AWS Global Infrastructure Overview

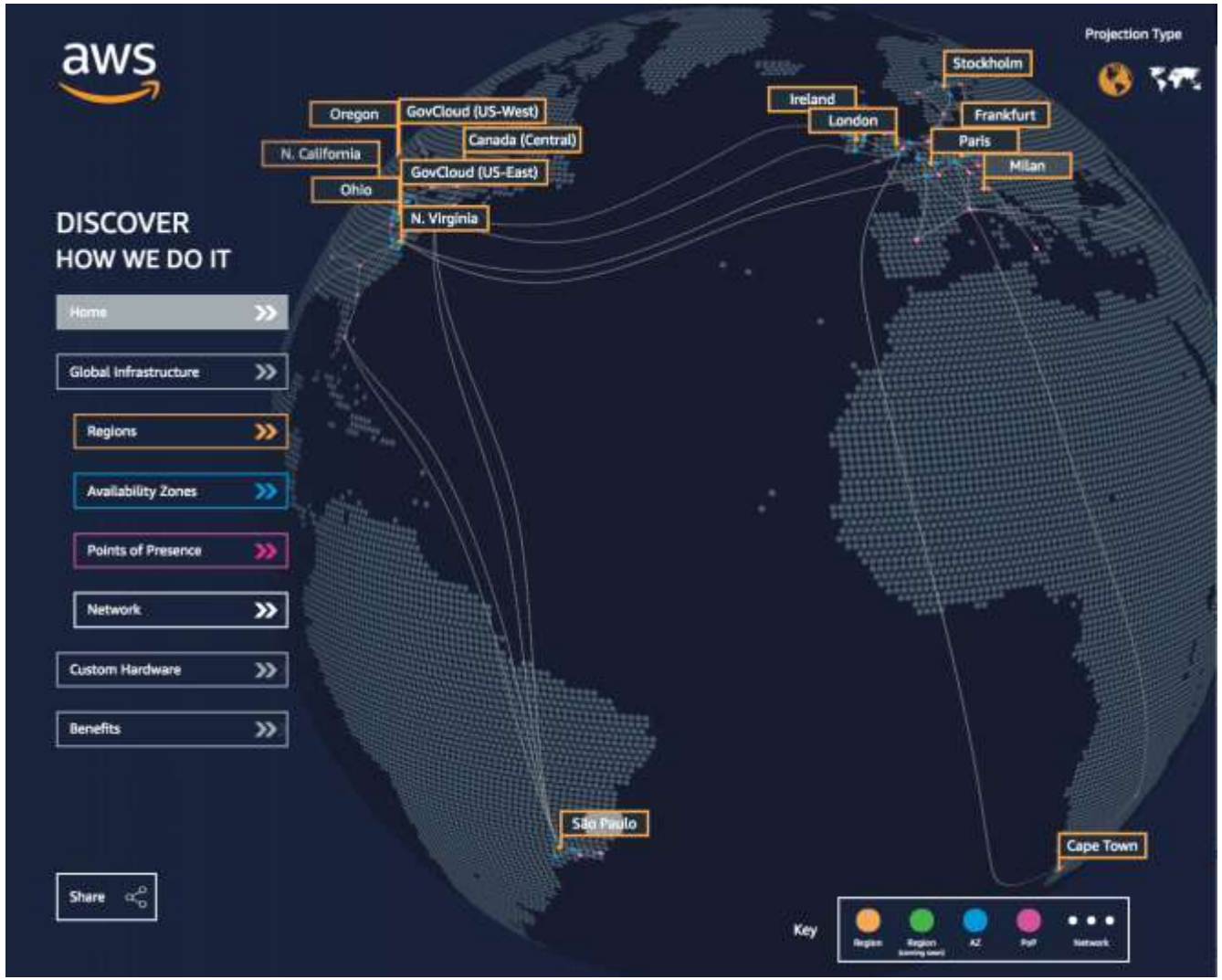
Section 1: AWS Global Infrastructure

AWS Global Infrastructure

innovate achieve lead

- The **AWS Global Infrastructure** is designed and built to deliver a **flexible**, **reliable**, **scalable**, and **secure** cloud computing environment with high-quality **global network performance**.
- This map from https://infrastructure.aws shows the current **AWS Regions** and more that are coming soon.

Educator-Led Demo: AWS Global Infrastructure Details



AWS Infrastructure

- Region
- Availability Zones
- Point of Presence
  - Edge locations
  - Edge Caches

AWS Regions

- An **AWS Region** is a geographical area.
  - Data replication across Regions is controlled by you.
  - Communication between Regions uses AWS backbone network infrastructure.
- Each Region provides full redundancy and connectivity to the network.
- A Region typically consists of two or more **Availability Zones**.

Example: London Region

Selecting a Region

Determine the right Region for your services, applications, and data based on these factors

Data governance, legal requirements

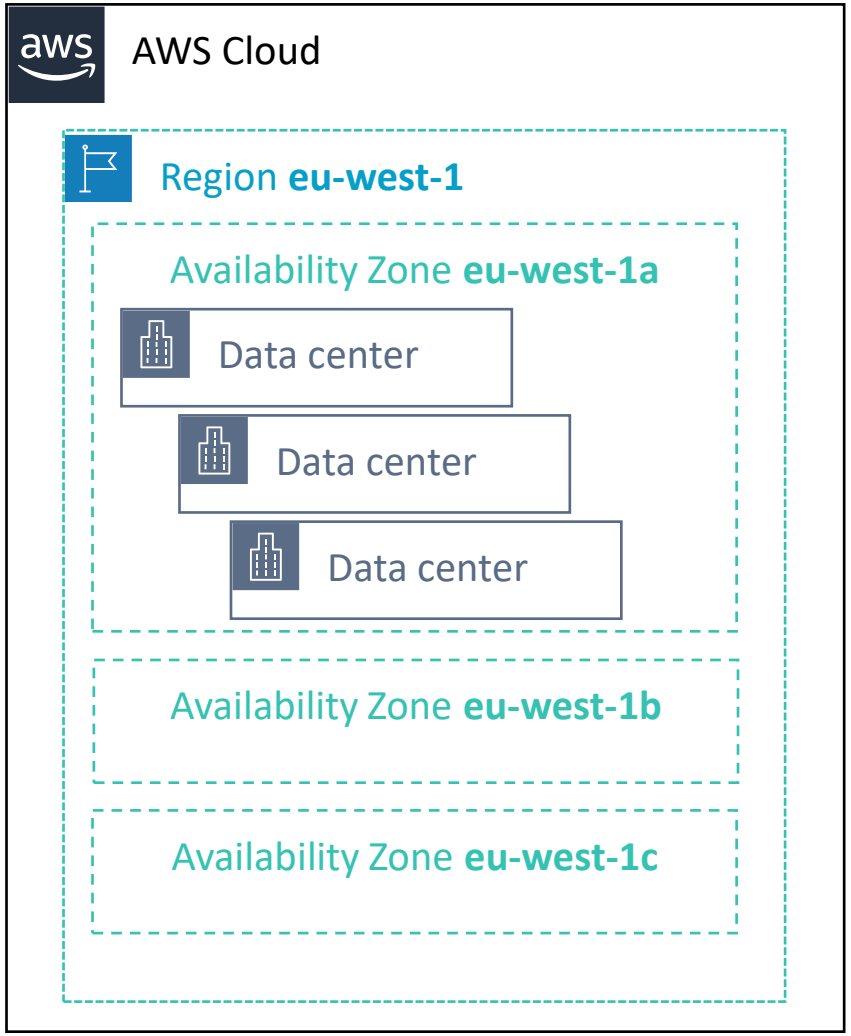
Proximity to customers (latency)

Services available within the Region

Costs (vary by Region)

Availability Zones

- Each **Region** has multiple Availability Zones.
- Each **Availability Zone** is a fully isolated partition of the AWS infrastructure.
  - There are currently 99 Availability Zones worldwide
  - Availability Zones consist of discrete **data centers**
  - They are designed for fault isolation
  - They are interconnected with other Availability Zones by using high-speed private networking
  - You choose your Availability Zones.
  - **AWS recommends replicating data and resources across Availability Zones** for resiliency.

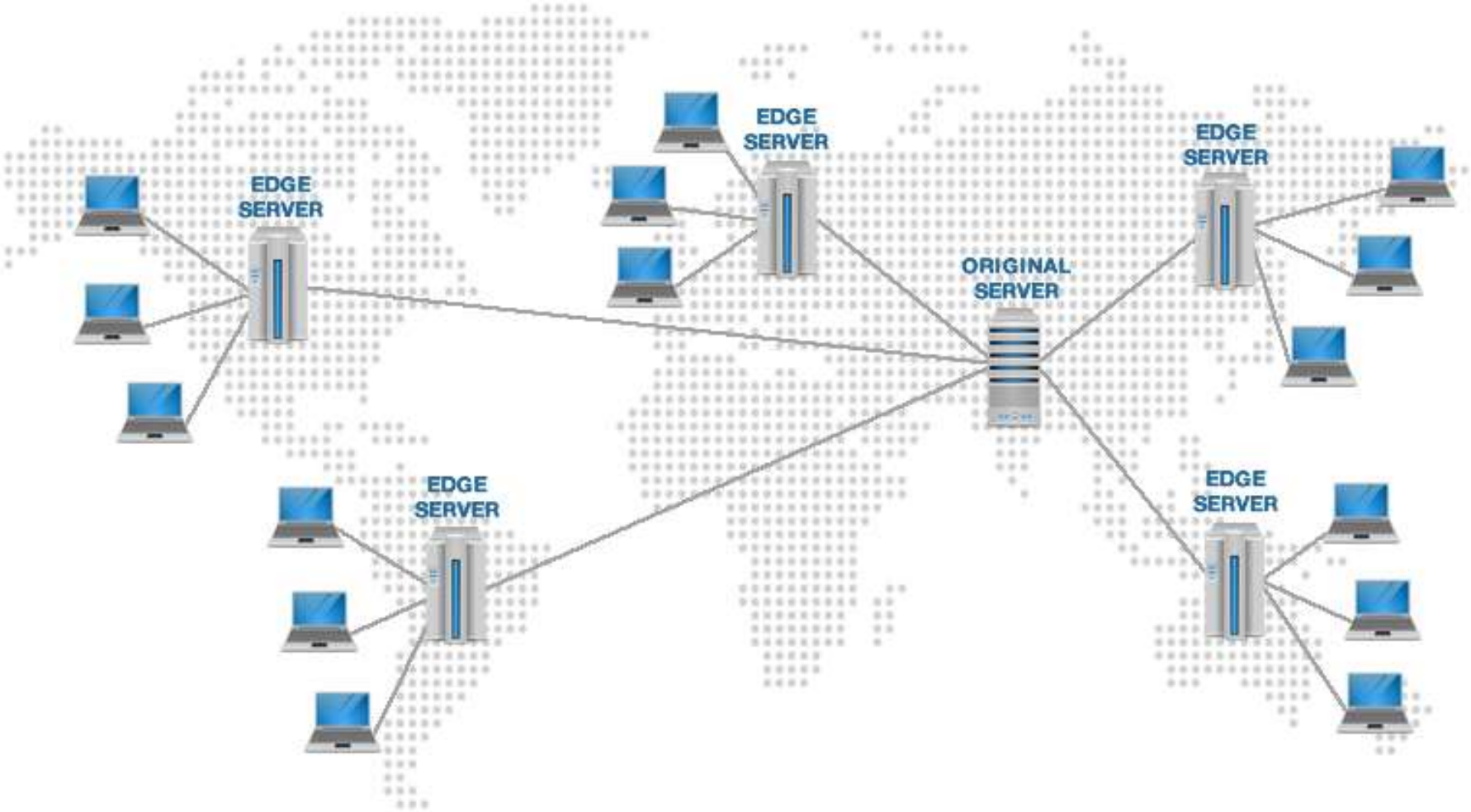


AWS data centers

- AWS data centers are designed for security.
- Data centers are where the data resides and data processing occurs.
- Each data center has redundant power, networking, and connectivity, and is housed in a separate facility.
- A data center typically has 50,000 to 80,000 physical servers.

Content Delivery Network (Traditional)

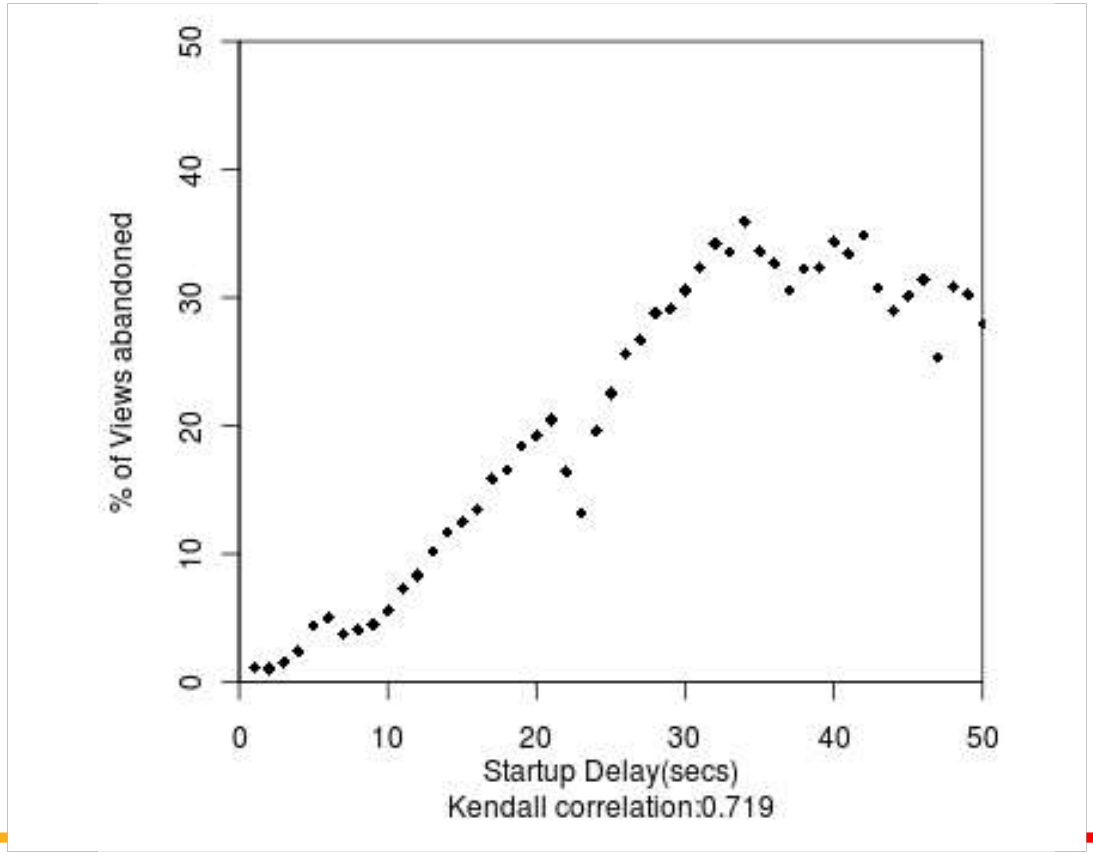
Content Delivery Network (Traditional)



Server placement

Why speed matters

- Impact on user experience
  - Users navigating away from pages
  - Video startup delay



Why speed matters

- Impact on user experience
  - Users navigating away from pages
  - Video startup delay
- Impact on revenue
  - Amazon: increased revenue 1% for every 100ms reduction in PLT
  - Shopzilla:12% increase in revenue by reducing PLT from 6 seconds to 1.2 seconds
- Ping from BOS to LAX: ~100ms

Strawman solution: Web caches

- ISP uses a middlebox that caches Web content
  - Better performance – content is closer to users
  - Lower cost – content traverses network boundary once
  - Does this solve the problem?

- No!
  - Size of all Web content is too large
    - Zipf distribution limits cache hit rate
  - Web content is **dynamic** and **customized**
    - Can't cache banking content
    - What does it mean to cache search results?

What is a CDN?

- Content Delivery Network
  - Also sometimes called Content Distribution Network
  - At least half of the world's bits are delivered by a CDN
    - Probably closer to 80/90%

- Primary Goals
  - Create replicas of content throughout the Internet
  - Ensure that replicas are always available
  - Directly clients to replicas that will give good performance

Key Components of a CDN

- Distributed servers
  - Usually located inside of other ISPs
  - Often located in IXPs (coming up next)
- High-speed network connecting them
- Clients (eyeballs)
  - Can be located anywhere in the world
  - They want fast Web performance
- Glue
  - Something that binds clients to “nearby” replica servers

Examples of CDNs

- Akamai
  - 147K+ servers, 1200+ networks, 650+ cities, 92 countries
- Limelight
  - Well provisioned delivery centers, interconnected via a private fiber-optic connected to 700+ access networks
- Edgecast
  - 30+ PoPs, 5 continents, 2000+ direct connections
- Others
  - Google, Facebook, AWS, AT&T, Level3, Brokers

Inside a CDN

- Servers are deployed in clusters for reliability
  - Some may be offline
    - Could be due to failure
    - Also could be “suspended” (e.g., to save power or for upgrade)
- Could be multiple clusters per location (e.g., in multiple racks)
- Server locations
  - Well-connected points of presence (PoPs)
  - Inside of ISPs

Mapping clients to servers

- CDNs need a way to send clients to the “best” server
  - The best server can change over time
  - And this depends on client location, network conditions, server load, …
  - What existing technology can we use for this?

- DNS-based redirection
  - Clients request www.foo.com
  - DNS server directs client to one or more IPs based on request IP
  - Use short TTL to limit the effect of caching

CDN redirection example

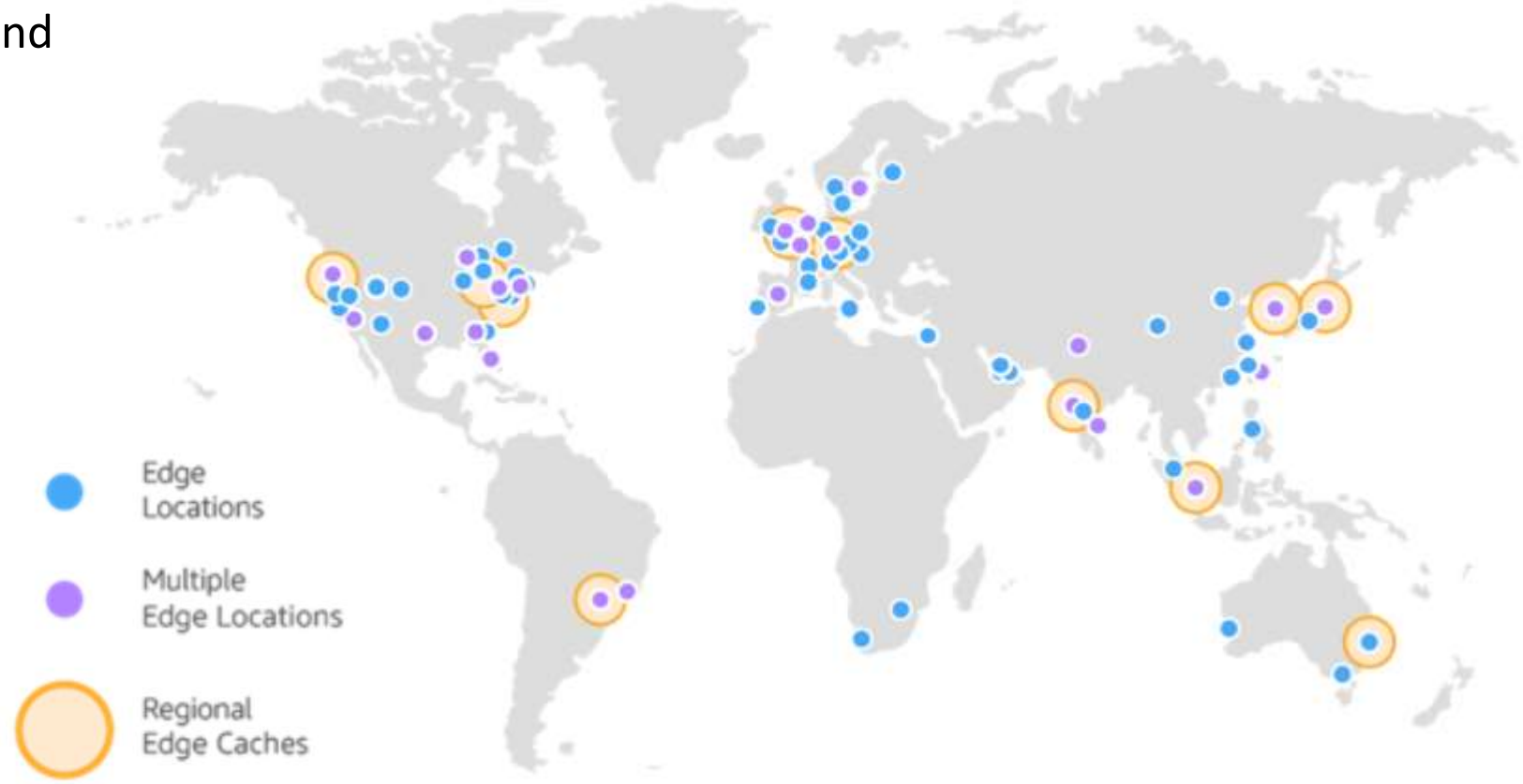

```
choffnes$ dig www.fox.com


;; ANSWER SECTION:
www.fox.com.                    510       IN        CNAME     www.fox-rma.com.edgesuite.net.
www.fox-rma.com.edgesuite.net. 5139 IN  CNAME     a2047.w7.akamai.net.
a2047.w7.akamai.net.            4         IN        A         23.62.96.128
a2047.w7.akamai.net.            4         IN        A         23.62.96.144
a2047.w7.akamai.net.            4         IN        A         23.62.96.193
a2047.w7.akamai.net.            4         IN        A         23.62.96.162
a2047.w7.akamai.net.            4         IN        A         23.62.96.185
a2047.w7.akamai.net.            4         IN        A         23.62.96.154
a2047.w7.akamai.net.            4         IN        A         23.62.96.169
a2047.w7.akamai.net.            4         IN        A         23.62.96.152
a2047.w7.akamai.net.            4         IN        A         23.62.96.186
```

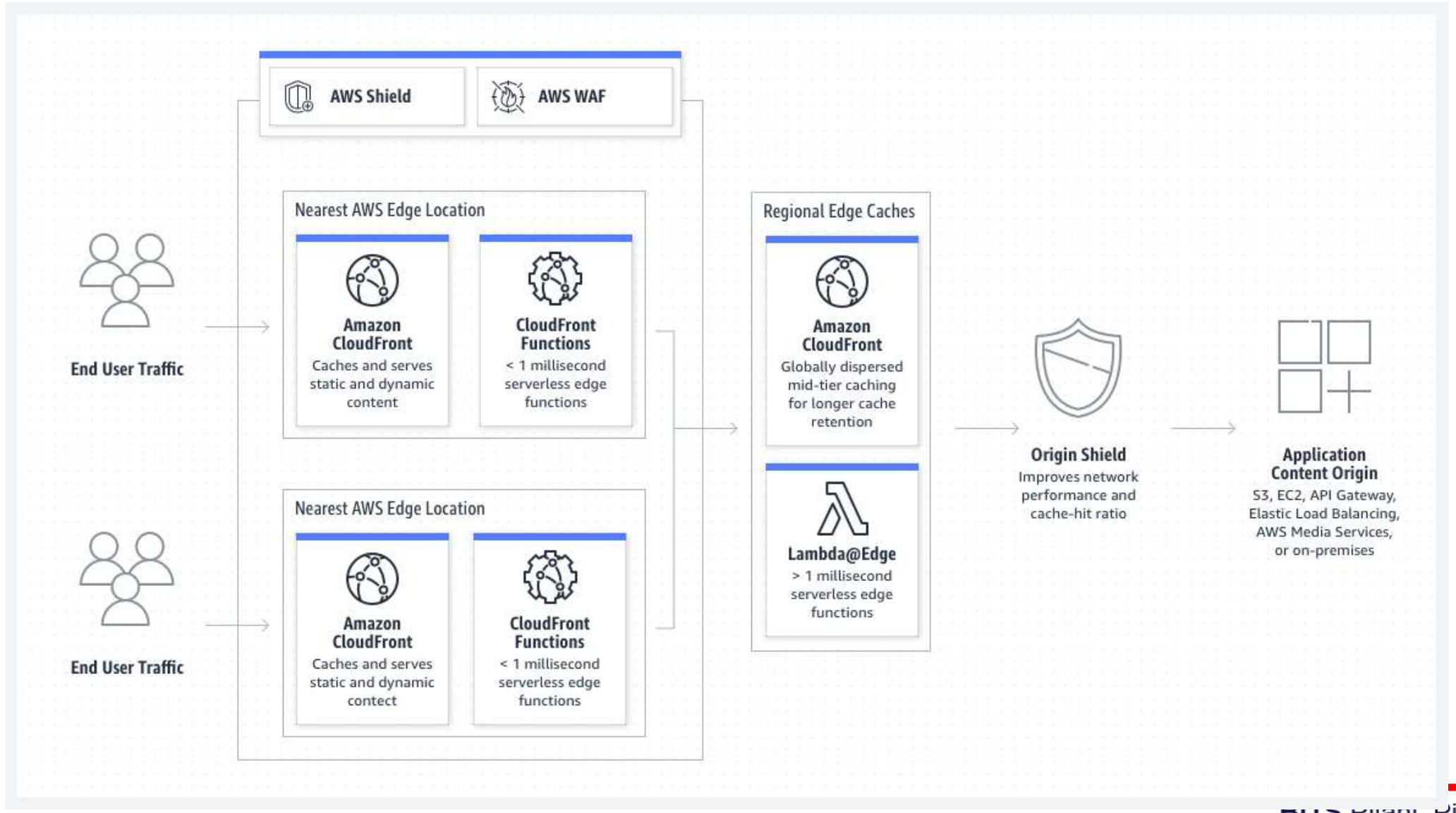
AWS CloudFront - AWS counterpart

Points of Presence (AWS version)

- AWS provides a global network of 410+ Points of Presence locations
- Consists of 400+ **edge locations** and 13 **Regional edge caches**
- Used with Amazon CloudFront
  - A global Content Delivery Network (CDN), that delivers content to end users with reduced latency
- Regional edge caches used for content with infrequent access.

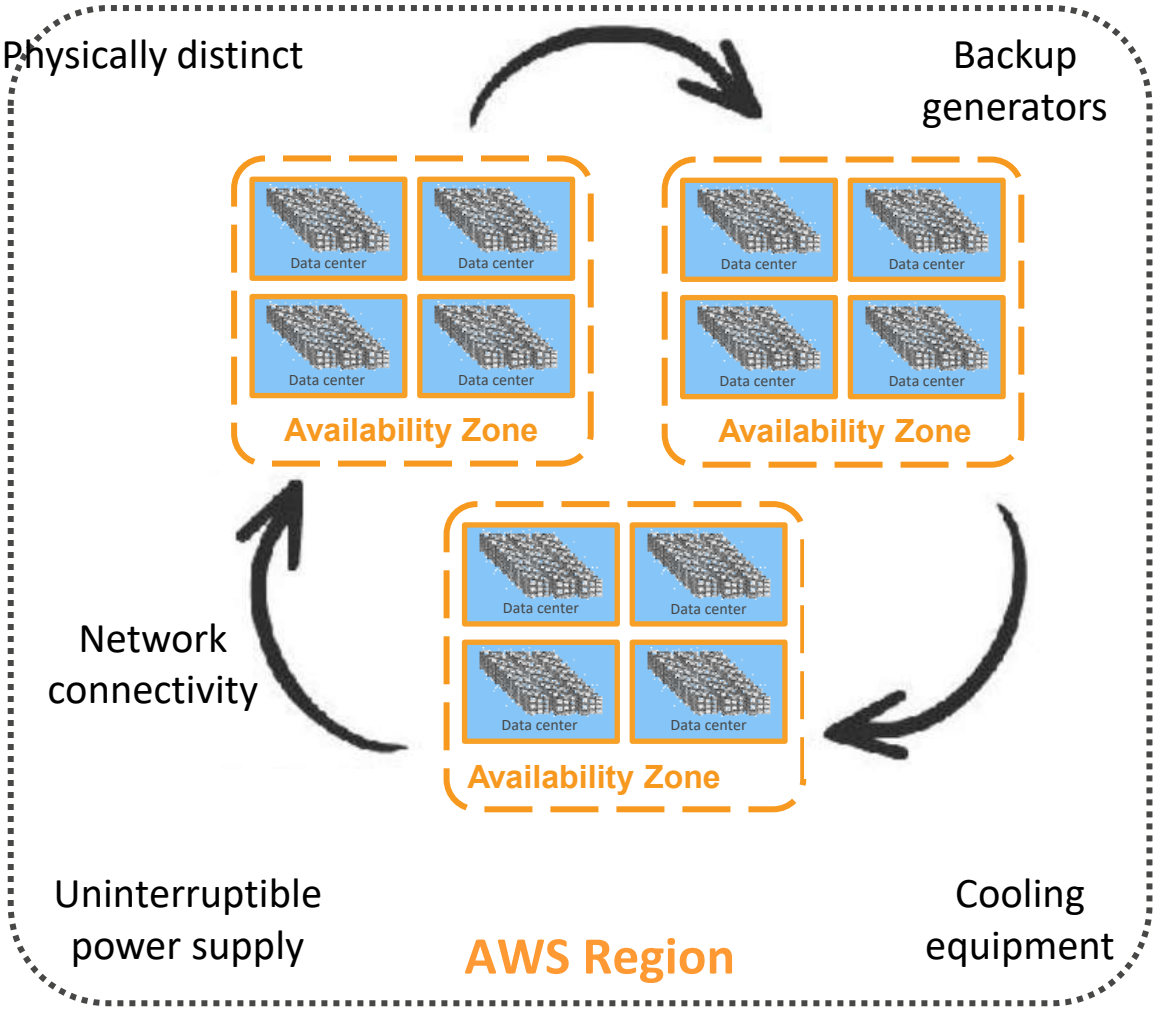


AWS CloudFront

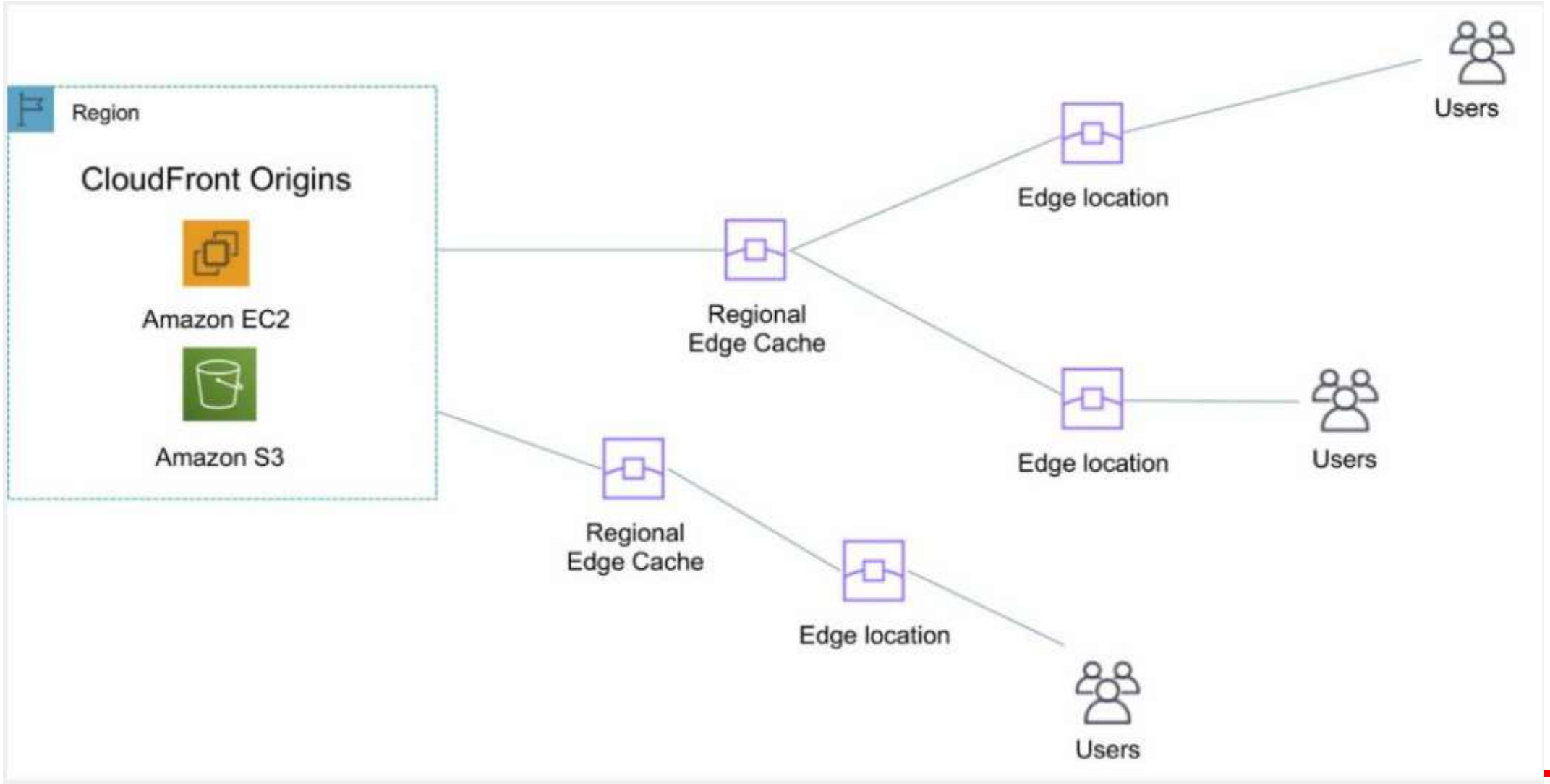


AWS infrastructure features

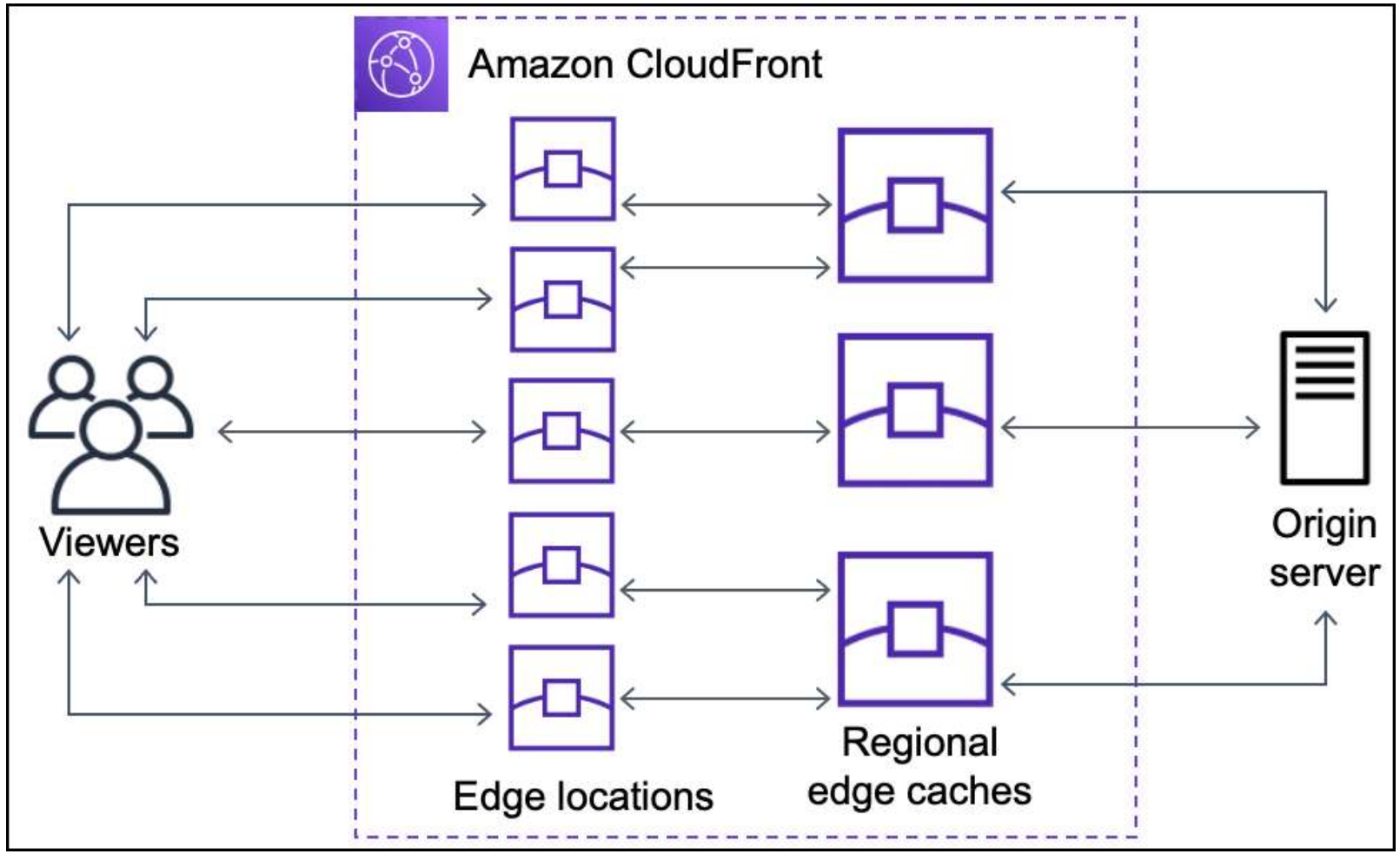
- Elasticity and scalability
  - Elastic infrastructure; dynamic adaption of capacity
  - Scalable infrastructure; adapts to accommodate growth
- Fault-tolerance
  - Continues operating properly in the presence of a failure
  - Built-in redundancy of components
- High availability
  - High level of operational performance
  - Minimized downtime
  - No human intervention



AWS Infrastructure - A Complete Picture



AWS Infrastructure - A Complete Picture



Key takeaways

- The **AWS Global Infrastructure** consists of Regions and Availability Zones.
- Your choice of a **Region** is typically based on compliance requirements or to reduce latency.
- Each **Availability Zone** is physically separate from other Availability Zones and has redundant power, networking, and connectivity.
- **Edge locations**, and **Regional edge caches** improve performance by caching content closer to users.

Section 2: AWS services and service category overview

AWS foundational services

Applications

 Virtual desktops

 Collaboration and sharing

Platform Services

Databases	Analytics	Application services	Deployment and management	Mobile Services
Relational	Cluster computing	Queuing	Containers	Identity
NoSQL	Real-time	Orchestration	DevOps tools	Sync
Caching	Data warehouse	App Streaming	Resource templates	Mobile Analytics
	Data workflows	Transcoding	Usage tracking	Notifications
		Email	Monitoring and logs	
		Search		

Foundation Services

 Compute (virtual, automatic scaling, and load balancing)

 Networking

 Storage (object, block, and archive)

Infrastructure

Regions

Availability Zones

 Edge locations

AWS categories of services


Analytics


Application Integration


AR and VR


Blockchain


Business Applications


Compute


Cost Management


Customer Engagement


Database


Developer Tools


End User Computing


Game Tech


Internet of Things


Machine Learning


Management and Governance


Media Services


Migration and Transfer


Mobile


Networking and Content Delivery


Robotics


Satellite


Security, Identity, and Compliance


Storage

Storage service category

Photo from https://www.pexels.com/photo/black-and-grey-device-159282/

AWS storage services

Amazon Simple Storage Service (Amazon S3)

Amazon Elastic Block Store (Amazon EBS)

Amazon Elastic File System (Amazon EFS)

Amazon Simple Storage Service Glacier

Compute service category

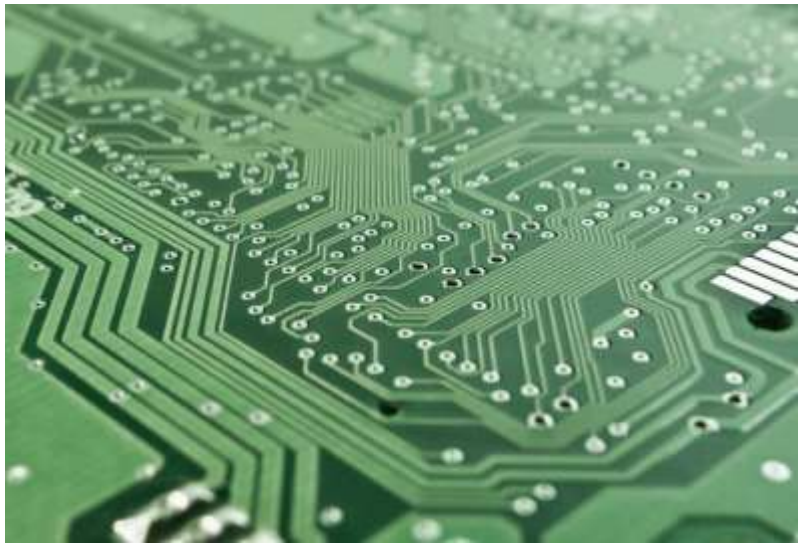


Photo from https://www.pexels.com/photo/technology-computer-lines-board-50711/

AWS Compute services

- Amazon EC2
- Amazon EC2 Auto Scaling
- Amazon Elastic Container Service (Amazon ECS)
- Amazon EC2 Container Registry
- AWS Elastic Beanstalk
- AWS Lambda
- Amazon Elastic Kubernetes Service (Amazon EKS)
- AWS Fargate

Database service category

Photo from https://aws.amazon.com/compliance/data-center/data-centers/

AWS Database services

Amazon Relational Database Service

Amazon Aurora

Amazon Redshift

Amazon DynamoDB

Networking and content delivery service category

Photo by Umberto on Unsplash

AWS networking and content delivery services

- Amazon VPC
- Elastic Load Balancing
- Amazon CloudFront
- AWS Transit Gateway
- Amazon Route 53
- AWS Direct Connect
- AWS VPN

Security, identity, and compliance service category

Photo by Paweł Czerwiński on Unsplash

AWS security, identity, and compliance services

- AWS Identity and Access Management (IAM)
- AWS Organizations
- Amazon Cognito
- AWS Artifact
- AWS Key Management Service
- AWS Shield

AWS cost management service category

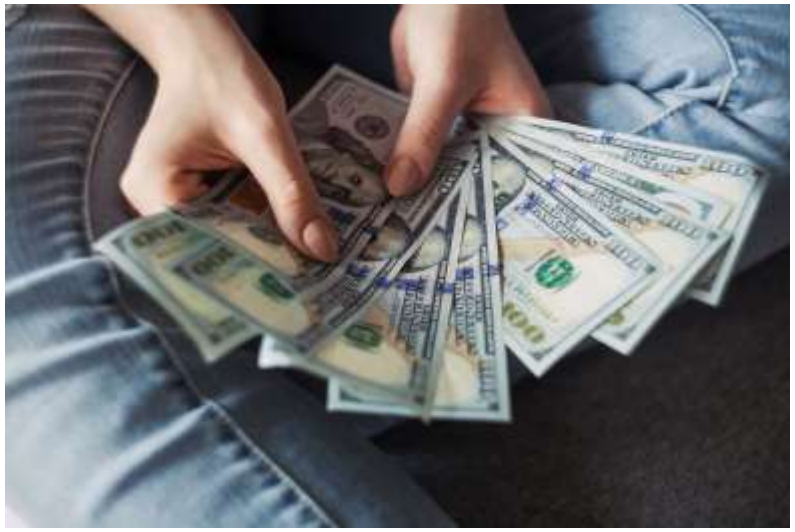


Photo by Alexander Mils on Unsplash

AWS cost management services

- AWS Cost and Usage Report
- AWS Budgets
- AWS Cost Explorer

Management and governance service category

Photo by Marta Branco from Pexels

AWS management and governance services

- AWS Management Console
- AWS Config
- Amazon CloudWatch
- AWS Auto Scaling
- AWS Command Line Interface
- AWS Trusted Advisor
- AWS Well-Architected Tool
- AWS CloudTrail

Activity: AWS Management Console clickthrough

Photo by Pixabay from Pexels.

Hands-on activity: AWS Management Console clickthrough

1. Launch the Sandbox hands-on environment and connect to the AWS Management Console**.**
2. Explore the AWS Management Console.
   A. Click the **Services** menu.
   B. Notice how services are grouped into service categories. For example, the **EC2** service appears in the **Compute** service category.
      Question #1: Under which service category does the **IAM** service appear?
      Question #2: Under which service category does the **Amazon VPC** service appear?
   C. Click the **Amazon VPC** service. Notice that the dropdown menu in the top-right corner displays an AWS Region (for example, it might display *N. Virginia*).
   D. Click the Region menu and switch to a different Region. For example, choose **EU (London)**.
   E. Click **Subnets** (on the left side of the screen). The Region has three subnets in it. Click the box next to one of the subnets. Notice that the bottom half of the screen now displays details about this subnet.
      Question #3: Does the subnet you selected exist at the level of the Region or at the level of the Availability Zone?
   F. Click **Your VPCs**. An existing VPC is already selected.
      Question #4: Does the VPC exist at the level of the Region or the level of the Availability Zone?
      Question #5: Which services are global instead of Regional? Check Amazon EC2, IAM, Lambda, and Route 53.

Activity answer key

- Question #1: Under which service category does the **IAM** service appear?
  - **Answer**: **Security, Identity, & Compliance**.
- Question #2: Under which service category does the **Amazon VPC** service appear?
  - **Answer**: **Networking & Content Delivery**
- Question #3: Does the subnet that you selected exist at the level of the Region or the level of the Availability Zone?
  - **Answer**: Subnets exist at the **level of the Availability Zone.**
- Question #4: Does the VPC exist at the level of the Region or the level of the Availability Zone?
  - **Answer**: VPCs exist at the **Region level.**
- Question #5: Which of the following services are global instead of Regional? Check Amazon EC2, IAM, Lambda, and Route 53.
  - **Answer: IAM and Route 53 are global**. Amazon EC2 and Lambda are Regional.

Additional resources

- AWS Global Infrastructure
- AWS Global Infrastructure Region Table
- AWS Cloud Products

Thank you