MPRASHANT

AWS ELB

Elastic Load Balancing

MPRASHANT

- ELB - Elastic Load Balancer
- ASG - Auto Scaling Groups

MPRASHANT

Let’s first understand the terms

- Scalability
- High Availability

MPRASHANT

Scalability

- Scalability means the ability to grow your system's resources when your application or website gets more traffic or more users.


WEBSITE
WEBSITE

MPRASHANT

Vertical Scalability (Scaling Up)

- Vertical Scalability means adding more power (CPU, RAM) to your existing server.
- Ex: t2.micro to m5.large

Horizontal Scalability (Scaling Out)

- Horizontal Scalability means adding more instances (servers) to distribute the load.
- You can add more EC2 instances behind a load balancer

MPRASHANT

High Availability (HA)

- High Availability (HA) means keeping your service up and running with minimal downtime, so it's always accessible to users.

- Ex: running resources in multiple AZs

MPRASHANT

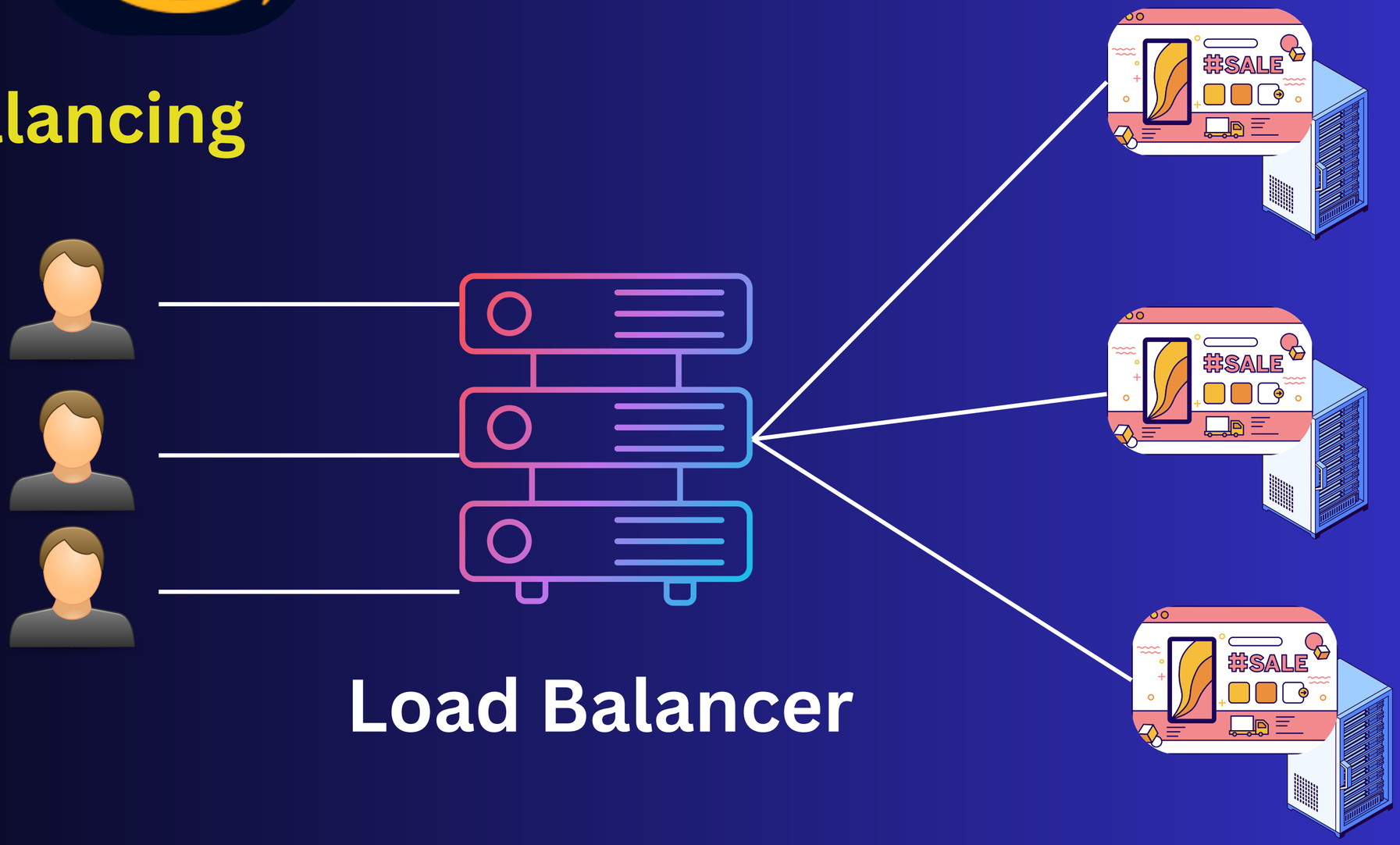
Elasticity

- Elasticity means the ability to automatically adjust resources as the demand changes—adding more when needed and removing when it’s no longer necessary.

- Ex: ASG

MPRASHANT

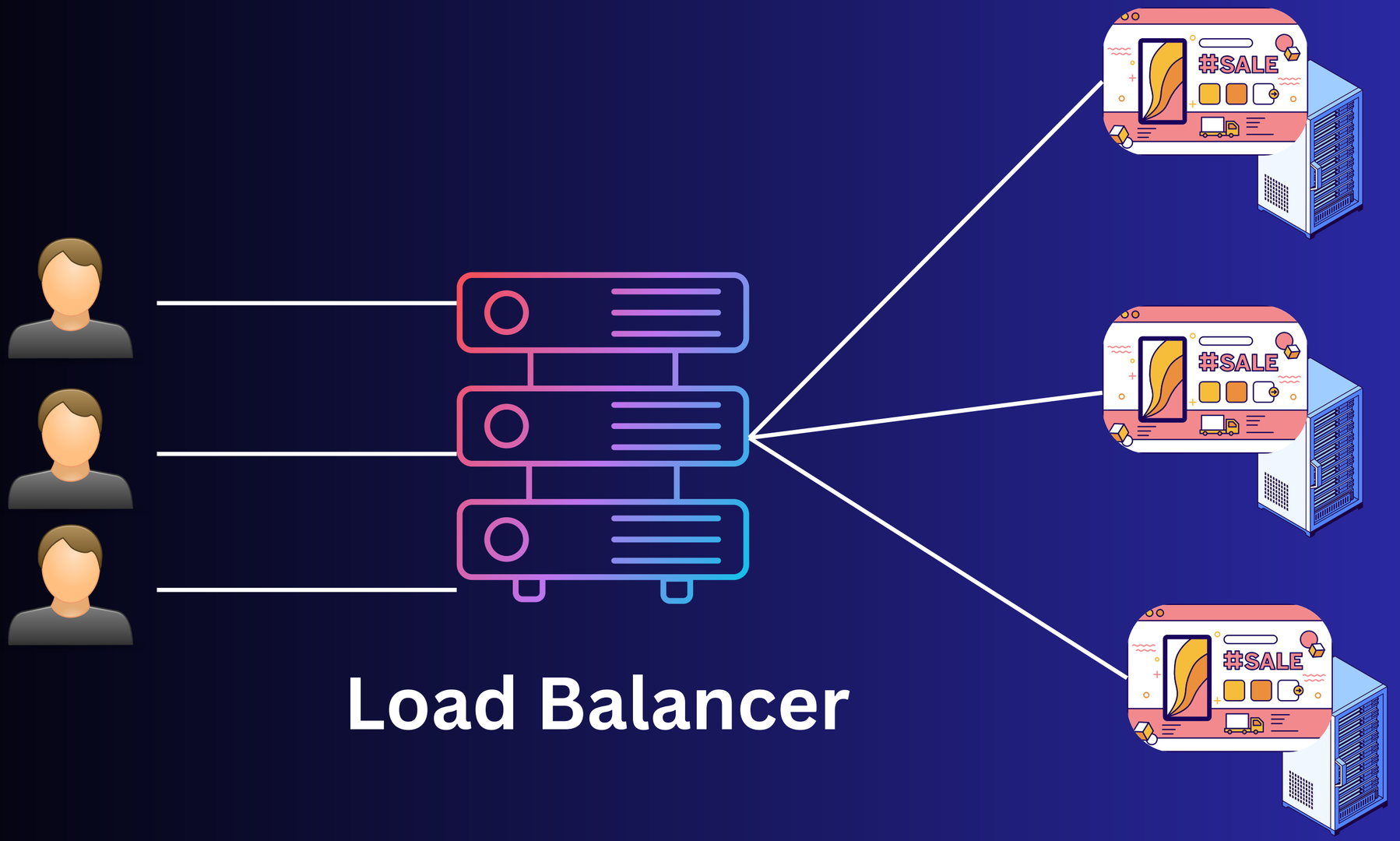
Load Balancing



MPRASHANT

aws

Mumbai



```
#!/bin/bash
sudo yum update -y

# Install Apache web server (httpd)
sudo yum install -y httpd
sudo systemctl start httpd
sudo systemctl enable httpd

# Create a simple HTML file to verify the web server is running, including dynamic hostname
echo "<html><h1>Welcome to Apache Web Server on Amazon Linux - $(hostname)!</h1></html>" > /var/www/html/index.html
```

MPRASHANT

Elastic Load Balancer Points

MPRASHANT

- **Distributes Traffic**: It splits incoming traffic across multiple servers so no single server gets overloaded.

- **Improves Availability**: If one server goes down, the load balancer automatically sends traffic to the working servers, ensuring your application stays available.

- **Scales Resources**: It helps manage high demand by adding more servers during peak times and distributing the load.

- Single point of Access need to be expose.
- HA across AZs

MPRASHANT

AWS offers different types of load balancers depending on your needs.

- **Application Load Balancer (ALB)** is perfect for web applications, handling complex HTTP and HTTPS requests (Layer 7)

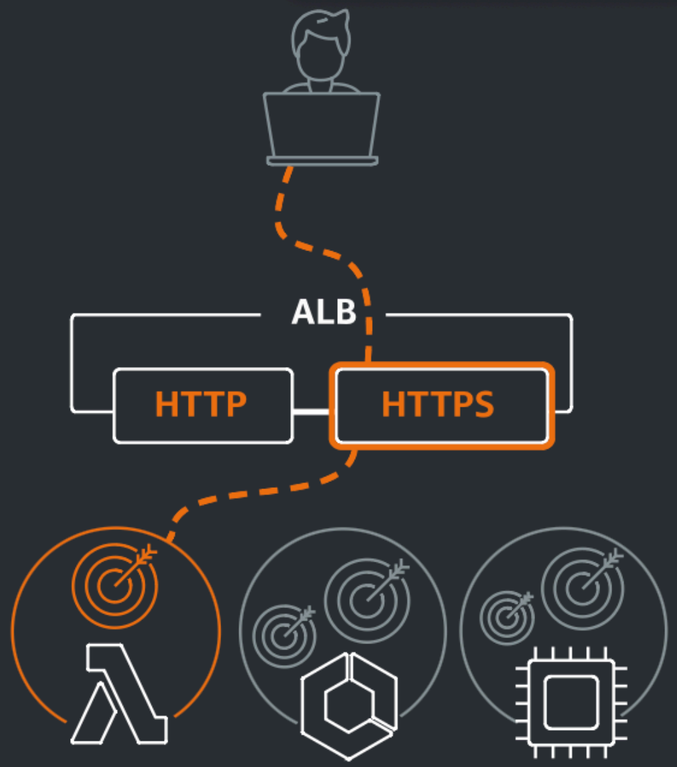
- **Network Load Balancer (NLB)** is designed for high-performance and low latency, perfect for TCP/UDP traffic (ex: gaming, financial apps) (Layer 4)

- **Gateway Load Balancer (GWLB)** helps deploy, scale, and manage third-party virtual appliances, such as firewalls and monitoring solutions.

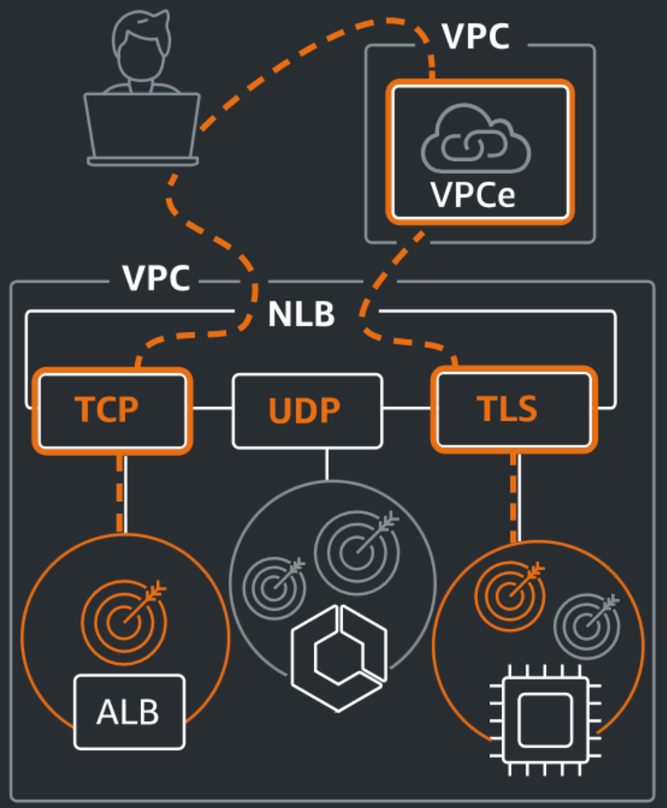
Load balancer types

Application Load Balancer Info

Compare and select load balancer



Network Load Balancer Info



Gateway Load Balancer Info

Practical: Creating ELB

- Set Up EC2 Instances: Create two or more EC2 instances, install a web server, and tag them for easy identification.
- Configure Security Groups: Set up a security group allowing HTTP and SSH access.
- Create the Load Balancer: Use the EC2 dashboard to create an Application Load Balancer and set it as internet-facing.
- Register Targets: Add EC2 instances to the target group and configure health checks.
- Test the Load Balancer: Access the DNS name of the load balancer and observe load balancing in action.
- Explain to Students: Highlight key concepts like traffic distribution, high availability, and scalability.

MPRASHANT

Auto Scaling Group (ASG)

AWS ASG (Auto Scaling Group) is a service that automatically adds or removes EC2 instances based on demand to ensure your application is always available.

It helps scale up when more capacity is needed and scale down during low usage to save costs, keeping the right number of servers running at all times.

Functions

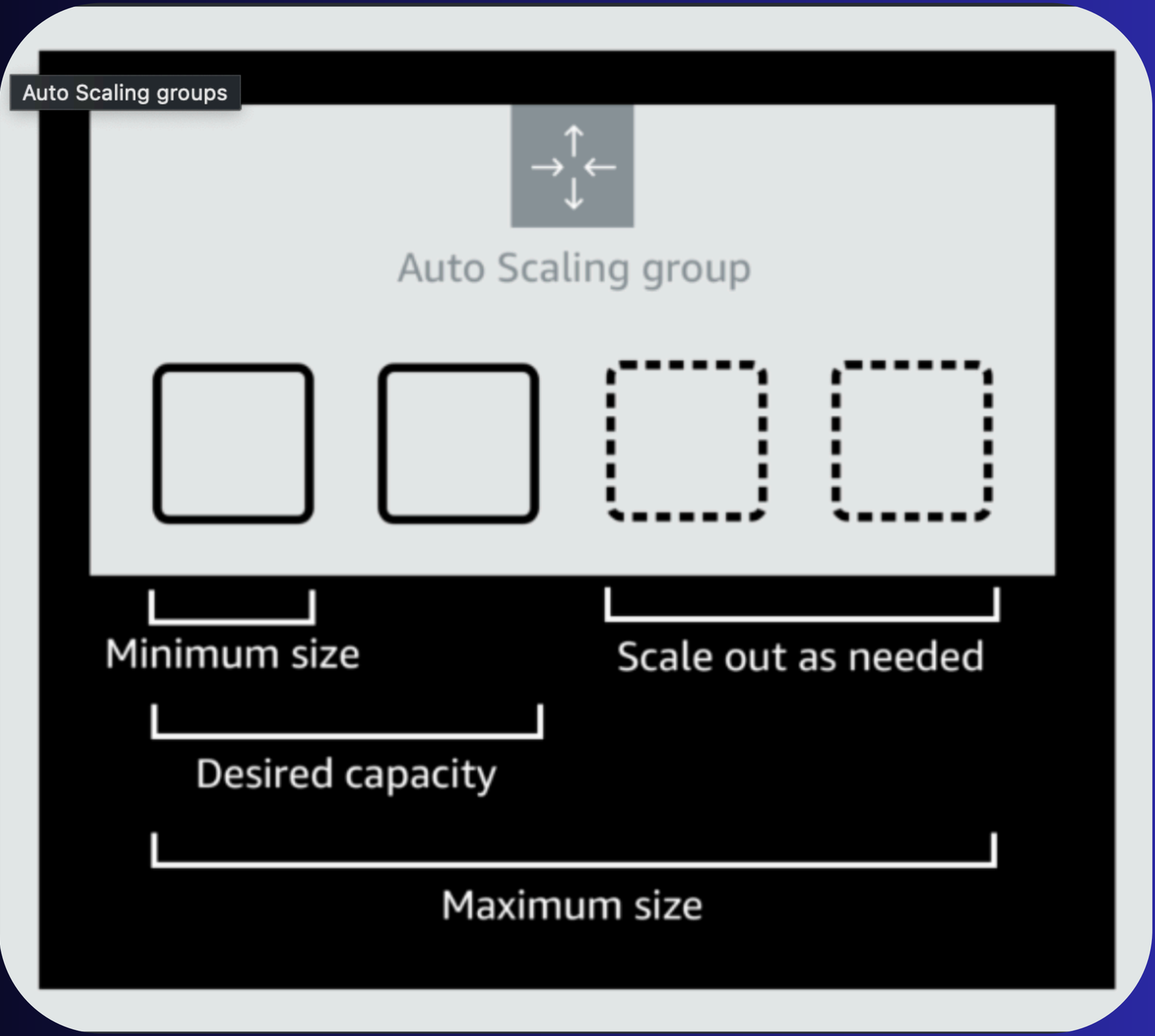
- **Automatic Scaling:** Scale the number of EC2 instances up or down based on demand.
- **Maintain Instance Health**: Replace unhealthy instances automatically to ensure reliability.

- **Use Scaling Policies:** Set rules for scaling based on metrics like CPU usage or request count.
- **Ensure Availability**: Always keep a defined number of instances running to meet application needs.

- **Schedule Scaling**: Pre-configure scaling activities for specific times (e.g., traffic peaks).

Functions

- **Distribute Instances:** Deploy instances across multiple Availability Zones for high availability.
- **Integrate with ELB:** Attach instances to an Elastic Load Balancer to automatically balance traffic.
- **Optimize Costs:** Scale down during low demand to save on infrastructure costs.


Auto Scaling groups
Auto Scaling group
Minimum size
Scale out as needed
Desired capacity
Maximum size

Steps to Create ASG

- Launch Template or Configuration
- Create Auto Scaling Group
- Select VPC and Subnets
- Attach Load Balancer (Optional)
- Configure Scaling Policies
- Health Checks
- Add Notifications (Optional)
- Review and Create