AutoScaling in AWS

By Bhupinder Rajput

AutoScaling

- Creating group of EC2 instances that can scale up or down depending on conditions we set.
- Scale out means increasing and Scale in means reducing
- Enable elasticity by scaling horizontally (same type and size scaling) through adding or terminating EC2 instances.
- Autoscaling ensures that we have the right number of AWS EC2 instances for our needs at all time.
- Autoscaling helps us save cost by cutting down the number of EC2 instances when not needed and scaling out to add more instances only when it is required.

Components of AutoScaling

Launch Configuration

- Like instance type, AMI, keypair, Security Group.

Autoscaling Group

- Group name, Group Size, VPC, Subnet, Health Check Period.

3.Scaling Policy

- Metric type(Like CPU Utilization),Target value.

Balancing

If autoscaling finds that the number of EC2 instances launched by ASG into subject AZ's is not balanced(EC2 instances are or not evenly distributed). Autoscaling do rebalancing activity by itself.

Autoscaling(AS) tries to balance the instances distribution across AZ's.

While Rebalancing, ASG launches new EC2 instances where there are less EC2 at present, and then terminates the instances from the AZ, that had more instances.

What causes Imbalancing of EC2

- If we add or remove same subnet/AZ from ASG.
- If we manually request for ec2 termination from our ASG.
- An AZ that did not have enough ec2 capacity now has enough capacity and it is one of our ASG AZ.

Attaching

We can attach a running EC2 instances to an autoscaling group by using AWS console or CLI, if the below conditions are met:

- Instances must be in running state (not stopped or terminated).
- AMI used to launch the EC2 still exist.
- Instance is not part of another ASG.
- Instance is in the same AZ of the same group.
- If the existing EC2 instances under the autoscaling group, plus the one to be needed, exceed the max capacity of the autoscaling group, the request will fail, EC2 instance would not be added.

Detaching

We can remove EC2 instances from an ASG using AWS console or CLI.

We can then manage the detached instance independently or attach it to another ASG.

When we detach an instance, we have the option to decrement the ASG desired capacity.

If we do not, the autoscaling group will launch another instances to replace the one detached.

When we delete an ASG, its parameters like max, min and desired capacity are all set to 0. Hence,it terminate all its EC2 instances.

If we want to keep the EC2 Istances and manage them independently, we can manually detach them first, then delete the ASG.

Working of Load Balancer

How load balancer work with ASG:

- We can attach one or more elastic load balancer to our autoscaling group.
- The elastic load balancer must be in the same region as the ASG.
- Once we do this, any EC2 instance existing or added to the ASG will be automatically registered with the ASG defined ELB.
- We do not need to register those instances manually on the ASG defined ELB.
- Instances and The ELB must be in the same VPC.

Health Check

ASG classifies its ec2 isntances health status as either healthy or unhealthy.

By default, AS uses EC2 status checks only to determine the health status of an instance.

When we have one or more ELB defined with the ASG, we can configure AS to use both the EC2 health check and the ELB health check to determine the instances health check.

Health check grace period is 300 sec by default.

If we set zero in grace period, the instance health is checked once it is in service.

Untill the grace period timer expires, any unhealthy status reported by EC2 status checks, or the ELB attached to the autoscaling group, will not be acted upon.

After grace period expires, ASG would consider an instances unhealthy in any of the following cases :

- EC2 status check report to ASG an instance status other than running.
- If ELB health check are configured to be used by the AS, the if the ELB report the instances as ‘out of service’.
- Unlike AZ rebalancing, termination of unhealthy instances happen first, then AS attempt to launch new instance to replace the ones terminated.
- Elastic IP and EBS volumes gets detached from the terminated instances, we need to manually attach them to the new instances.

Sending Notification

On four situation, ASG sends a SNS email Notification

- An instance is launched.
- An instance is terminated.
- An instance fails to launch.
- An instance fails to terminate.

Merging AutoScaling Groups

Can only be done from the CLI (not AWS console).

We can merge multiple, single AZ ASG into single, one multi-AZ ASG.

Scale out means launching more EC2 instances.

Scale in means terminating one or more instances by scaling policy.

It is always recommended to create a scale-in event for each scale-out event we create.

Launch config cant be edited,only copied or deleted.

Monitoring

AWS EC2 services sends EC2 metrics to cloudwatch about the ASG instances :

- Basic monitoring (Every 300 secs enabled by default & free of cost).
- We can enable detailed (every 60 sec-chargeable).

When the launch config is done by AWS CLI, detailed monitoring for EC2 instances is enabled by default.

Standby State

We can manually move an instance from an ASG and put it in standby state.

Instances in standby state are still managed by autoscaling.

Instances in standby state are charged as normal in service instances.

They do not count towards available EC2 instances for workload/APP use.

AS does not perform health check instances in standby state.

Auto-Scaling Policies

Manual or Dynamic

- Define how much we want to scale based on defined conditions.
- ASG uses alarms and polices to determine scaling.
- For simple or step scaling, a scaling adjustment can't change the capacity of the group above the max group size or below the min group size.

Predictive/Scheduled Scaling

- It looks at historic pattern and forecast them into the future to schedule change in the no of EC2 instances. It uses machine learning model to forecast daily and weekly pattern.

Target Tracking Policies

- Increase or decrease the current capacity of the group based on a target value for a specific metric. This is similar to the way that our thermostat maintain the temp of our home.

Step Scaling

- Increase or decrease the current capacity of the group based on a set of scaling adjustment, known as step adjustment, that vary based on the size of the alarm reach.
- Does not support/wait for a cool-down times.
- Support a warm-up timer time taken by newly launch instance to be ready and contribute to the watched metric.

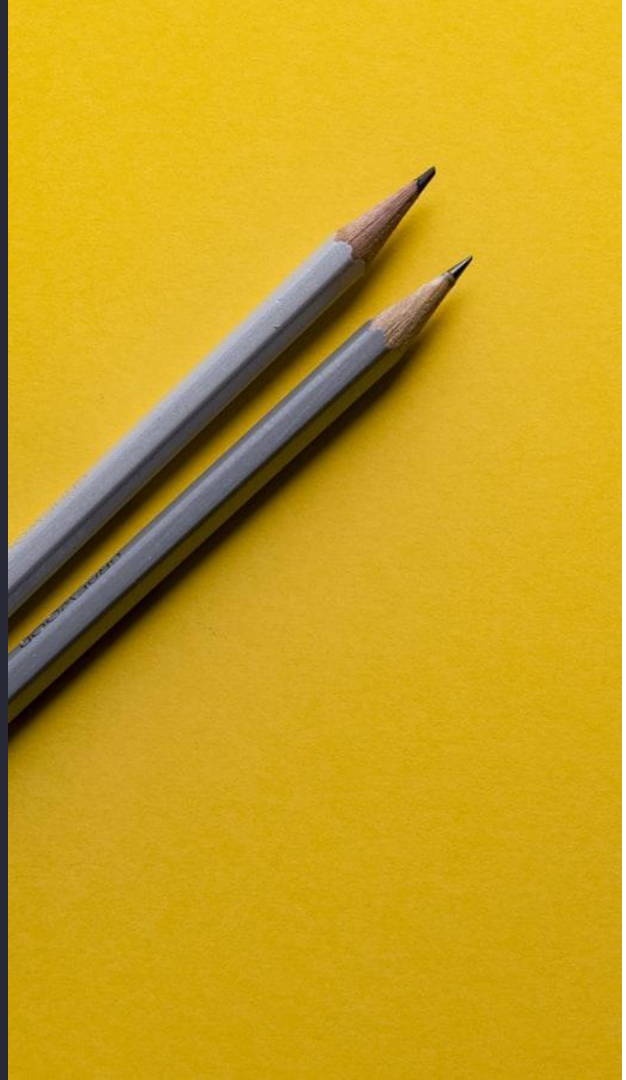
Auto-Scaling Policies (Contd...)

Simple Scaling

- Single adjustment (up or down) in response to an alarm (cooldown timer)-300 sec.

Scheduled Scaling

- Use for predictable change.
- We need to configure a schedule action for a scale out at a specific date/time and to a required capacity.
- A schedule action must have a unique date/time.
- We cannot configure two scheduled activities at the same time/date.



Thanks!

Any questions?

You can find me at:

technical Guftgu

info@technicalguftgu.in