Course and Labs

AWS Autoscaling

Mohanraj Shanmugam

Autoscaling

- 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.

Autoscaling

- 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.

Benefits of Autoscaling

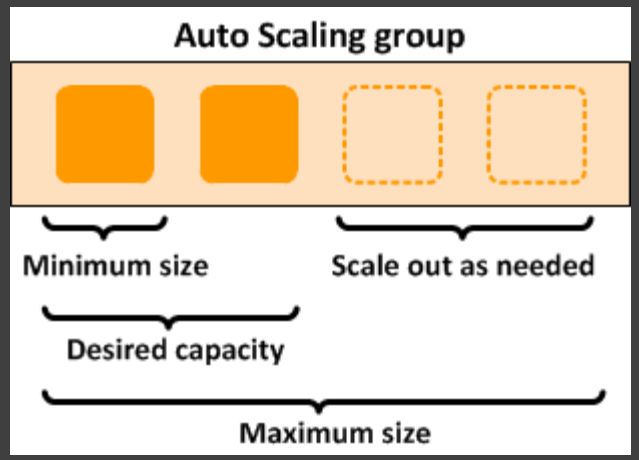
- Maintain your Amazon EC2 instance availability
  - Whether you are running one Amazon EC2 instance or thousands, you can use Auto Scaling to detect impaired Amazon EC2 instances and unhealthy applications, and replace the instances without your intervention.
  - This ensures that your application is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

Benefits of Autoscaling

- 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.

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.



Auto Scaling Components

- **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.

Auto Scaling Components

- **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.