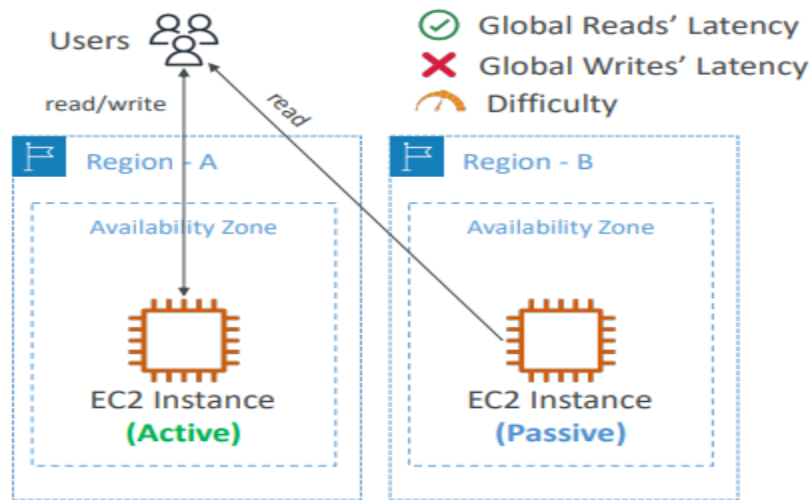- ✗ Global Latency
- ◠ Difficulty

Region

Availability Zone

EC2 Instance

Single Region, Multi AZ

- ✓ High Availability
- ✗ Global Latency
- ◠ Difficulty

Region

Availability Zone - A

EC2 Instance

Availability Zone - B

EC2 Instance

Global Applications Architecture
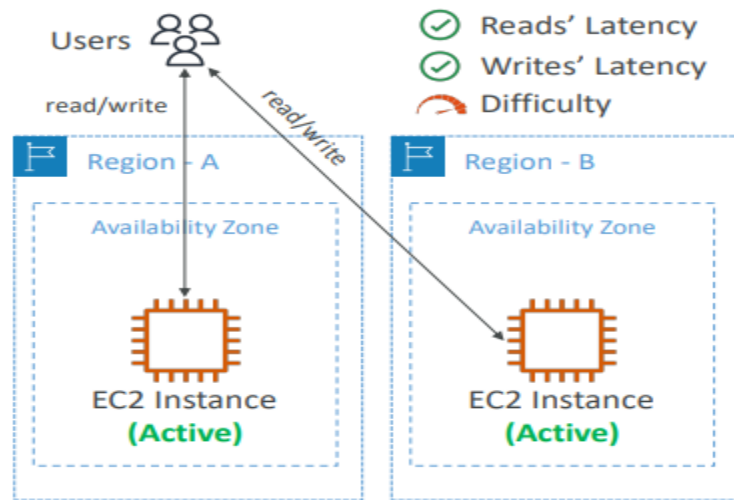
Global Applications in AWS - Summary

- **Global DNS: Route 53**
  - Great to route users to the closest deployment with least latency
  - Great for disaster recovery strategies
- **Global Content Delivery Network (CDN): CloudFront**
  - Replicate part of your application to AWS Edge Locations – decrease latency
  - Cache common requests – improved user experience and decreased latency
- **S3 Transfer Acceleration**
  - Accelerate global uploads & downloads into Amazon S3
- **AWS Global Accelerator**
  - Improve global application availability and performance using the AWS global network

Global Applications in AWS - Summary

- **AWS Outposts**
  - Deploy Outposts Racks in your own Data Centers to extend AWS services
- **AWS WaveLength**
  - Brings AWS services to the edge of the 5G networks
  - Ultra-low latency applications
- **AWS Local Zones**
  - Bring AWS resources (compute, database, storage, …) closer to your users
  - Good for latency-sensitive applications