

Scalability

Thao Huy Vu

Maharishi International University - Fairfield, Iowa



All rights reserved. No part of this slide presentation may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying or recording, or by any information storage and retrieval system, without permission in writing from Maharishi International University (MIU).

Agenda

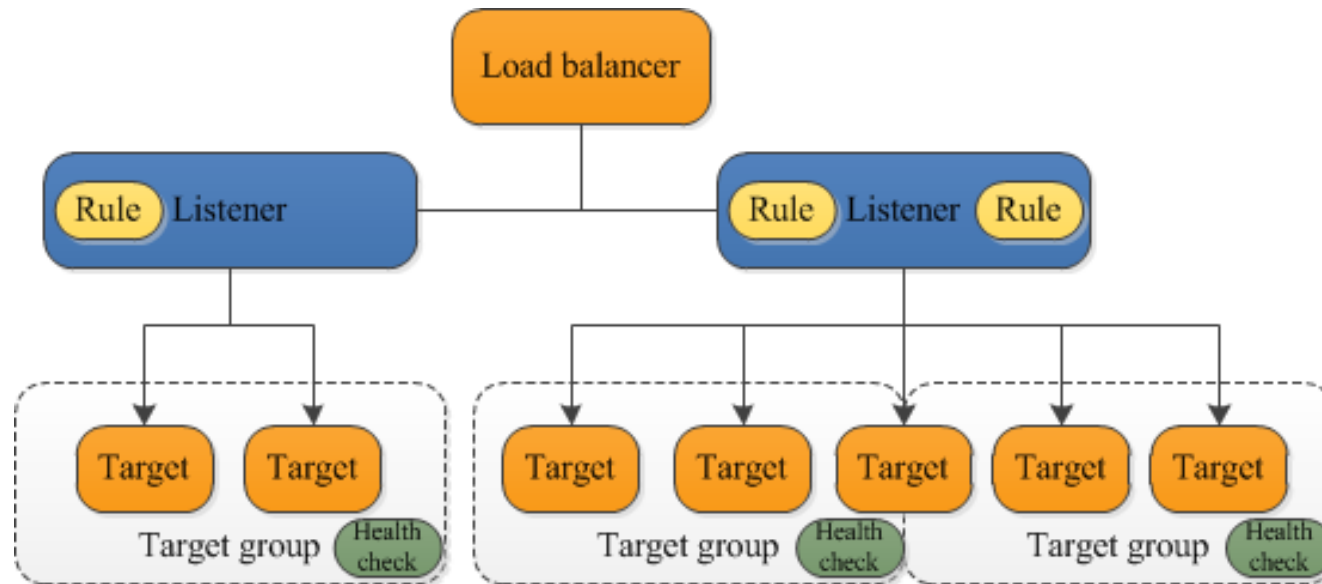
- Elastic Load Balancer (ELB)
- Auto Scaling Group (ASG)
- EBS
- EFS
- S3
- CloudFront

Elastic Load Balancer (ELB)

- **Distributes** incoming web **traffic** (visitors to a web site) and equally across multiple EC2 instances running the same app.
- **Fault tolerance:** Seamlessly providing the required amount of load-balancing capacity needed to route application traffic.
- **Reliability:** Prevent one server from being overloaded while another server can handle more visitors.
- Handle **millions** of transactions in a second.
- **Types of Load Balancer:**
 - Application Load Balancer
 - Network Load Balancer
 - Classic Load Balancer

Application Load Balancer (ALB)

- A **Listener** is a process that **checks for incoming connection requests** on a specified **port and protocol**.
- A **rule** in an AWS ALB **evaluates incoming requests** and applies an **action** if the request **matches a condition**.



ALB Listener Actions

- **Actions** define **what happens to requests** that match a Listener rule.
 - Forward: Send traffic to one or more target groups.
 - Redirect: Redirect the request to another URL or protocol.
 - Fixed Response: Respond with a fixed HTTP status code and optional message or body.

Actions

Action types

Routing actions

☒ Forward to target groups

☐ Redirect to URL

☐ Return fixed response

Forward to target group [Info](#)

Choose a target group and specify routing weight or [Create target group](#).

Target group

Select a target group ▼

[Add target group](#)

You can add up to 4 more target groups.



Weight

1

0-999

Percent

100%


ALB Target Groups

- A logical grouping of targets (e.g. EC2, ECS tasks, Lambda) that receive traffic from the ALB based on rules.
- Key features:
 - **Targets:** EC2 instances, ECS tasks, or IP addresses that receive traffic.
 - **Health Checks:** ALB monitors target health and stops traffic to unhealthy targets, ensuring reliability.
 - **Routing:** Enables traffic routing based on listener rules, such as path patterns, hostnames, or query parameters.
 - **Port Configuration:** Each target group maps to a specific port on the targets.

Target Group

Choose a target type

☒ Instances

- Supports load balancing to instances within a specific VPC.
- Facilitates the use of [Amazon EC2 Auto Scaling](#)  to manage and scale your EC2 capacity.

☐ IP addresses

- Supports load balancing to VPC and on-premises resources.
- Facilitates routing to multiple IP addresses and network interfaces on the same instance.
- Offers flexibility with microservice based architectures, simplifying inter-application communication.
- Supports IPv6 targets, enabling end-to-end IPv6 communication, and IPv4-to-IPv6 NAT.

☐ Lambda function

- Facilitates routing to a single Lambda function.
- Accessible to Application Load Balancers only.

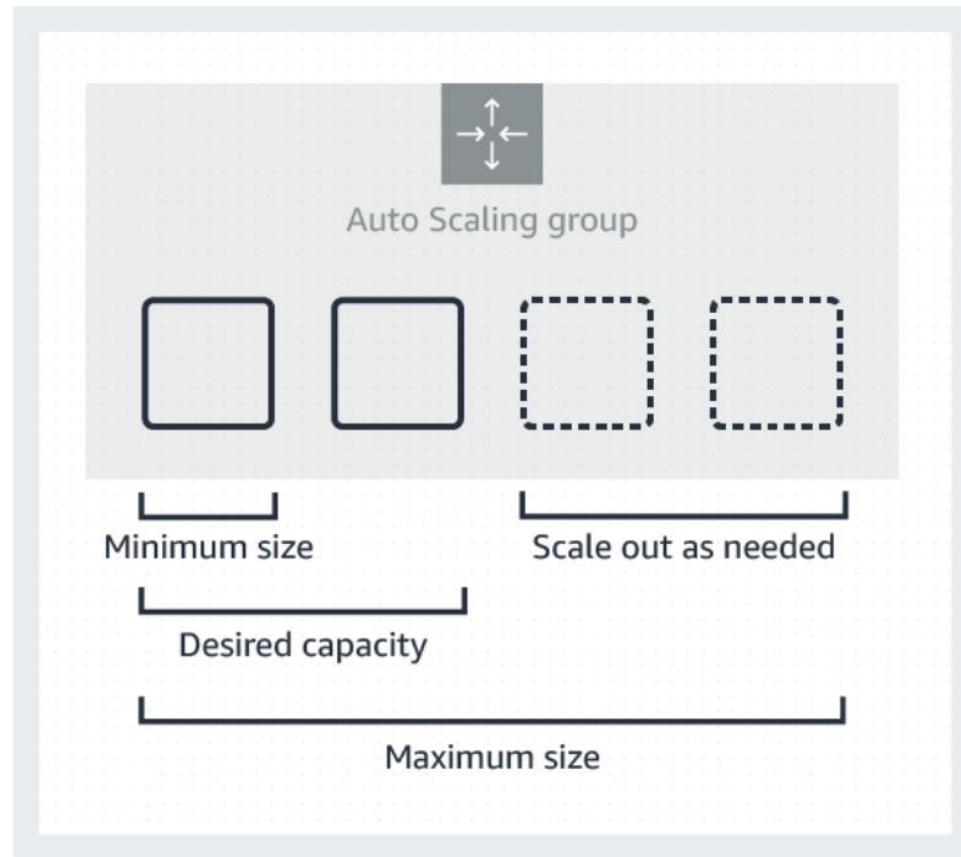
☐ Application Load Balancer

- Offers the flexibility for a Network Load Balancer to accept and route TCP requests within a specific VPC.
- Facilitates using static IP addresses and PrivateLink with an Application Load Balancer.

Auto Scaling Group (ASG)

- Auto Scaling Groups (ASG) automatically scale **EC2 instances** in and out based on scaling policies.
- **Scaling Out (Up)**: Adds more instances to meet increased demand.
- **Scaling In (Down)**: Removes instances to reduce costs during lower demand.
- ASG adjusts the **desired count**, ensuring the optimal number of instances matches the workload.

Auto Scaling Group



ASG & ELB

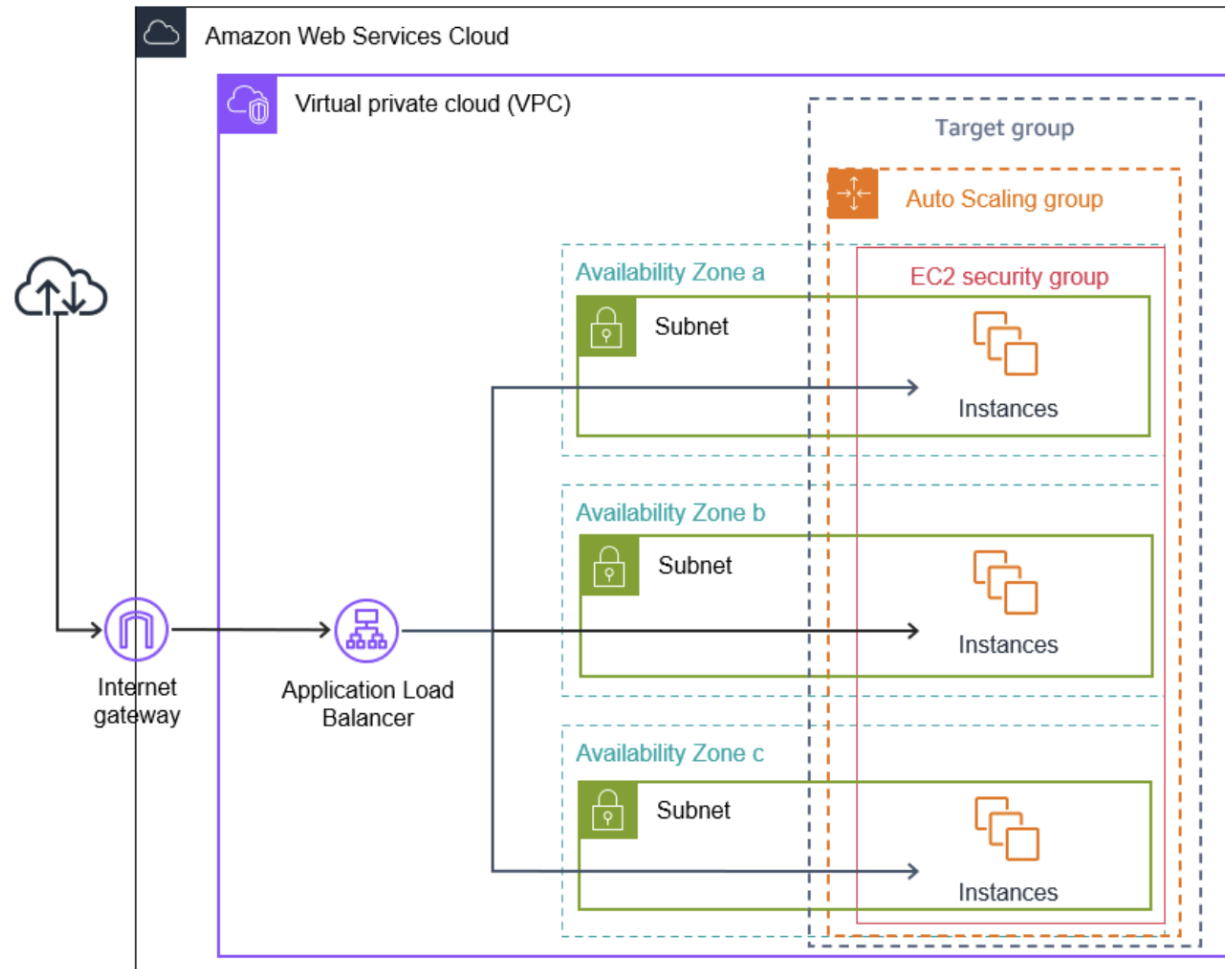
- **Auto Scaling:**

- Automatically adjusts the number of **EC2 instances** or other resources based on predefined metrics (e.g., CPU utilization, SQS queue length).
- Ensures the application scales **out** (adds instances) during high demand and **in** (removes instances) during low demand.
- Helps handle varying workloads efficiently while optimizing cost.

- **Elastic Load Balancing (ELB):**

- Automatically distributes incoming application traffic across multiple targets (e.g., EC2 instances, containers) in one or more **Availability Zones (AZs)**.
- Improves fault tolerance, availability, and application performance by spreading traffic evenly.

Auto Scaling Group



ASG Component

- Auto Scaling Group: Maintain the health and desired count of instances
 - Networking
 - Scaling policies
 - Load Balancer
 - Health check configuration
- Launch Templates: Define the instance configuration for use in an Auto Scaling Group
 - AMI
 - instance type
 - IAM profile
 - User Data – Startup script. It will be executed when a EC2 starts.
 - SG and so on.

Amazon Elastic Block Storage (EBS)

- **Definition:** Amazon EBS is a high-performance, block storage service for Amazon EC2 instances, providing persistent and reliable storage.
- **Durability:**
 - Data is automatically replicated within the same Availability Zone (AZ).
 - Ensures data persistence even if the associated EC2 instance is stopped or terminated.
- **Performance:**
 - Provides low-latency storage for mission-critical applications.
 - Supports various volume types for different workloads (e.g., general-purpose SSD, provisioned IOPS SSD, HDD).
- **Scalability:** Volumes can be resized dynamically to accommodate growing storage needs.
- **Backup and Recovery:**
 - Supports snapshots to create backups of volumes.
 - Snapshots are stored in Amazon S3 and can be used to create new volumes.
- **Encryption:** Supports encryption at rest and in transit for secure data storage.

Amazon Elastic File System

- **Definition:** A fully managed, scalable file storage service designed for use with AWS compute resources like EC2, Lambda, and containers.
- **Scalability:** Automatically scales storage capacity up or down as files are added or removed.
- **Shared Access:** Provides concurrent access for multiple instances and applications across Availability Zones.
- **Performance Modes:**
 - **General Purpose:** Low latency for latency-sensitive applications.
 - **Max I/O:** High throughput for large-scale, data-heavy workloads.
- **Storage Classes:**
 - **Standard:** For frequently accessed data.
 - **Infrequent Access (IA):** Lower-cost option for infrequently accessed files.
- **High Durability and Availability:** Stores data across multiple Availability Zones for fault tolerance.
- **POSIX (Portable Operating System Interface) Compliance:** Supports standard file system semantics (e.g., file locking, directories) across different systems.

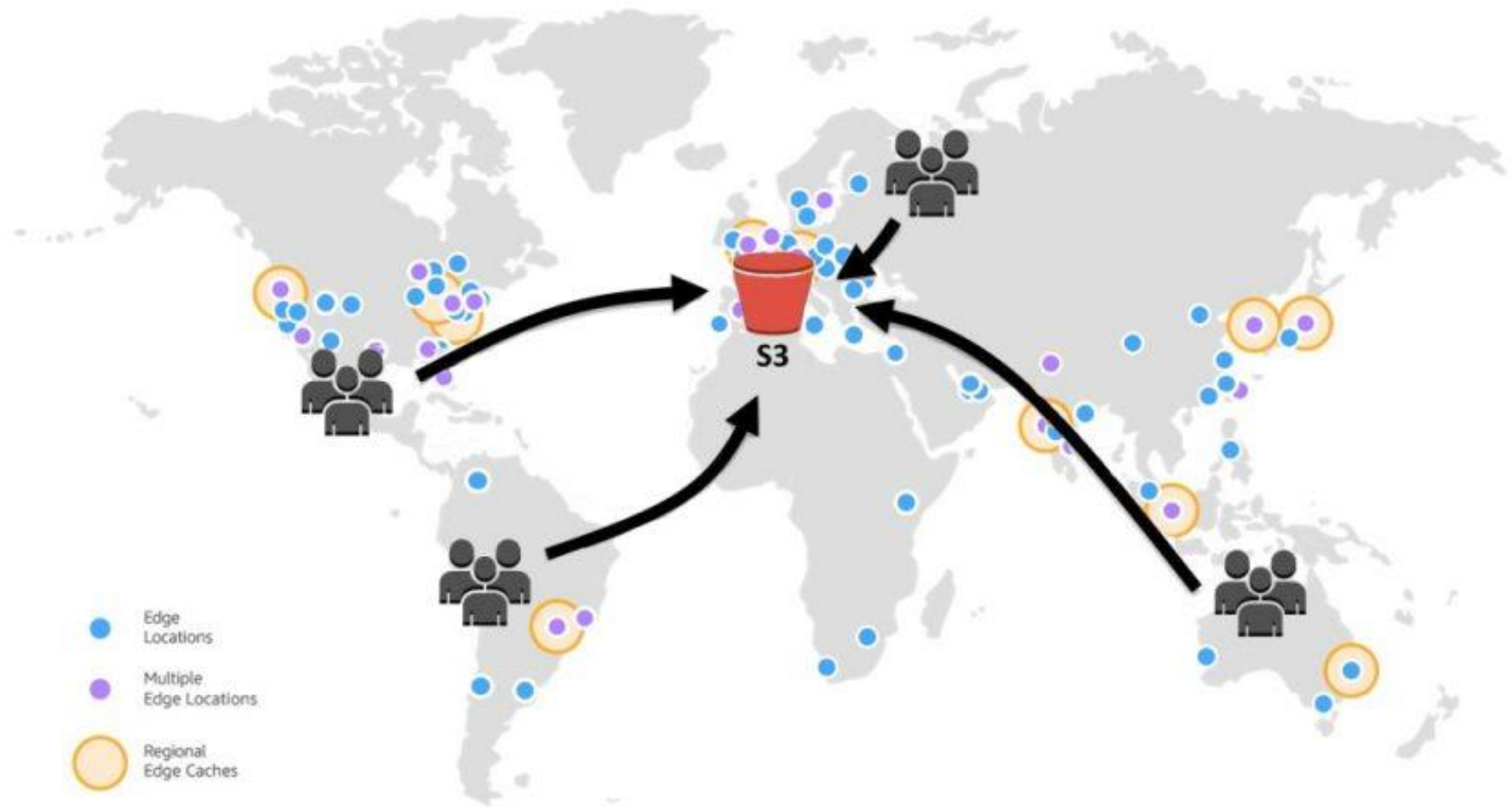
Amazon S3

- **Definition:** A scalable, durable, and secure object storage service provided by AWS.
- **Storage Type:** Stores data as objects in buckets. Each object consists of data, metadata, and a unique key.
- **Key Features:**
 - **Scalability:** Automatically scales to handle unlimited data.
 - **Durability:** 99.999999999% (11 nines) durability.
 - **Accessibility:** Access via the web, AWS CLI, SDKs, or APIs.
 - **Security:** Supports encryption, IAM policies, bucket policies, and ACLs.
- **Use Cases:**
 - Data backups, archives, and disaster recovery.
 - Hosting static websites.
 - Storing media files, logs, or application data.

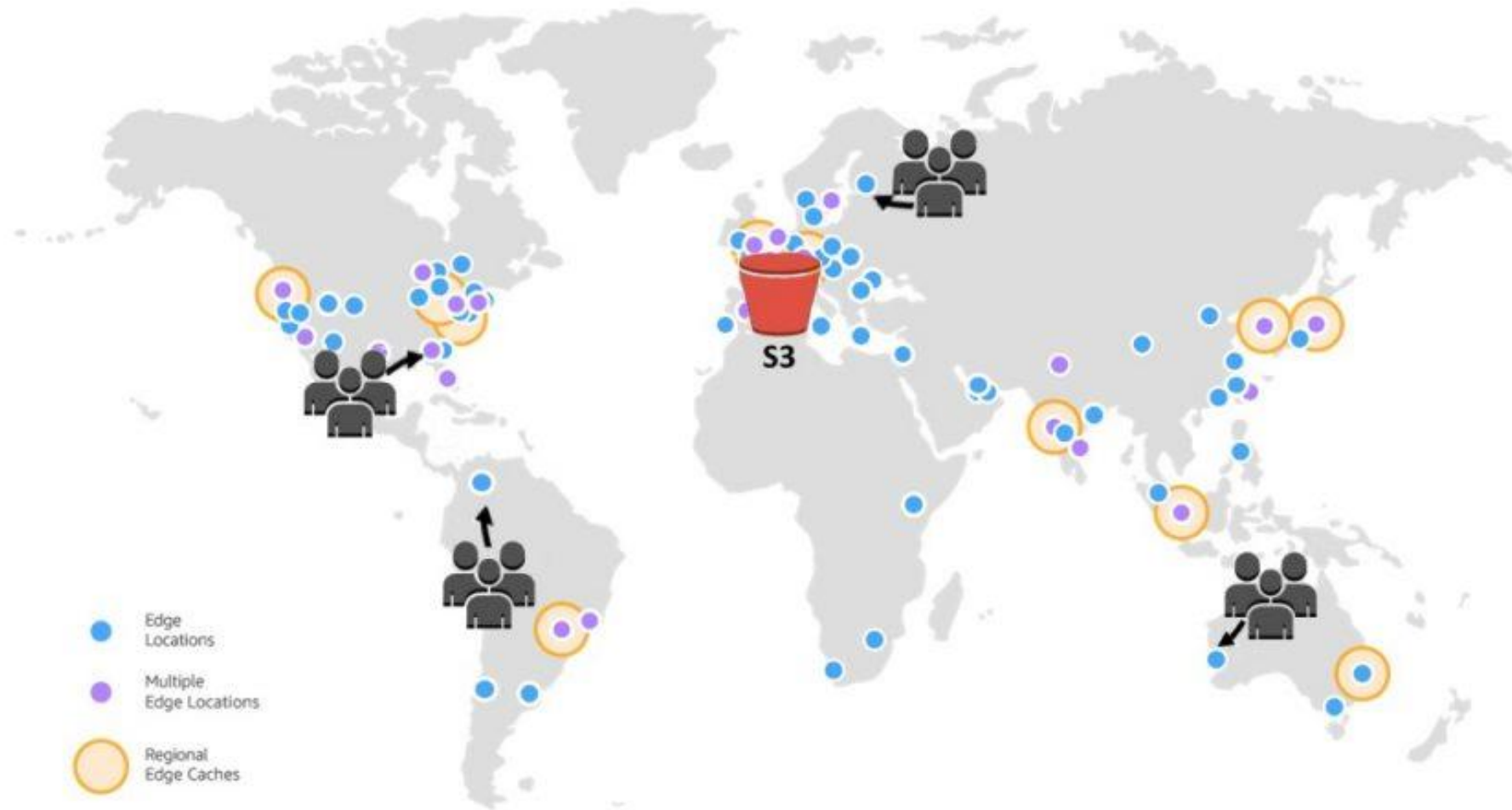
S3 with CloudFront

- **Static Content Hosting:** Traditionally, static content like images, videos, or scripts (.css, .js) was hosted on a single server, serving all users globally from one location, resulting in higher latency for distant users.
- **Better Solution:**
 - Host static content on **Amazon S3**.
 - Use **Amazon CloudFront** (a Content Delivery Network) to distribute content globally.

Hosting static contents on S3 without CloudFront



Hosting static contents on S3 with CloudFront



Reference

- AWS: <https://docs.aws.amazon.com>
- ChatGPT: <https://chatgpt.com>
- Google AI: <https://gemini.google.com>
- Practical Tutorials: <https://thaovu.org>