ARYATECHNO

# AWS Cloud EC2 Scaling

**Topics :** AWS
**Written on** December 01, 2023

Scaling in Amazon EC2 (Elastic Compute Cloud) refers to adjusting the compute capacity of your instances to accommodate changes in demand or workload. There are two primary types of scaling: horizontal scaling (adding or removing instances) and vertical scaling (resizing instances). AWS provides several services and features to help you scale your EC2 instances efficiently:

1. **Auto Scaling:**
    - **Horizontal Scaling:** AWS Auto Scaling allows you to automatically adjust the number of EC2 instances in a group based on predefined conditions. You can set up scaling policies to increase or decrease the number of instances in response to metrics like CPU utilization or network traffic.
    - **Integration with Load Balancers:** Auto Scaling can work in conjunction with Elastic Load Balancing (ELB) to distribute incoming traffic across multiple instances. This helps improve fault tolerance and ensures that instances can be added or removed dynamically as demand changes.
2. **Manual Scaling:**
    - You can manually scale your EC2 instances by launching or terminating instances based on your current requirements. This can be done through the AWS Management Console, AWS Command Line Interface (CLI), or SDKs.
3. **Elastic Load Balancing (ELB):**
    - ELB distributes incoming application traffic across multiple EC2 instances to ensure even load distribution. This improves fault tolerance and enables better responsiveness to changing workloads.
    - ELB can be configured to automatically register and deregister instances based on health checks, allowing for dynamic scaling.
4. **Amazon EC2 Auto Scaling Groups:**
    - Auto Scaling Groups allow you to group a collection of EC2 instances and define scaling policies for the group. Instances in the group can be automatically launched or terminated based on demand.
    - You can configure Auto Scaling Groups to use multiple Availability Zones for increased fault tolerance.
5. **Spot Instances for Bursting:**

  - Spot Instances allow you to take advantage of spare EC2 capacity at a lower cost. This can be useful for burstable workloads that can tolerate interruptions, as Spot Instances may be terminated if the capacity is needed elsewhere.

6. **Reserved Instances:**

   - Reserved Instances provide cost savings over On-Demand pricing in exchange for a commitment to a one- or three-year term. They are particularly useful for applications with predictable workloads.

7. **AWS Auto Scaling Policies:**

   - Auto Scaling policies define the conditions for scaling actions, including scaling out (adding instances) and scaling in (removing instances). Policies can be based on metrics such as CPU utilization, network traffic, or custom metrics.

8. **Scheduled Scaling:**

   - Scheduled Scaling allows you to plan ahead for predictable changes in demand by defining scheduled actions to increase or decrease the desired capacity of your Auto Scaling Group.