Amazon EC2 Auto Scaling

If you’ve tried to access a website that wouldn’t load and frequently timed out, the website might have received more requests than it was able to handle. This situation is similar to waiting in a long line at a coffee shop, when there is only one barista present to take orders from customers

Amazon EC2 **Auto Scaling enables you to automatically add or remove Amazon EC2 instances in response to changing application demand**. By automatically scaling your instances in and out as needed, you are able to maintain a greater sense of application availability.

Within Amazon EC2 Auto Scaling, you can use two approaches: dynamic scaling and predictive scaling.

 *Dynamic scaling* responds to changing demand.

 *Predictive scaling* automatically schedules the right number of Amazon EC2 instances based on predicted demand.

Example: Amazon EC2 Auto Scaling

**In the cloud, computing power is a programmatic resource**, so you can take a more flexible approach to the issue of scaling. By adding Amazon EC2 Auto Scaling to an application, you can **add new instances to the application when necessary** and **terminate them when no longer needed.**

Suppose that you are preparing to launch an application on Amazon EC2 instances. When configuring the size of your Auto Scaling group, you might set the minimum number of Amazon EC2 instances at one. This means that at all times, there must be at least one Amazon EC2 instance running.

When you create an **Auto Scaling group**, you can set the minimum number of Amazon EC2 instances. The **minimum capacity** is the number of Amazon EC2 instances that launch immediately after you have created the Auto Scaling group. In this example, the Auto Scaling group has a minimum capacity of one Amazon EC2 instance.

Elastic Load Balancing

**Elastic Load Balancing** (ELB) is the AWS service that automatically **distributes incoming application traffic across multiple resources**, such as Amazon EC2 instances.

A **load balancer acts as a single point of contact for all incoming web traffic** to your Auto Scaling group. This means that as you add or remove Amazon EC2 instances in response to the amount of incoming traffic, these requests route to the load balancer first. Then, the requests spread across multiple resources that will handle them. For example, if you have multiple Amazon EC2 instances, Elastic Load Balancing distributes the workload across the multiple instances so that no single instance has to carry the bulk of it.

It is engineered to address the undifferentiated heavy lifting of load balancing.

**Elastic Load Balancing is a Regional construct**, and we'll explain more of what that means in later videos. But the key value for you is that because it **runs at the Region level** rather than on individual EC2 instances, the service is **automatically highly available** with no additional effort on your part.

**ELB is automatically scalable**. As your traffic grows, ELB is designed to handle the additional throughput with no change to the hourly cost. When your EC2 fleet auto-scales out, as each instance comes online, the auto-scaling service just lets the Elastic Load Balancing service know that it's ready to handle the traffic, and off it goes.