

Auto Scaling Solutions

- Overview
- Configuration
- Groups
- Termination Policies
- Elastic Load Balancing
- Load Balancer Concepts

Auto Scaling Overview

EPISODE 9.01

Auto Scaling

- Monitors applications
- Adjusts capacity
- Manages costs

Auto Scaling Functionality



Scalable AWS Resources

- EC2 Auto Scaling groups
- Aurora DB clusters
- DynamoDB global secondary indexes
- DynamoDB tables
- Elastic Container Service (ECS) services
- Spot Fleet requests

Auto Scaling Costs

- Free to use
- Results of use may cost:
 - More instances
 - CloudWatch
 - ELB load balancers

Auto Scaling Groups

EPISODE 9.02

Auto Scaling Groups

- Collection of instances with similar characteristics
 - Can be scaled based on criteria
 - Unhealthy instances can be auto-replaced
 - Any state other than “Running” is unhealthy

Group Considerations

- Time to launch and configure a server
- Relevant metrics to your application
 - CPU utilization
 - Network throughput
 - Free memory

Group Considerations

- What AZs should the Auto Scaling group span?
- Scale to increase or decrease capacity?
- Specify min number of instances always running



Termination Policies

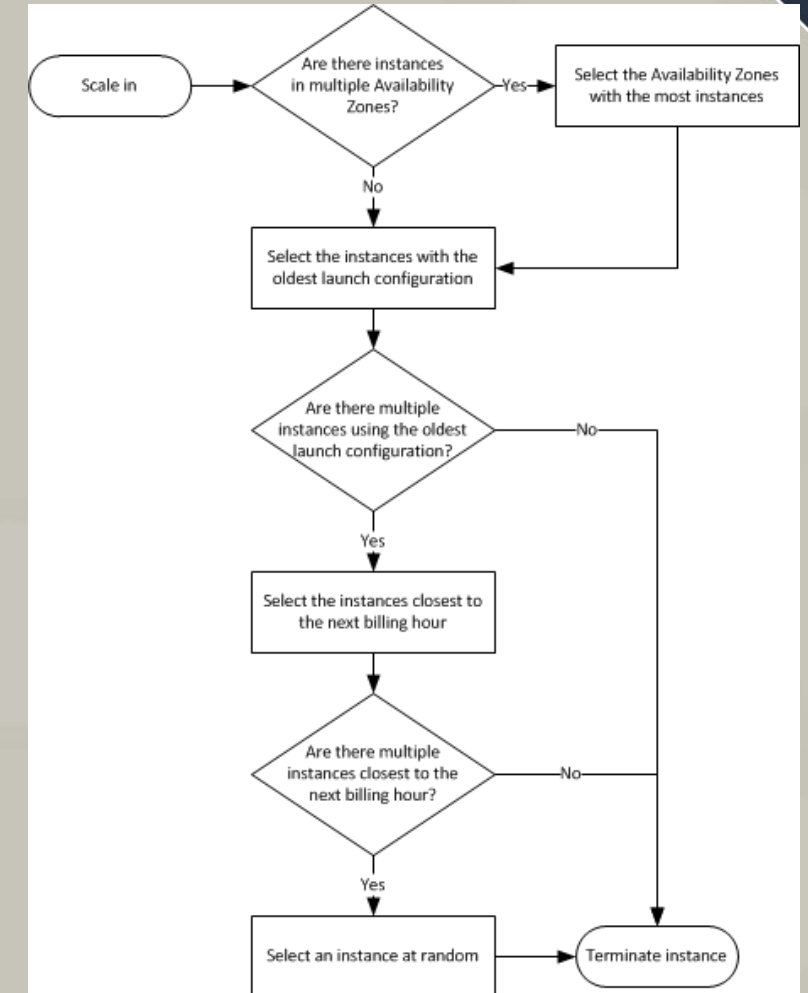
EPISODE 9.03

Scaling Out and Scaling In

- Scaling out - adding instances
- Scaling in - removing instances

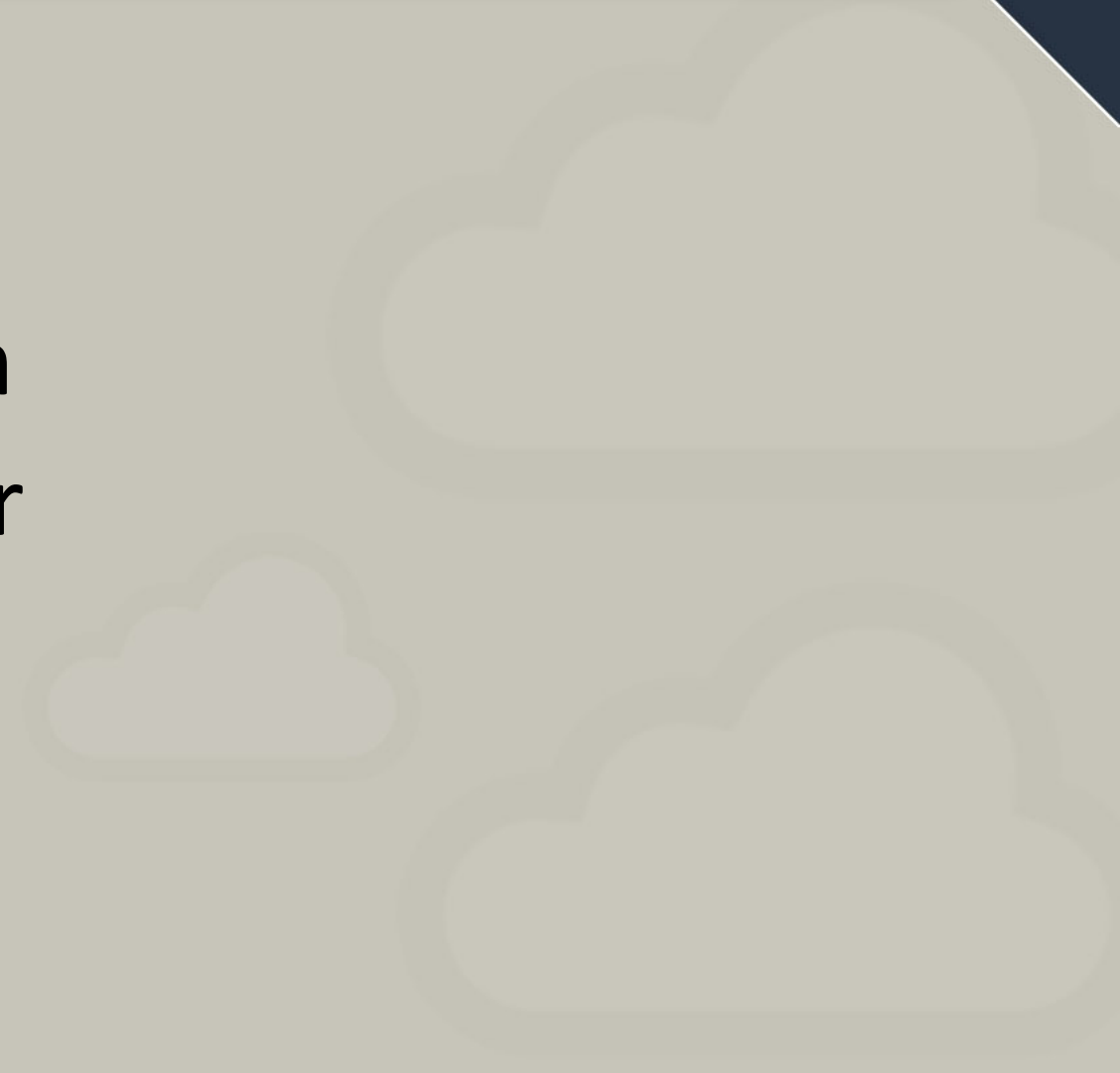
Default Termination Policy

- If there are instances in multiple Availability Zones, select the Availability Zone with the most instances and at least one instance that is not protected from scale in. If there is more than one Availability Zone with this number of instances, select the Availability Zone with the instances that use the oldest launch configuration.
- Determine which unprotected instances in the selected Availability Zone use the oldest launch configuration. If there is one such instance, terminate it.
- If there are multiple instances that use the oldest launch configuration, determine which unprotected instances are closest to the next billing hour. (This helps you maximize the use of your EC2 instances and manage your Amazon EC2 usage costs.) If there is one such instance, terminate it.
- If there is more than one unprotected instance closest to the next billing hour, select one of these instances at random.



Custom Termination Policies



- OldestInstance
 - NewestInstance
 - OldestLaunchConfiguration
 - ClosestToNextInstanceHour
 - Default
- 

Auto Scaling Configuration Lab

Episode 9.04

DEMO

- Working with *AWS* Auto Scaling

Auto Scaling Launch Methods

Episode 9.05

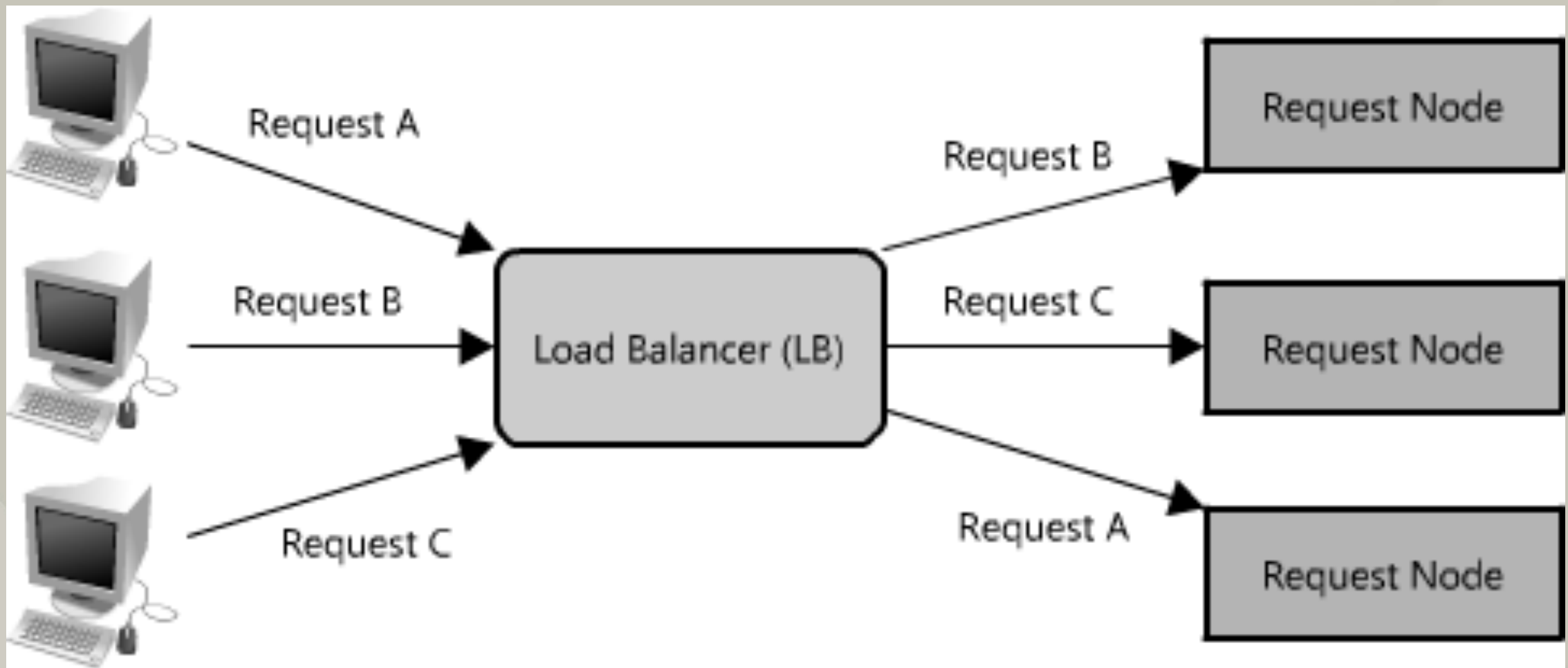
DEMO

- Creating an Auto Scaling group from a template
- Creating an Auto Scaling group from a launch configuration
- Creating an Auto Scaling group using an EC2 instance
- Creating an Auto Scaling group with the EC2 launch wizard

Load Balancer Concepts

EPISODE 9.06

Load Balancing Defined



Load Balancing Categories

- Sender initiated
 - Sender locates best target
- Receiver initiated
 - Receiver selects best target

Static Load Balancing

- Multi-tier application
 - Specific actions are assigned to specific servers/resources
 - Actions always processed on assigned target
 - No scalability

Dynamic Load Balancing

- True load balancing
 - Actions dynamically assigned
 - Scalability is provided
- Used by AWS Elastic Load Balancing (ELB)

Load Balancing Algorithms

- Round Robin
- Randomized
- Centrally Managed
- Threshold-Based



Elastic Load Balancing (ELB)

EPISODE 9.07

ELB Benefits

- Highly available
- Secure
- Flexible
- Monitoring and auditing included
- Elastic
- Hybrid

ELB Types DEMO

- Application load balancer
- Network load balancer
- Classic load balancer

<https://aws.amazon.com/elasticloadbalancing/?nc=sn&loc=0>

Supported Services

- EC2
- ECS
- Auto Scaling
- CloudWatch
- Route 53

DEMO

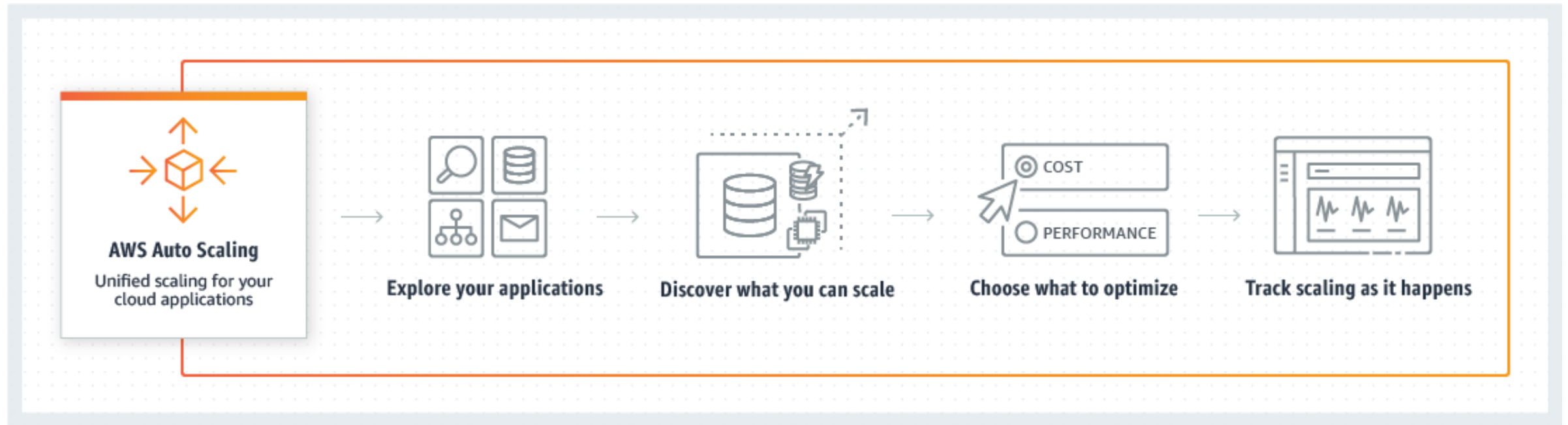
- ELB Features

- EPISODE 9.01
- Auto Scaling Overview

Auto Scaling

- Monitors applications
- Adjusts capacity
- Manages costs

Auto Scaling Functionality



Scalable AWS Resources

- EC2 Auto Scaling groups
- Aurora DB clusters
- DynamoDB global secondary indexes
- DynamoDB tables
- Elastic Container Service (ECS) services
- Spot Fleet requests

Auto Scaling Costs

- Free to use
- Results of use may cost:
 - More instances
 - CloudWatch
 - ELB load balancers

- EPISODE 9.03
- Groups

Auto Scaling Groups

- Collection of instances with similar characteristics
 - Can be scaled based on criteria
 - Unhealthy instances can be auto-replaced
- Any state other than “Running” is unhealthy

Group Considerations

- Time to launch and configure a server
- Relevant metrics to your application
 - CPU utilization
 - Network throughput
 - Free memory

Group Considerations

- What AZs should the Auto Scaling group span?
- Scale to increase or decrease capacity?
- Specify min number of instances always running

DEMO

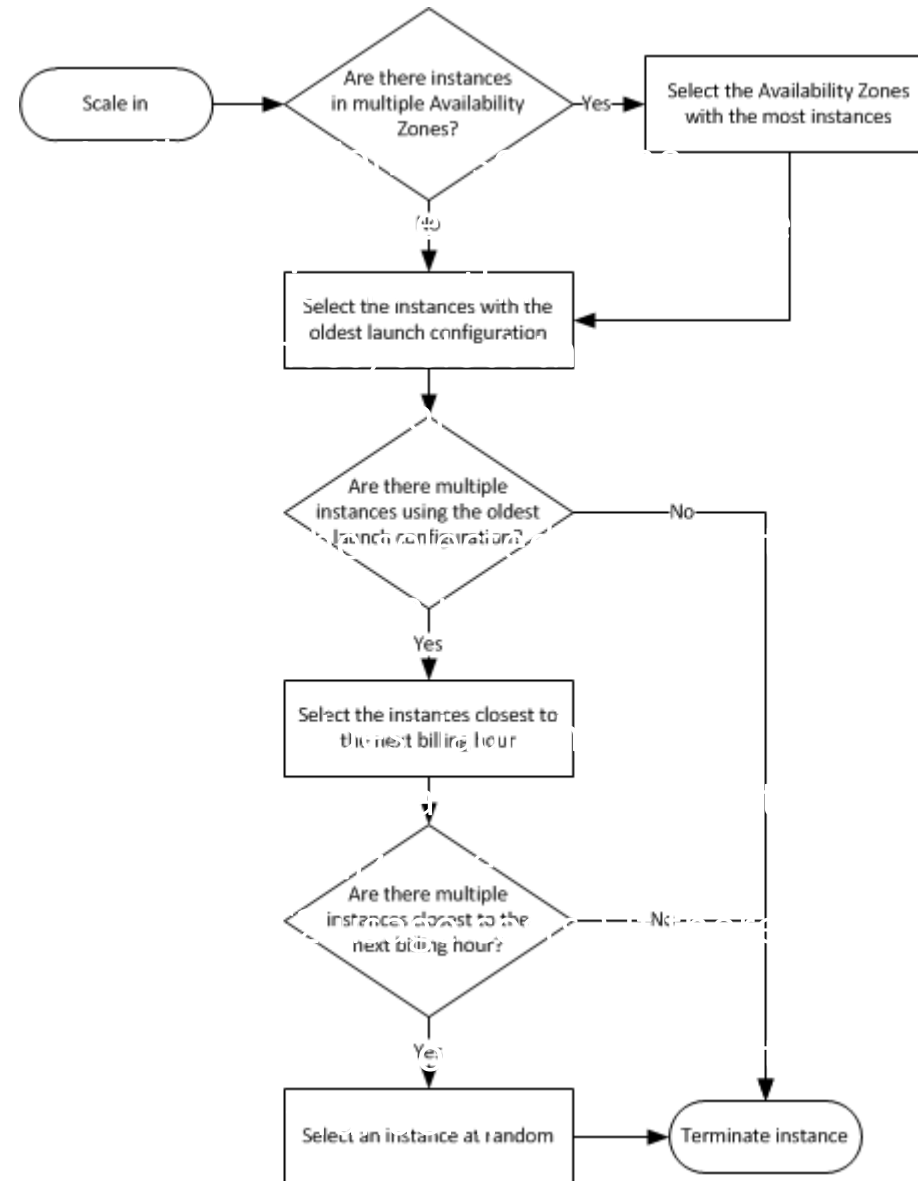
- Creating an Auto Scaling group from a template
- Creating an Auto Scaling group from a launch configuration
- Creating an Auto Scaling group using an EC2 instance
- Creating an Auto Scaling group with the EC2 launch wizard

- EPISODE 9.04
- Termination Policies

Scaling Out and Scaling In

- Scaling out - adding instances
- Scaling in - removing instances

Default Termination Policy



Custom Termination Policies

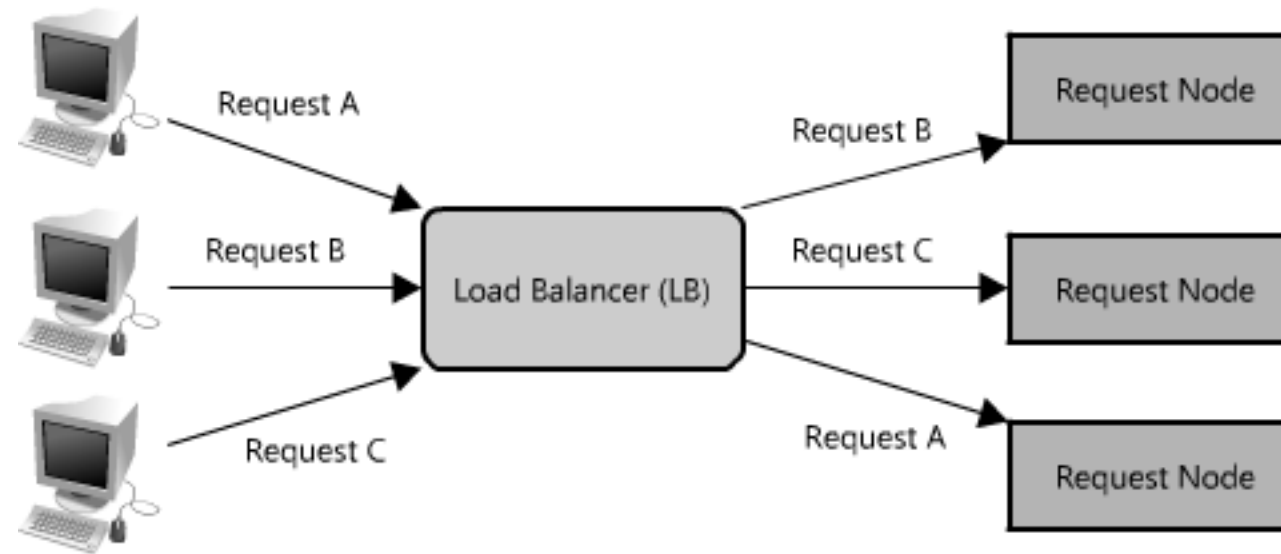
- OldestInstance
- NewestInstance
- OldestLaunchConfiguration
- ClosestToNextInstanceHour
- Default

DEMO

- Termination Policy Review

- EPISODE 9.06
- Load Balancer Concepts

Load Balancing Defined



Load Balancing Categories

- Sender initiated
 - Sender locates best target
- Receiver initiated
 - Receiver selects best target

Static Load Balancing

- Multi-tier application
 - Specific actions are assigned to specific servers/resources
 - Actions always processed on assigned target
 - No scalability

Dynamic Load Balancing

- True load balancing
 - Actions dynamically assigned
 - Scalability is provided
- Used by AWS Elastic Load Balancing (ELB)

Load Balancing Algorithms

- Round Robin
- Randomized
- Centrally Managed
- Threshold-Based

- EPISODE 9.05
- Elastic Load Balancing
(ELB)

ELB Benefits

- Highly available
- Secure
- Flexible
- Monitoring and auditing included
- Elastic
- Hybrid

ELB Types DEMO

- Application load balancer
- Network load balancer
- Classic load balancer

<https://aws.amazon.com/elasticloadbalancing/?nc=sn&loc=0>

Supported Services

- EC2
- ECS
- Auto Scaling
- CloudWatch
- Route 53

DEMO

- ELB Features

DEMO

- Creating a Classic Load Balancer