

AUTO-SCALING PRAC

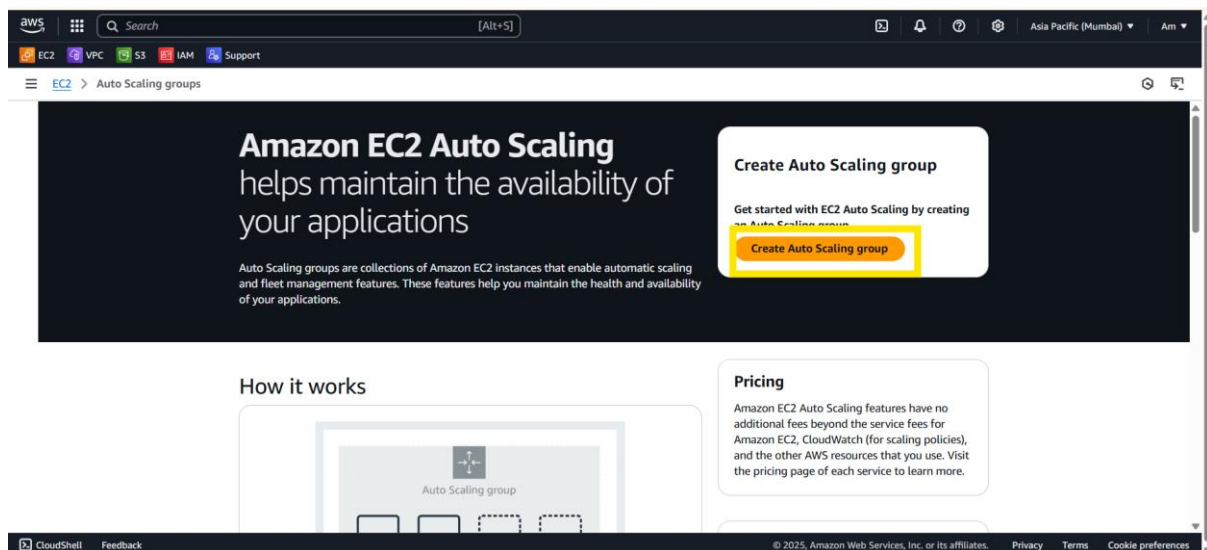
Define Auto-scaling:

system that automatically adjusts your cloud resources (like servers or databases) to match your application's needs.

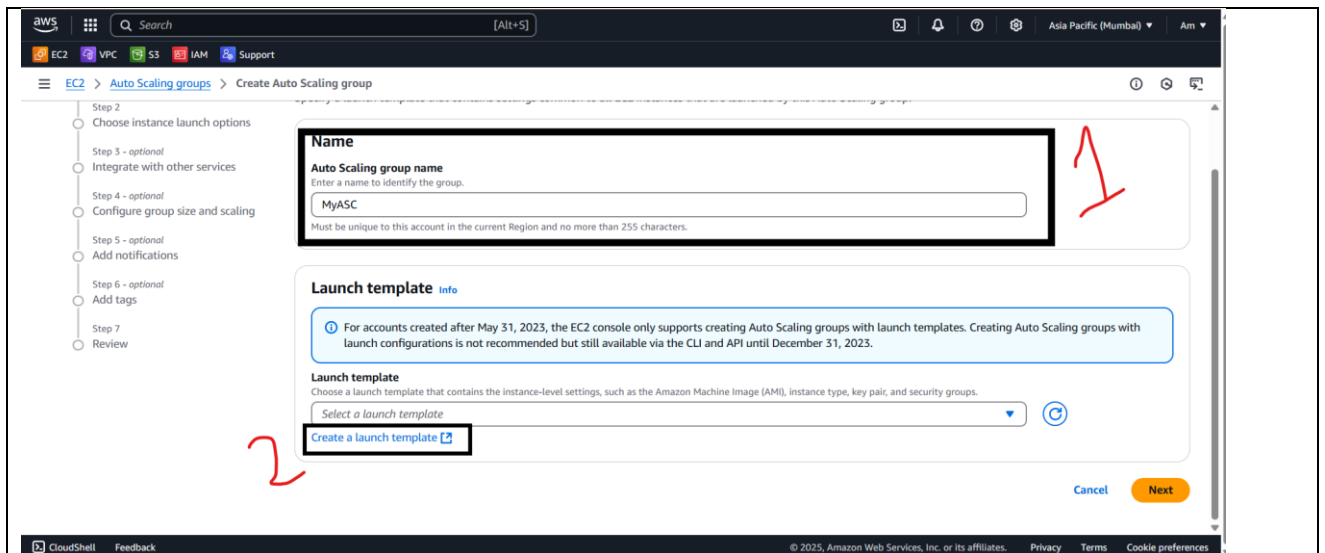
Why use Auto-scaling:

automatically adjusts your resources to match the demands of your application, ensuring optimal performance and cost efficiency.

Click on create autoscaling group:-



STEP 1: name it and create launch template to launch instance through launch template.



STEP 2: all other things will be same as we create a launch template I have only added a new SG name all traffic, other details can be seen in SC.

aws

Search

[Alt+S]

EC2

VPC

S3

IAM

Support

EC2 > Launch templates > Create launch template

Security group name - required

alltraffic

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 access to developers

VPC | Info

vpc-0077c0b100707fd7e
172.31.0.0/16

(default)

Inbound Security Group Rules

▼ Security group rule 1 (All, All, 0.0.0.0/0)

Remove

Type | Info

All traffic

Protocol | Info

All

Port range | Info

All

Source type | Info

Anywhere

Source | Info

0.0.0.0/0

Description - optional | Info

e.g. SSH for admin desktop

▼ Security group rule 2 (TCP, 22, 0.0.0.0/0)

Remove

Type | Info

ssh

Protocol | Info

TCP

Port range | Info

22

Source type | Info

Anywhere

Source | Info

0.0.0.0/0

Description - optional | Info

e.g. SSH for admin desktop

STEP 3: Adding script to install nginx and apache.

aws | Search [Alt+S]

EC2 VPC S3 IAM Support

EC2 > Launch templates > Create launch template

User data - optional | Info

Upload a file with your user data or enter it in the field.

[Choose file](#)

```
#!/bin/bash

sudo apt update -y
sudo apt install nginx -y
systemctl start nginx
systemctl enable nginx

sudo install apache2
systemctl start apache2
systemctl enable apache2
```

☐ User data has already been base64 encoded

USING THIS SCRIPT SO THAT, WHEN THE INSTANCE IS LAUNCHED WE CAN DIRECTLY ACCESS BOTH THESE PROXY SERVERS BY USING PUBLIC IP OF THE INSTANCE.

STEP 4: Choosing the instance requirements of how many instances to be launched at the initial start of the instance.

aws | Search [Alt+S]

EC2 VPC S3 IAM Support

EC2 > Auto Scaling groups > Create Auto Scaling group

Step 1: Choose launch template
Step 2: **Choose instance launch options**
Step 3 - optional: Integrate with other services
Step 4 - optional: Configure group size and scaling
Step 5 - optional: Add notifications
Step 6 - optional: Add tags
Step 7: Review

Choose instance launch options | Info

Choose the VPC network environment that your instances are launched into, and customize the instance types and purchase options.

[Reset to launch template](#)

☒ **Specify instance attributes**
Provide your compute requirements. We fulfill your desired capacity with matching instance types based on your allocation strategy selection.

☐ **Manually add instance types**
Add one or more instance types. Any of the instance types may be launched to fulfill your desired capacity based on your allocation strategy selection.

Required instance attributes
Enter your compute requirements in virtual CPUs (vCPUs) and memory.

vCPUs
Enter the minimum and maximum number of vCPUs per instance.

0 minimum 1 maximum

☐ No minimum ☐ No maximum
Maximum vCPUs is required and must be greater than 0.

Memory (GiB)
Enter the minimum and maximum Gigs of memory per instance.

0 minimum 1 maximum

☐ No minimum ☐ No maximum
Maximum memory is required and must be greater than 0.

Additional instance attributes - optional
Add instance attributes to further limit which instance types may be used to fulfill your desired capacity.

CloudShell Feedback

© 2024 Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

STEP 5: NETWORK CONFIG—Selecting all the AZ's

quickly.

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

vpc-0077c0b100707fd7e
172.31.0.0/16 Default

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

ap-south-1a | subnet-0b4a2735ea77b0881
172.31.32.0/20 Default

ap-south-1b | subnet-057511227b8b41e9b
172.31.0.0/20 Default

ap-south-1c | subnet-022b390090e2a32de
172.31.16.0/20 Default

Create a subnet

Availability Zone distribution - new
Auto Scaling automatically balances instances across Availability Zones. If launch failures occur in a zone, select a strategy.

☒ **Balanced best effort**
If launches fail in one Availability Zone, Auto Scaling will attempt to launch in another healthy Availability Zone.

☐ **Balanced only**
If launches fail in one Availability Zone, Auto Scaling will continue to attempt to launch in the unhealthy Availability Zone to preserve balanced distribution.

Cancel Skip to review Previous Next

STEP 6: In health checks we give 120 seconds.

Explain why 120 seconds?

- 120 seconds mean the servers we have taken nginx and apache will refresh after 120 seconds.
- If the load increases on one primary instance, autoscaling will re-direct the traffic to other instance.

Application Recovery Controller (ARC) zonal shift - new Info
During an Availability Zone impairment, target instance launches towards other healthy Availability Zones.

☐ **Enable zonal shift**
New instance launches will be retargeted towards healthy Availability Zones until the zonal shift is canceled.

Health checks
Health checks increase availability by replacing unhealthy instances. When you use multiple health checks, all are evaluated, and if at least one fails, instance replacement occurs.

EC2 health checks
[Always enabled](#)

Additional health check types - optional Info

☐ **Turn on Elastic Load Balancing health checks**
Elastic Load Balancing monitors whether instances are available to handle requests. When it reports an unhealthy instance, EC2 Auto Scaling can replace it on its next periodic check.

☐ **Turn on VPC Lattice health checks**
VPC Lattice can monitor whether instances are available to handle requests. If it considers a target as failed a health check, EC2 Auto Scaling replaces it after its next periodic check.

☐ **Turn on Amazon EBS health checks**
EBS monitors whether an instance's root volume or attached volume stalls. When it reports an unhealthy volume, EC2 Auto Scaling can replace the instance on its next periodic health check.

Health check grace period Info
This time period delays the first health check until your instances finish initializing. It doesn't prevent an instance from terminating when placed into a non-running state.

120 seconds

Cancel Skip to review Previous Next

CloudShell Feedback © 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

STEP 7: Set desired capacity and max. desired capacity.

EC2

VPC

S3

IAM

Support

Search

[Alt+S]

Asia Pacific (Mumbai)

Am

EC2

> Auto Scaling groups

> Create Auto Scaling group

Step 1

Choose launch template

Step 2

Choose instance launch options

Step 3 - optional

Integrate with other services

Step 4 - optional

Configure group size and scaling

Step 5 - optional

Add notifications

Step 6 - optional

Add tags

Step 7

Review

Configure group size and scaling - optional

info

Define your group's desired capacity and scaling limits. You can optionally add automatic scaling to adjust the size of your 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.

1

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

Equal or less than desired capacity

Max desired capacity

3

Equal or greater than desired capacity

Automatic scaling - optional

CloudShell

Feedback

© 2025, Amazon Web Services, Inc. or its affiliates.

Privacy

Terms

Cookie preferences