A CLOUD GURU

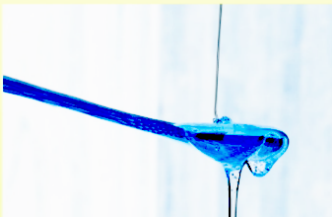
Advanced Load Balancer Theory

Ryan Kroonenburg
AWS COMMUNITY HERO & ALEXA CHAMPION
FOUNDER OF A CLOUD GURU

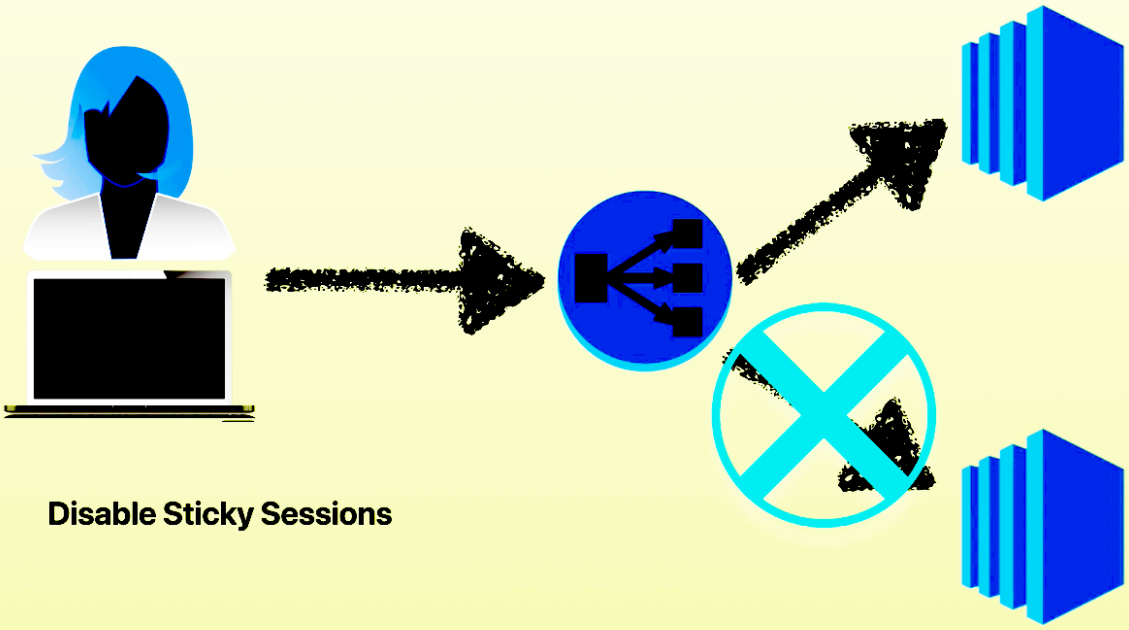
What Are Sticky Sessions?

Classic Load Balancer routes each request independently to the registered EC2 instance with the smallest load.
Sticky sessions allow you to bind a user's session to a specific EC2 instance. This ensures that all requests from the user during the session are sent to the same instance.

You can enable Sticky Sessions for Application Load Balancers as well, but the traffic will be sent at the Target Group Level.

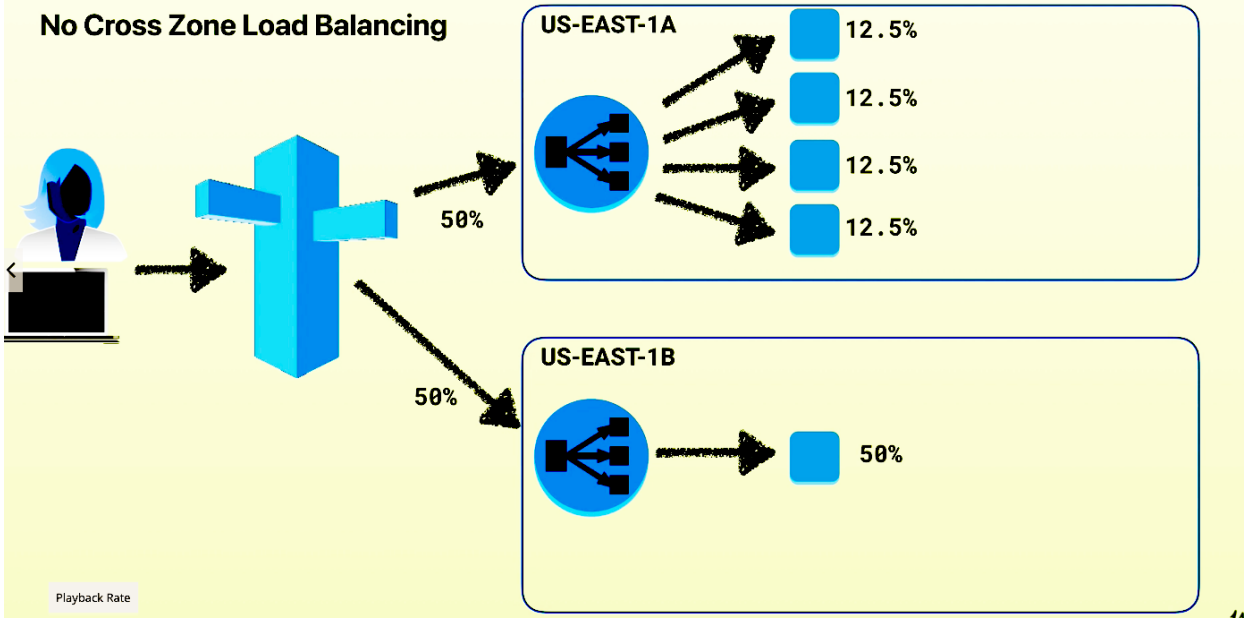


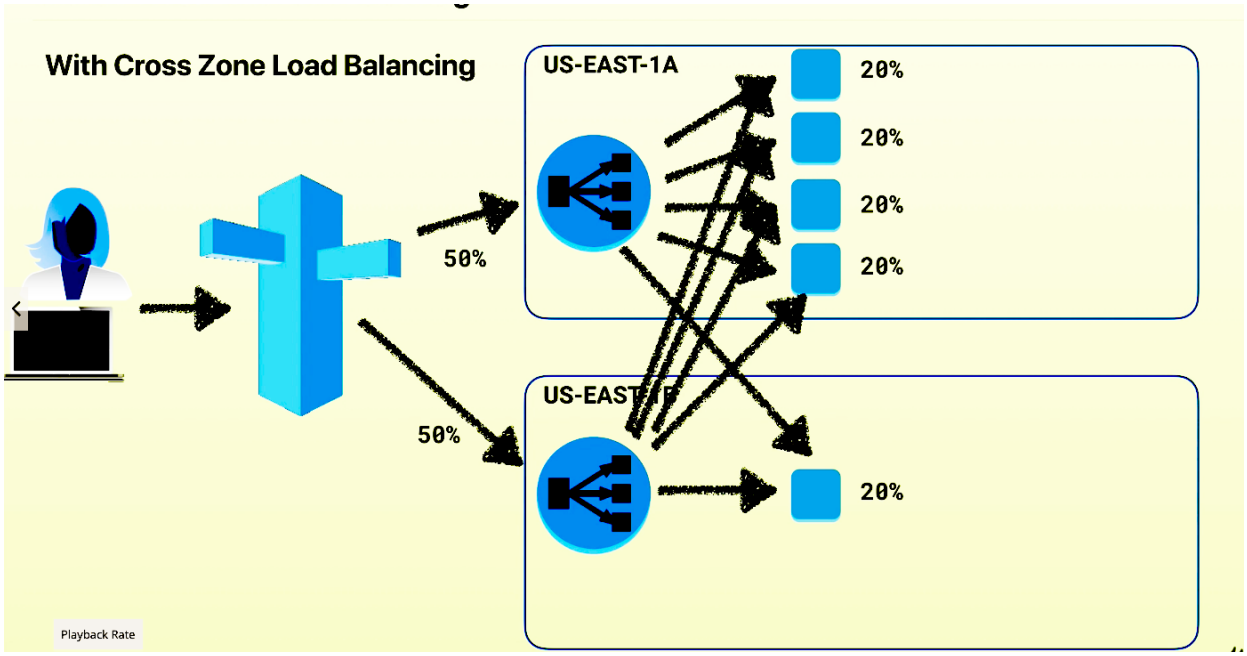
Sticky Sessions

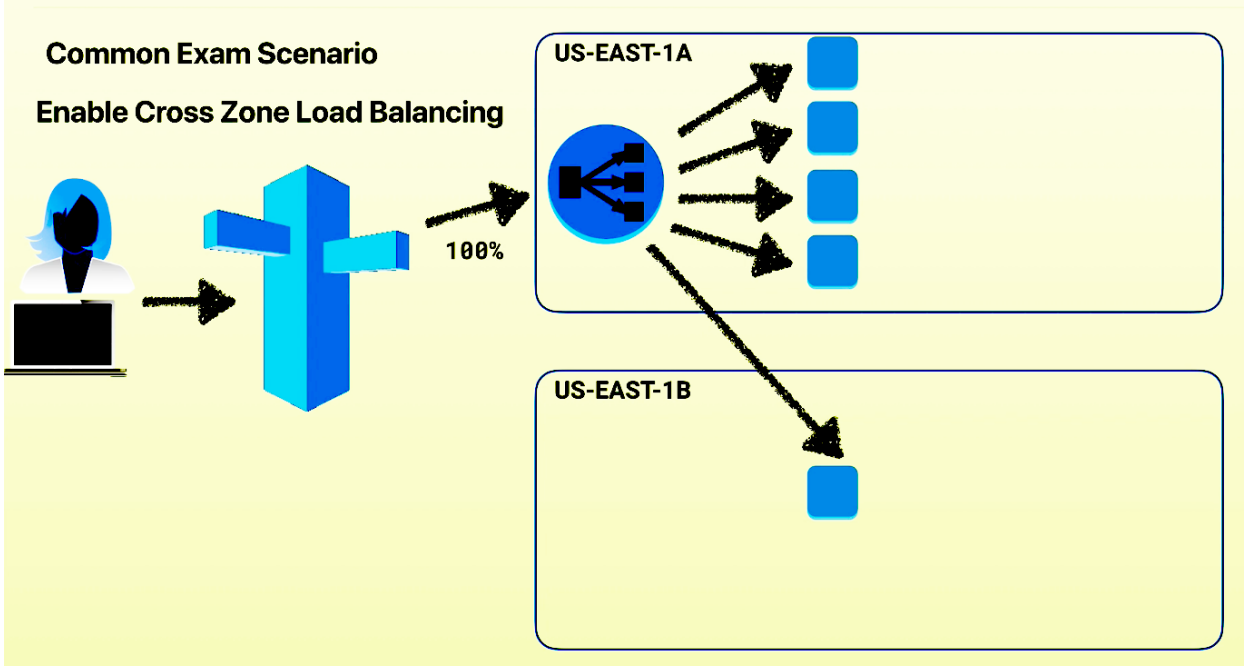


Disable Sticky Sessions

No Cross Zone Load Balancing

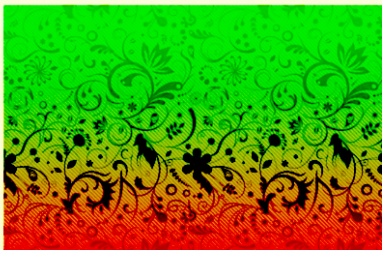



With Cross Zone Load Balancing
US-EAST-1A
20%
20%
20%
20%
50%
50%
US-EAST-1B
20%
Playback Rate


Common Exam Scenario
Enable Cross Zone Load Balancing
US-EAST-1A
100%
US-EAST-1B

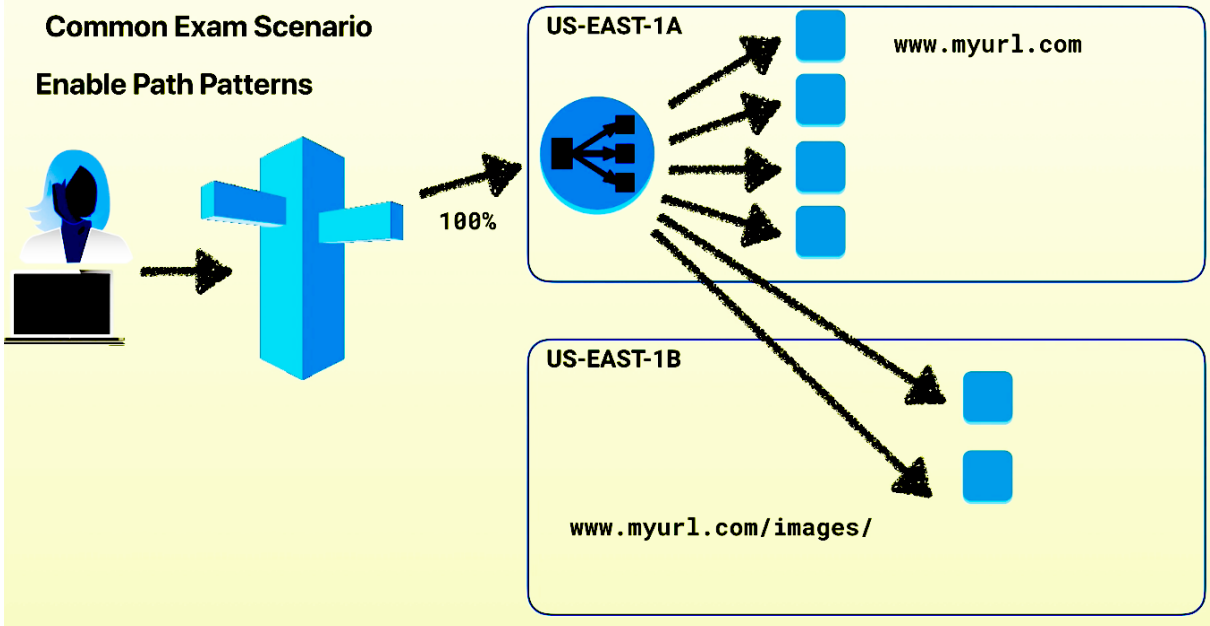
What Are Path Patterns?

You can create a listener with rules to forward requests based on the URL path. This is known as path-based routing. If you are running microservices, you can route traffic to multiple back-end services using path-based routing. For example, you can route general requests to one target group and requests to render images to another target group.



Common Exam Scenario

Enable Path Patterns



Advanced Load Balancer Theory

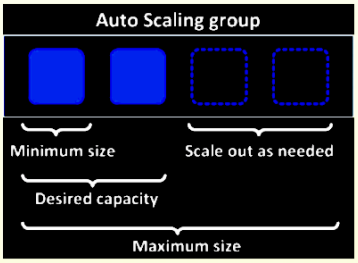
- Sticky Sessions enable your users to stick to the same EC2 instance. Can be useful if you are storing information locally to that instance.
- Cross Zone Load Balancing enables you to load balance across multiple availability zones.
- Path patterns allow you to direct traffic to different EC2 instances based on the URL contained in the request.

A CLOUD GURU

Auto Scaling

Ryan Kroonenburg
AWS COMMUNITY HERO & ALEXA CHAMPION
FOUNDER OF A CLOUD GURU

Auto Scaling Has 3 Components



1. **Groups**
   Logical component. Webserver group or Application group or Database group etc.

2. **Configuration Templates**
   Groups uses a launch template or a launch configuration as a configuration template for its EC2 instances. You can specify information such as the AMI ID, instance type, key pair, security groups, and block device mapping for your instances.

3. **Scaling Options**
   Scaling Options provides several ways for you to scale your Auto Scaling groups. For example, you can configure a group to scale based on the occurrence of specified conditions (dynamic scaling) or on a schedule.

What are my scaling options?

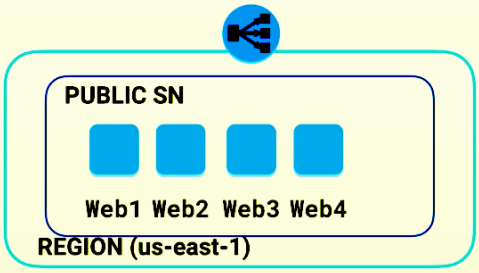
- Maintain current instance levels at all times
- Scale manually
- Scale based on a schedule
- Scale based on demand
- Use predictive scaling

Maintain current instance levels at all times

You can configure your Auto Scaling group to maintain a specified number of running instances at all times.

To maintain the current instance levels, Amazon EC2 Auto Scaling performs a periodic health check on running instances within an Auto Scaling group.

When Amazon EC2 Auto Scaling finds an unhealthy instance, it terminates that instance and launches a new one.



Scale manually

Manual scaling is the most basic way to scale your resources, where you specify only the change in the maximum, minimum, or desired capacity of your Auto Scaling group.

Amazon EC2 Auto Scaling manages the process of creating or terminating instances to maintain the updated capacity.

Scale based on a schedule

Scaling by schedule means that scaling actions are performed automatically as a function of time and date.

This is useful when you know exactly when to increase or decrease the number of instances in your group, simply because the need arises on a predictable schedule.

Scale based on demand

Most Popular Option!!!

A more advanced way to scale your resources - using scaling policies - lets you define parameters that control the scaling process.

For example, let's say that you have a web application that currently runs on two instances and you want the CPU utilization of the Auto Scaling group to stay at around 50 percent when the load on the application changes. This method is useful for scaling in response to changing conditions, when you don't know when those conditions will change. You can set up Amazon EC2 Auto Scaling to respond for you. We will do this in the next lab.

Use predictive scaling

You can also use Amazon EC2 Auto Scaling in combination with AWS Auto Scaling to scale resources across multiple services.

AWS Auto Scaling can help you maintain optimal availability and performance by combining predictive scaling and dynamic scaling (proactive and reactive approaches, respectively) to scale your Amazon EC2 capacity faster.