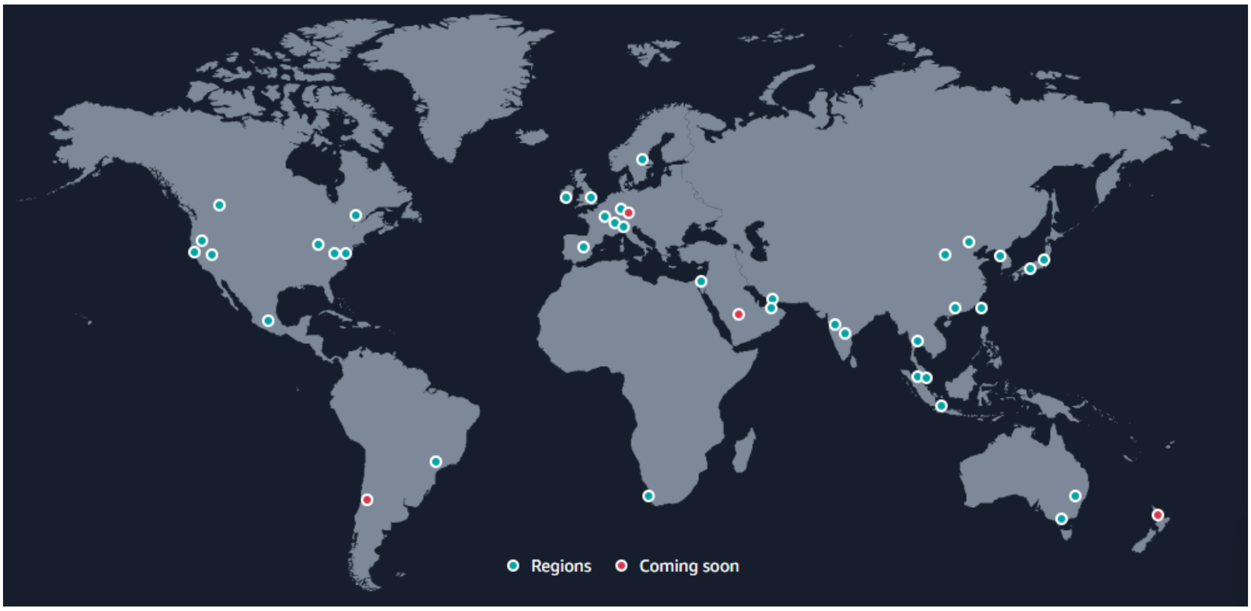
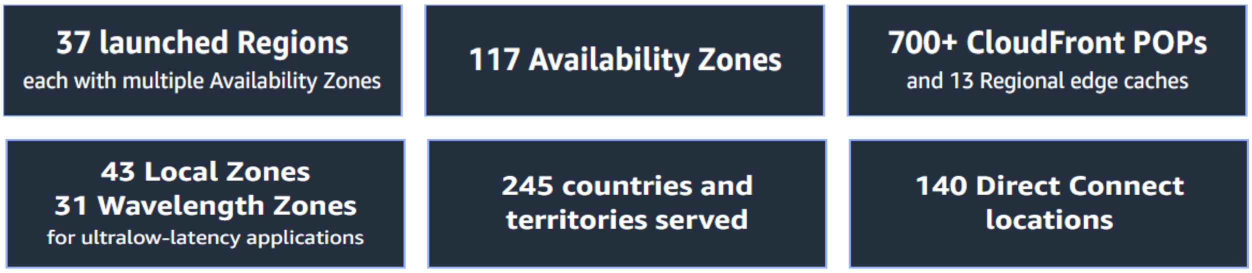
AWS Global Infrastructure



Currently AWS have:



37 launched Regions
each with multiple Availability Zones

117 Availability Zones

700+ CloudFront POPs
and 13 Regional edge caches

43 Local Zones
31 Wavelength Zones
for ultralow-latency applications

245 countries and territories served

140 Direct Connect locations

What is AWS and AWS Global Infrastructure?

- **AWS (Amazon Web Services)** is a cloud platform that provides computing, storage, networking, and many other services over the internet on a pay-as-you-go basis. It allows businesses and individuals to build and run applications without needing physical servers.
- **AWS Global Infrastructure** refers to the global network of data centers, servers, and connectivity (like Regions, Availability Zones, and Edge Locations) that power these cloud services.
- This infrastructure is designed for high availability, scalability, and low latency. In short, AWS provides the tools, and its infrastructure is the strong foundation that makes those tools reliable and globally accessible.

Let's understand the concepts…

AWS Regions:

- AWS services are hosted in multiple locations world-wide. This include above given components.

Let's see…

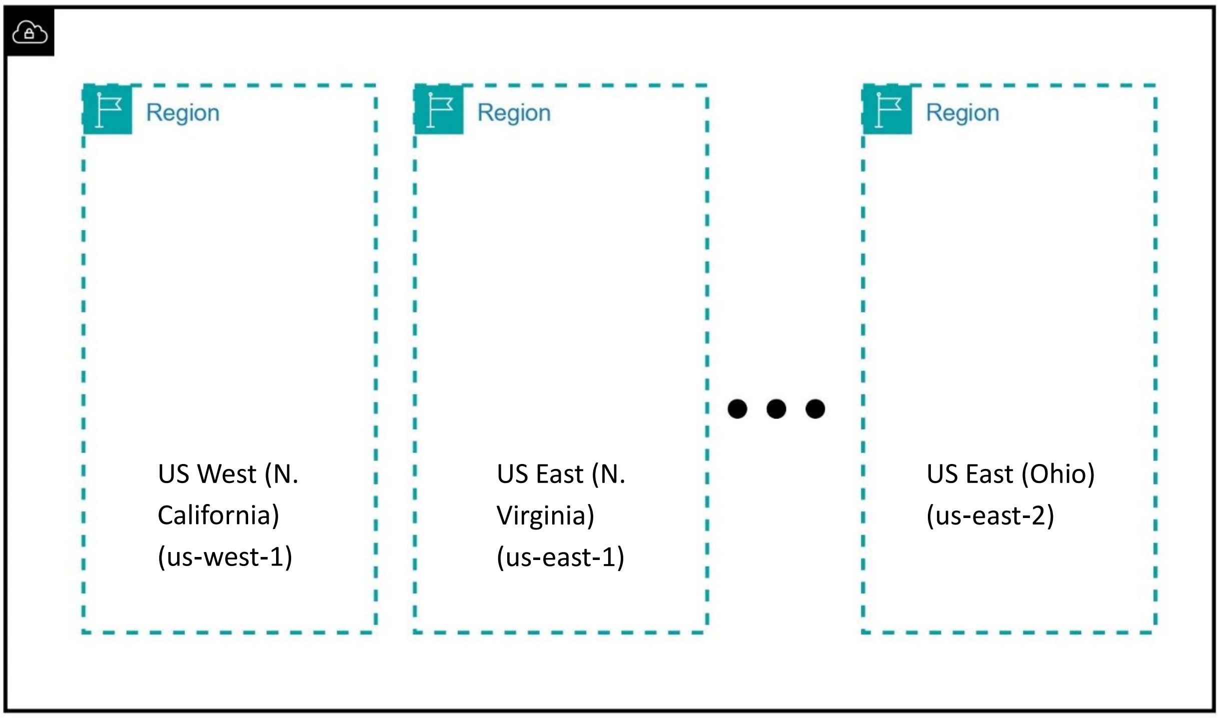
Region: Each Region is a separate geographic area.

- **37** launched regions
- Each region contains **availability zones** (minimum 3)
- **Region is isolated** from others **for fault tolerance and stability.**
- Most AWS services are **Regional**

→ *resources are tied to the Region where you create them.*

- You can copy some AWS resources or infrastructure (like S3 buckets or RDS databases) to other Regions, but you must **do it yourself** — AWS does **not do it automatically**.

Let's take an example of USA country, AWS has total 6 regions in the whole USA. See diagram, there are different regions across the country, isolated from each other.


Region
US West (N. California) (us-west-1)
Region
US East (N. Virginia) (us-east-1)
Region
US East (Ohio) (us-east-2)

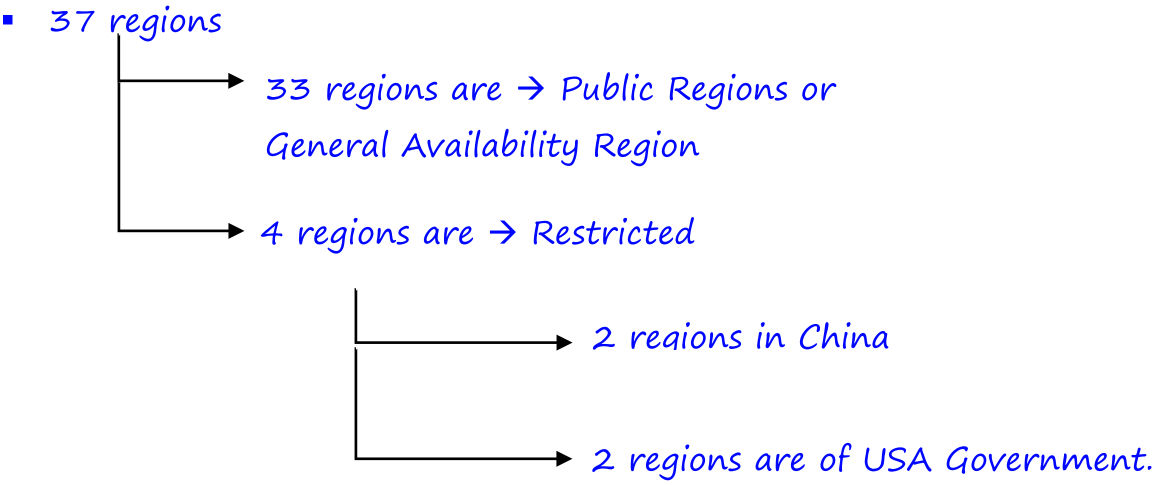
** Note: AWS has 6 Regions in the USA. This diagram shows only 3 for illustration.*

Public Regions or General Availability (GA) Regions

Geographic Region Grouping	Sr. No.	Location	Region Name (code)	Availability Zones
Asia Pacific and china	1	Hong Kong	🇭🇰 Asia Pacific (Hong Kong SAR) (ap-east-1)	3
	2	India	🇮🇳 Asia Pacific (Hyderabad) (ap-south-2)	3
	3	Indonesia	🇮🇩 Asia Pacific (Jakarta) (ap-southeast-3)	3
	4	Malaysia	🇲🇾 Asia Pacific (Malaysia) (ap-southeast-5)	3
	5	Australia	🇦🇺 Asia Pacific (Melbourne) (ap-southeast-4)	3
	6	India	🇮🇳 Asia Pacific (Mumbai) (ap-south-1)	3
	7	Japan	🇯🇵 Asia Pacific (Osaka) (ap-northeast-3)	3
	8	S Korea	🇰🇷 Asia Pacific (Seoul) (ap-northeast-2)	4
	9	Singapore	🇸🇬 Asia Pacific (Singapore) (ap-southeast-1)	3
	10	Australia	🇦🇺 Asia Pacific (Sydney) (ap-southeast-2)	3
	11	Taiwan	🇹🇼 Asia Pacific (Taipei) (ap-east-2)	3
	12	Thailand	🇹🇭 Asia Pacific (Thailand) (ap-southeast-7)	3
	13	Japan	🇯🇵 Asia Pacific (Tokyo) (ap-northeast-1)	4
North America	14	Canada	🇨🇦 Canada (Central) (ca-central-1)	3
	15	Canada	🇨🇦 Canada West (Calgary) (ca-west-1)	3
	16	Mexico	🇲🇽 Mexico (Central) (mx-central-1)	3
	17	USA	🇺🇸 US West (N. California) (us-west-1)	3
	18	USA	🇺🇸 US East (N. Virginia) (us-east-1)	6
	19	USA	🇺🇸 US East (Ohio) (us-east-2)	3
	20	USA	🇺🇸 US West (Oregon) (us-west-2)	4
South America	21	Brazil	🇧🇷 South America (São Paulo) (sa-east-1)	3
Europe / Middle East / Africa	22	South Africa	🇿🇦 Africa (Cape Town) (af-south-1)	3
	23	Sweden	🇸🇪 Europe (Stockholm) (eu-north-1)	3
	24	Germany	🇩🇪 Europe (Frankfurt) (eu-central-1)	3
	25	Ireland	🇮🇪 Europe (Ireland) (eu-west-1)	3
	26	UK	🇬🇧 Europe (London) (eu-west-2)	3
	27	Italy	🇮🇹 Europe (Milan) (eu-south-1)	3
	28	France	🇫🇷 Europe (Paris) (eu-west-3)	3
	29	Spain	🇪🇸 Europe (Spain) (eu-south-2)	3
	30	Switzerland	🇨🇭 Europe (Zurich) (eu-central-2)	3
	31	Israel	🇮🇱 Israel (Tel Aviv) (il-central-1)	3
	32	UAE	🇦🇪 Middle East (UAE) (me-central-1)	3
	33	Bahrain	🇧🇭 Middle East (Bahrain) (me-south-1)	3

Restricted / Permission-based Regions					
Asia Pacific and china	34	China	🇨🇳	China (Beijing) (cn-north-1)	3
	35	China	🇨🇳	China (Ningxia) (cn-northwest-1)	3
North America	36	USA	🇺🇸	AWS GovCloud (US-East) (us-gov-east-1)	3
	37	USA	🇺🇸	AWS GovCloud (US-West) (us-gov-west-1)	3
Total Regions	✓ 37		Total Availability Zones		✓ 117

- As of today (June 14th, 2025), AWS has 37 Regions and 117 Availability Zones worldwide; and it's continuously expanding, with new Regions and availability zones.



- 37 regions
  - 33 regions are → Public Regions or General Availability Region
  - 4 regions are → Restricted
    - 2 reaions in China
    - 2 regions are of USA Government.

Availability Zones AZs

- **Availability Zones** are isolated locations within each Region.
- Each Region has multiple, independent locations known as Availability Zones, minimum 3 AZs are there per region.
- Availability Zones

  Connected → low-latency, high-bandwidth, highly-redundant networking, over dedicated metro **fiber cable.**

- Each **Availability Zone (AZ)** = **one or more** separate **data centers**

  Why? = **In case of failures (like fire, flooding, tornado)** affect only one AZ
  Other AZs will be working.

- Each **Availability Zone (AZ)** has → Redundant (- Power, - networking, - connectivity)
- Some AWS services use **zonal resources:** → A zonal resource is specific to the Availability Zone in which you create it.
- If you want to keep your **application continue running, without downtime and interruption**, it is suggested to **deploy** your application **in multiple Availability Zones**, so that your application remains available even if one Availability Zone fails.
- Let's see an example, in India there are 2 regions, **Mumbai** and **Hyderabad**. In below given diagram, it's virtual representation of AWS Mumbai region. Mumbai region has 3 AZs, **ap-south-1a, ap-south-1b** and **ap-south-1c.**

E.g. **ap-south-1a** → Asia Pacific (**ap-south-1** is region code for Mumbai)

ap: Asia Pacific

south: **South** subregion (specifically: India)

1: **First** AWS Region in this area

a: A specific Availability Zone (AZ) in that Region (mapped per account)

** You and someone else may see different physical locations mapped to 1a, 1b, or 1c — AWS randomly maps AZ names per account to ensure load balancing.*

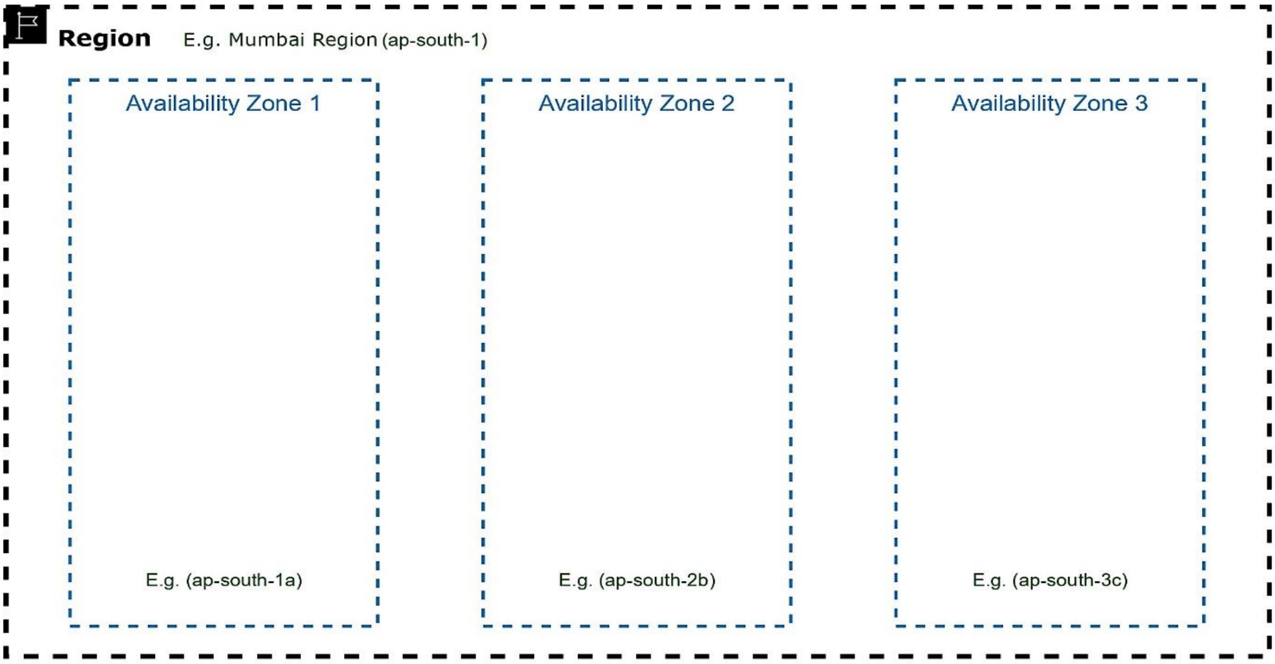


Diagram 1: Region

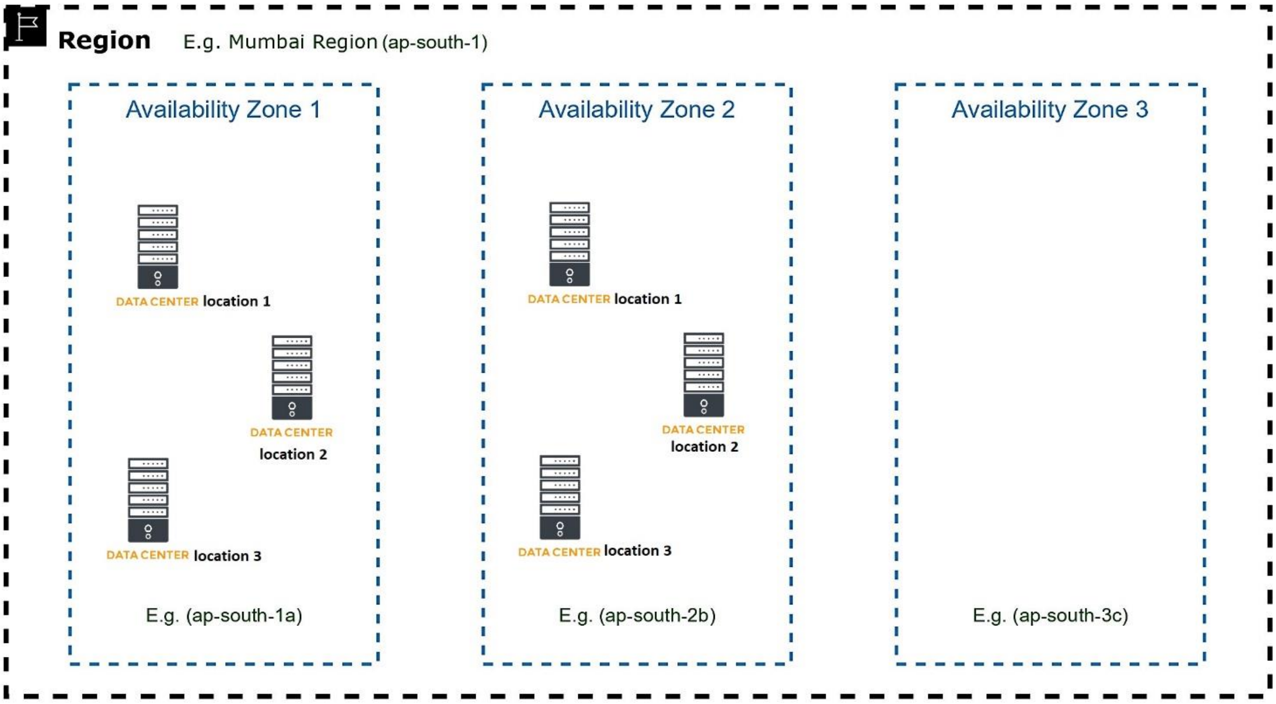


Diagram 2: Each Availability Zone (AZ) has one or more separate data centers.

● Note: **AWS does *not* publicly disclose** the exact physical locations of its data centers or Availability Zones (AZs).

Why?

- *Security reasons*
- *Operational confidentiality*
- *Compliance and risk control*

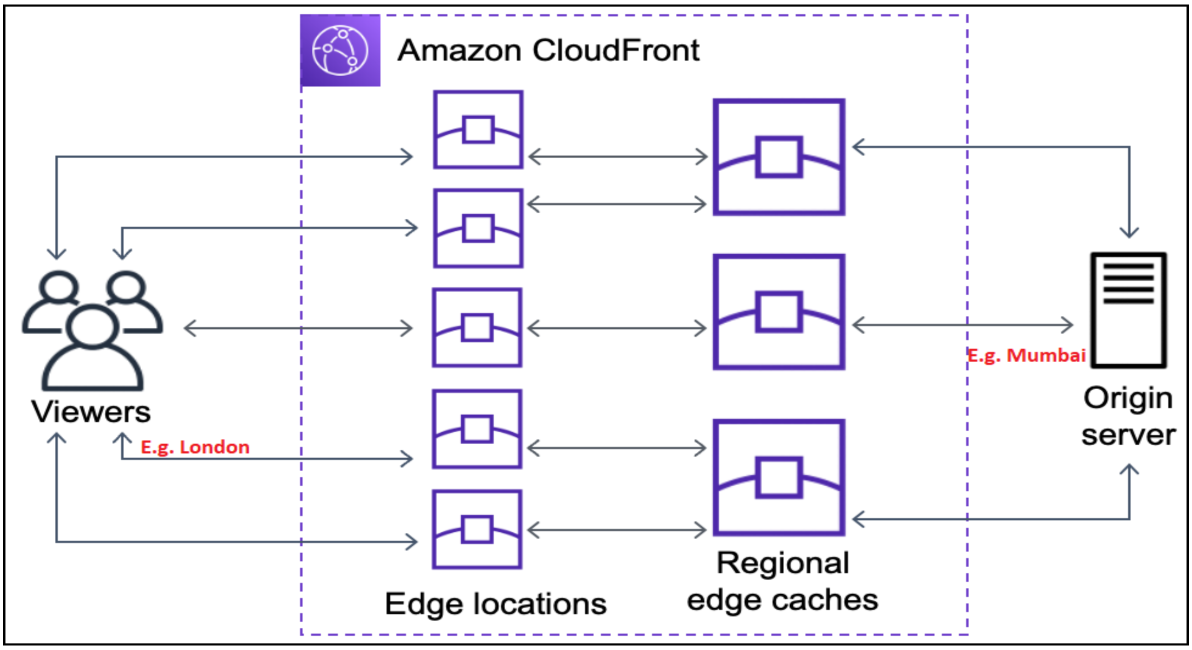
700 CloudFront POPs and 13 regional edge caches

- **CloudFront:** CloudFront is CDN (Content Delivery Network) service of AWS.
  - it creates cache of content and stores it in POPs near to user so user will get fastest delivery of content.
  - it helps to reduce load on origin server.
  - 2 level caching system (regional cache, edge location cache POPs)

- **POPs:** Point of Presence
  - Located in major cities worldwide (Delhi, New York, London, etc)
  - Serve user requests fast, close to user
  - Used by CloudFront, Route 53, AWS Shield etc.
  - 700+ worldwide
  - Also called **Edge location**
  - Can't select manually, AWS will select automatically.
  - It's a physical location with CloudFront servers, but not necessary every time it will be AWS own data center, it can be anywhere.
  - Only **used for fast content delivery**, nothing else

- **13 regional edge caches:**
  - globally total 13 regional edge caches
  - intermediate level caching layer
  - can store large cache memory

Let's understand the concept, see diagram carefully:



- Suppose an image is stored in an S3 bucket in the Mumbai region.
  Now, a user from London requests that image.
- **Step 1: Edge Location (POP)**
  - The request first goes to the nearest Edge Location (Point of Presence) in or near London.
  - If the image is already cached there, it is immediately delivered to the user.
  - This gives ultra-fast response.
- **Step 2**: Regional Edge Cache
  - If the image is not found at the Edge Location, the request is forwarded to the nearest Regional Edge Cache.
  - These caches have larger memory and store content longer.
  - If found here, it is delivered to the user.
- **Step 3:** Origin Server (S3 Mumbai)
  - If the image is not in the Regional Edge Cache, the request finally goes to the origin server in Mumbai (S3).
  - The image is fetched and delivered to the user.
- **Smart Caching**
  - After that, CloudFront automatically stores the image at the Edge Location or Regional Edge Cache.
  - So next time, if any other user (even from a different location) requests the same image, it will be delivered faster from the cache.

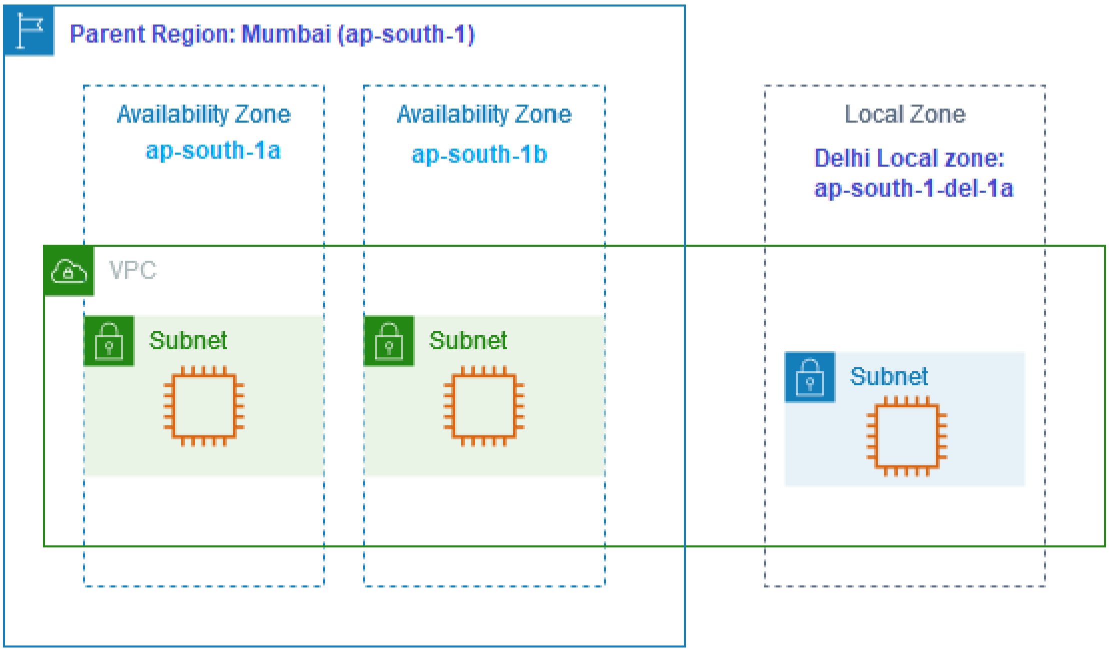
You can check all the **cities** from this
https://aws.amazon.com/cloudfront/features/?ams%23interactive-card-vertical%23pattern-data.filter=%257B%2522filters%2522%253A%255B%255D%257D

(P.T.O.)

43 Local Zones and 31 Wavelength zones:

- **Local zones:**
  - AWS Local Zones places compute, storage, database, and other select AWS resources close to large population and industry centers. You can use Local Zones to provide your users with low-latency access to your applications.
  - are not full-fledged data centers like the 117 AZs.
  - **not a part of 117 availability zones (AZs), they are separate zones, total 43 as of now.**
  - Local Zones = Mini data centers placed in cities **outside the main AWS Region**, but controlled by the Region
  - Associated with a parent Region (e.g., *San Diego (us-west-2-san-1a)* is tied to *Oregon (us-west-2)*).
  - Used for latency-sensitive applications like real-time gaming, media & entertainment, hybrid migrations, ML inference.
  - Services supported: EC2, EBS, VPC, ELB, FSx, ECS/EKS, etc. (though fewer than full regions).

Let's see an example with diagram



Imagine you're running an online video editing service , hosted in AWS Mumbai Region (ap-south-1). Users upload videos, edit them, and download them, all in real-time.

→ Everything works well until…
Your service becomes popular in North India (Delhi, Noida, Gurugram). Now users from Delhi start complaining:

☹ *"It's slow when I upload or preview a video!"*

That's because your servers are only in Mumbai, which is ~1,400 km away. Network latency slows things down for North Indian users.

→ So here local zones will help. Use AWS Local Zone in Delhi!
AWS has a Local Zone in Delhi (ap-south-1-del-1a), which is:

- **Physically** located **in Delhi**
- Logically **connected to Mumbai Region** (ap-south-1)
- Allows you to run compute (like EC2), storage (EBS), containers (ECS/EKS), etc.
- Accessed using the same Mumbai Region in AWS Console; just opt-in to use it.

In the Diagram:
You see a VPC (virtual private cloud) created in Mumbai Region.
It has two subnets in AZs (ap-south-1a, ap-south-1b) – traditional Mumbai infrastructure.
And a third subnet inside Delhi Local Zone (ap-south-1-del-1a) — closer to North India users.
Now you can deploy latency-sensitive apps in Delhi zone while managing it all from Mumbai! And user in Delhi will get fastest response.

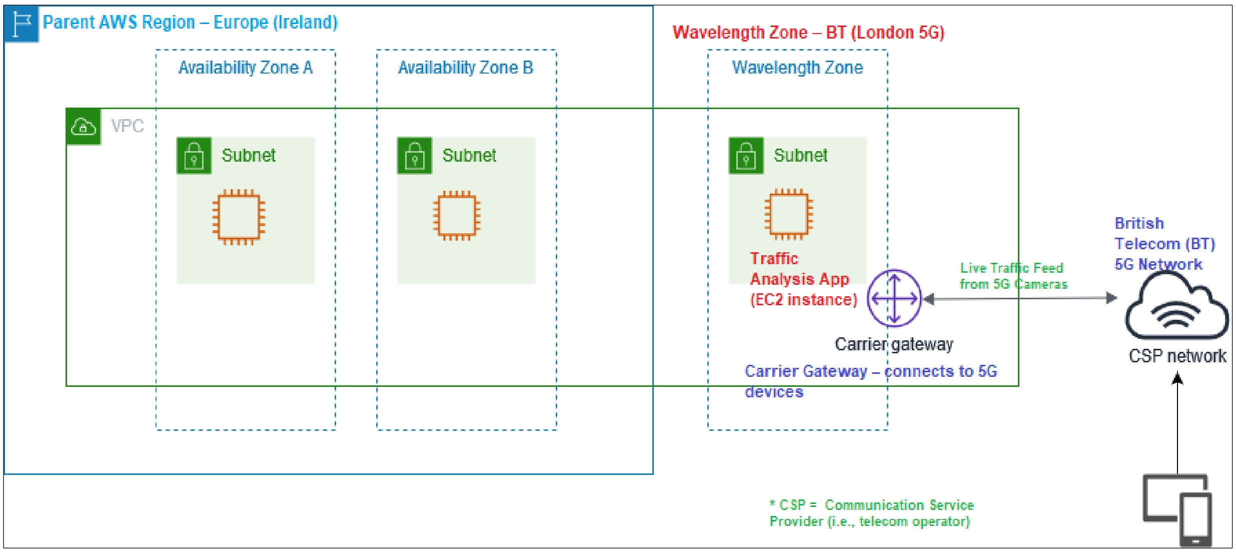
Here you will find list of all local zones
https://docs.aws.amazon.com/local-zones/latest/ug/available-local-zones.html

(P.T.O.)

- **Wavelength zones:**
  - Run applications using AWS Infrastructure **and services in AWS telco partners' data centers** to meet your low latency, data residency, and resiliency needs.
  - Same like local zone, but deployed servers inside telecom data center. (Jio, airtel, etc.)
  - services like **EC2, EBS, VPC, ECS, EKS, Load Balancers** are available within the Wavelength Zone.
  - Connected via **high-speed fiber** to a parent AWS Region for broader service access.
  - As of now 31 wavelength zones are worldwide. (Not in India)

  You can see all the wavelength zone locations here https://aws.amazon.com/wavelength/locations/

Let's see an example, see the diagram


Parent AWS Region – Europe (Ireland)
Availability Zone A
Availability Zone B
VPC
Subnet
Subnet
Wavelength Zone – BT (London 5G)
Wavelength Zone
Subnet
Traffic Analysis App (EC2 instance)
Carrier gateway
Carrier Gateway – connects to 5G devices
Live Traffic Feed from 5G Cameras
British Telecom (BT) 5G Network
CSP network
* CSP = Communication Service Provider (i.e., telecom operator)

Suppose, the London Traffic Department wants to monitor road traffic using **5G-connected cameras** across the city. They need:

- Live video analysis
- Low latency (fast response)
- Fast decisions (e.g., traffic light changes, alerts)

How **AWS Wavelength** Helps:

- 5G Cameras are installed at traffic signals and intersections.
- These cameras stream live video data over BT's 5G network.
- The **5G feed is sent to an AWS EC2 instance in the Wavelength Zone** (London BT); not far away, so very low latency.

- The EC2 instance runs a traffic analysis app that:
    - Detects congestion
    - Counts vehicles
    - Sends alerts to traffic control teams
- The analyzed data is optionally stored or backed up in Availability Zone A or B of AWS Europe (Ireland) Region for reporting later.

Why Not Use Only the Region?

- If they used just the Ireland main region, 5G camera data would travel hundreds of kilometers; too slow for real-time alerts.
- **Wavelength Zones solve** that by **bringing AWS services closer to users, at the telecom's 5G network edge** (BT in this case).

In short, Think of AWS Wavelength Zone as a mini-AWS datacentre inside your city's 5G network. It helps apps like live traffic monitoring work super-fast — without sending data far away!

(P.T.O.)

245 countries and territories served

- AWS provides access to its cloud services (like EC2, S3, Lambda, etc.) to **customers in 245 countries and territories**, **even if** there is **no AWS Region physically located** there.
- How do they deliver services to all 245?

  Through nearby Regions & AZs, Edge Locations (PoPs), Local Zones, Wavelength Zones and Global Network.

Let's see an example…

Even without a region, AWS users in:

1. Nepal
2. Bangladesh
3. Sri Lanka
4. Belize

can still use AWS through nearby regions (e.g., Mumbai, Singapore) and Edge locations.

(P.T.O.)

140 Direct Connect locations

- A **physical AWS facility or partner colocation** where you can connect your **network to AWS** using a **fiber/cable line**.

  Think like

  *You're plugging a **LAN cable directly into AWS** from your company building; instead of sending your traffic over the shared internet.*

- Why use it?
  - Private connection
  - Low latency and Consistent speed
  - Cost saving

  You can see all the Direct Connect location list here
  https://aws.amazon.com/directconnect/locations/

  Let's see an example, suppose Hospital Chain Using AWS Direct Connect
  A large hospital chain with branches across Delhi, Mumbai, and Bangalore uses a medical record system hosted on AWS.
  They want:
  - Fast, secure access to patient data
  - No delays during emergencies (latency could cost lives!)
  - Full control over data movement, not relying on public internet

  So here, the hospital partners with a colocation provider like **GPX Mumbai (a Direct Connect location).**
  They set up AWS Direct Connect between:
  - Their data center (on-premises)
  - And AWS Mumbai Region

  So that their medical record system communicates with AWS **over a private fiber line.**
  Doctors across India get **faster and more reliable** access to patient data.
  The system is HIPAA-compliant, secure, and not affected by public internet issues.
  You can think like, "It's like setting up a private express highway from the hospital's servers to AWS — no internet traffic, no noise, just fast and secure access."