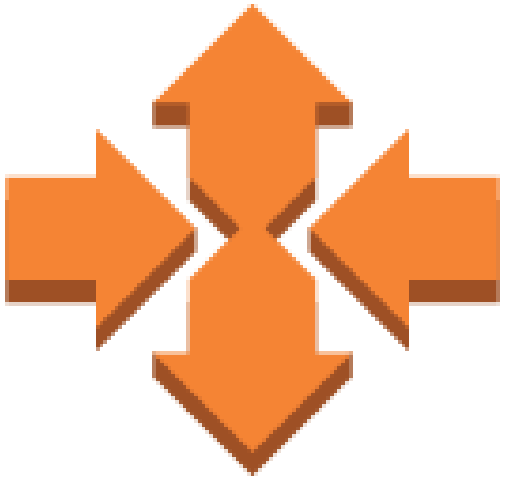
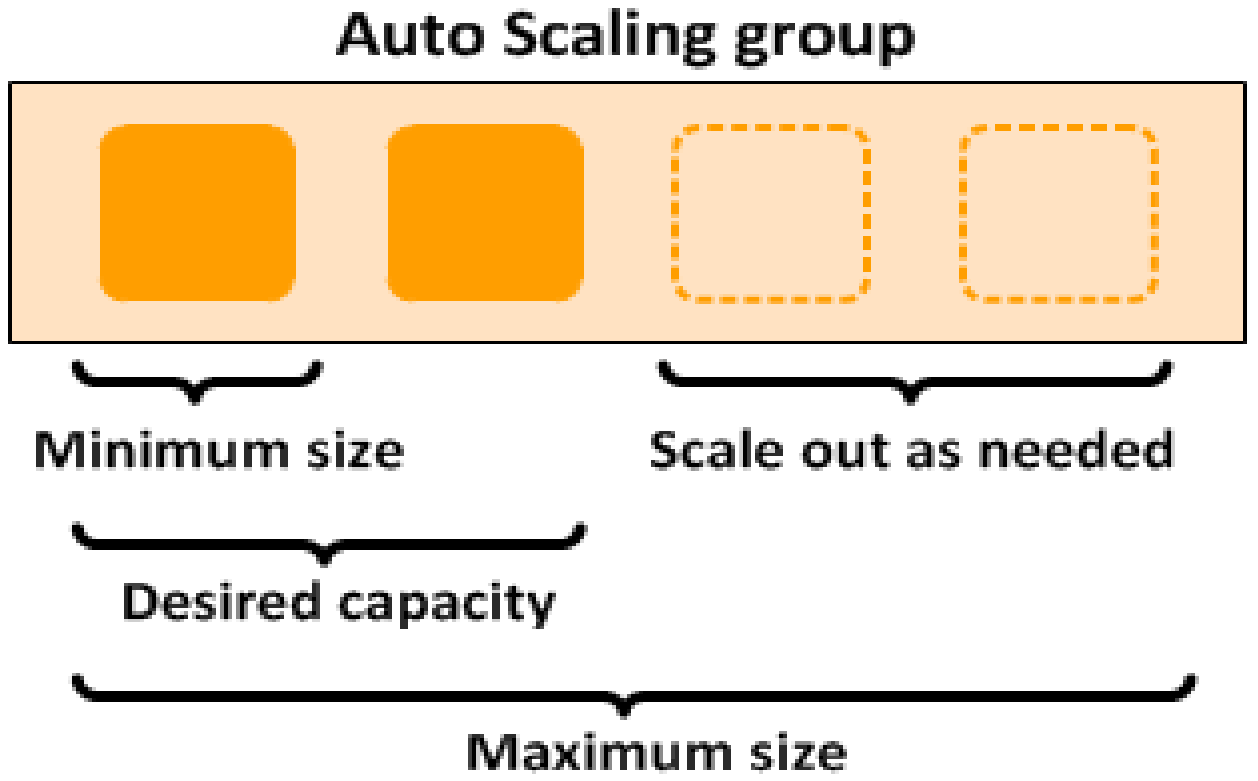


Auto Scaling

What is Autoscaling ?

Amazon EC2 Auto Scaling helps you ensure that you have the correct number of Amazon EC2 instances available to handle the load for your application.

Autoscaling scales up and down a group of servers based on computing or traffic demand by provisioning new services.


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

Auto Scaling Groups

- With AWS Autoscaling, we create collections of EC2 instances, called **Auto Scaling groups (ASG)**.
- You can specify the **minimum number** of instances in each Auto Scaling group, and Amazon EC2 Auto Scaling ensures that your group never goes below this size.
- You can specify the **maximum number** of instances in each Auto Scaling group, and Amazon EC2 Auto Scaling ensures that your group never goes above this size
- If you specify the **desired capacity**, either when you create the group or at any time thereafter, Amazon EC2 Auto Scaling ensures that your group has this many instances.
- If you specify scaling policies, then Amazon EC2 Auto Scaling can launch or terminate instances as demand on your application increases or decreases.

Auto Scaling Components

The key components of Amazon EC2 Auto Scaling.

Groups

- Your EC2 instances are organized in to *groups* so that they can be treated as a logical unit for the purposes of scaling and management. When you create a group, you can specify its **minimum**, **maximum**, and, desired number of EC2 instances

Configuration templates

- Your group 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.

Scaling options

- Amazon EC2 Auto Scaling 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 or Manual

Scaling the Size of Your Auto Scaling Group

Scaling is the ability to increase or decrease the compute capacity of your application. Scaling starts with an event, or scaling action, which instructs an Auto Scaling group to either launch or terminate Amazon EC2 instances.

Scaling Options

- Maintain current instance levels at all times
- Manual scaling
- Scale based on a schedule
- Scale based on demand

Scaling Options

Amazon EC2 Auto Scaling provides several ways for you to scale your Auto Scaling group.

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

Manual scaling

- 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.

Scaling Options

Scale based on a schedule (Schedule Scaling)

- 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 (Dynamic Scaling)

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

Life Cycle Hooks

Lifecycle hooks enable you to perform custom actions by *pausing* instances as an Auto Scaling group launches or terminates them.

When an instance is paused, it remains in a wait state until either you complete the lifecycle action using the complete-lifecycle-action CLI command or CompleteLifecycleAction API action, or the timeout period ends (one hour by default).

Life Cycle Hooks