

Guided Lab: Scaling EC2 instances using Auto-Scale Group (ASG)

Description

EC2 Auto Scaling ensures that you maintain application availability and lets you scale Amazon EC2 capacity up or down automatically according to the conditions you define. This represents horizontal scaling, as opposed to increasing the specifications of a single instance (vertical scaling).

Prerequisites

This lab assumes you have experience creating EC2 instances and are familiar with its basic components.

If you find any gaps in your knowledge, consider taking the following labs:

- Creating an Amazon EC2 instance (Linux)

Objectives

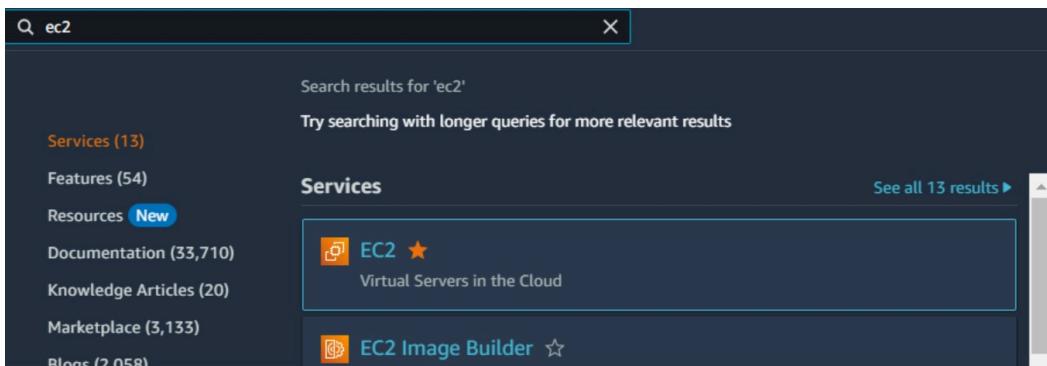
In this lab, you will:

- Understand the importance of EC2 Auto Scaling.
- Learn to set up Launch Templates.
- Experience the dynamic adjustment of EC2 capacity with Auto Scaling Groups.

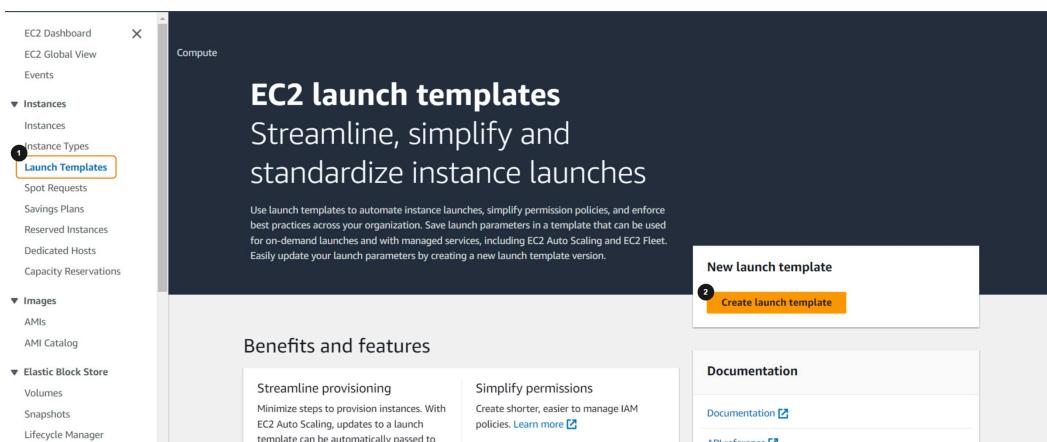
Lab Steps

Creating a Launch template

1. Search 'ec2' in the AWS Management Console search bar. Click **EC2** on the search results.



2. On the left window, under **Instances**, select **Launch Templates**, then click **Create launch template**.



3. Enter '**asg-template**' for the template name and tick the **Auto Scaling guidance** checkbox.

Launch template name and description

Launch template name - *required*

1 asg-template

Must be unique to this account. Max 128 chars. No spaces or special characters like '&', '*', '@'.

Template version description

A prod webserver for MyApp

Max 255 chars

Auto Scaling guidance [Info](#)

2 Select this if you intend to use this template with EC2 Auto Scaling

Provide guidance to help me set up a template that I can use with EC2 Auto Scaling

► Template tags

► Source template

4. Under Application and OS Images, click the default Amazon Linux AMI.

▼ Application and OS Images (Amazon Machine Image) - *required* [Info](#)

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. Search or Browse for AMIs if you don't see what you are looking for below

 Search our full catalog including 1000s of application and OS images

Recents

Quick Start



Browse more AMIs

Including AMIs from AWS, Marketplace and the Community

Amazon Machine Image (AMI)

Amazon Linux 2023 AMI

ami-0dbc3d7bc646e8516 (64-bit (x86)) / ami-055859c8e0f361065 (64-bit (Arm))
Virtualization: hvm ENA enabled: true Root device type: ebs

Free tier eligible

Description

Amazon Linux 2023 AMI 2023.2.20231018.2 x86_64 HVM kernel-6.1

Architecture

64-bit (x86)

AMI ID

ami-0dbc3d7bc646e8516

Verified provider

5. Under the Instance Type section, select t2.micro.

▼ Instance type [Info](#)

Instance type

t2.micro

Free tier eligible

Family: t2 1 vCPU 1 GiB Memory Current generation: true
On-Demand Windows base pricing: 0.0162 USD per Hour
On-Demand SUSE base pricing: 0.0116 USD per Hour
On-Demand RHEL base pricing: 0.0716 USD per Hour
On-Demand Linux base pricing: 0.0116 USD per Hour

All generations

[Compare instance types](#)

[Additional costs apply for AMIs with pre-installed software](#)

6. Under the Key Pair section, click Create new key pair.

▼ Key pair (login) [Info](#)

You can use a key pair to securely connect to your instance. Ensure that you have access to the selected key pair before you launch the instance.

Key pair name - required

Select

 [Create new key pair](#)

7. Enter 'asg-demo-key' for the key pair name and follow the configurations below. Finally, click Create key pair.

Create key pair X

Key pair name

1 Key pairs allow you to connect to your instance securely.

asg-demo-key

The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.

Key pair type

RSA

RSA encrypted private and public key pair

ED25519

ED25519 encrypted private and public key pair

Private key file format

.pem

3

For use with OpenSSH

.ppk

For use with PuTTY

 When prompted, store the private key in a secure and accessible location on your computer. You will need it later to connect to your instance. [Learn more](#)



4

[Cancel](#)

 [Create key pair](#)

8. Under Network settings, follow the configurations shown in the screenshot below.

Don't include in launch template

Create new subnet

When you specify a subnet, a network interface is automatically added to your template.

Firewall (security groups) Info

A security group is a set of firewall rules that control the traffic for your instance. Add rules to allow specific traffic to reach your instance.

Select existing security group

Create security group

Security group name - *required*

asg-demo-sg

This security group will be added to all network interfaces. The name can't be edited after the security group is created. Max length is 255 characters. Valid characters: a-z, A-Z, 0-9, spaces, and ._-:/()#,@[]+=;&{}!\$*

Description - *required* Info

Allows SSH from my computer

VPC - *required* Info

vpc-080e9d01d1e9f8245
192.168.5.0/26

Inbound Security Group Rules

▼ Security group rule 1 (TCP, 22, 49.150.93.115/32)

Remove

Type Info

ssh

Protocol Info

TCP

Port range Info

22

Source type Info

My IP

Name Info

Description - *optional* Info

e.g. SSH for admin desktop

49.150.93.115/32 X

9. Scroll down to the bottom page and click the **Advanced details** dropdown menu.

► Advanced details Info

10. Select **Enable** on the dropdown menu of the **Detailed CloudWatch monitoring** option.

Stop protection | Info

Don't include in launch template

Detailed CloudWatch monitoring | Info

Enable

Additional charges apply

Elastic GPU | Info

Don't include in launch template

By default, Amazon EC2 sends metrics to CloudWatch every 5 minutes. This includes metrics like CPU utilization, disk reads/writes, and network packets in/out. For more granular insight, you can opt for detailed monitoring. This captures the same metrics but updates

them every 1 minute. In this lab, we'll use Detailed Monitoring to get quicker feedback, enabling our Auto Scaling Group to scale in/out faster during our tests.

11. On the right window, click the **Create launch template** button.

▼ Summary

Software Image (AMI)

Amazon Linux 2023 AMI 2023.2.2...[read more](#)
ami-0dbc3d7bc646e8516

Virtual server type (instance type)

t2.micro

Firewall (security group)

New security group

Storage (volumes)

1 volume(s) - 8 GiB



Free tier: In your first year includes 750 hours of t2.micro (or t3.micro in the Regions in which t2.micro is unavailable) instance usage on free tier AMIs per month, 30 GiB of EBS storage, 2 million IOs, 1 GB of snapshots, and 100 GB of bandwidth to the internet.

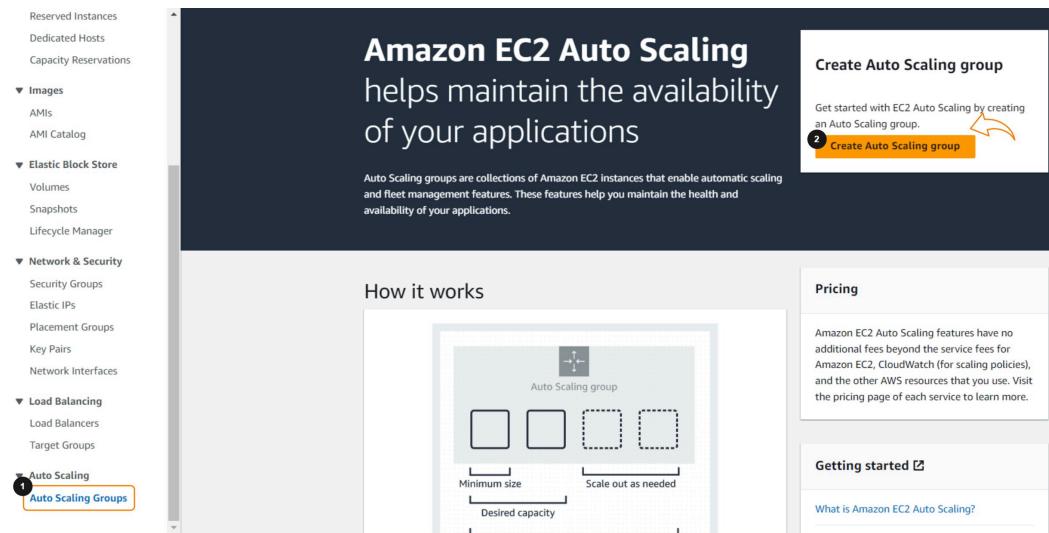


Cancel

Create launch template

Creating an Auto Scaling Group

12. In the Amazon EC2 Console, under the **Auto Scaling** menu, click **Auto Scaling Groups**, then click the **Create Auto Scaling group** button.



13. Enter 'asg-lab' into the **Auto Scaling group name** field.

Choose launch template or configuration Info

Specify a launch template that contains settings common to all EC2 instances that are launched by this Auto Scaling group. If you currently use launch configurations, you might consider migrating to launch templates.

Name
Auto Scaling group name Enter a name to identify the group. <input type="text" value="asg-lab"/> Must be unique to this account in the current Region and no more than 255 characters.

14. In the Launch template section, choose the '*asg-template*' that we created in the **Creating a Launch template** section.

Launch template [Info](#)

[Switch to launch configuration](#)

Launch template

Choose a launch template that contains the instance-level settings, such as the Amazon Machine Image (AMI), instance type, key pair, and security groups.

1 asg-template

[Create a launch template](#)

Version

Default (1)



[Create a launch template version](#)

Description

-

Launch template

asg-template

lt-012f50d9f64ae7ae9

Instance type

t2.micro

AMI ID

ami-0dbc3d7bc646e8516

Security groups

-

Request Spot Instances

No

Key pair name

asg-demo-key

Security group IDs

sg-05b5515d336e64f17

Additional details

Storage (volumes)

-

Date created

Fri Oct 27 2023 22:42:05 GMT+0800
(Singapore Standard Time)

[Cancel](#)

[Next](#)

15. In the **Network** section, select the default VPC in the VPC dropdown menu. From the **Availability Zones and subnets** dropdown menu, select all the default subnets. Click **Next**.

Network [Info](#)

For most applications, you can use multiple Availability Zones and let EC2 Auto Scaling balance your instances across the zones. The default VPC and default subnets are suitable for getting started quickly.

VPC

Choose the VPC that defines the virtual network for your Auto Scaling group.

1 vpc-080e9d01d1e9f8245
192.168.5.0/26



[Create a VPC](#)

Availability Zones and subnets

Define which Availability Zones and subnets your Auto Scaling group can use in the chosen VPC.

Select Availability Zones and subnets



2 us-east-1a | subnet-0658d5a6c6531580b X
192.168.5.0/28

us-east-1b | subnet-080e4a793c23366be X
192.168.5.16/28

us-east-1c | subnet-06198423051f6c3b6 X
192.168.5.32/28

[Create a subnet](#)

[Next](#)

16. Scroll down to the bottom of **Configure advanced options – optional** page, and click **Next**.

In this lab, we'll focus on the basics of setting up an Auto Scaling Group. Therefore, we'll skip the "Configure advanced options" step. Our primary goal is to demonstrate how scaling works without diving into the more intricate configurations.

17. For the Group Size, set the following configurations:

- Desired Capacity: 2
- Minimum Capacity: 1
- Maximum Capacity: 3

The screenshot shows two configuration panels for an Auto Scaling group.

Group size (Info): Set the initial size of the Auto Scaling group. After creating the group, you can change its size to meet demand, either manually or by using automatic scaling.

Desired capacity type: Choose the unit of measurement for the desired capacity value. vCPUs and Memory(GiB) are only supported for mixed instances groups configured with a set of instance attributes.

Units (number of instances) ▾

Desired capacity: Specify your group size.

2

Scaling (Info): You can resize your Auto Scaling group manually or automatically to meet changes in demand.

Scaling limits: Set limits on how much your desired capacity can be increased or decreased.

Min desired capacity: 1
Max desired capacity: 3

Equal or less than desired capacity Equal or greater than desired capacity

Automatic scaling - optional: Choose whether to use a target tracking policy (Info). You can set up other metric-based scaling policies and scheduled scaling after creating your Auto Scaling group.

No scaling policies: Your Auto Scaling group will remain at its initial size and will not dynamically resize to meet demand.

Target tracking scaling policy: Choose a CloudWatch metric and target value and let the scaling policy adjust the desired capacity in proportion to the metric's value.

We start with a **Desired Capacity** of 2, meaning that initially, we want two instances up and running. However, as demands shift and the need for resources changes, the ASG can adjust the number of instances. It will always keep at least 1 instance active, which is our **Minimum Capacity**. However, even during high demand, it won't spin up more than 3 instances, which is set by **Maximum Capacity**.

Amazon's Auto Scaling Groups (ASG) offers the following scaling policies to accommodate a variety of loads:

- Target tracking scaling – Increase and decrease the current capacity of the group based on an Amazon CloudWatch metric

and a target value. It works similarly to the way that your thermostat maintains the temperature of your home—you select a temperature, and the thermostat does the rest.

- Step scaling – Increase and decrease the current capacity of the group based on a set of scaling adjustments, known as step adjustments, that vary based on the size of the alarm breach.
- Simple scaling – Increase and decrease the current capacity of the group based on a single scaling adjustment, with a cooldown period between each scaling activity.

18. In the **Scaling policies** section, we'll set our choice to **None** for now and configure a Simple scaling policy later. To skip the succeeding optional steps, click the **Skip to review** button.

Scaling Info
You can resize your Auto Scaling group manually or automatically to meet changes in demand.

Scaling limits
Set limits on how much your desired capacity can be increased or decreased.

Min desired capacity 1	Max desired capacity 3
Equal or less than desired capacity	Equal or greater than desired capacity

Automatic scaling - optional
Choose whether to use a target tracking policy | [Info](#)
You can set up other metric-based scaling policies and scheduled scaling after creating your Auto Scaling group.

No scaling policies
Your Auto Scaling group will remain at its initial size and will not dynamically resize to meet demand.

Target tracking scaling policy
Choose a CloudWatch metric and target value and let the scaling policy adjust the desired capacity in proportion to the metric's value.

Instance maintenance policy - new [Info](#)
Control your Auto Scaling group's availability during instance replacement events. This includes health checks, instance refreshes, maximum instance lifetime features and events that happen automatically to keep your group balanced, called rebalancing events.

Control availability and cost during replacement events
An instance maintenance policy determines how much availability your application has when EC2 Auto Scaling replaces instances. It also establishes guardrails that limit the amount of capacity that can be added or removed when replacing instances.

Choose a replacement behavior depending on your availability requirements

19. Click **Create Auto Scaling Group**.

Instance scale-in protection

Instance scale-in protection
Enable instance protection from scale in

Step 5: Add notifications

Notifications
No notifications

Step 6: Add tags

Tags (0)
Key Value Tag new instances
No tags

Cancel Previous **Create Auto Scaling group**

20. On the ASG dashboard. Click on the name of your ASG. This will take you to the **details page** for your Auto Scaling Group.

Auto Scaling groups (1) Info

Search your Auto Scaling groups

Name	Launch template/configuration	Instances	Status	Desired capacity	Min	Max	Availability Zones
asg-lab	asg-template Version Default	2	-	2	1	3	us-east-1a, us-east-1b, us-east-1c

Configuring the Auto Scaling Group with a Simple Scaling Policy

21. Click on the **Automatic scaling** tab, then click **Create dynamic scaling policy** button.

EC2 > Auto Scaling groups > asg-lab

asg-lab

Activity **Automatic scaling** Instance management Monitoring Instance refresh

Scaling policies resize your Auto Scaling group to meet changes in demand. With reactive dynamic scaling policies, you can track specific CloudWatch metrics and take action when the CloudWatch alarm threshold is met. Use predictive scaling policies along with dynamic scaling policies in the following situations: when your application demand changes quickly, but with a recurring pattern, or when your EC2 instances require more time to initialize.

Dynamic scaling policies (0) Info Actions Create dynamic scaling policy

We will set up two distinct scaling policies: one for scaling out and another for scaling in. The scale-out policy is what prompts ASG to add more instances when demands surge, such as during high CPU utilization. Conversely, the scale-in policy reduces the number of instances when demand subsides.

Let's start by setting up the scale-out policy.

22. Create a scale-out policy by filling out the Scaling Policy fields with the following values.

Create dynamic scaling policy

Policy type
1 Simple scaling

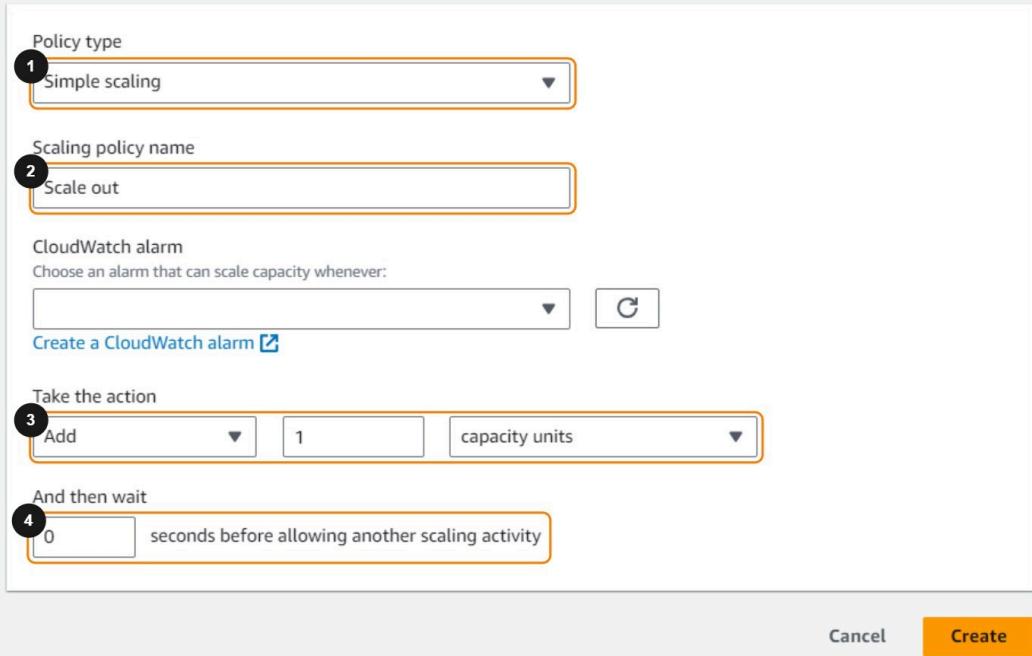
Scaling policy name
2 Scale out

CloudWatch alarm
Choose an alarm that can scale capacity whenever:
Create a CloudWatch alarm [Create](#)

Take the action
3 Add 1 capacity units

And then wait
4 0 seconds before allowing another scaling activity

[Cancel](#) [Create](#)



*Note: We're setting the **seconds before allowing another scaling activity** to 0 so ASG can respond immediately for demonstration purposes. In a real-world scenario, you'd typically use a longer cooldown to account for the time it takes an instance to start up and begin handling traffic.

23. Click **Create a CloudWatch alarm**. This will open a new tab on the CloudWatch Console.

Create dynamic scaling policy

Policy type
Simple scaling

Scaling policy name
Scale out

CloudWatch alarm
Choose an alarm that can scale capacity whenever:
[Create a CloudWatch alarm](#)

Take the action
Add 1 capacity units

And then wait
0 seconds before allowing another scaling activity

[Cancel](#) [Create](#)

CloudWatch alarms play a pivotal role in the operations of ASGs.

They monitor specific metrics, like CPU utilization, and when thresholds are breached, they signal ASG to scale in or out.

For this lab, we'll be creating two alarms based on CPU utilization: The first alarm will signal ASG to scale out if the CPU usage exceeds 30% over a 1-minute span. Conversely, the second alarm will signal ASG to scale in if the CPU usage dips below 30% for the same duration.

24. On **Specify metric and conditions**, click the **Select metric** button.

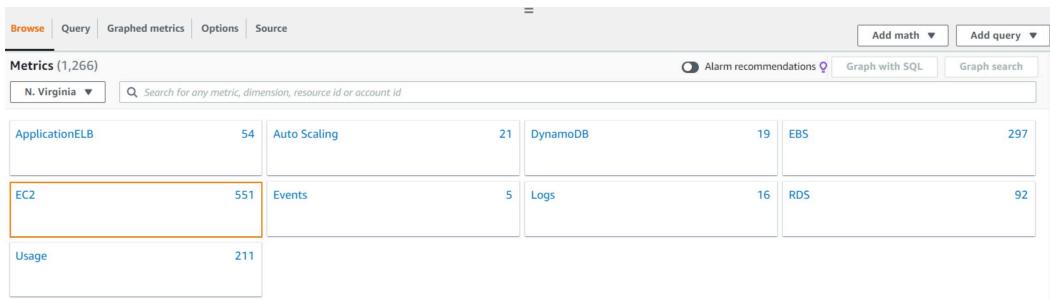
Specify metric and conditions

Metric

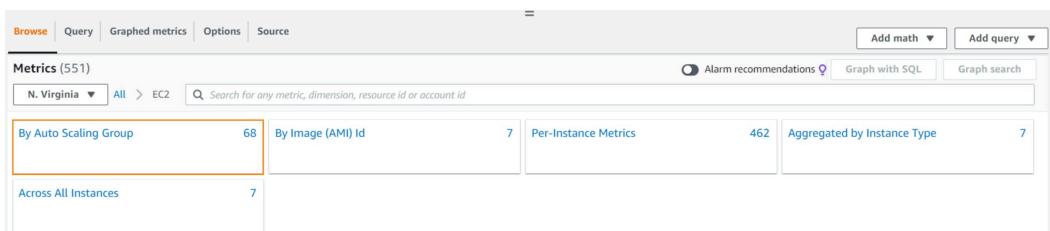
Graph
Preview of the metric or metric expression and the alarm threshold.
[Select metric](#)

[Cancel](#) [Next](#)

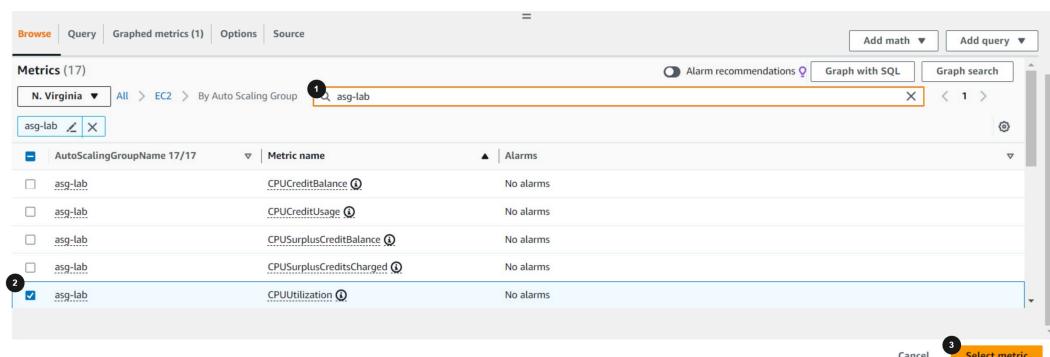
25. Select **EC2**.



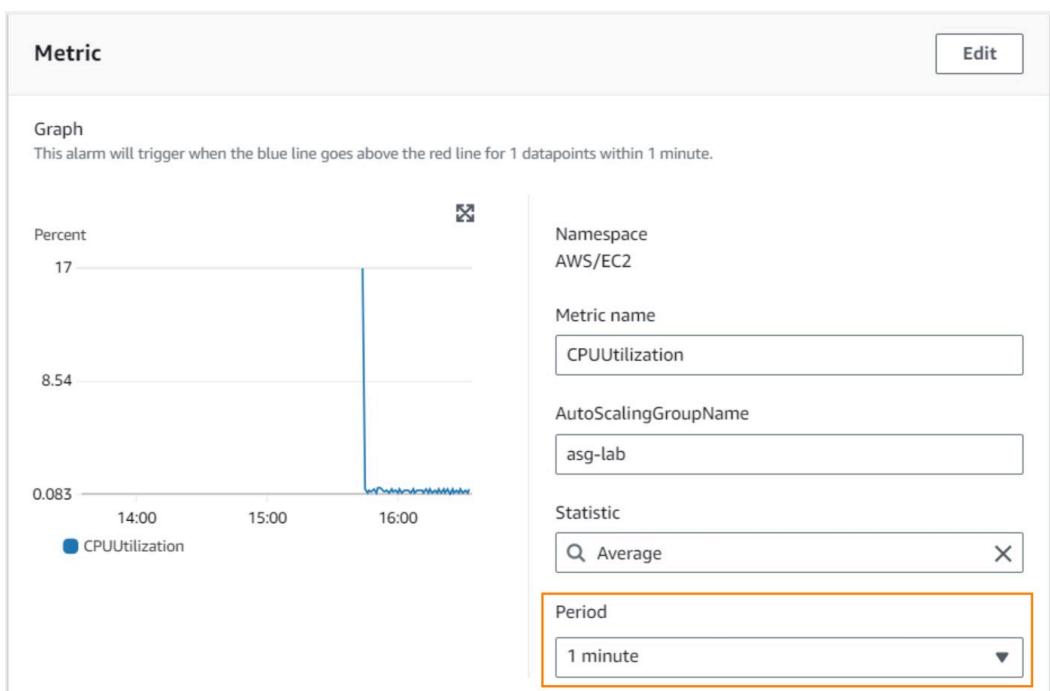
26. Select By Auto Scaling Group.



27. Enter 'asg-lab' into the search field. Toggle the checkbox for **CPUUtilization metric**. Then, click **Select Metric**.



28. In the **Metric** menu, change the **Period** to **1 minute**.



29. In the **Conditions** menu, fill out the fields with the following configuration values. Then, click **Next**.

Conditions

Threshold type

1 Static

Use a value as a threshold

Anomaly detection

Use a band as a threshold

Whenever CPUUtilization is...

Define the alarm condition.

Greater
> threshold

2 Greater/Equal
>= threshold

Lower/Equal
<= threshold

Lower
< threshold

than...

Define the threshold value.

3 30

Must be a number

▶ Additional configuration



4

Cancel

Next

30. Remove Notifications from Configure actions. Then, scroll down and click Next.

Configure actions

Notification

Alarm state trigger

Define the alarm state that will trigger this action.

In alarm

The metric or expression is outside of the defined threshold.

OK

The metric or expression is within the defined threshold.

Insufficient data

The alarm has just started or not enough data is available.

Send a notification to the following SNS topic

Define the SNS (Simple Notification Service) topic that will receive the notification.

Select an existing SNS topic

Create new topic

Use topic ARN to notify other accounts

Send a notification to...

Only email lists for this account are available.

31. Enter 'asg-scale-out-alarm' for the Alarm name. Click "Next", then, click "Create alarm".

Name and description

Alarm name
asg-scale-out-alarm

Alarm description - optional [View formatting guidelines](#)

Edit | **Preview**

```
# This is an H1
**double asterisks will produce strong character**
This is [an example](https://example.com/) inline link.
```

Up to 1024 characters (0/1024)

Info Markdown formatting is only applied when viewing your alarm in the console. The description will remain in plain text in the alarm notifications.

Cancel **Previous** **Next**

32. Go back to the ASG tab. Click the reload icon and select the newly created 'asg-scale-out-alarm' from the dropdown. Then, click "Create".

[EC2](#) > [Auto Scaling groups](#) > asg-lab

Create dynamic scaling policy

Policy type
Simple scaling

Scaling policy name
Scale out

CloudWatch alarm
Choose an alarm that can scale capacity whenever:

And then wait
0 seconds before allowing another scaling activity

Cancel **Create**

To create the scale in policy, follow the instructions between **Steps 22 – 32**, but **with the following changes**:

Scale in policy:

Create dynamic scaling policy

Policy type
Simple scaling

Scaling policy name ①
Scale in

CloudWatch alarm
Choose an alarm that can scale capacity whenever:
Create a CloudWatch alarm

Take the action ②
Remove 1 capacity units

And then wait
300 seconds before allowing another scaling activity

Cancel Create

CloudWatch alarm condition:

Conditions

Threshold type
 Static Use a value as a threshold Anomaly detection Use a band as a threshold

Whenever CPUUtilization is...
Define the alarm condition.
 Greater > threshold Greater/Equal \geq threshold Lower/Equal \leq threshold Lower < threshold

than...
Define the threshold value.
30
Must be a number

► Additional configuration

For alarm name, enter 'asg-scale-in-alarm'

Testing the Scale in Policy

33. Go back to the ASG details page. Click on the **Instance Management** tab. This area provides insights into the status and health of your instances being managed by ASG. You'll notice that ASG initially creates two instances. This is because we set the desired capacity to 2.

Since we've configured a scale-in policy that removes 1 instance when CPU utilization is below 30 percent within 1 minute, expect the instances to be reduced to 1 after a minute or so.

34. In the **Activity** tab, you'll see a notification indicating that the scale-in policy was triggered. Please be patient and wait for this notification to appear. To ensure you're seeing the most recent updates, consider clicking the refresh icon once in a while. You may also monitor the number of instances in your ASG under the **Instance Management** tab.

Now let's test our scale-out policy by simulating a high CPU load using a tool called '**stress**'.

35. SSH into your EC2 instance and run the following commands.

```
#This updates the package list in your system
sudo yum update -y
```

```
#This installs stress
sudo yum install stress -y
```

```
#Spawns 50 workers for 5 minutes
stress --cpu 50 --timeout 5m
```

36. Monitor the **Activity** for scaling notifications or check the **Instance Management** tab to observe changes in the number of instances. Please be patient to wait, as the scaling won't happen instantaneously.

 Successful	Launching a new EC2 instance: i-0ea4bbdc4b22a032f	At 2023-10-27T17:26:45Z a monitor alarm asg-scale-out-alarm in state ALARM triggered policy Scale out changing the desired capacity from 2 to 3. At 2023-10-27T17:26:56Z an instance was started in response to a difference between desired and actual capacity, increasing the capacity from 2 to 3.
 Successful	Launching a new EC2 instance: i-0d16ab5d0b72d9880	At 2023-10-27T17:24:45Z a monitor alarm asg-scale-out-alarm in state ALARM triggered policy Scale out changing the desired capacity from 1 to 2. At 2023-10-27T17:24:57Z an instance was started in response to a difference between desired and actual capacity, increasing the capacity from 1 to 2.

Congratulations! You've successfully set up and experienced Auto Scaling in action. While we tackled a straightforward scaling scenario, remember there are more advanced scaling policies that cater to dynamic and complex workloads. This was just the starting point, and there's so much more to explore and learn. Keep up the momentum, and happy scaling!