# AWS Global Infrastructure

**What is the AWS Global Infrastructure?**

The AWS Global Infrastructure is **globally distributed hardware and datacenters** that **are physically networked together** to act as one large resource for the end customer.

The AWS Global Infrastructure is made up of the following resources:

- **25** Launched Regions
- **81** Availability Zones
- **108** Direct Connection Locations
- **275+** Points of Presence
- **11** Local Zone
- **17** Wavelength Zones



AWS has **millions** of active customers and **tens of thousands** of partners globally

# Global Infrastructure – Regions

Regions are **geographically distinct locations** consisting of one or more Availability Zones.

Every region is **physically isolated** from and independent of every other region in terms of **location, power, water supply**
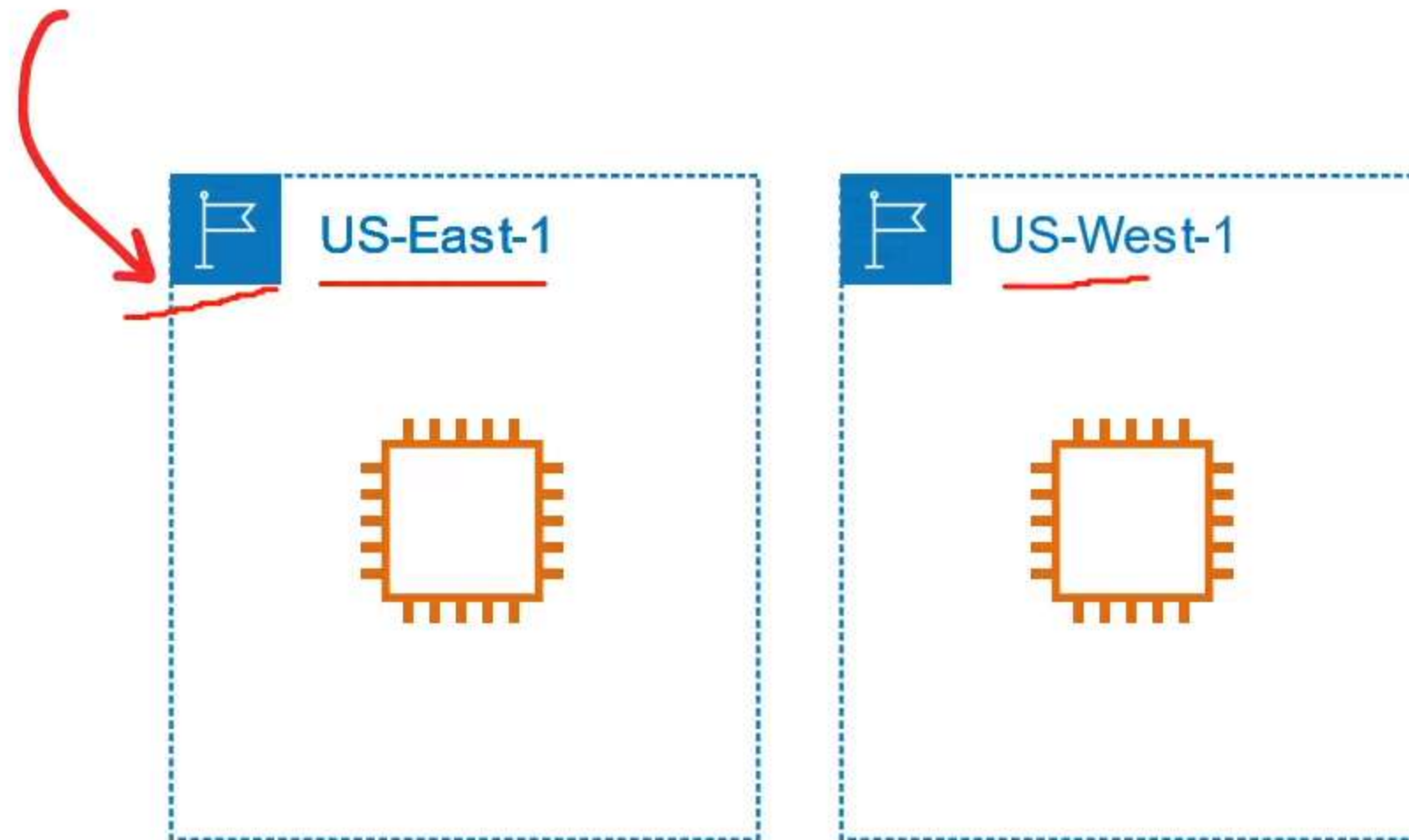
🔵 Launch Regions

🟠 Coming Soon

AWS First Region *2006*
**US-East 1** (Northern Virginia)

This is what a **region** will look like
represented in an architectural diagram.

# Global Infrastructure – Regions

Each region generally has three Availability Zones

- Some new users are limited to two eg. US-West

New services almost always become available first in **US-EAST**
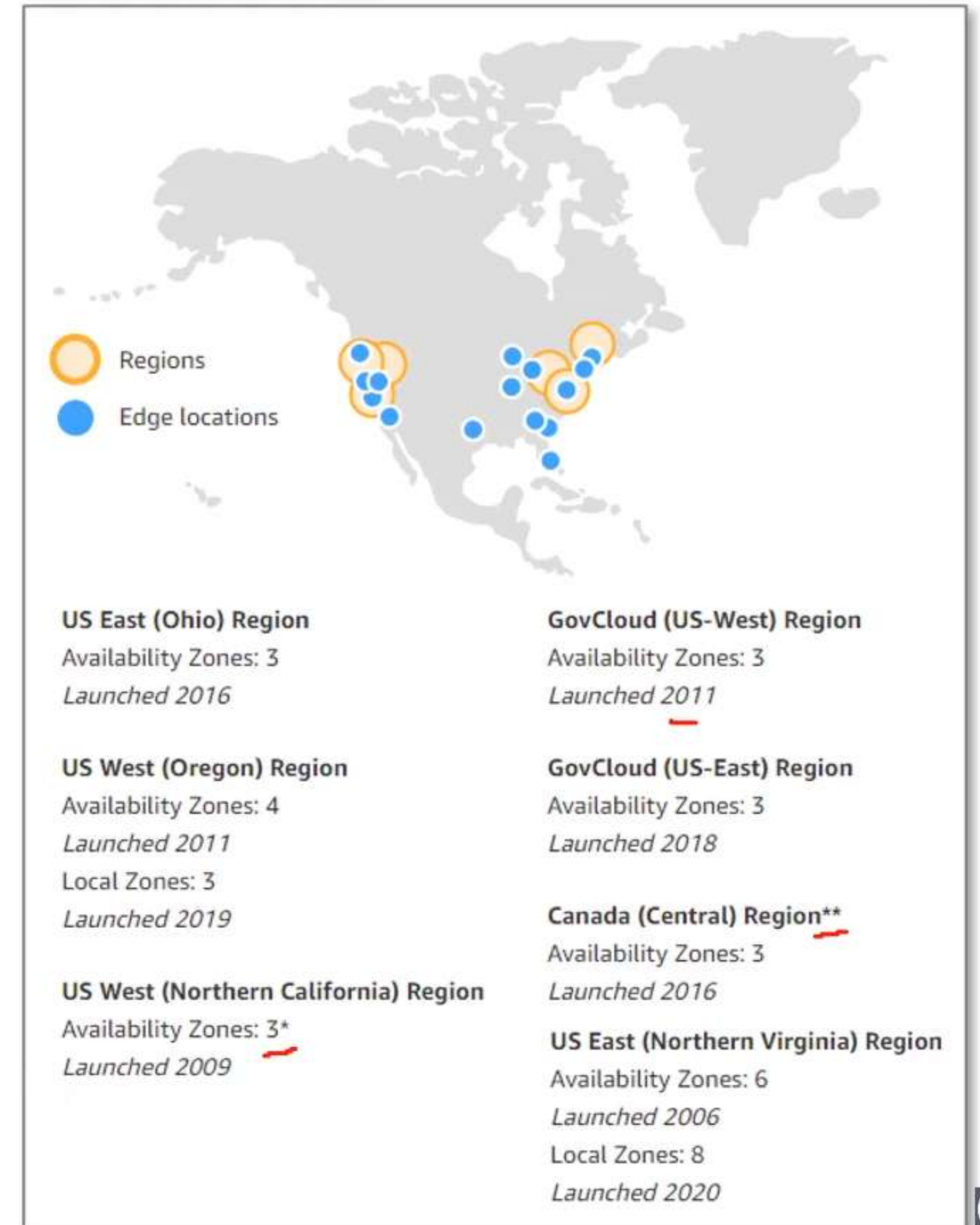
Not all AWS Services are available in all regions

All your billing information appears in **US-EAST-1** (North Virginia)

The cost of AWS services vary per region

When you choose a region there are four factors you need to consider:
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?



Regions
Edge locations

**US East (Ohio) Region**
Availability Zones: 3
*Launched 2016*

**US West (Oregon) Region**
Availability Zones: 4
*Launched 2011*
Local Zones: 3
*Launched 2019*

**US West (Northern California) Region**
Availability Zones: 3*
*Launched 2009*

**GovCloud (US-West) Region**
Availability Zones: 3
*Launched 2011*

**GovCloud (US-East) Region**
Availability Zones: 3
*Launched 2018*

**Canada (Central) Region****
Availability Zones: 3
*Launched 2016*

**US East (Northern Virginia) Region**
Availability Zones: 6
*Launched 2006*
Local Zones: 8
*Launched 2020*

# Global Infrastructure – Regional vs Global Services
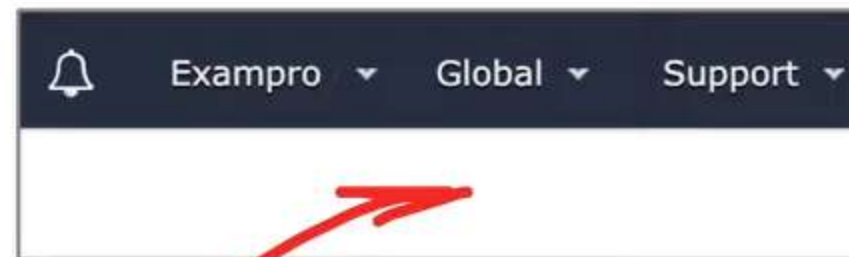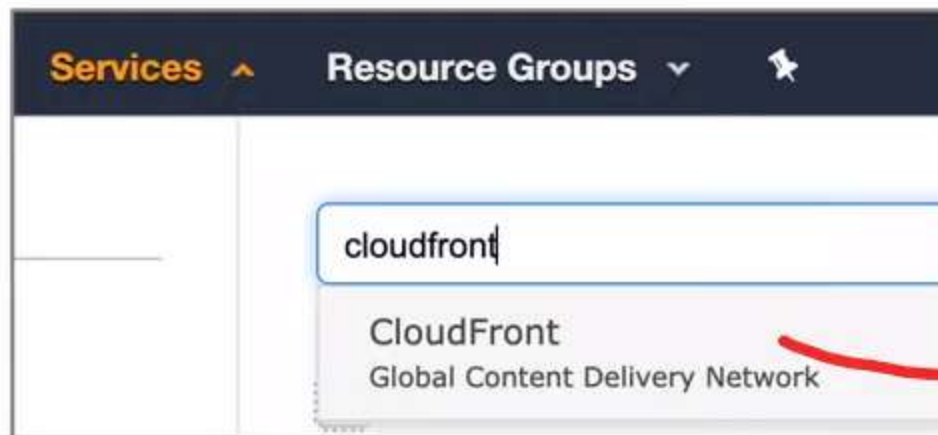
## Regional Services

AWS **scopes** their AWS Management Console on a selected Region.
This will determine <u>where</u> an AWS service will be launched and what will be seen within an AWS Service's console.
You generally don't 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 fixed to "Global"
E.g. Amazon S3, CloudFront, Route53, IAM

For these global services at the time of creation:
- There is no concept of region.                    eg. IAM User
- A single region must be explicitly chosen    eg. S3 Bucket
- A group of regions are chosen                     eg. CloudFront Distribution
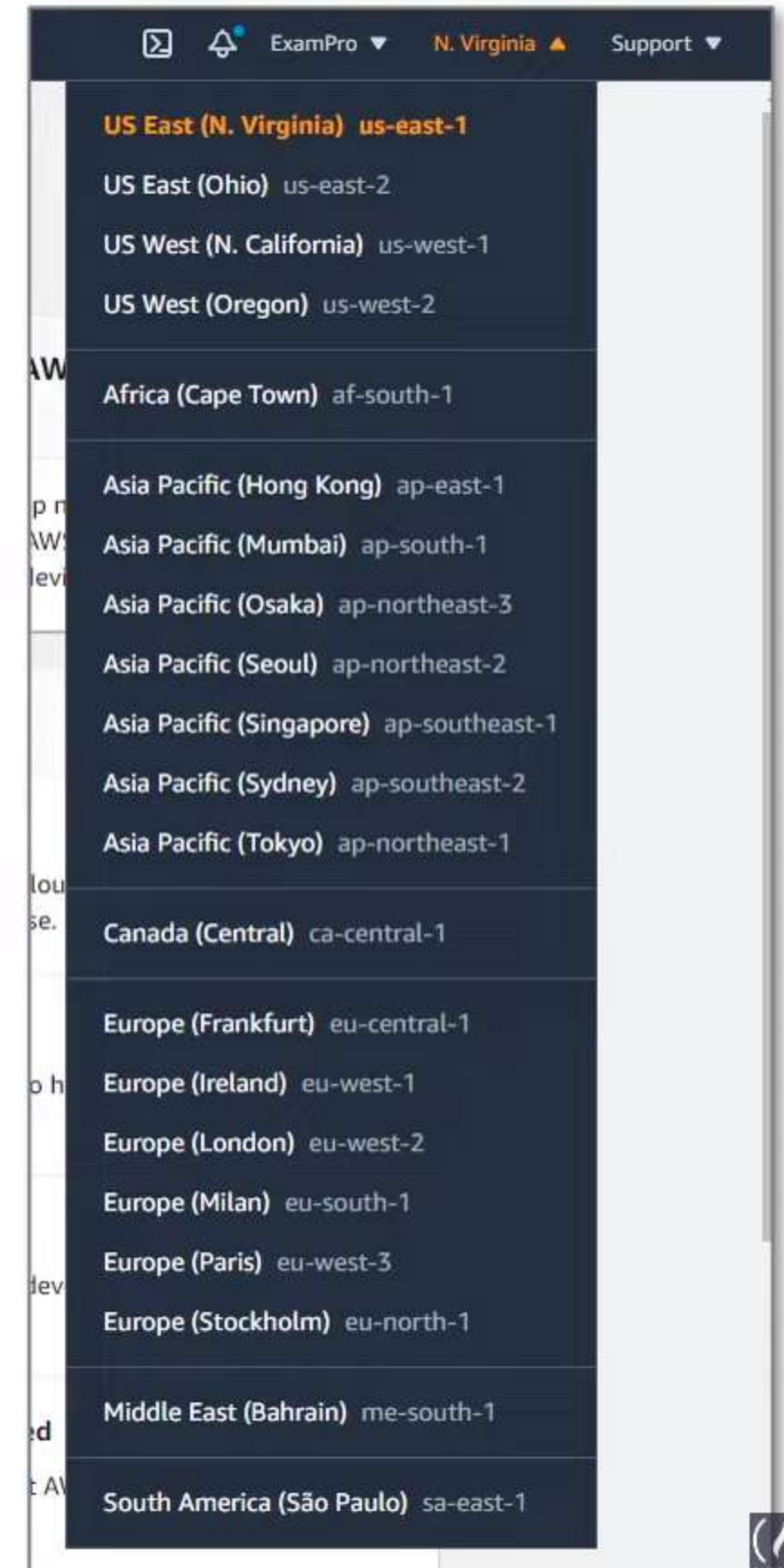
# Global Infrastructure – Availability Zones

An **Availability Zone** (AZ) is physical location made up of one or more datacenter.

A datacenter is a secured building that contains hundreds of thousands of computers.

A region will *generally contain 3 Availability Zones

Datacenters within a region will be isolate from each other (different buildings). But they will be close enough to provide low-latency (< 10ms).

Its common practice to run workloads in at least 3 AZs to ensure services remain available in case one or two datacenters fail. (High Availability)

AZs are represented by a Region Code, followed by a letter identifier eg. **us-east-1a**
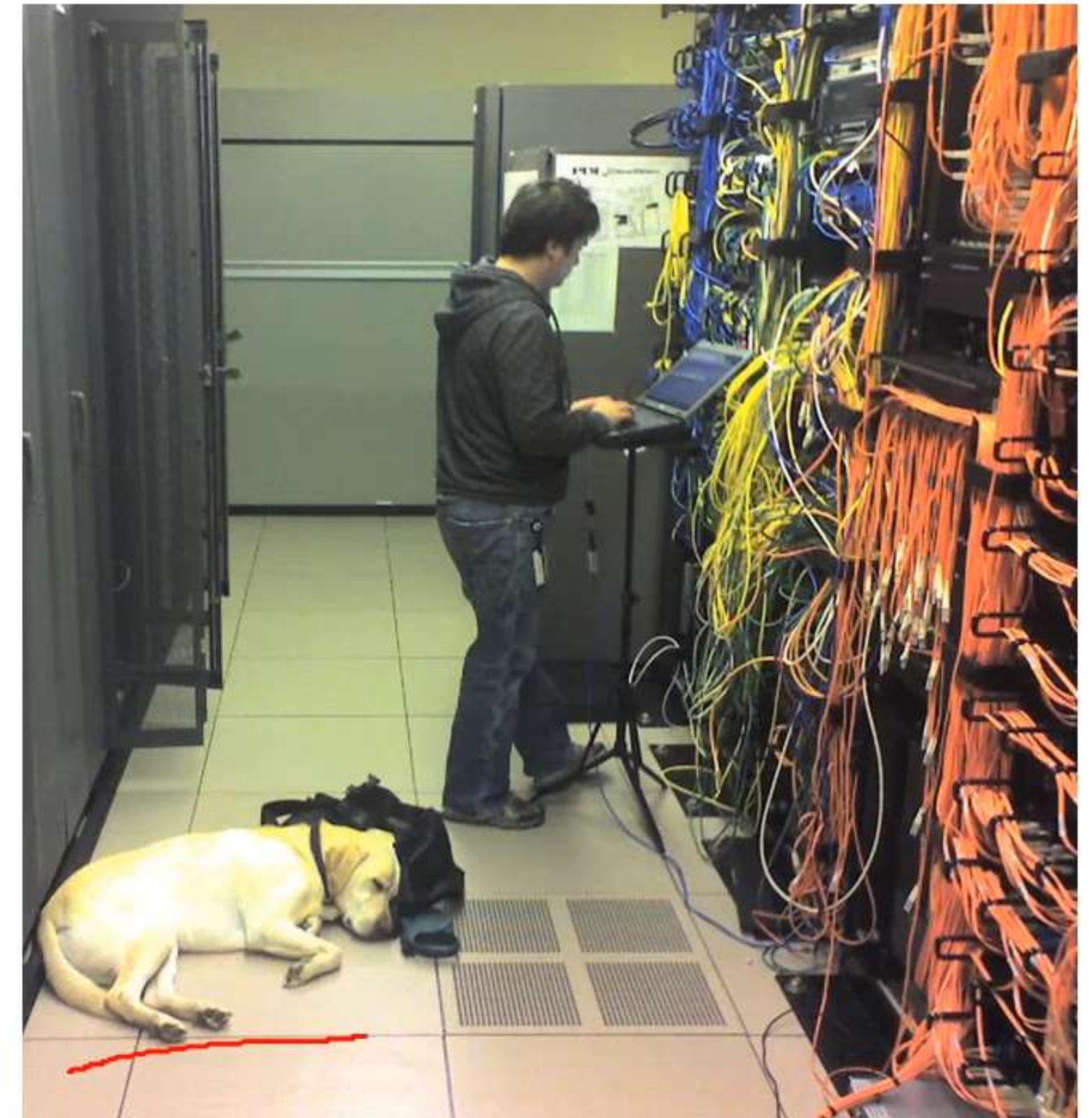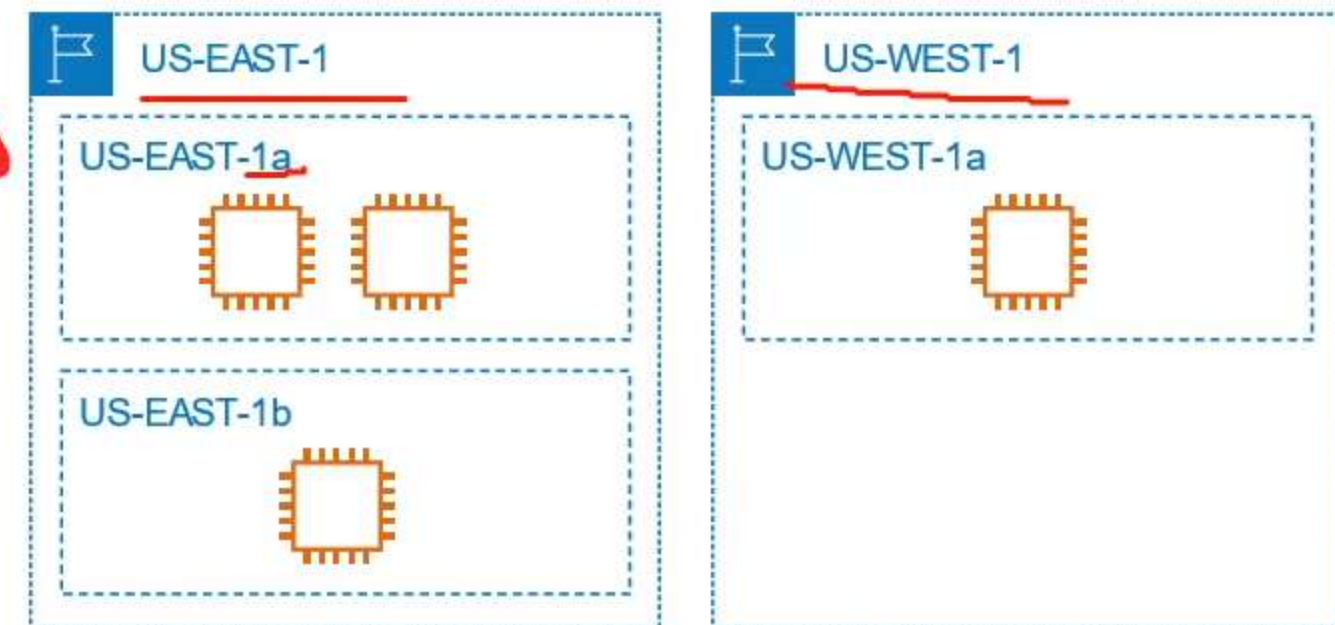
# Global Infrastructure – Availability Zones

A Subnet is associated with an Availability Zone.

You never choose the AZ when launching resources.
You choose the Subnet which is associated to the AZ.

| Subnet ⓘ | ✓ No preference (default subnet in any Availability Zone) |
| --- | --- |
| | subnet-d9de91f7 \| Default in us-east-1c |
| Public IP ⓘ | subnet-d0c28f8c \| Default in us-east-1a |
| | subnet-349fdf53 \| Default in us-east-1b |
| | subnet-a8c2f8a7 \| Default in us-east-1f |
| ent group ⓘ | subnet-b9db4c87 \| Default in us-east-1e |
| servation ⓘ | subnet-13869659 \| Default in us-east-1d |

Example of an architectural diagram, representing
two AZs, the Subnets associated with those AZs, and
EC2 instances (Virtual Machines) launched in those subnets

US-EAST-1
US-EAST-1a
US-EAST-1b

US-WEST-1
US-WEST-1a

The US-EAST-1 region has 6 AZs
(the most Availability Zones of any region)
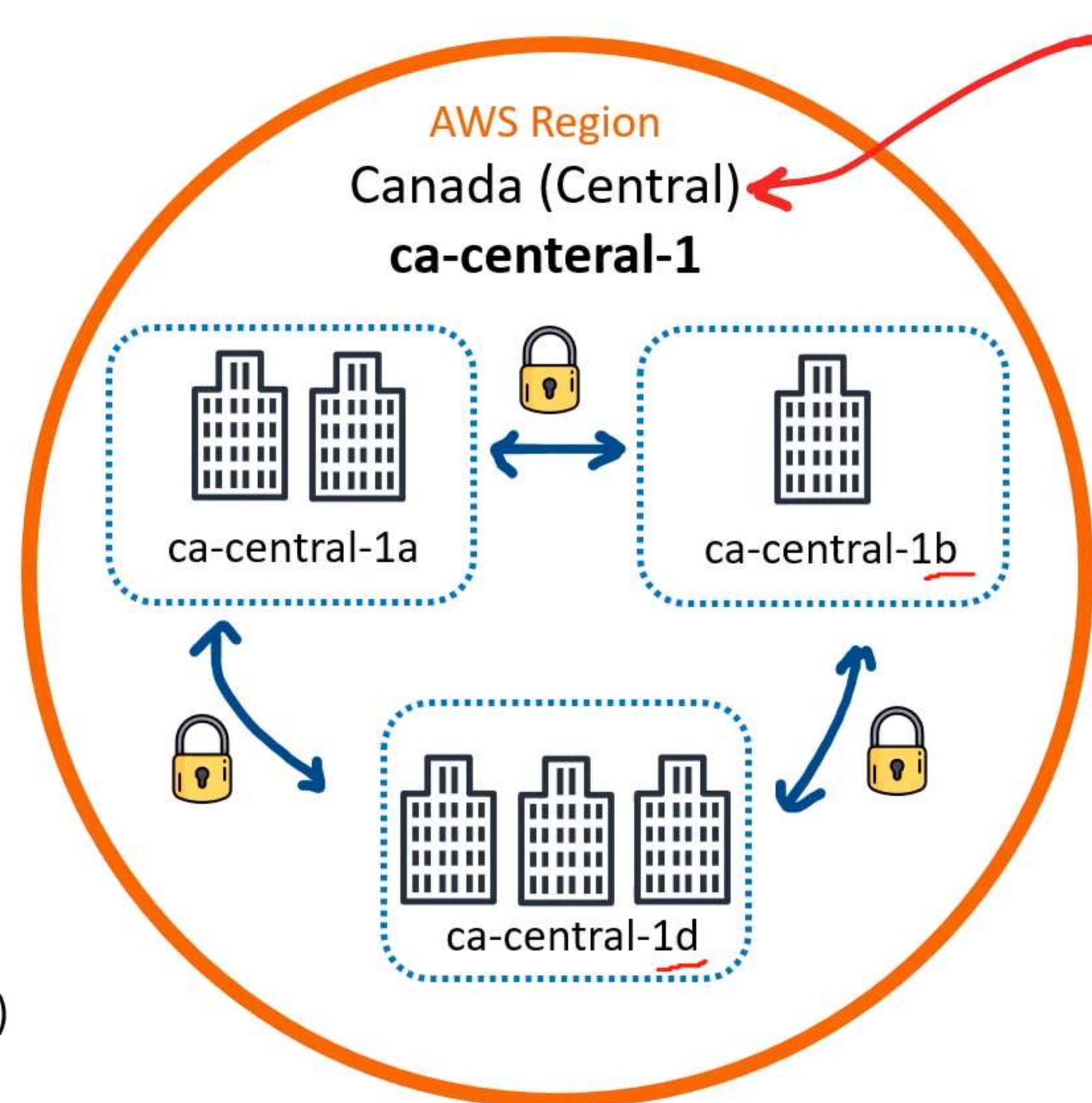
# Global Infrastructure – Availability Zones

A region has multiple Availability Zones

An Availability Zone is made up of **one or more** datacenters

All AZs in an AWS Region are interconnected with high-bandwidth, low-latency networking, over fully redundant, dedicated metro fiber providing high-throughput, low-latency networking between

All traffic between AZs is encrypted

AZs are within 100 km (60 miles) of each other.

**Montreal**

AWS Region
Canada (Central)
**ca-centeral-1**

ca-central-1a

ca-central-1b

ca-central-1d

# Global Infrastructure – Fault Tolerance

**Fault Domain**

**What is a fault domain?**

A fault domain is a section of a network that is vulnerable to damage if a critical device or system fails. The purpose of a fault domain is that if a failure occurs **it will not cascade outside that domain**, limiting the damage possible.

You can have fault domains nested inside fault domains.

**What is a fault level?**

A fault level is a collection of fault domains.

The scope of a fault domain could be:
- specific servers in a rack
- an entire rack in a datacenter
- an entire room in a datacenter
- the entire data center building

Its up to the Cloud Service Provider (CSPs) to define the boundaries of a domain

An AWS Region would be a **Fault Level** ➡️

| Fault Level us-east-1 (Region) | |
|---|---|
| Fault Domain us-east-1a (AZ) | Fault Domain us-east-1b (AZ) |

A Availability Zone would be a **Fault Domain** ➡️

# Global Infrastructure – Fault Tolerance

Each Amazon Region is designed to be completely **isolated** from the other Amazon Regions.
- This achieves the greatest possible fault tolerance and stability

Each Availability Zone is **isolated**, but the Availability Zones in a Region are connected through low-latency links

Each Availability Zone is designed as an **independent failure zone**

- *A "Failure Zone" is AWS describing a Fault Domain.*

**Failure Zone**
- Availability Zones are physically separated within a typical metropolitan region and are located in lower risk flood plains
- discrete uninterruptible power supply (UPS) and onsite backup generation facilities
- data centers located in different Availability Zones are designed to be supplied by independent substations to reduce the risk of an event on the power grid impacting more than one Availability Zone.
- Availability Zones are all redundantly connected to multiple tier-1 transit providers

**Multi-AZ for High Availability**
If an application is partitioned across AZs, companies are better isolated and protected from issues such as **power outages, lightning strikes, tornadoes, earthquakes**, and more.
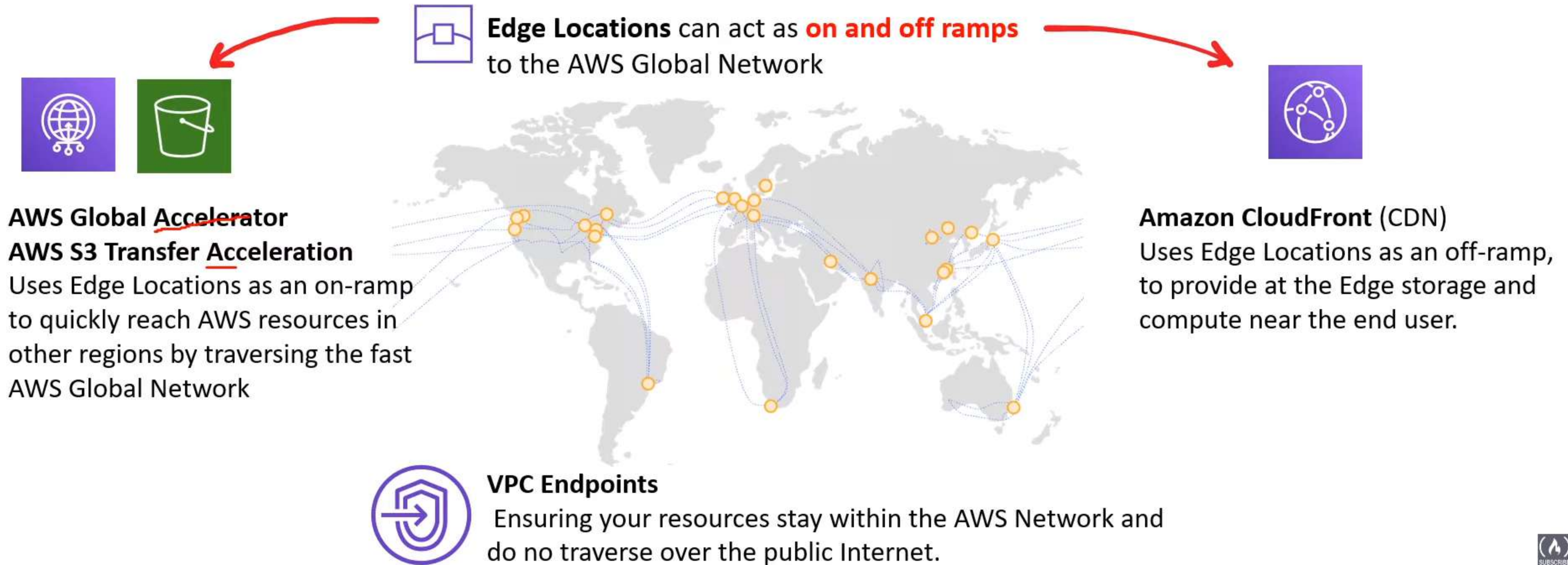
# AWS Global Network

The AWS Global Network represent the **interconnections between AWS Global Infrastructure**.
Commonly referred to as the "The Backbone of AWS".
Think of it as private expressway, where things can move very fast between datacenters.

**Edge Locations** can act as **on and off ramps** to the AWS Global Network

**AWS Global Accelerator**
**AWS S3 Transfer Acceleration**
Uses Edge Locations as an on-ramp to quickly reach AWS resources in other regions by traversing the fast AWS Global Network

**Amazon CloudFront** (CDN)
Uses Edge Locations as an off-ramp, to provide at the Edge storage and compute near the end user.

**VPC Endpoints**
 Ensuring your resources stay within the AWS Network and do no traverse over the public Internet.

# Global Infrastructure – Point of Presence (PoP)

**Points of Presence (PoP)** is an intermediate location between an AWS Region and the end user, and this location could be a datacenter or collection of hardware.

For **AWS** a Point of Presence is a data center **owned by AWS or a trusted partner** that is utilized by AWS Services related **for content delivery or expediated upload.**

**PoP** resources are:
* **Edge Locations**
* **Regional Edge Caches**

**Edge Locations** are datacenters that hold cached (copy) on the most popular files (eg. web pages,images and videos) so that the delivery of distance to the end users are reduce

**Regional Edge Locations** are datacenters that hold much larger caches of less-popular files to reduce a full round trip and also to reduce the cost of transfer fees.
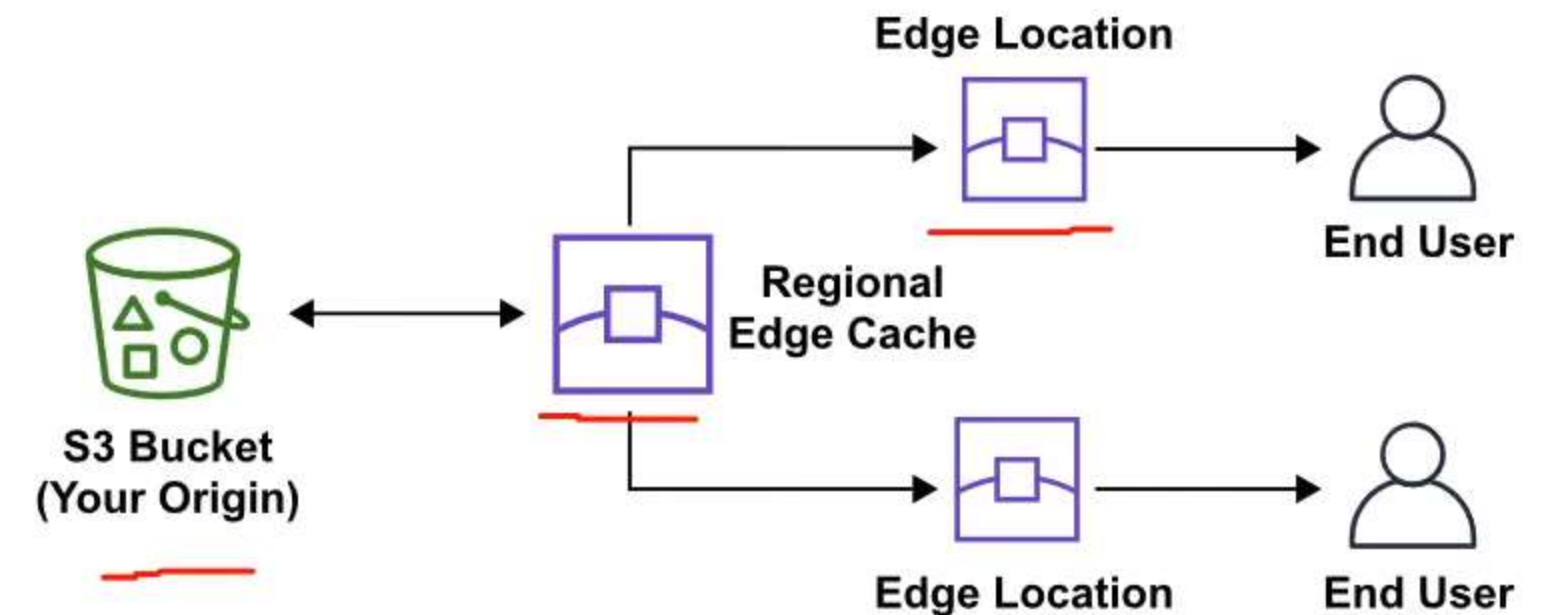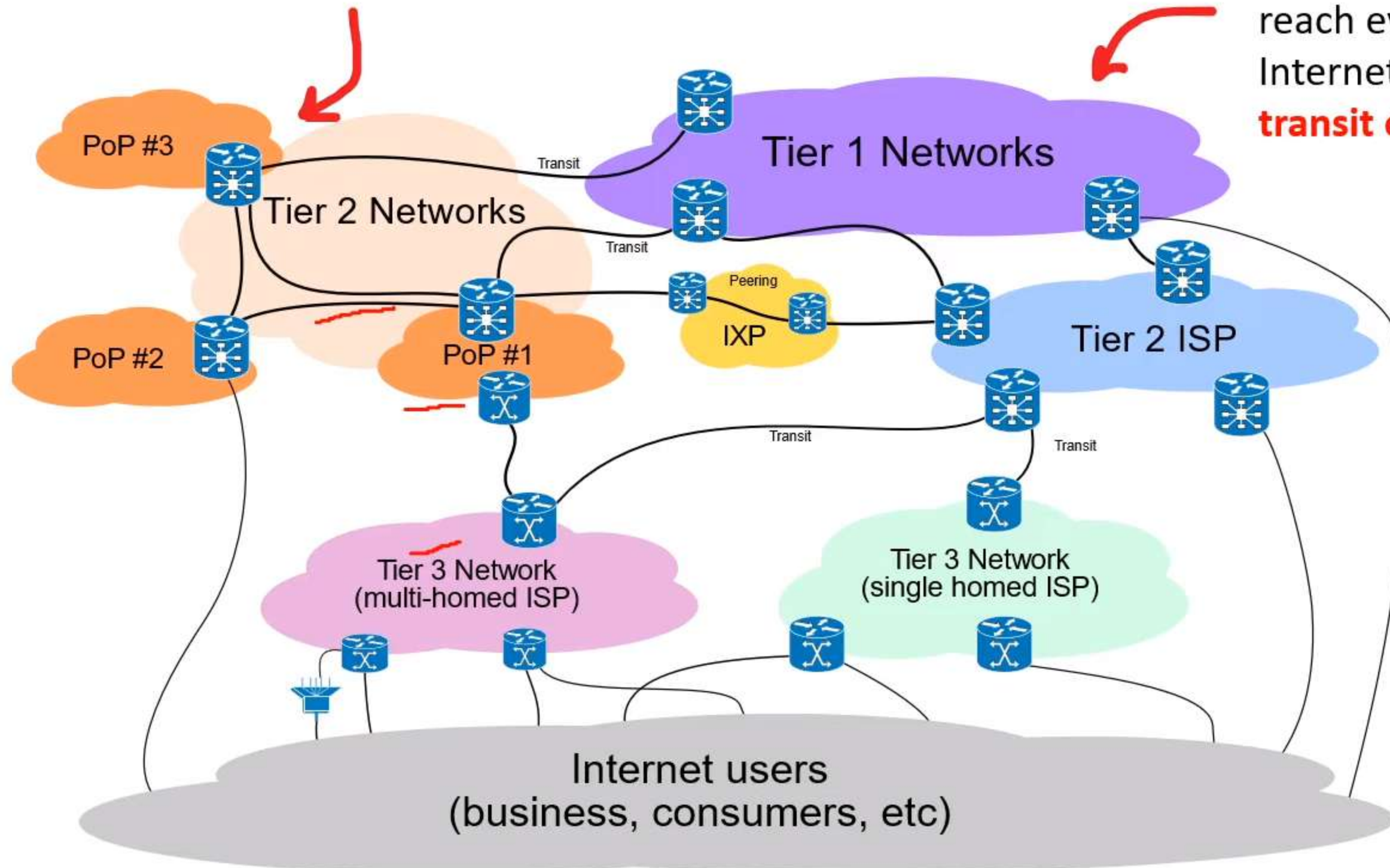
S3 Bucket (Your Origin) — Regional Edge Cache — Edge Location — End User

# Global Infrastructure – Point of Presence (PoP)

PoPs live at the edge/**intersection** of two networks

**Tier 1 network** is a network that can reach every other network on the Internet **without purchasing IP transit or paying for peering**.

AWS Availability Zones are all redundantly connected to multiple **tier-1 transit providers**



PoP #3

PoP #2

PoP #1

Tier 2 Networks

Transit

Transit

Tier 1 Networks

Peering

IXP

Tier 2 ISP

Transit

Transit

Tier 3 Network
(multi-homed ISP)

Tier 3 Network
(single homed ISP)

Internet users
(business, consumers, etc)

# Global Infrastructure – Point of Presence (PoP)

The following AWS Services use PoPs **for content delivery or expediated upload.**

**Amazon CloudFront** is a **Content Delivery Network (CDN) service** that:
- You point your website to CloudFront so that it will route requests to nearest Edge Location cache
- allows you to choose an **origin** (such as a web-server or storage) that will be source of cached
- caches the contents of what origin would returned to various Edge Locations around the world

**Amazon S3 Transfer Acceleration** allows you to generate a special URL that can be used by end users to upload files to a nearby Edge Location. Once a file is uploaded to an Edge Location, it can move much faster within the AWS Network to reach S3.

**AWS Global Accelerator** can find the optimal path from the end user to your web-servers. Global Accelerator are deployed within Edge Locations so you send user traffic to an Edge Location instead of directly to your web-application.
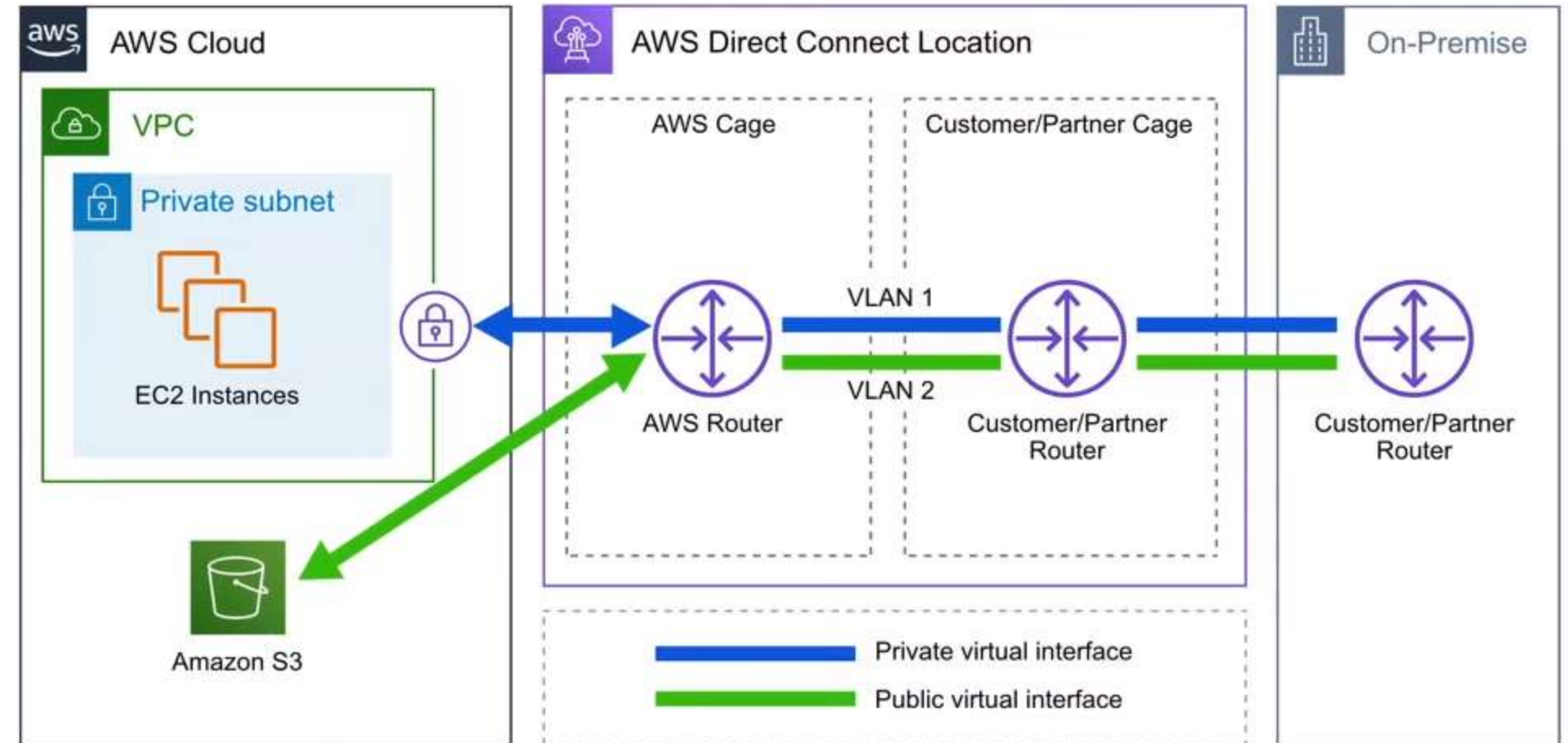
# AWS Direct Connect

AWS Direct Connect is a **private/dedicated connection between your datacenter, office, co-location and AWS**.

**Direct Connect** has two **very-fast network** connection options:

1. Lower Bandwidth **50MBps-500MBps**
2. Higher Bandwidth **1GBps** or **10GBps**

A co-location (aka carrier-hotel) is a data center where equipment, space, and bandwidth are available for rental to retail customers



✓ Helps **reduce network costs** and **increase bandwidth throughput**. (great for high traffic networks)

✓ Provides a **more consistent network experience** than a typical internet-based connection. (reliable and secure)

# Global Infrastructure – Direct Connect Locations

**Direct Connect Locations** are **trusted partnered datacenters** that you can establish ==a dedicated high speed, low-latency connection from your on-premise to AWS.==

**A partnered datacenter in Toronto** 🇨🇦



**Allied Data Centres**
250 Front Street West Toronto

You would use the ==**AWS Direct Connect**== service to order and establish a connection

# Global Infrastructure – Local Zones

**Local Zones** are datacenters located very close to a densely populated area to provide single-digit millisecond low latency performance (eg. 7ms) for that area.

- **Los Angeles, California** was the first Local Zone to be deployed
  - It is a logical extension of the US-West Region
  - The Identifier looks like the following: **us-west-2-lax-1a**
- Only specific AWS Services have been made available
  - EC2 Instance Types (T3, C5, R5, R5d, I3en, G4)
  - EBS (io1 and gp2)
  - Amazon FSx
  - Application Load Balancer
  - Amazon VPC

The purpose of Local Zone is the support highly-demanding applications sensitive to latencies:
- Media & Entertainment
- Electronic Design Automation
- Ad-Tech
- Machine Learning

To use Local Zones you need to Opt-In

# Global Infrastructure – Wavelength Zones

AWS Wavelength Zones allows for **edge-computing on 5G Networks**. Applications will have **ultra-low latency** being as close as possible to the users
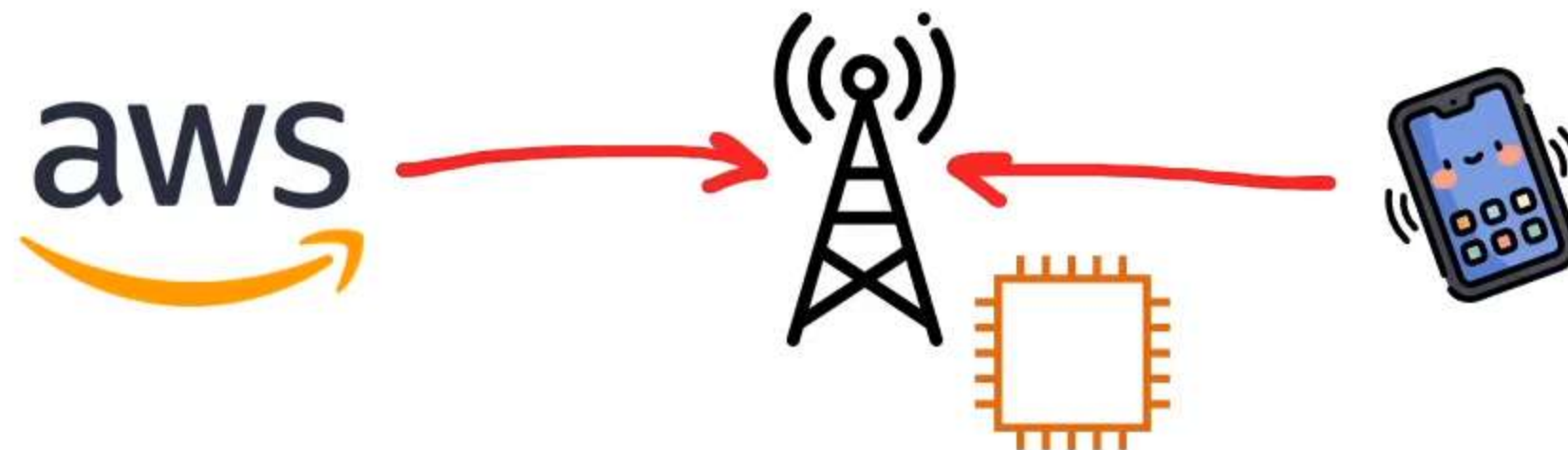
AWS has partnered with various Telecom companies to utilize their 5G networks

You create a Subnet tied to a Wavelength Zone and then you can launch Virtual Machines (VMs) to the edge of the targeted 5G Networks.

# Global Infrastructure – Data Residency

**What is Data Residency?**
The physical or geographic location of where an organization or cloud resources reside.

**What is Compliance Boundaries?**
A regulatory compliance (legal requirement) by a government or organization that describes where data and cloud resources are allowed to reside

**What is Data Sovereignty?**
Data Sovereignty is the jurisdictional control or legal authority that can be asserted over data because it's physical location is within jurisdictional boundaries

For workloads that need to meet compliance boundaries strictly defining the data residency of data and cloud resources in AWS you can use:

**AWS Config** is a Policy as Code service.
You can create rules to continuous check AWS resources configuration. If they deviate from your expectations you are alerted or AWS Config can in some cases auto-remediate.

**IAM Policies** can be written explicitly deny access to specific AWS Regions. A **Service Control Policy** (SCP) are permissions applied organization wide.

**AWS Outposts** is physical rack of servers that you can put in your data center. Your data will reside whenever the Outpost Physically resides

# Global Infrastructure – AWS for government

**What is Public Sector?**

Public sectors include public goods and governmental services such as:

- military
- law enforcement
- infrastructure
- public transit
- public education
- health care
- the government itself

AWS can be utilized by public sector or organizations developing cloud workloads for the public sector.

AWS achieves this by meeting **regulatory compliance programs** along with specific governance and security controls

AWS has special regions for US regulation called **GovCloud**