Step Scaling vs Simple Scaling Policies in Amazon EC2

Amazon's EC2 Auto Scaling provides an effective way to ensure that your infrastructure is able to dynamically respond to changing user demands. For example, to accommodate a sudden traffic increase on your web application, you can set your Auto Scaling group to automatically add more instances. And when traffic is low, have it automatically reduce the number of instances. This is a cost-effective solution since it only provisions EC2 instances when you need them. EC2 Auto Scaling provides you with several dynamic scaling policies to control the scale-in and scale-out events.

In this article, we'll discuss the differences between a simple scaling policy and step scaling policy. And we'll show you how to create an Auto Scaling group with step scaling policy applied.

Simple Scaling

Simple scaling relies on a metric as a basis for scaling. For example, you can set a CloudWatch alarm to have a CPU Utilization threshold of 80%, and then set the scaling policy to add 20% more capacity to your Auto Scaling group by launching new instances. Accordingly, you can also set a CloudWatch alarm to have a CPU utilization threshold of 30%. When the threshold is met, the Auto Scaling group will remove 20% of its capacity by terminating EC2 instances.

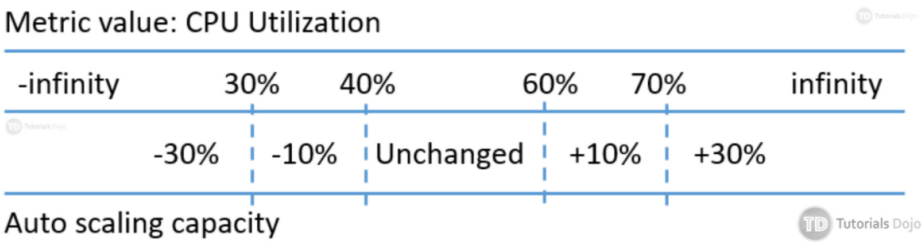
When EC2 Auto Scaling was first introduced, this was the only scaling policy supported. It does not provide any fine-grained control to scaling in and scaling out.

Step Scaling

Step Scaling further improves the features of simple scaling. Step scaling applies "step adjustments" which means you can set multiple actions to vary the scaling depending on the size of the alarm breach.

When a scaling event happens on simple scaling, the policy must wait for the health checks to complete and the cooldown to expire before responding to an additional alarm. This causes a delay in increasing capacity especially when there is a sudden surge of traffic on your application. With step scaling, the policy can continue to respond to additional alarms even in the middle of the scaling event.

Here is an example that shows how step scaling works:



In this example, the Auto Scaling group maintains its size when the CPU utilization is between 40% and 60%. When the CPU utilization is greater than or equal to 60% but less than 70%, the Auto Scaling group increases its capacity by an additional 10%. When the utilization is greater than 70%, another step in scaling is done and the capacity is increased by an additional 30%. On the other hand, when the overall CPU utilization is less than or equal to 40% but greater than 30%, the Auto Scaling group decreases the capacity by 10%. And if utilization

equal to 40% but greater than 30%, the Auto Scaling group decreases the capacity by 10%. And if utilization further dips below 30%, the Auto Scaling group removes 30% of the current capacity.

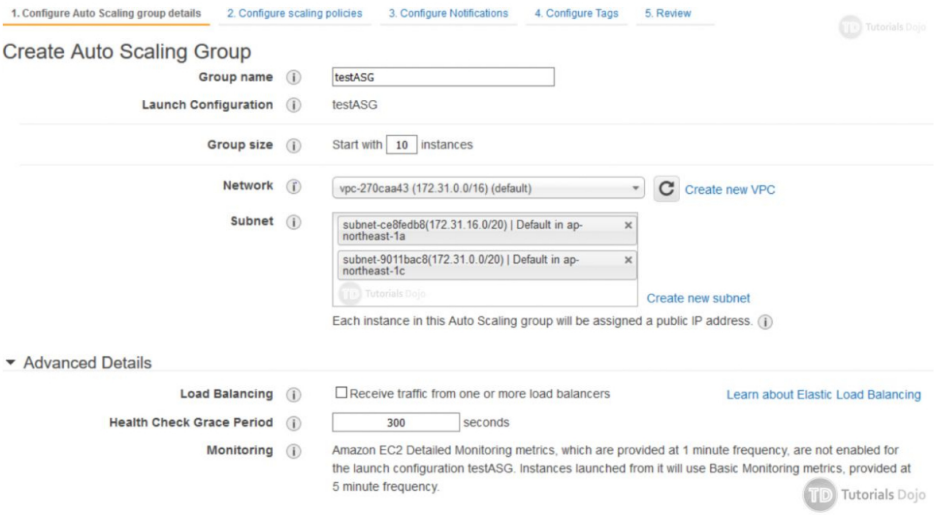
This effectively provides multiple steps in scaling policies that can be used to fine-tune your Auto Scaling group response to dynamically changing workload.

Creating a Step Scaling Policy for an Auto Scaling Group

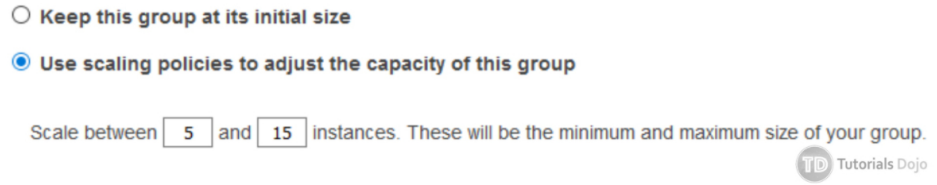
Based on the step scaling policy described above, the following guide will walk you through the process of applying this policy when creating your Auto Scaling group.

1. First, create your Launch Configuration for your EC2 instances. Check this guide if you haven't created one yet.
2. Go to **EC2** > **Auto Scaling Groups** > **Create Auto Scaling group**
3. Select your **Launch Configuration** and click **Next Step**.
4. Configure details for your Auto Scaling group.

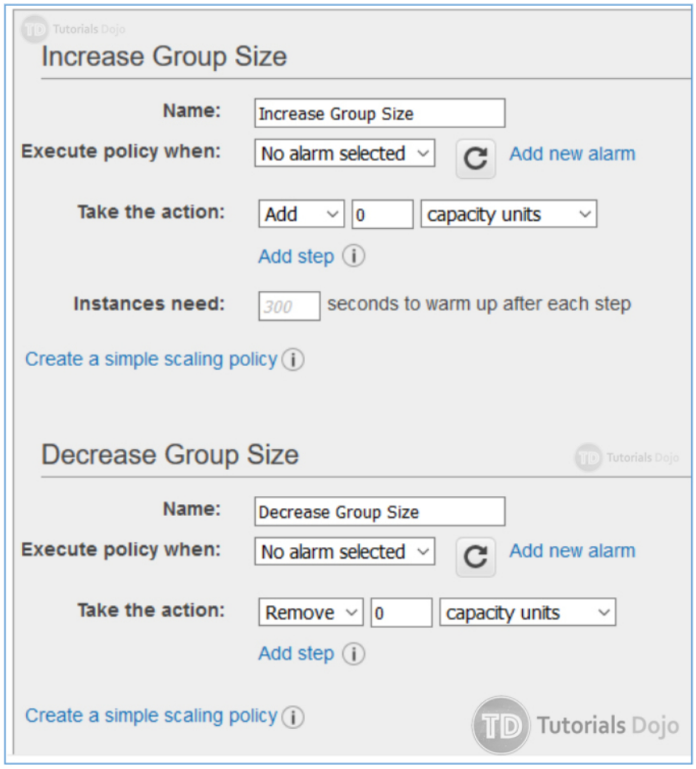
    a. **Group name** – descriptive name for this ASG.
    b. **Group size** – the initial size of your ASG. Let's set this to 10 for this example.
    c. **Network** – the VPC to use for your ASG.
    d. **Subnet** – the subnets in the VPC on where to place the EC2 instances. It's recommended to select subnet in multiple availability zones to improve the fault tolerance of your ASG.
    e. **Advanced Details** – in this section, you can check the **Load Balancing** option to select which load balancer to use for your ASG. (We won't configure a load balancer for this example). You can also set the **Health Check Grace Period** in this section. This is the length of time that Auto Scaling waits before checking the instance's health status. We'll leave the default to 300 seconds but you can adjust this if you know your EC2 instances need more or less than 5 minutes before they become healthy.



5. Click **Next: Configure scaling policies** to proceed.
6. Here, we'll configure the step scaling policy. Select the "**Use scaling policies to adjust the capacity of this group**" option and this will show an additional section for defining scaling policy. For this example, let's set 5 and 15 as the minimum and maximum size for this Auto Scaling group.



7. On the Scale Group Size section, you will be able to set the scaling policy for the group. But this is only for simple scaling so you have to click the **"Scale the Auto Scaling group using step or simple scaling policies"** link to show more advanced options for step scaling. You should see the **Increase Group Size** and **Decrease Group Size** section after clicking it.

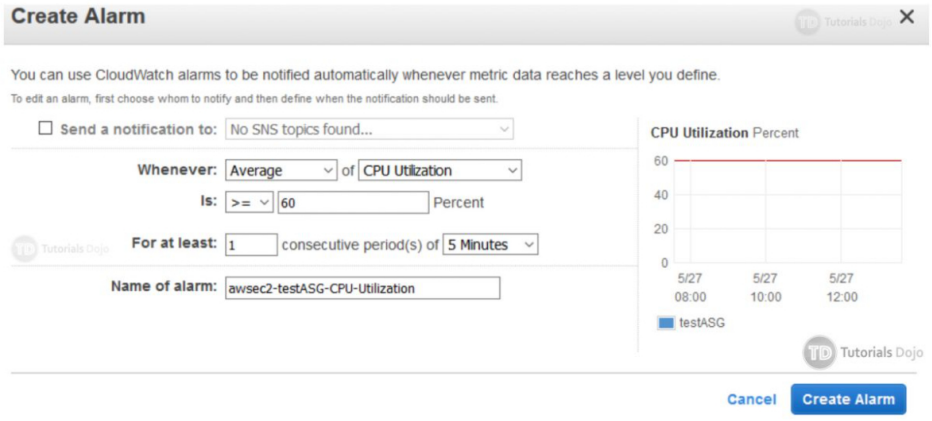


8. Now, we can set the step scaling policy for scaling out.

a. Set a name for your "**Increase Group Size**" policy. Click "**Add a new alarm**" to add a CloudWatch rule on when to execute the policy.

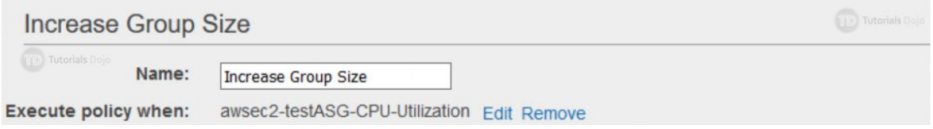
b. On the **Create Alarm** box, you can set an SNS notification. (We won't add it for this example).

c. Create a rule for whenever the **Average CPU Utilization** is greater than or equal to 60 percent for at least 1 consecutive period of 5 minutes. Set a name for your alarm. Click **Create Alarm**.



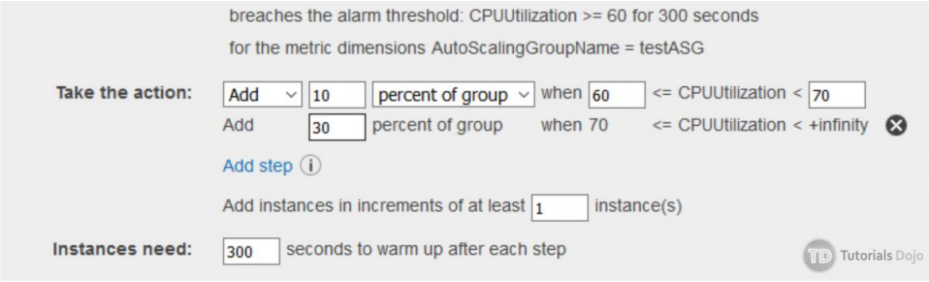
d. For the "**Take the action**" setting, we'll **Add 10 percent** of the group when CPU Utilization is greater than or equal to **60 and less than 70 percent**.

e. Click "Add Step" to add another action, we'll **Add 30 percent** of the group when CPU Utilization is **greater than or equal to 70 percent**.



Recent Posts

- Amazon Aurora Machine Learning
- Amazon Comprehend
- AWS DeepLens



f. Set 1 for "**Add instances in increments of at least**". This will ensure that at least 1 instance is added when the threshold is reached.

g. Set instances need **300 seconds to warm up** after each step.

Instance warmup – this specifies the timeout before the instance's own metric can be added to the group. Until the warmup time expires, the instance metric (CPU utilization in this case) is not counted toward the aggregated metric of the whole Auto Scaling group.

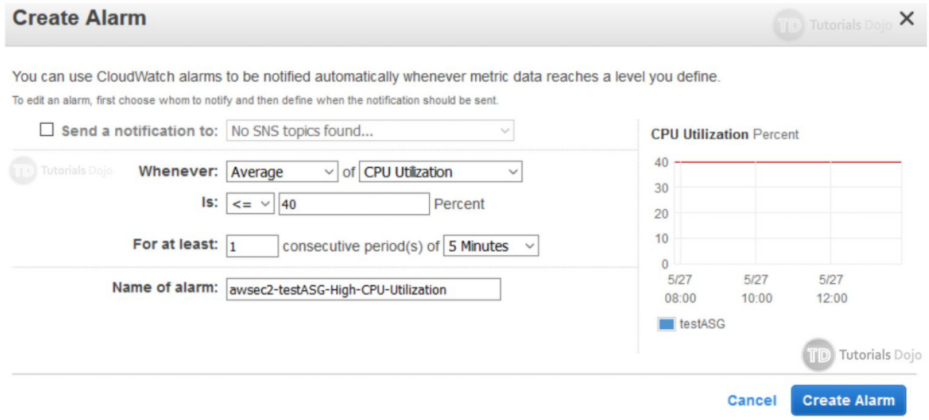
While scaling in, instances that are terminating are considered as part of the current capacity of the group. Therefore, it won't remove more instances from the Auto Scaling group than necessary.

9. Next, we can set the step scaling policy for the scaling in.

a. Set a name for your "**Decrease Group Size**" policy. Click "**Add a new alarm**" to add a CloudWatch rule on when to execute the policy.

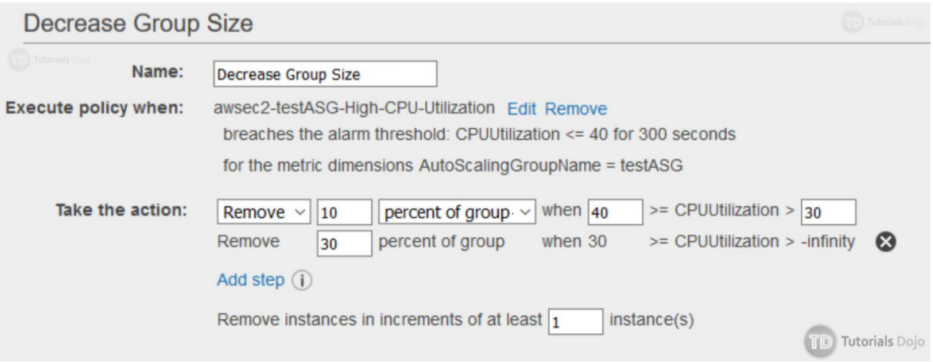
b. On the **Create Alarm** box, you can set an SNS notification. (We won't add it for this example).

c. Create a rule for whenever the **Average CPU Utilization** is less than or equal to 40 percent for at least 1 consecutive period of 5 minutes. Set a name for your alarm. Click **Create Alarm**.



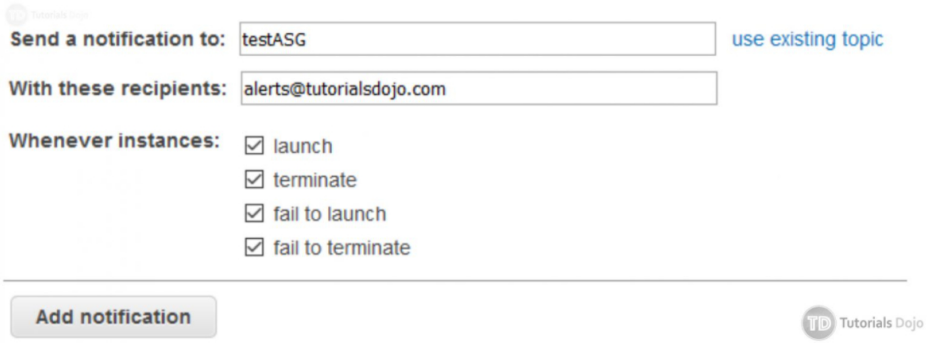
d. For the "**Take the action**" setting, we'll **remove 10 percent** of the group when CPU Utilization is less than or equal to **40 and greater than 30**.

e. Click "Add Step" to add another action, we'll **remove 30 percent** of the group when CPU Utilization is **less than or equal to 30 percent**.



f. Set 1 for "**Remove instances in increments of at least**". This will ensure that at least 1 instance is removed when the threshold is reached.

10. Click **Next: Configure Notifications** to proceed. On this part, you can click "**Add notification**" so that you can receive an email whenever a specific event occurs. Here's an example:



11.Click **Next: Configure Tags**. Create tags for instances in your Auto Scaling group.

12. Click **Review** to get to the review page.

13. After reviewing the details, click **Create Auto Scaling group**.

Your Auto Scaling group with step scaling policies should now be created. Remember, the initial desired size is 10, with a minimum of 5 and a maximum of 15.

The scale-out rule will have a step scaling policy, a 10% increase if CPU utilization is 60 – 70%, and will add 30% more instance if utilization is more than 70%.

The scale-in rule will have a step scaling policy, a 10% decrease if CPU utilization is 30 – 40%, and will remove 30% more instances if the utilization is less than 30%.

Sources:
https://docs.aws.amazon.com/autoscaling/ec2/userguide/as-scaling-simple-step.html
https://docs.aws.amazon.com/autoscaling/ec2/userguide/Cooldown.html
https://docs.aws.amazon.com/autoscaling/ec2/userguide/GettingStartedTutorial.html
https://docs.aws.amazon.com/autoscaling/ec2/userguide/create-launch-config.html

Submit

We respect your email privacy

Related Posts

Amazon Aurora Machine Learning

November 30th, 2020 | 0 Comments

Amazon Comprehend

November 25th, 2020 | 0 Comments

AWS DeepLens

November 25th, 2020 | 0 Comments

Amazon SageMaker

November 25th, 2020 | 0 Comments

Amazon Lex

November 25th, 2020 | 0 Comments