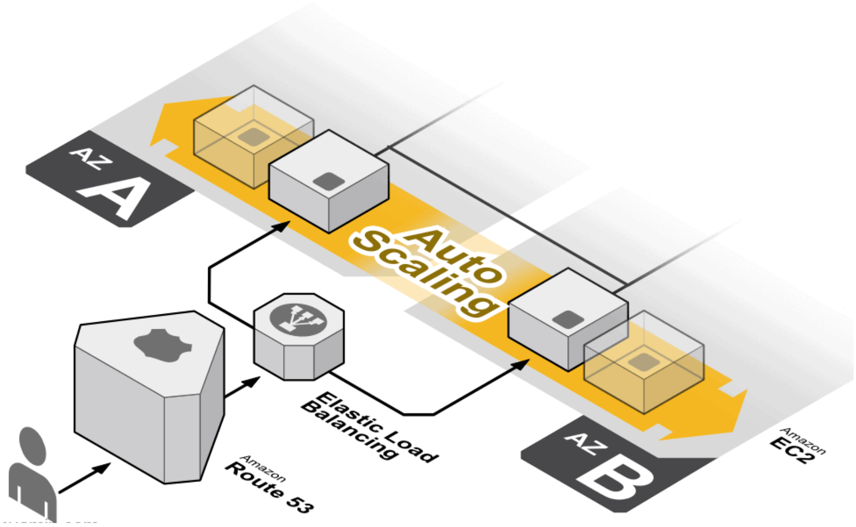
AWS Auto-Scaling: Overview, Features, Benefits & Configuration

AWS Auto-Scaling is a mechanism that automatically permits you to increase or decrease your resources to meet demand based on custom-defined metrics and thresholds. Through Auto-scaling, it's simple to set up application scaling for multiple resources across multiple services in minutes.

What is Auto-Scaling?

AWS Auto-Scaling monitors your applications continuously and adjusts the capacity automatically to take care of steady, predictable performance at the lowest possible cost. It gives the ability to ensure a correct number of EC2 instances are running every time to handle the load of the application. It helps to accomplish better availability and cost management.



Features of Auto-Scaling

- **Unified Scaling:** Through Auto-Scaling, we can configure automatic scaling for all of the scalable resources powering your application from one unified interface, including:
  - **Amazon EC2:** Launch or terminate EC2 instances in an AWS Auto Scaling group.
  - **Amazon EC2 Spot Fleets:** Launch or terminate EC2 instances from an Amazon EC2 Spot Fleet, or replace the instances automatically which get interrupted for price or capacity reasons.
  - **Amazon ECS:** Adjust ECS service desired count up or right down to answer load variations.
  - **Amazon DynamoDB:** Allows a DynamoDB table or a worldwide secondary index to extend its provisioned read and write capacity to handle sudden increases in traffic without throttling.
  - **Amazon Aurora:** Dynamically adjust the amount of Aurora Read Replicas provisioned for an Aurora DB cluster to handle sudden increases in the active connections or workload.
- **Automatic Resource Discovery:** AWS Auto-Scaling scans your environment and finds out the scalable cloud resources automatically underlying your application, so we don't need to manually identify these resources.

- **Built-in scaling strategies:** Through AWS Auto-Scaling, we can select one of three predefined optimization strategies designed to optimize performance, optimize costs, or balance the two. We can also set our own target resource utilization.
- **Predictive Scaling:** It predicts future traffic, including continuously occurring spikes, and provisions the right number of EC2 instances beforehand of predicted changes.
- **Fully managed:** It automatically creates target-tracking scaling policies for all of the resources in our scaling plan, using our selected scaling strategy to line the target values for every metric.
- **Smart scaling policies:** It continually calculates the acceptable scaling adjustments and immediately adds and removes capacity as required to stay your metrics on target.

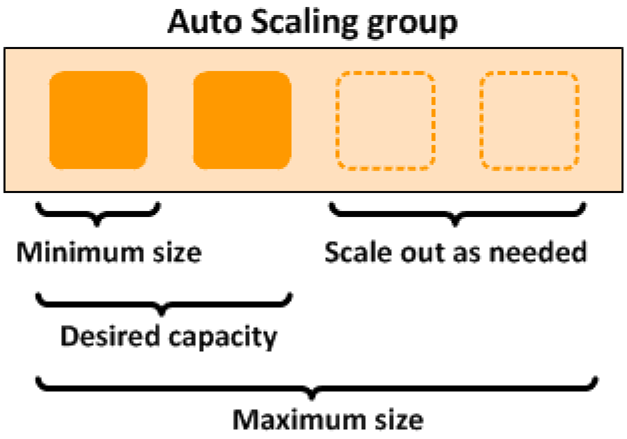
Benefits of Auto-Scaling

- **Setup Scaling Quickly:** It allows you to set target utilization levels for multiple resources during a single, intuitive interface. we can quickly see the typical utilization of all of your scalable resources without having to navigate to other consoles.
- **Make Smart Scaling Decisions:** It enables you to create scaling plans that automate how groups of varied resources answer changes in demand.
- **Automatically Maintain Performance:** It maintains flawless application performance and availability, even when workloads are periodic, unpredictable, or continuously changing. It continually monitors your applications to make sure that they're operating at your required performance levels.
- **Pay Only For What You Need:** It helps you to optimize your utilization and cost efficiencies when consuming AWS services so you simply pay for the resources you really need.

Elastic:
Automatically adapt capacity to demand
Reliable:
Counteract failures of instances or AZs
Customizable:
With bootstrapping & lifecycle hooks

EC2 Auto-Scaling

Amazon EC2 Auto-Scaling helps us to take care of our application availability and allows us to add or remove EC2 instances automatically according to the conditions defined. With the help of fleet management features of EC2 Auto-Scaling to maintain the health and availability of our fleet. We can also use the dynamic and predictive scaling features of EC2 Auto-Scaling to feature or remove EC2 instances.

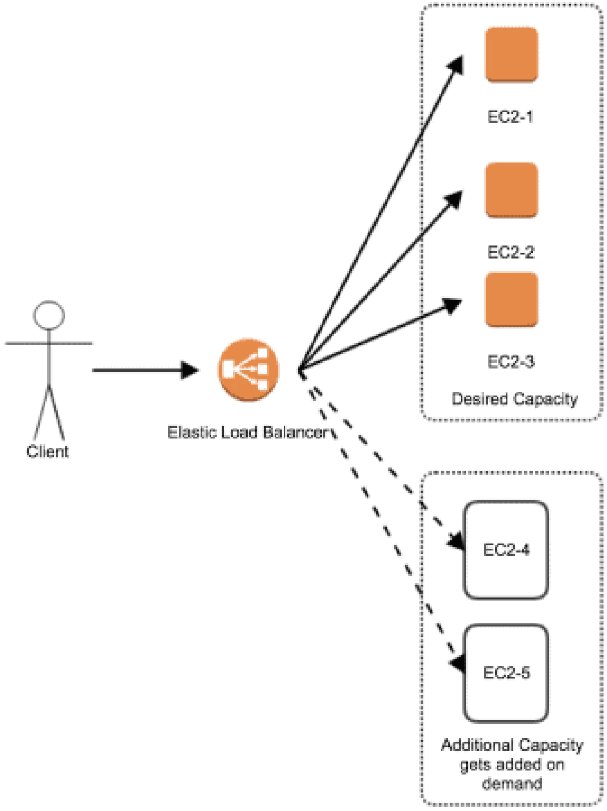


Application Auto-Scaling

Application Auto-Scaling is a web service for developers and system administrators who need a solution for consequently scaling their scalable resources for individual AWS services past Amazon EC2. It enables us to configure automatic scaling for the subsequent resources:

- Amazon ECS services
- Spot Fleet requests
- Amazon EMR clusters
- AppStream 2.0 fleets
- DynamoDB tables and global secondary indexes
- Aurora replicas
- Amazon SageMaker endpoint variants

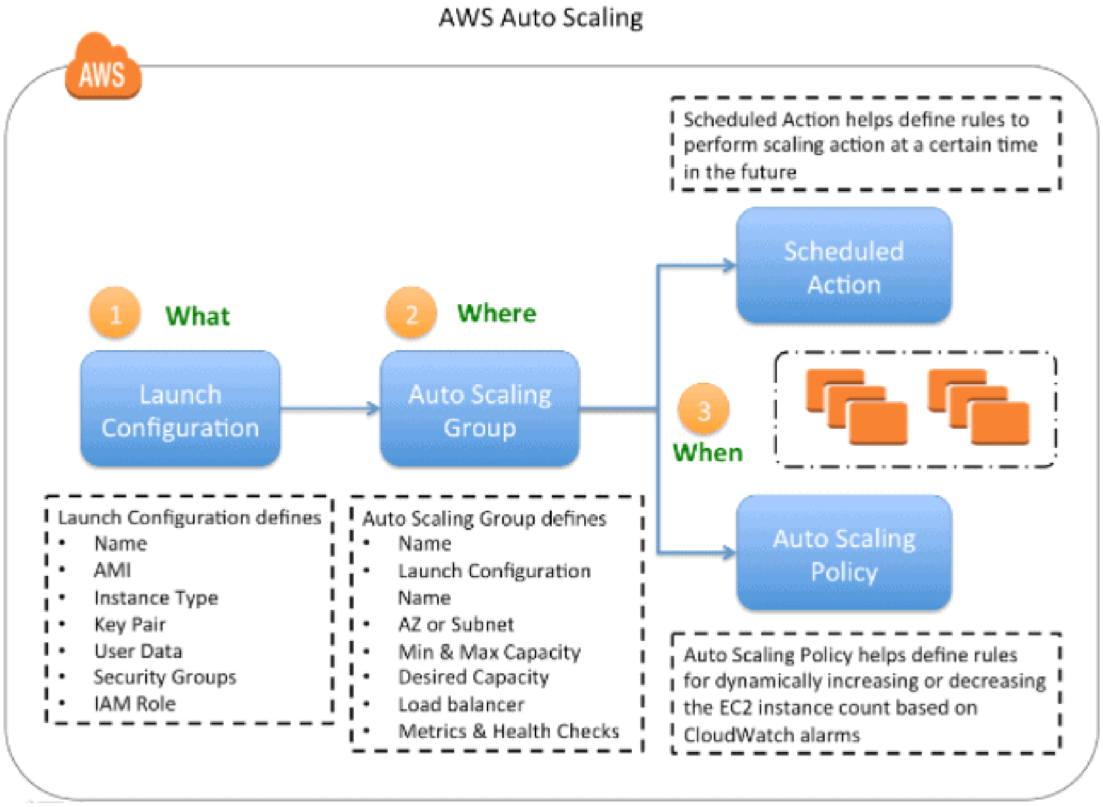
- Custom resources provided by your applications



Auto-Scaling Policy Types

- **Target tracking scaling:** Increase or decrease the present capacity of the group based on a target value for a selected metric.
- **Step scaling:** Increase or decrease the present capacity of the group based on a set of scaling adjustments, known as step adjustments, that change based on the size of the alarm breach.
- **Simple scaling:** Increase or decrease the present capacity of the group based on a single scaling adjustment.

Auto-Scaling Configuration



Scheduled Action helps define rules to perform scaling action at a certain time in the future

Launch Configuration defines
- Name
- AMI
- Instance Type
- Key Pair
- User Data
- Security Groups
- IAM Role

Auto Scaling Group defines
- Name
- Launch Configuration Name
- AZ or Subnet
- Min & Max Capacity
- Desired Capacity
- Load balancer
- Metrics & Health Checks

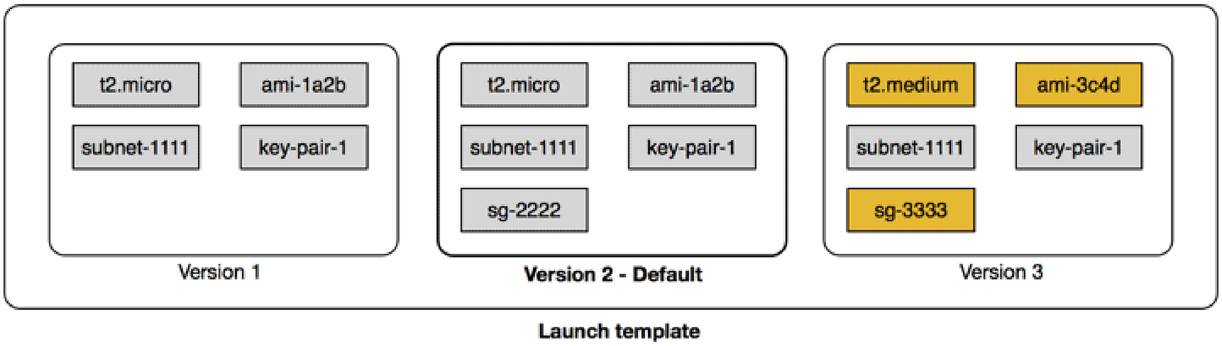
Auto Scaling Policy helps define rules for dynamically increasing or decreasing the EC2 instance count based on CloudWatch alarms

Launch Configuration

- Launch Configuration is an **instance configuration** template used by the Auto-Scaling group to launch Amazon EC2 instances.
- A launch configuration is similar to an EC2 configuration and **involves the selection of the Amazon Machine Image (AMI)**, the instance type, a key pair, one or more security groups, and a block device mapping.
- Launch configurations are often related to **multiple Auto-Scaling groups.**
- Launch configuration **can't be modified** after creation.
- Basic or detailed monitoring for the instances within the Auto-Scaling group is often enabled when a launch configuration is created.
- Basic **monitoring is enabled by default** when you create the launch configuration utilizing the AWS Management Console, and detailed monitoring is permitted when you create the launch configuration using the AWS CLI or an API.

Launch Template

- A Launch Template identical to a launch configuration, with additional features.
- Launch Template **allows multiple versions** of a template to be defined.
- With versioning, a subset of the complete set of parameters is often created and then reused to create other templates or template versions.
- Launch Template enables the selection of both Spot and On-Demand Instances or multiple instance types.
- Launch templates support EC2 Dedicated Hosts. Dedicated Hosts are physical servers with EC2 instance capacity that are dedicated to your use.



Auto-Scaling Group

- An auto-Scaling group is a group of Amazon EC2 instances that Auto-Scaling Manages. While creating an Auto-Scaling group, you must first specify either the launch configuration or the Launch template you created.
- You must also specify how many **running instances you want Autoscaling** to provision and maintain using the launch configuration or template you have created.
- Specifically, you've got to specify the minimum and maximum size of the Auto-Scaling group.
- You also have the option to set the desired number of instances you want autoscaling to provision and maintained.