AWS Auto Scaling Cheat Sheet

by Veronique Robitaille (v@indalit.com)

What is Auto Scaling?

A method by which you can have instances added/removed on the fly.

Configuration Elements of Auto Scaling

Minimum Size

- Minimum number of instances that should always be running.

Maximum Size

- Maximum number of running instances.

Desired Capacity

- Number of instances that should be running when there are no scaling activites.

Characteristics of Auto Scaling

Lifecycle Hooks

- They perform custom actions at launch and termination of an instance.

Termination Protection

- Instances can be protected from a scale-in activity to keep them running.

Automatic Balancing of instances

- If you have more than one AZ then Auto Scaling will automatically balance them.

Monitoring

1) Health checks that verify the status of the instances. This is done by EC2 status checks, ELB & custom health checks.

2) CloudWatch Metrics to monitor the instances and tell Auto Scaling when an instance is unhealthy.

3) CloudWatch Events which monitors specific events like launching or terminating an instance and when Lifecycle Actions occur.

Limits

Launch Configurations per region	100
Auto Scaling Groups per region	20
Scaling Policies per group	50
Scheduled Actions per group	125
Lifecycle Hooks per group	50
Step Adjustments per Scaling Policy	20
Launch Configuration per group	1

Types of Auto Scaling

Fleet Management	Keeps a fixed number of instances running. Does not react to spikes in traffic.
Manual Scaling	Change manually the number of desired instances.
Schedule Scaling	Scaling-in and scaling-out is configured at specific times and dates.
Dynamic Scaling	Scaling-in and scaling-out happens based on configured rules called policies.
- *Simple Scaling Policy*	Uses a metric like CPU utlisation to scale.
- *Step Scaling Policy*	Like Simple Scaling but has multiple steps to scale. These steps are called Step Adjustments.
- *Target Tracking Scaling Policy*	Keeps instance at a fixed level using, for example, CPU utilization.

Components of Auto Scaling

Launch Configuration	Template used to launch a new instance. Is composed of the instance type, OS, userdata... Cannot make modifications to one.
Auto Scaling Group	Group of instances that share a Launch Configuration and scale based on policies.
Scaling Policy	Rule or rules that start a scaling activity (adding or removing instances from an Auto Scaling Group). Can have more than one policy per Auto Scaling Group.

- When more than one Scaling Policy for an Auto Scaling Group, Auto Scaling will choose a policy with the highest impact on the number of instances.

Custom Termination Policy

First, it will make sure instances are balanced between AZs If not, then it will choose the AZ with the most instances.

Then choose from the following:

- OldestInstance
- NewestInstance
- OldestLaunchConfiguration
- Default AWS Termination Policy

Pricing

This service is part of EC2 and is provided for free. You pay for the instances you use.