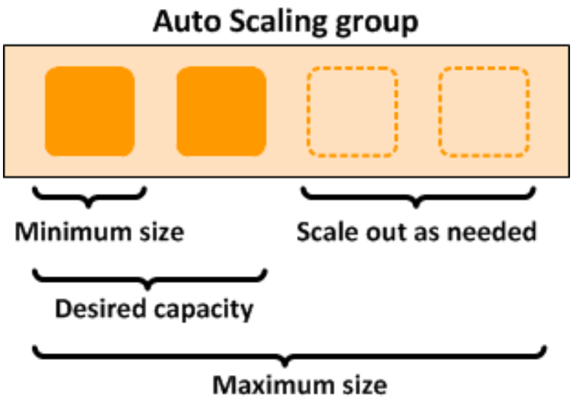
Auto Scaling

Ponnam Phani Krishna
PONNAM.PHANI@GMAIL.COM

Auto Scaling

AWS Auto Scaling monitors your applications and automatically adjusts capacity to maintain steady, predictable performance at the lowest possible cost. Using AWS Auto Scaling, it's easy to setup application scaling for multiple resources across multiple services in minutes.

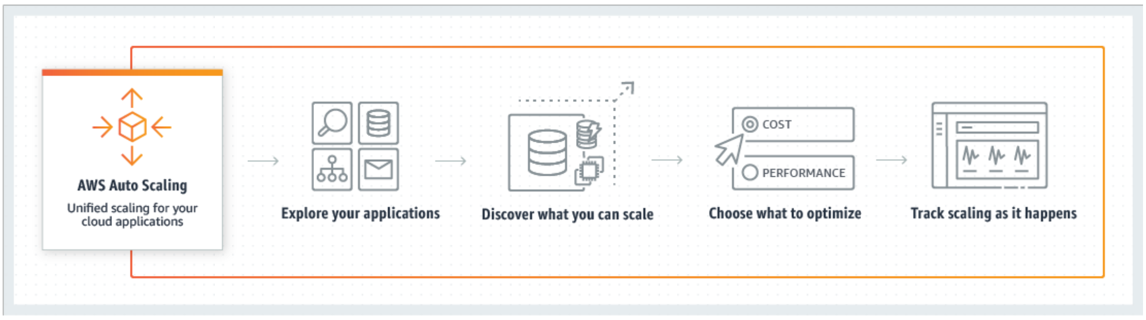
Amazon EC2 Auto Scaling helps you ensure that you have the correct number of Amazon EC2 instances available to handle the load for your application. You create collections of EC2 instances, called *Auto Scaling groups*. You can specify the minimum number of instances in each Auto Scaling group, and Amazon 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.



Benefits:

- Setup Scaling Quickly
- Improve fault Tolerance
- Increase Application Availability
- Automatically Maintain performance
- Lower Costs

How It Works:


AWS Auto Scaling
Unified scaling for your cloud applications
Explore your applications
Discover what you can scale
COST
PERFORMANCE
Choose what to optimize
Track scaling as it happens

Auto Scaling Components:

The following are the key components of amazon EC2 Auto Scaling:

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.

Configuration Templates: Your group uses a *launch template*, or a *launch configuration* (not recommended, offers fewer features), 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.

An auto scaling group starts by launching enough instances to meet its desired capacity. It maintains the number of instances by performing periodic health checks on the instances in the group. The auto scaling group continues to maintain a fixed number of instances even if an instance becomes unhealthy.

You can use scaling policies to increase or decrease the number of instances in your group dynamically to meet changing conditions. When the scaling policy is in effect, the Auto Scaling group adjusts the desired capacity of the group, between the minimum and maximum capacity values that you specify and launches or terminates the instances as needed. You can also scale on a schedule.

When instances are launched, if you specified multiple Availability Zones, the desired capacity is distributed across these Availability Zones. If a scaling action occurs, Amazon EC2 Auto Scaling automatically maintains balance across all of the Availability Zones that you specify.

Scale-in: Removes an existing instance from the auto scaling group

Scale-Out: Add a new instance to the auto scaling group.

We integrate Load Balancer with autoscaling group to distribute the load across all EC2 instances in the Auto Scaling group.

Auto Scaling Lab:

ToDO List 1:

1. Create an AMI to use with Auto scaling (a simple webserver)
2. Create an Application Load Balancer and target group with no Targets registered
3. Create Launch Configuration for auto scaling
4. Create Auto scaling group and include the created Target group created.
5. Test the autoscaling by simulating the conditions created in the auto scaling group.