Autoscaling

- Introduction to Amazon Auto Scaling
- Auto Scaling Components
- Features and Benefits
- Auto Scaling Basic Lifecycle
- Auto Scaling Instance States
- Auto Scaling Limits
- Vertical and Horizontal Scaling
- Pricing

Amazon Auto Scaling

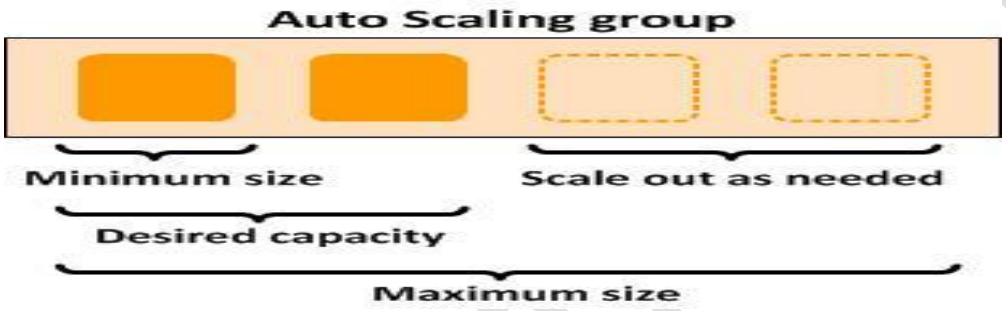
- Auto Scaling helps you maintain application availability and allows you to scale your **Amazon EC2** capacity up or down automatically according to conditions you define.
- You can use Auto Scaling to help ensure that you are running your desired number of **Amazon EC2 instances**.
- Auto Scaling can also automatically **increase** the number of Amazon EC2 instances during demand spikes to maintain performance and **decrease** capacity during lulls to reduce costs.
- Auto Scaling is well suited both to applications that have stable demand patterns or that experience hourly, daily, or weekly variability in usage

What Is Auto Scaling?

- Auto Scaling helps you ensure that you have the correct number of EC2 instances available to handle the load for your application. You create collections of EC2 instances, called **Auto Scaling groups**. You can specify the **minimum** number of instances in each Auto Scaling group, and 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 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, Auto Scaling ensures that your group has this many instances. If you specify scaling policies, then Auto Scaling can **launch** or **terminate** instances as demand on your application increases or decreases.

For example, the following Auto Scaling group has a minimum size of 1 instance, a desired capacity of 2 instances, and a maximum size of 4 instances. The scaling policies that you define adjust the number of instances, within your minimum and maximum number of instances, based on the criteria that you specify.



Auto Scaling Components

Groups

- Your EC2 instances are organized into *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

Launch configurations

- Your group uses a *launch configuration* as a template for its EC2 instances. When you create a launch configuration, you can specify information such as the AMI ID, instance type, key pair, security groups, and block device mapping for your instances.

Scaling plans

- A *scaling plan* tells Auto Scaling when and how to scale. For example, you can base a scaling plan on the occurrence of specified conditions (dynamic scaling) or on a schedule

Maintain your Amazon EC2 instance availability

- Whether you are running one Amazon EC2 instance or thousands, you can use Auto Scaling to detect impaired Amazon EC2 instances and unhealthy applications, and replace the instances without your intervention. This ensures that your application is **getting the compute capacity that you expect**

Automatically Scale Your Amazon EC2 Fleet

- Auto Scaling enables you to follow the **demand curve** for your applications closely, reducing the need to manually provision Amazon EC2 capacity in advance. For example, you can set a condition to add new Amazon EC2 instances in increments to the Auto Scaling group when the average **utilization** of your Amazon EC2 fleet is **high**; and similarly, you can set a condition to **remove** instances in the same increments when **CPU utilization is low**.

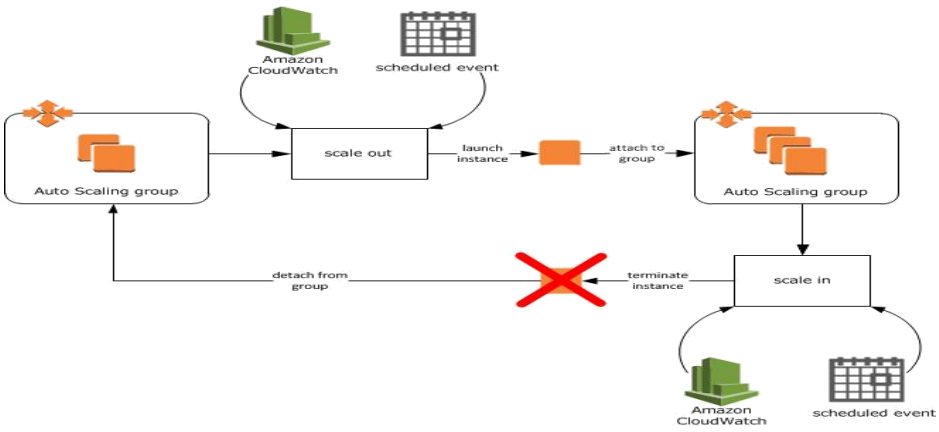
- If you have **predictable load changes**, you can set a schedule through Auto Scaling to plan your scaling activities. You can use Amazon **CloudWatch** to **send alarms** to trigger scaling activities and **Elastic Load Balancing** to help distribute traffic to your instances within Auto Scaling groups. Auto Scaling enables you to run your Amazon EC2 fleet at optimal utilization

Features and Benefits

- Scale out Amazon EC2 instances seamlessly and automatically when demand increases.
- Shed unneeded Amazon EC2 instances automatically and save money when demand subsides.
- Scale dynamically based on your Amazon CloudWatch metrics, or predictably according to a schedule that you define.
- Replace unhealthy or unreachable instances to maintain higher availability of your applications.
- Receive notifications via Amazon Simple Notification Service (Amazon SNS) to be alerted when you use Amazon CloudWatch alarms to initiate Auto Scaling actions, or when Auto Scaling completes an action.
- Run On-Demand or Spot Instances, including those inside your virtual private cloud (VPC) or high performance computing (HPC) clusters.
- **If you're sig ed up for the A azo EC2 service, you're already registered to use Auto Scali g a d c**an begin using the feature via the API or command line interface.

Auto Scaling Basic Lifecycle

The following illustration shows the basic lifecycle of instances within an Auto Scaling group. The Auto Scaling group has a desired capacity of two instances, a CloudWatch alarm that can trigger scaling events, and policies that scale the group at specific dates and times



State	Action
Pending	Installing Software to Pending Instances
	Filling a Cache of Servers
InService	Updating or Modifying Instances in an Auto Scaling Group
	Troubleshooting Instances in an Auto Scaling Group
Terminating	Analyzing an Instance Before Termination
	Retrieving Logs from Terminating Instances

State	Action
Pending	Installing Software to Pending Instances
	Filling a Cache of Servers

InService	Updating or Modifying Instances in an Auto Scaling Group
	Troubleshooting Instances in an Auto Scaling Group
Terminating	Analyzing an Instance Before Termination
	Retrieving Logs from Terminating Instances

* Note that you can attach or detach at most 10 load balancers at a time.

If you reach the default limit for an AWS resource, you can request a limit increase

- **Horizontal scalability** is the ability to increase capacity by connecting multiple hardware or software entities so that they work as a single logical unit
- When servers are clustered, the original server is being scaled out horizontally. If a cluster requires more resources to improve performance and provide high availability (HA), an administrator can scale out by adding more servers to the cluster.
- An important advantage of horizontal scalability is that it can provide administrators with the ability to increase capacity on the fly.
- **Vertical scalability,** on the other hand, increases capacity by adding more resources, such as more memory or an additional CPU, to a machine. Scaling vertically, which is also called scaling up, usually requires downtime while new resources are being added and has limits that are defined by hardware

Pricing

Auto Scaling Pricing

- Auto Scaling is enabled by Amazon CloudWatch and carries **no additional fees**.
- Amazon EC2 and Amazon CloudWatch service fees apply and are billed separately. Partial hours are billed as full hours