Tasks To Be Performed:

1. Manage the scaling requirements of the company by:

    a. Deploying multiple compute resources on the cloud as soon as the load increases and the CPU utilization exceeds 80%

    b. Removing the resources when the CPU utilization goes under 60%

2. Create a load balancer to distribute the load between compute resources.

3. Route the traffic to the company’s domain

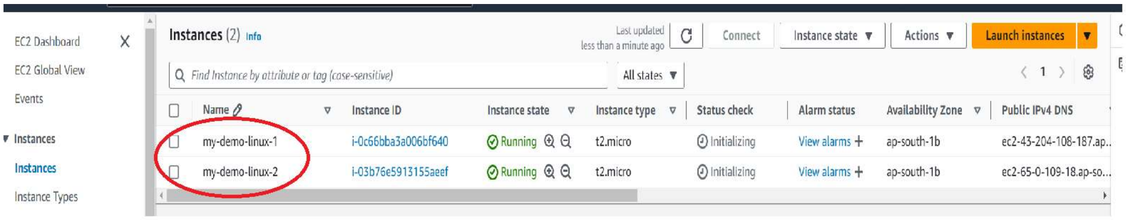
Solution:

Creating a load balancer.

1) First create 2 instances with the following user data.

```
#!/bin/bash
# Use this for your user data (script from top to bottom)
# install httpd (Linux 2 version)
yum update -y
yum install -y httpd
systemctl start httpd
systemctl enable httpd
echo "<h1>Hello World from $(hostname -f)</h1>" >
/var/www/html/index.html
```

2) Both of the instances are created.

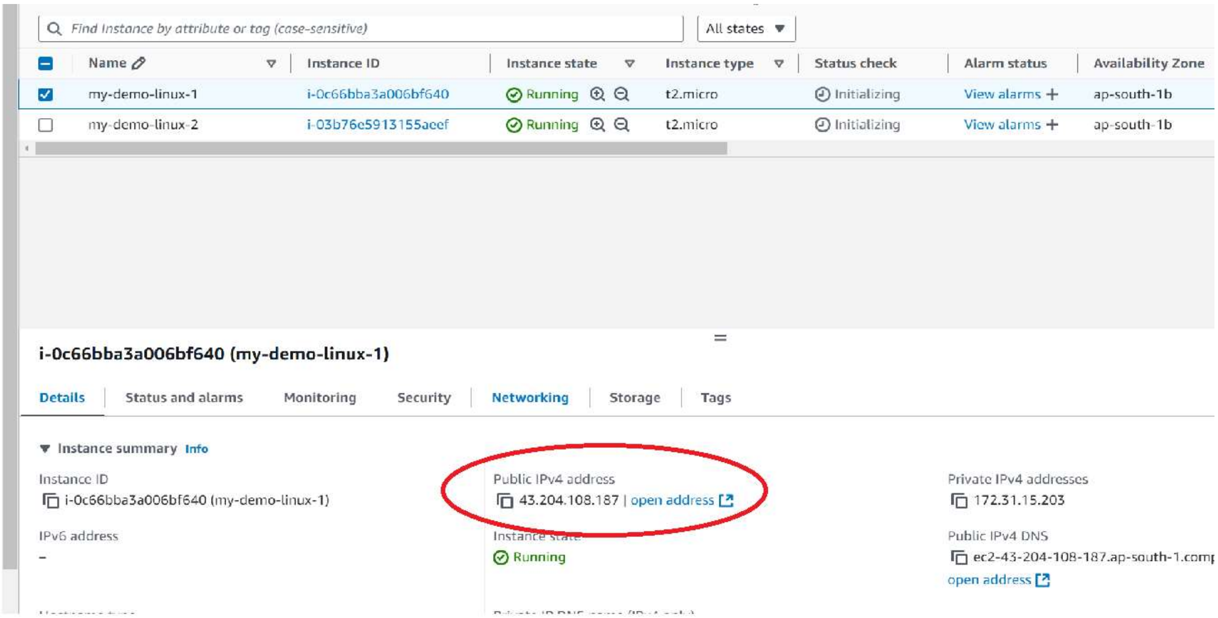


Instances (2) Info

Last updated less than a minute ago

	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS
□	my-demo-linux-1	i-0c66bba3a006bf640	Running	t2.micro	Initializing	View alarms +	ap-south-1b	ec2-43-204-108-187.ap...
□	my-demo-linux-2	i-03b76e5913155aeef	Running	t2.micro	Initializing	View alarms +	ap-south-1b	ec2-65-0-109-18.ap-so...

3) To check the user data select the instance and copy the public IPV4 address and paste it in the browser and we can see the user data running successfully.



	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone
☑	my-demo-linux-1	i-0c66bba3a006bf640	Running	t2.micro	Initializing	View alarms +	ap-south-1b
□	my-demo-linux-2	i-03b76e5913155aeef	Running	t2.micro	Initializing	View alarms +	ap-south-1b

i-0c66bba3a006bf640 (my-demo-linux-1)

▾ Instance summary Info

Instance ID
i-0c66bba3a006bf640 (my-demo-linux-1)

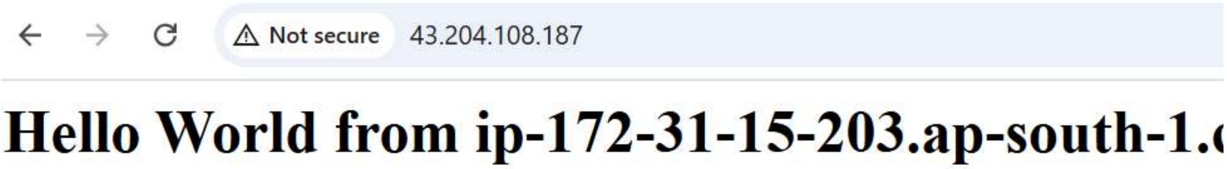
Public IPv4 address
43.204.108.187 | open address

Private IPv4 addresses
172.31.15.203

IPv6 address
–

Instance state
Running

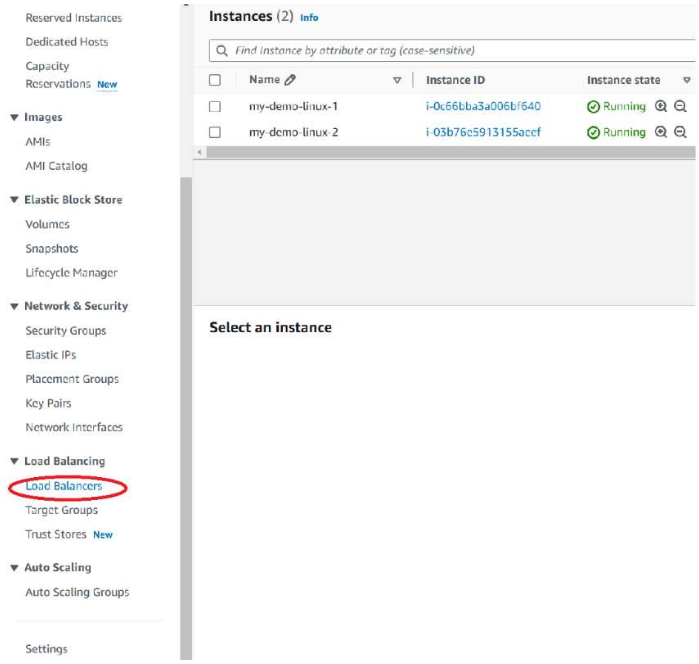
Public IPv4 DNS
ec2-43-204-108-187.ap-south-1.comp
open address



43.204.108.187

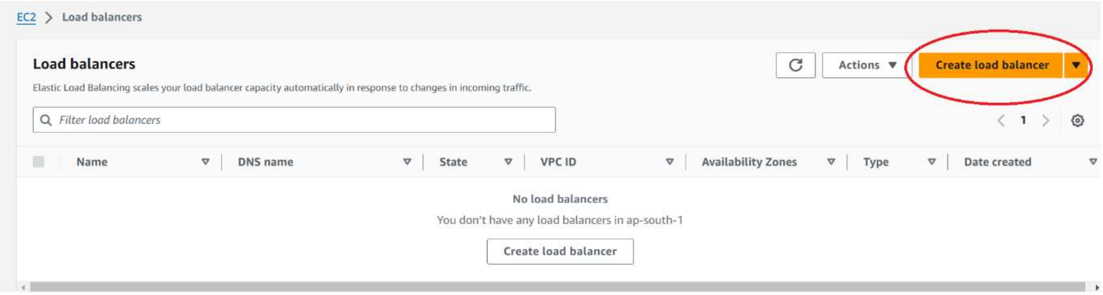
Hello World from ip-172-31-15-203.ap-south-1.c

4) On the left side of EC2 console click load balancer and click create load balancer.



□	Name	Instance ID	Instance state
□	my-demo-linux-1	i-0c66bba3a006bf640	Running
□	my-demo-linux-2	i-03b76c5913155accf	Running

Select an instance

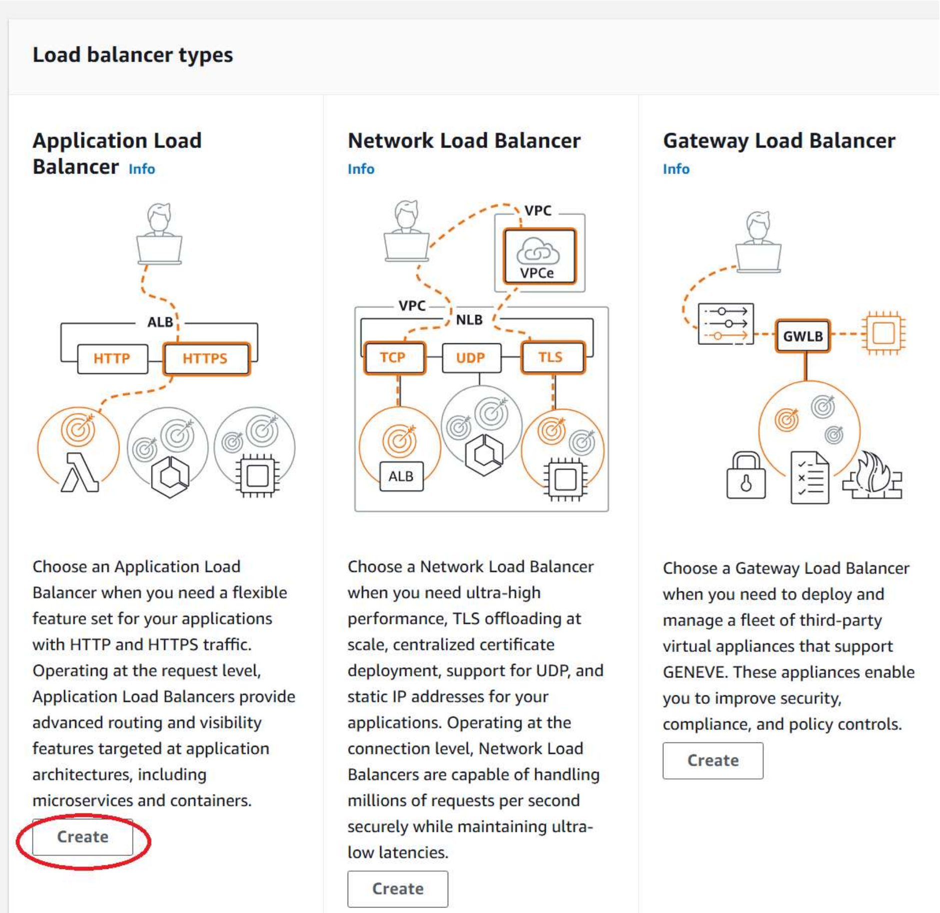

EC2 > Load balancers
Load balancers
Elastic Load Balancing scales your load balancer capacity automatically in response to changes in incoming traffic.
Actions ▼
Create load balancer
Filter load balancers
1

□	Name	DNS name	State	VPC ID	Availability Zones	Type	Date created

No load balancers

You don't have any load balancers in ap-south-1

5) There are 3 types of load balancer. So we choose according to our requirements. In this case we choose application load balancer and click create.



Load balancer types

Application Load Balancer Info

Choose an Application Load Balancer when you need a flexible feature set for your applications with HTTP and HTTPS traffic. Operating at the request level, Application Load Balancers provide advanced routing and visibility features targeted at application architectures, including microservices and containers.

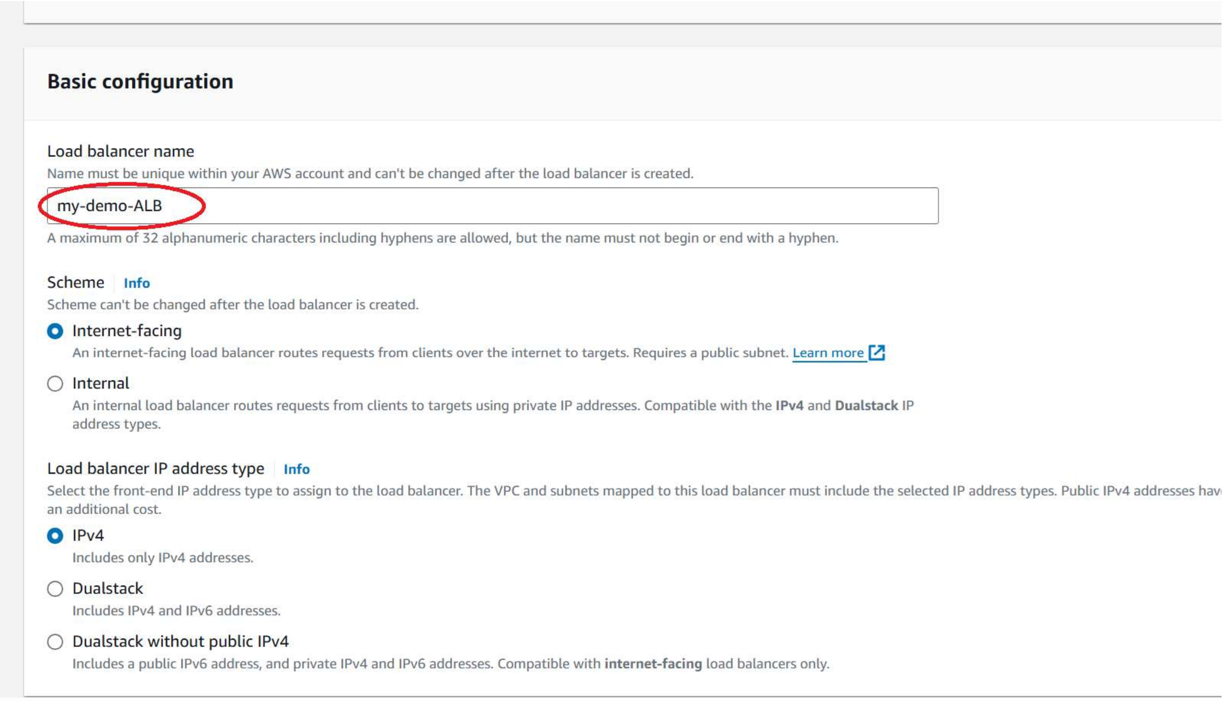
Network Load Balancer Info

Choose a Network Load Balancer when you need ultra-high performance, TLS offloading at scale, centralized certificate deployment, support for UDP, and static IP addresses for your applications. Operating at the connection level, Network Load Balancers are capable of handling millions of requests per second securely while maintaining ultra-low latencies.

Gateway Load Balancer Info

Choose a Gateway Load Balancer when you need to deploy and manage a fleet of third-party virtual appliances that support GENEVE. These appliances enable you to improve security, compliance, and policy controls.

6) Under basic configuration give load balancer name, choose scheme and IP address type.



Basic configuration

Load balancer name

Name must be unique within your AWS account and can't be changed after the load balancer is created.

my-demo-ALB

A maximum of 32 alphanumeric characters including hyphens are allowed, but the name must not begin or end with a hyphen.

Scheme Info

Scheme can't be changed after the load balancer is created.

● Internet-facing

An internet-facing load balancer routes requests from clients over the internet to targets. Requires a public subnet. Learn more

○ Internal

An internal load balancer routes requests from clients to targets using private IP addresses. Compatible with the **IPv4** and **Dualstack** IP address types.

Load balancer IP address type Info

Select the front-end IP address type to assign to the load balancer. The VPC and subnets mapped to this load balancer must include the selected IP address types. Public IPv4 addresses hav an additional cost.

● IPv4

Includes only IPv4 addresses.

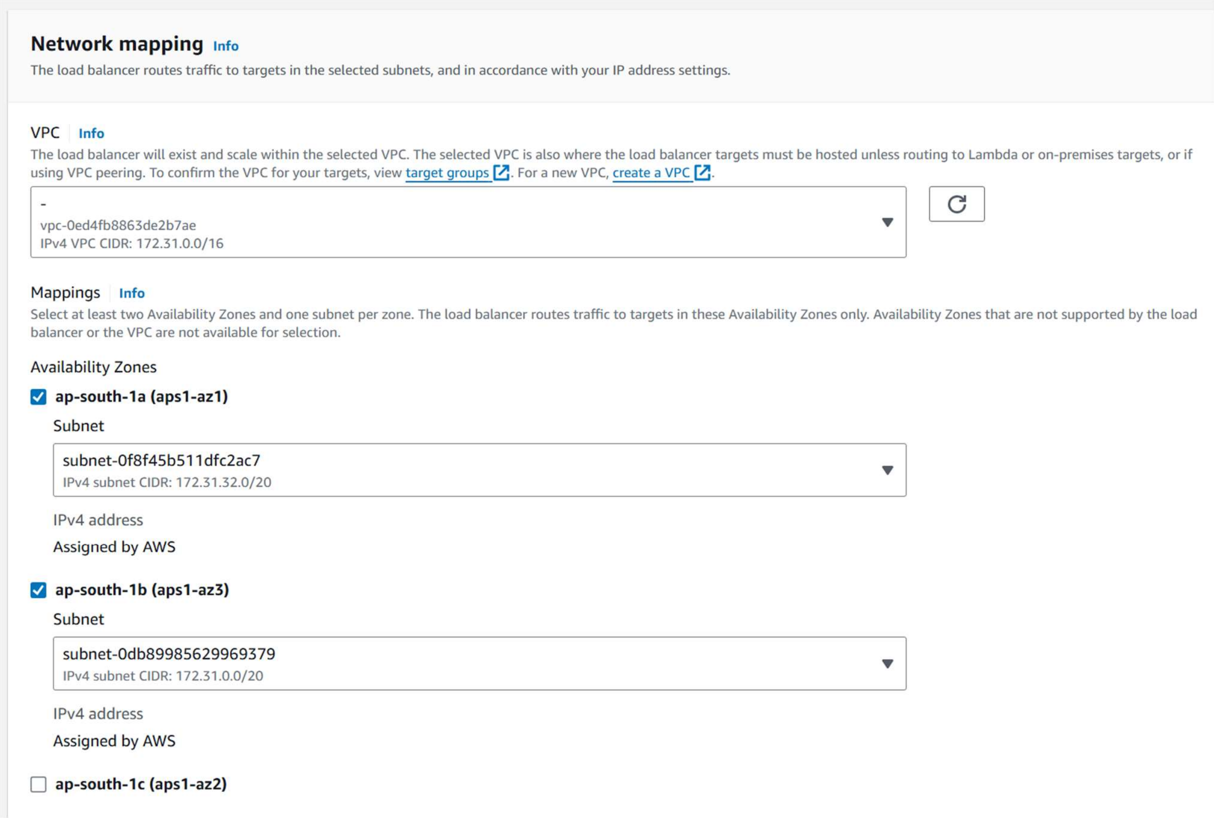
○ Dualstack

Includes IPv4 and IPv6 addresses.

○ Dualstack without public IPv4

Includes a public IPv6 address, and private IPv4 and IPv6 addresses. Compatible with **internet-facing** load balancers only.

7) Under network mapping choose the VPC and the AZ s where we want our load balancer to manage the traffic.



Network mapping Info

The load balancer routes traffic to targets in the selected subnets, and in accordance with your IP address settings.

VPC | Info

The load balancer will exist and scale within the selected VPC. The selected VPC is also where the load balancer targets must be hosted unless routing to Lambda or on-premises targets, or if using VPC peering. To confirm the VPC for your targets, view target groups. For a new VPC, create a VPC.

-
vpc-0ed4fb8863de2b7ae
IPv4 VPC CIDR: 172.31.0.0/16

Mappings | Info

Select at least two Availability Zones and one subnet per zone. The load balancer routes traffic to targets in these Availability Zones only. Availability Zones that are not supported by the load balancer or the VPC are not available for selection.

Availability Zones

☑ **ap-south-1a (aps1-az1)**

Subnet

subnet-0f8f45b511dfc2ac7
IPv4 subnet CIDR: 172.31.32.0/20

IPv4 address
Assigned by AWS

☑ **ap-south-1b (aps1-az3)**

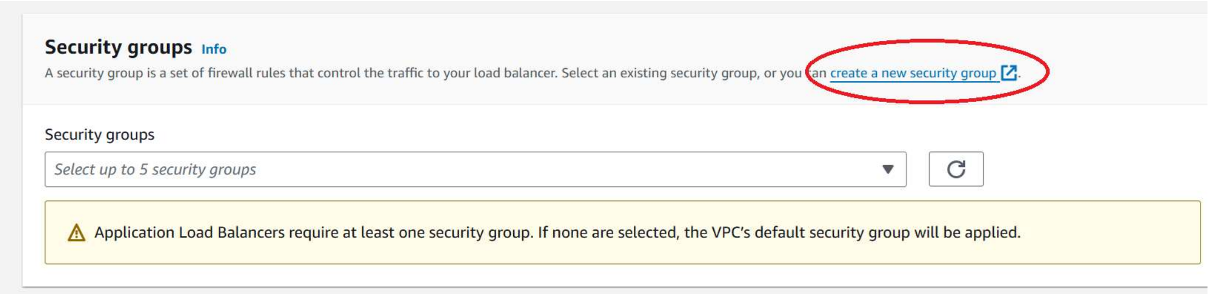
Subnet

subnet-0db89985629969379
IPv4 subnet CIDR: 172.31.0.0/20

IPv4 address
Assigned by AWS

□ **ap-south-1c (aps1-az2)**

8) Under security group create a security group with basic details for the load balancer with inbound rules allowing only HTTP traffic from anywhere. And outbound rules allowing traffic from ALB to go to the security group of our EC2 instances.



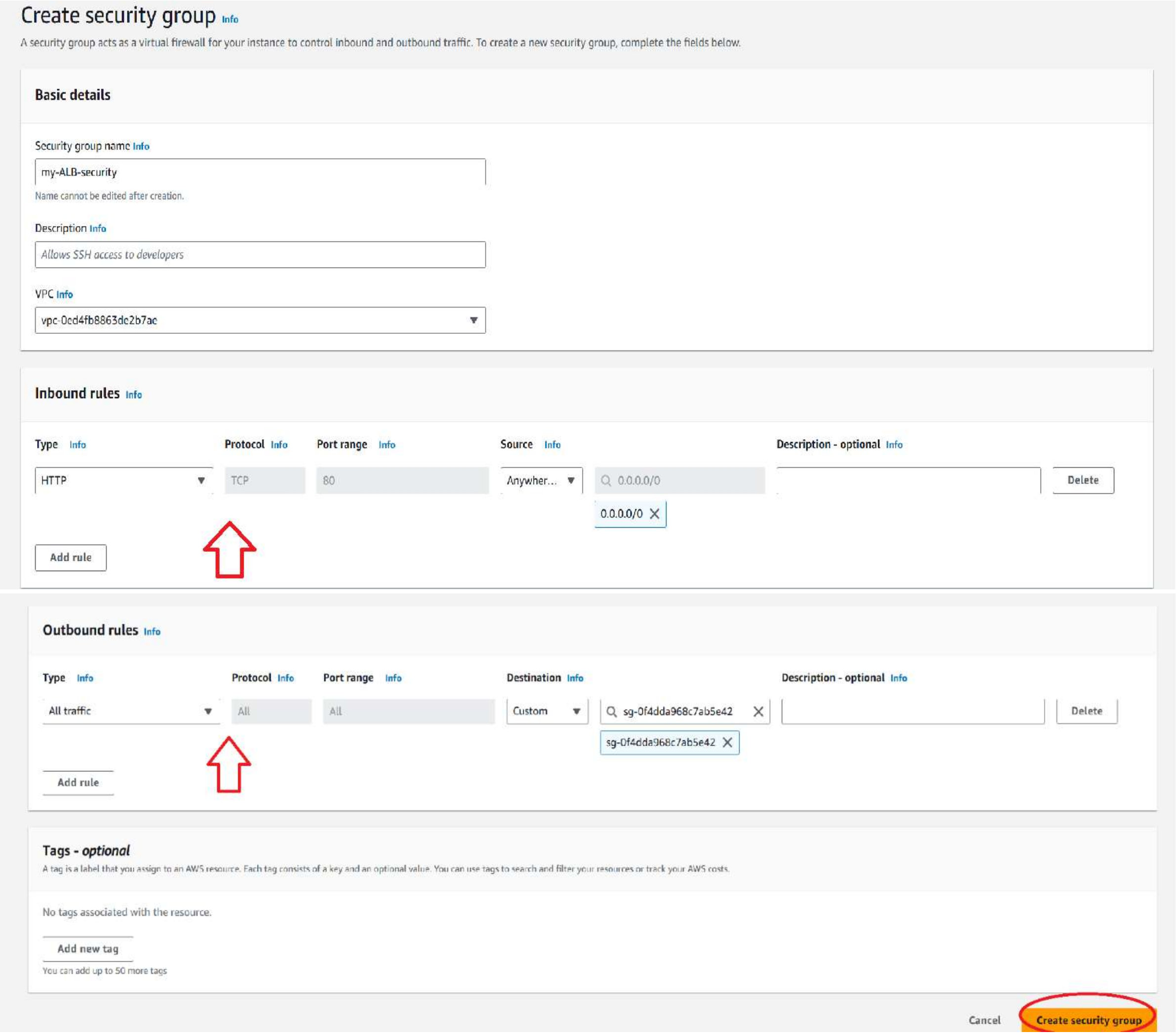
Security groups Info

A security group is a set of firewall rules that control the traffic to your load balancer. Select an existing security group, or you can create a new security group.

Security groups

Select up to 5 security groups

⚠ Application Load Balancers require at least one security group. If none are selected, the VPC's default security group will be applied.



Create security group Info

A security group acts as a virtual firewall for your instance to control inbound and outbound traffic. To create a new security group, complete the fields below.

Basic details

Security group name Info

my-ALB-security

Name cannot be edited after creation.

Description Info

Allows SSH access to developers

VPC Info

vpc-0cd4fb8863dc2b7ac

Inbound rules Info

Type	Protocol	Port range	Source		Description - optional
HTTP	TCP	80	Anywher...	0.0.0.0/0 0.0.0.0/0	

Outbound rules Info

Type	Protocol	Port range	Destination		Description - optional
All traffic	All	All	Custom	sg-0f4dda968c7ab5e42 sg-0f4dda968c7ab5e42	

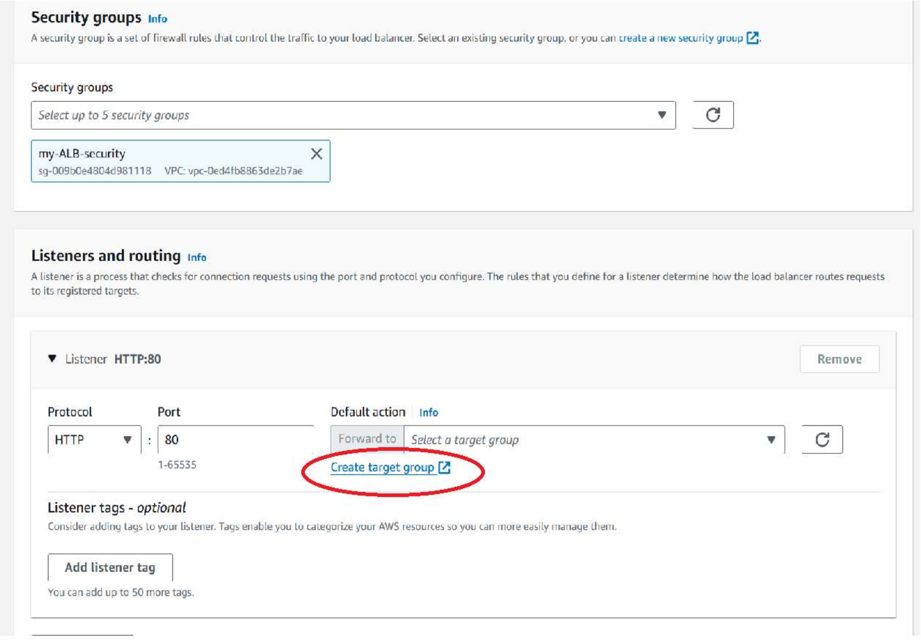
Tags - *optional*

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

No tags associated with the resource.

You can add up to 50 more tags

9) Under listeners and routing create a target group .



Security groups Info

A security group is a set of firewall rules that control the traffic to your load balancer. Select an existing security group, or you can create a new security group.

Security groups

Select up to 5 security groups

my-ALB-security
sg-009b0e4804d981118 VPC: vpc-0ed4fb8363de2b7ae

Listeners and routing Info

A listener is a process that checks for connection requests using the port and protocol you configure. The rules that you define for a listener determine how the load balancer routes requests to its registered targets.

▼ Listener **HTTP:80**

Remove

Protocol	Port	Default action Info
HTTP	80	Forward to: Select a target group
	1-65535	Create target group

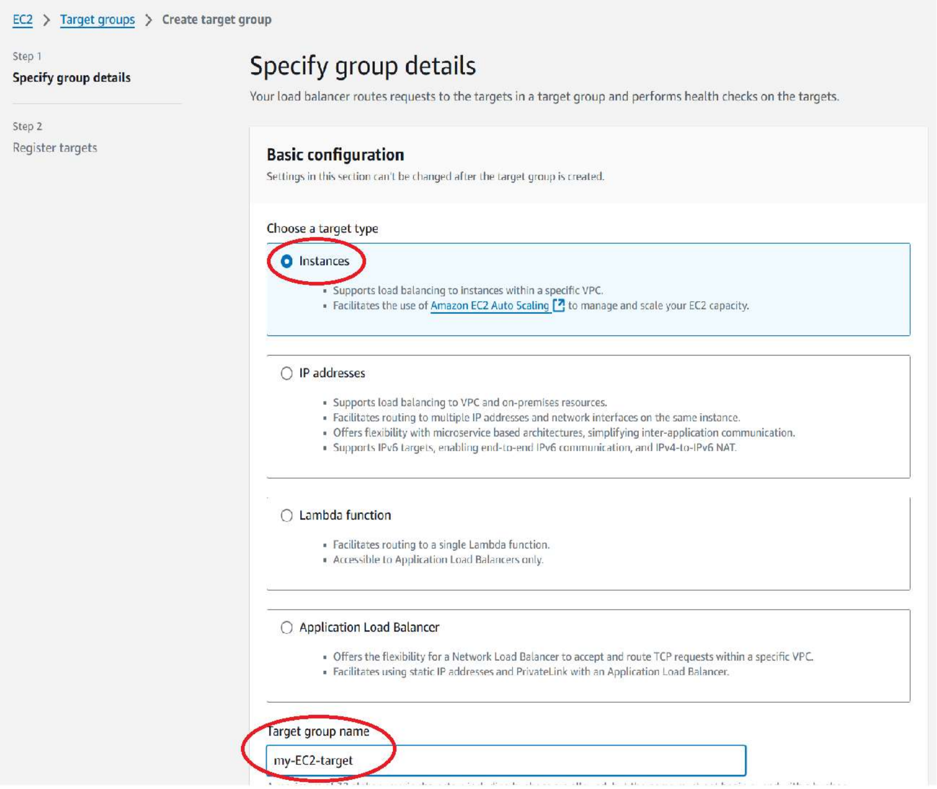
Listener tags - *optional*

Consider adding tags to your listener. Tags enable you to categorize your AWS resources so you can more easily manage them.

Add listener tag

You can add up to 50 more tags.

10) In group details choose the targets as instances and give a name.



EC2 > Target groups > Create target group

Step 1
Specify group details

Step 2
Register targets

Specify group details

Your load balancer routes requests to the targets in a target group and performs health checks on the targets.

Basic configuration

Settings in this section can't be changed after the target group is created.

Choose a target type

● Instances
- Supports load balancing to instances within a specific VPC.
- Facilitates the use of Amazon EC2 Auto Scaling to manage and scale your EC2 capacity.

○ IP addresses
- Supports load balancing to VPC and on-premises resources.
- Facilitates routing to multiple IP addresses and network interfaces on the same instance.
- Offers flexibility with microservice based architectures, simplifying inter-application communication.
- Supports IPv6 targets, enabling end-to-end IPv6 communication, and IPv4-to-IPv6 NAT.

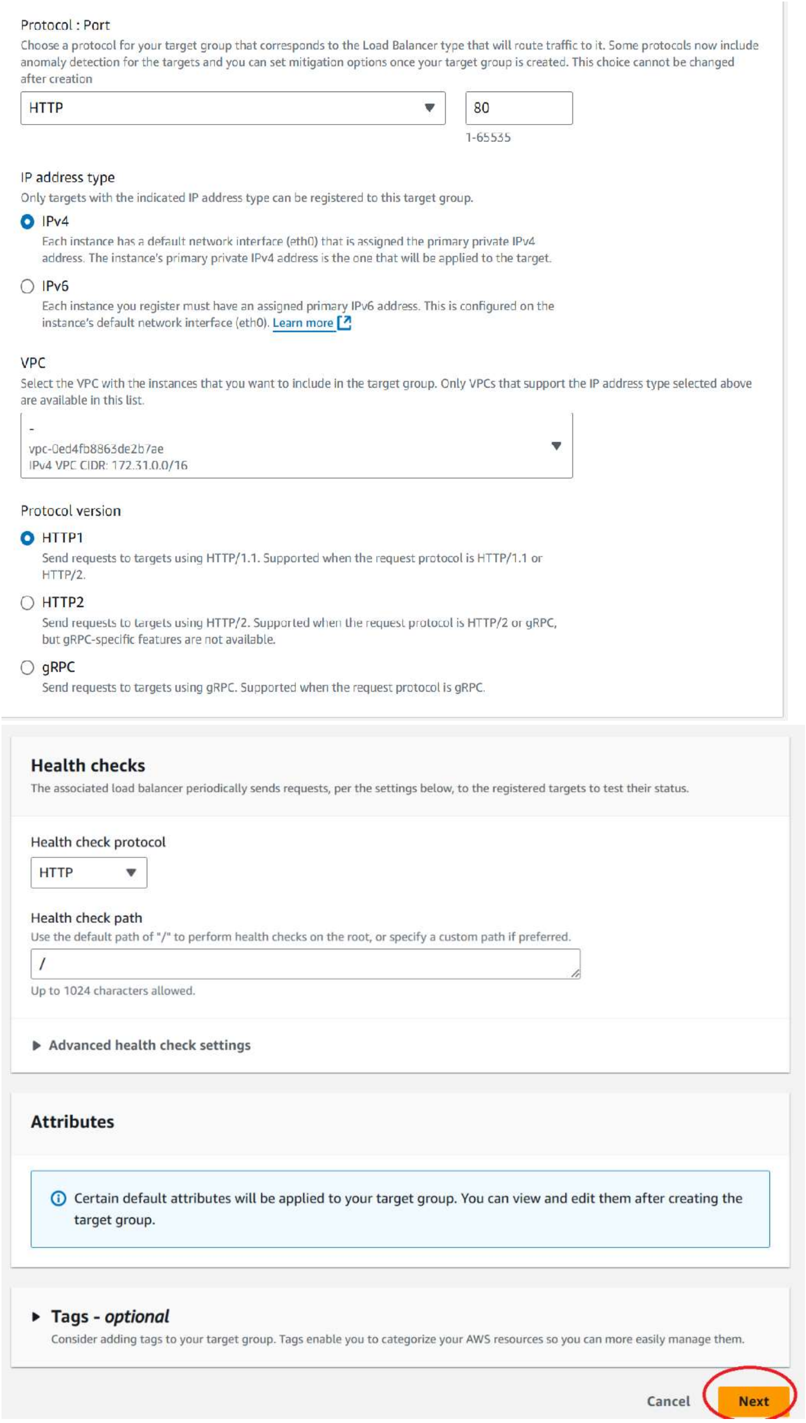
○ Lambda function
- Facilitates routing to a single Lambda function.
- Accessible to Application Load Balancers only.

○ Application Load Balancer
- Offers the flexibility for a Network Load Balancer to accept and route TCP requests within a specific VPC.
- Facilitates using static IP addresses and PrivateLink with an Application Load Balancer.

Target group name

my-EC2-target

11) Choose the protocol, IP address type, VPC, health checks and then click next.



Protocol : Port

Choose a protocol for your target group that corresponds to the Load Balancer type that will route traffic to it. Some protocols now include anomaly detection for the targets and you can set mitigation options once your target group is created. This choice cannot be changed after creation

HTTP ▼ | 80

1-65535

IP address type

Only targets with the indicated IP address type can be registered to this target group.

● IPv4

Each instance has a default network interface (eth0) that is assigned the primary private IPv4 address. The instance's primary private IPv4 address is the one that will be applied to the target.

○ IPv6

Each instance you register must have an assigned primary IPv6 address. This is configured on the instance's default network interface (eth0). Learn more

VPC

Select the VPC with the instances that you want to include in the target group. Only VPCs that support the IP address type selected above are available in this list.

-
vpc-0ed4fb8863de2b7ae
IPv4 VPC CIDR: 172.31.0.0/16 ▼

Protocol version

● HTTP1

Send requests to targets using HTTP/1.1. Supported when the request protocol is HTTP/1.1 or HTTP/2.

○ HTTP2

Send requests to targets using HTTP/2. Supported when the request protocol is HTTP/2 or gRPC, but gRPC-specific features are not available.

○ gRPC

Send requests to targets using gRPC. Supported when the request protocol is gRPC.

Health checks

The associated load balancer periodically sends requests, per the settings below, to the registered targets to test their status.

Health check protocol

HTTP ▼

Health check path

Use the default path of "/" to perform health checks on the root, or specify a custom path if preferred.

/

Up to 1024 characters allowed.

▶ Advanced health check settings

Attributes

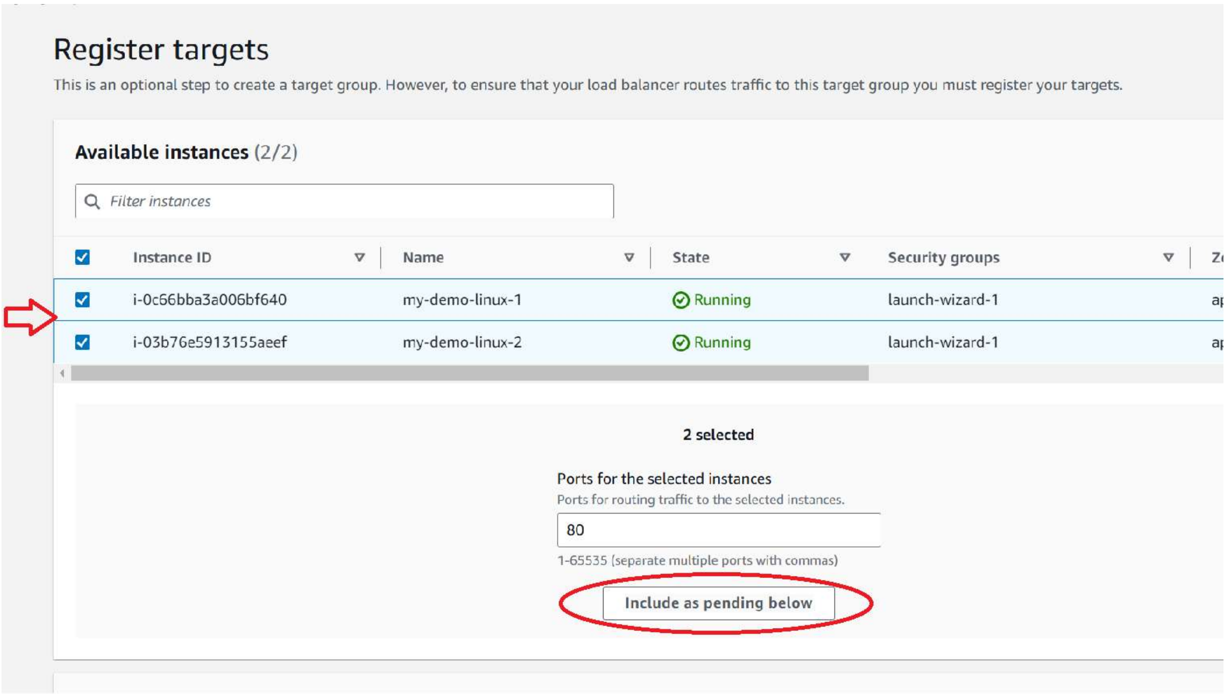
ⓘ Certain default attributes will be applied to your target group. You can view and edit them after creating the target group.

▶ **Tags - *optional***

Consider adding tags to your target group. Tags enable you to categorize your AWS resources so you can more easily manage them.

Cancel | Next

12) Under register targets choose the 2 instances we created earlier and click include as pending below.



Register targets

This is an optional step to create a target group. However, to ensure that your load balancer routes traffic to this target group you must register your targets.

Available instances (2/2)

Filter instances

☑	Instance ID	Name	State	Security groups	Z
☑	i-0c66bba3a006bf640	my-demo-linux-1	✓ Running	launch-wizard-1	ap
☑	i-03b76e5913155aeef	my-demo-linux-2	✓ Running	launch-wizard-1	ap

2 selected

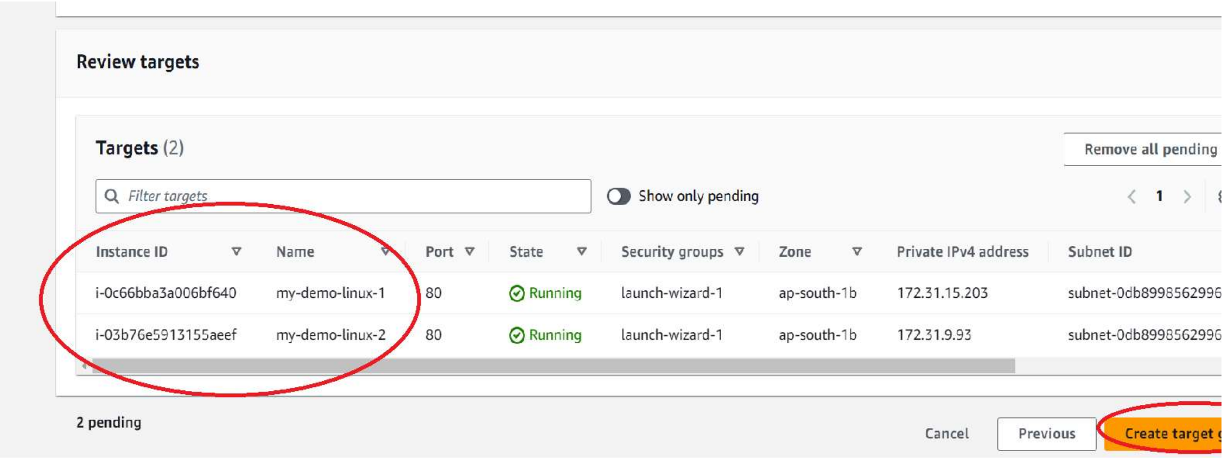
Ports for the selected instances

Ports for routing traffic to the selected instances.

80

1-65535 (separate multiple ports with commas)

13) Under review targets we can see both the instances and click create target group.



Review targets

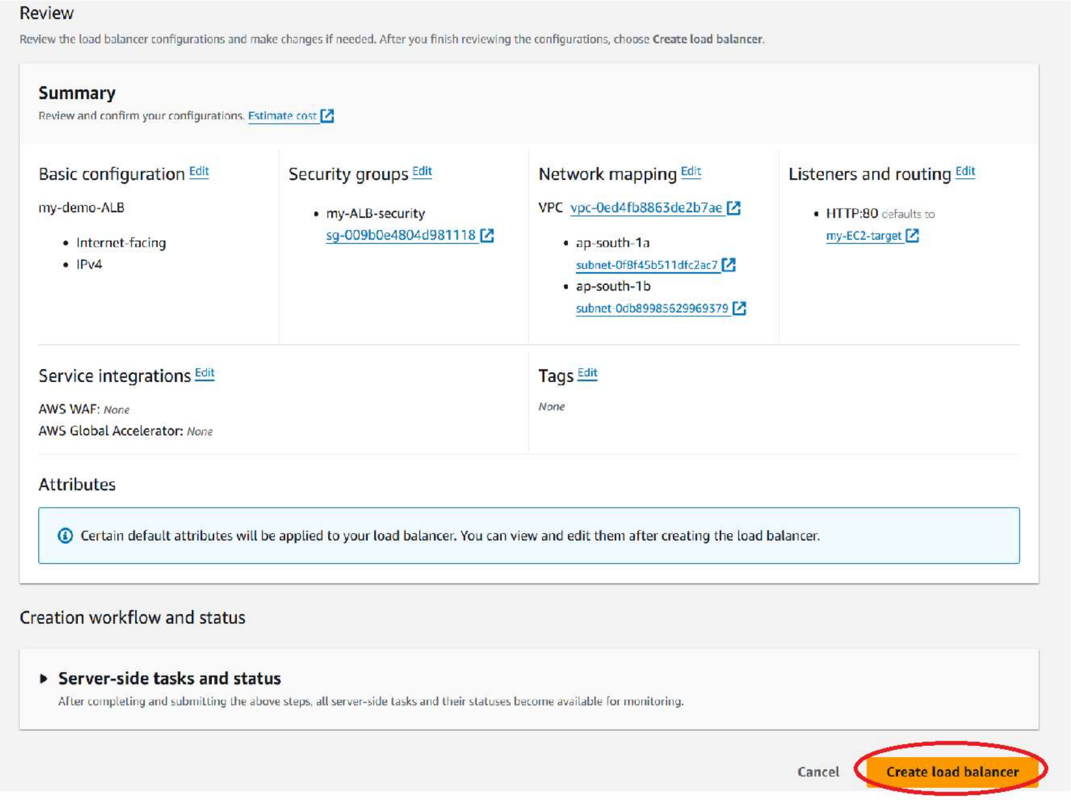
Targets (2)

Filter targets

Instance ID	Name	Port	State	Security groups	Zone	Private IPv4 address	Subnet ID
i-0c66bba3a006bf640	my-demo-linux-1	80	✓ Running	launch-wizard-1	ap-south-1b	172.31.15.203	subnet-0db8998562996
i-03b76e5913155aeef	my-demo-linux-2	80	✓ Running	launch-wizard-1	ap-south-1b	172.31.9.93	subnet-0db8998562996

2 pending

14) Review all the specifications of ALB and click create load balancer.



Review

Review the load balancer configurations and make changes if needed. After you finish reviewing the configurations, choose **Create load balancer**.

Summary

Review and confirm your configurations. Estimate cost

Basic configuration Edit

my-demo-ALB

- Internet-facing
- IPv4

Security groups Edit

- my-ALB-security sg-009b0e4804d981118

Network mapping Edit

VPC vpc-0ed4fb8863de2b7ae

- ap-south-1a subnet-0f8f45b511dfc2ac7
- ap-south-1b subnet-0db89985629969379

Listeners and routing Edit

- HTTP:80 defaults to my-EC2-target

Service integrations Edit

AWS WAF: *None*

AWS Global Accelerator: *None*

Tags Edit

None

Attributes

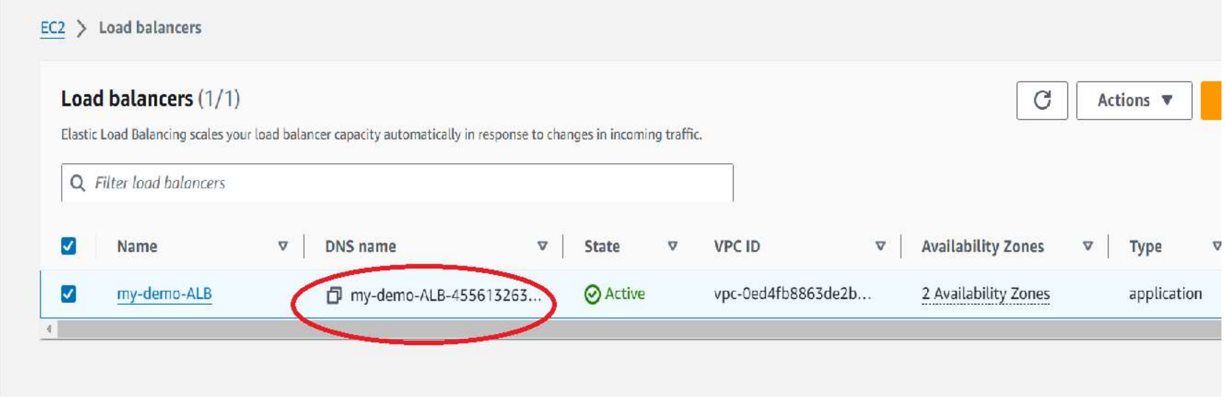
Certain default attributes will be applied to your load balancer. You can view and edit them after creating the load balancer.

Creation workflow and status

Server-side tasks and status

After completing and submitting the above steps, all server-side tasks and their statuses become available for monitoring.

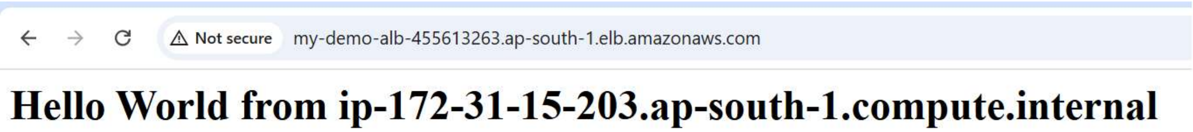
15) Our ALB is active and it got a DNS. If we copy the DNS and paste it in the browser we can see the traffic is equally distributed between both the EC2 instances.(round robin algorithm).



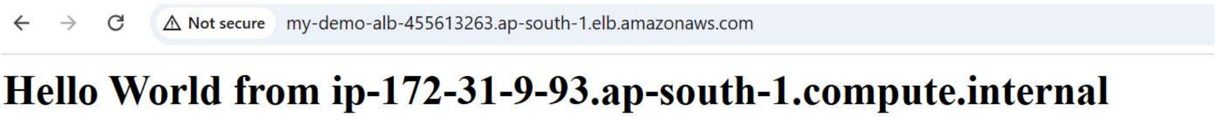
Load balancers (1/1)

Elastic Load Balancing scales your load balancer capacity automatically in response to changes in incoming traffic.

☑	Name	DNS name	State	VPC ID	Availability Zones	Type
☑	my-demo-ALB	my-demo-ALB-455613263...	Active	vpc-0ed4fb8863de2b...	2 Availability Zones	application



Hello World from ip-172-31-15-203.ap-south-1.compute.internal



Hello World from ip-172-31-9-93.ap-south-1.compute.internal

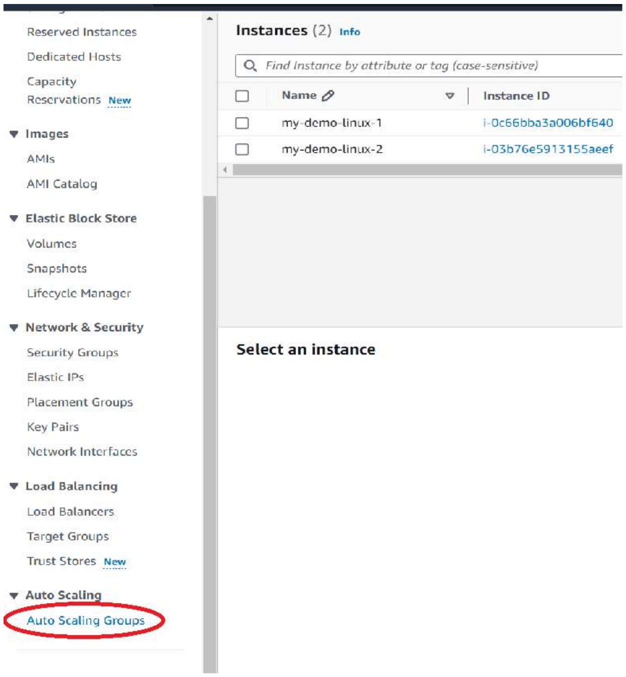
Manage the scaling requirements of the company by:

a. Deploying multiple compute resources on the cloud as soon as the load increases and the CPU utilization exceeds 80%

b. Removing the resources when the CPU utilization goes under 60%

solution:

1) Click on autoscaling groups left side of EC2 console.



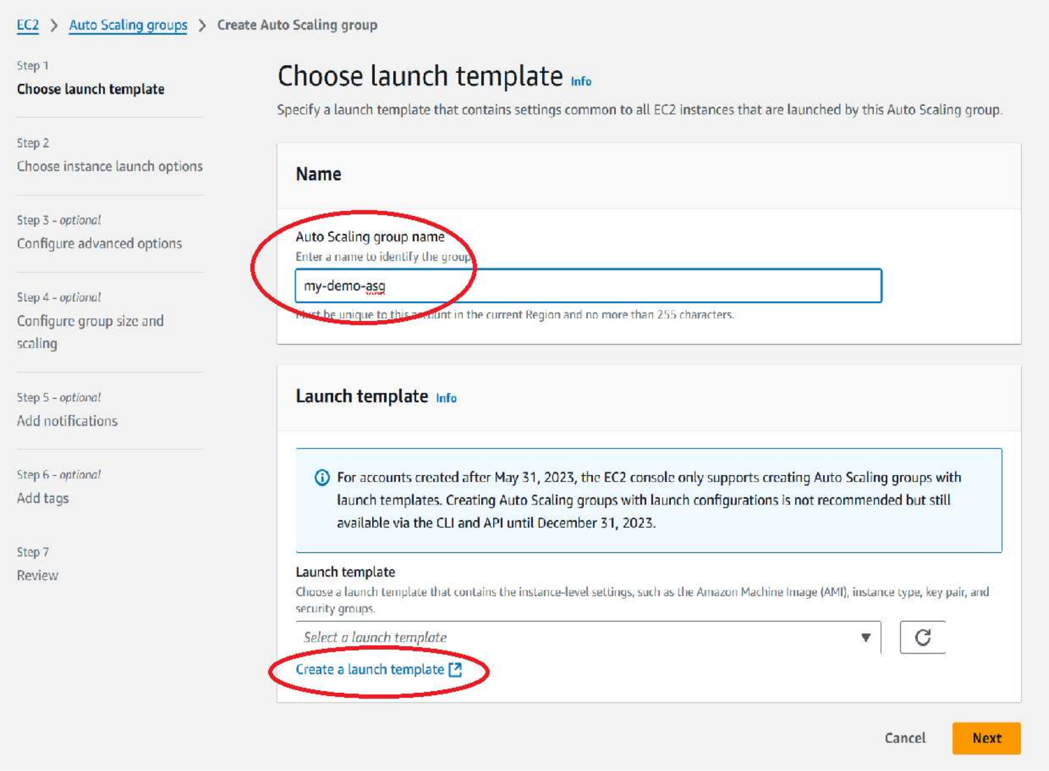
Instances (2) Info

Find instance by attribute or tag (case-sensitive)

□	Name	Instance ID
□	my-demo-linux-1	i-0c66bba3a006bf640
□	my-demo-linux-2	i-03b76e5913155aeef

Select an instance

2) Give a autoscaling group name and click create launch template.


EC2 > Auto Scaling groups > Create Auto Scaling group
Step 1
Choose launch template
Step 2
Choose instance launch options
Step 3 - optional
Configure advanced options
Step 4 - optional
Configure group size and scaling
Step 5 - optional
Add notifications
Step 6 - optional
Add tags
Step 7
Review

Choose launch template Info

Specify a launch template that contains settings common to all EC2 instances that are launched by this Auto Scaling group.

Name

Auto Scaling group name

Enter a name to identify the group.

my-demo-asg

Must be unique to this account in the current Region and no more than 255 characters.

Launch template Info

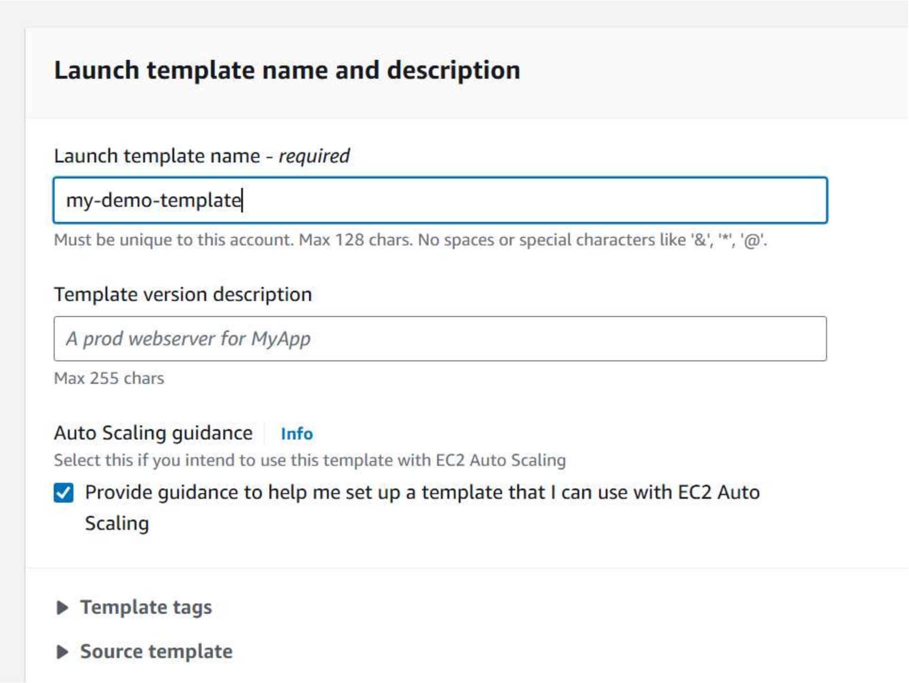
For accounts created after May 31, 2023, the EC2 console only supports creating Auto Scaling groups with launch templates. Creating Auto Scaling groups with launch configurations is not recommended but still available via the CLI and API until December 31, 2023.

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.

Select a launch template

3) Give a template name and as we create EC2 instances all the steps are same.



Launch template name and description

Launch template name - *required*

my-demo-template

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

Template version description

A prod webserver for MyApp

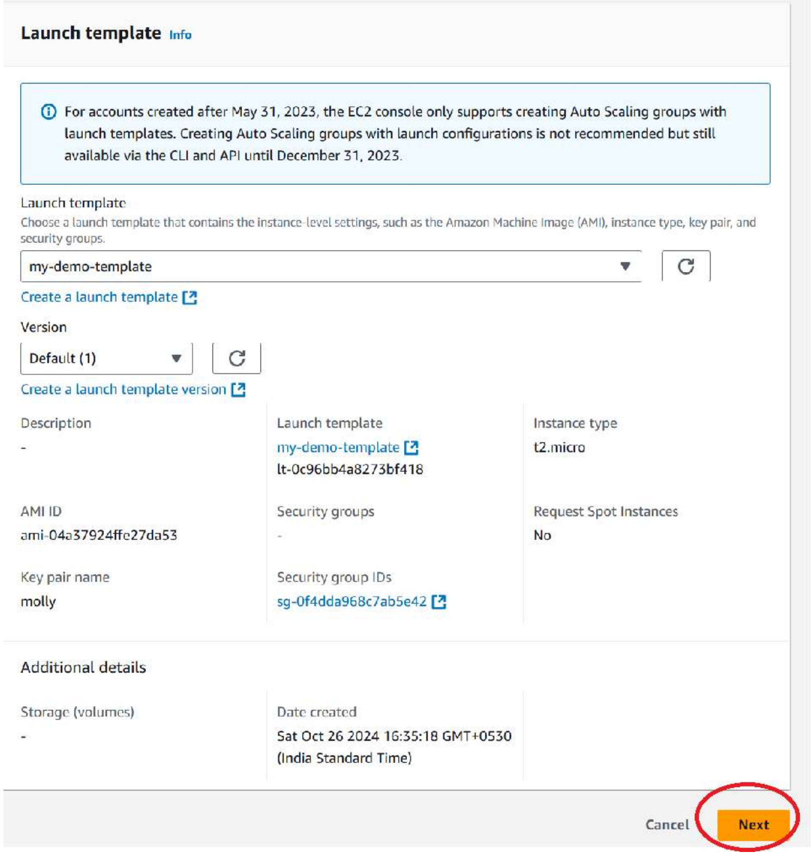
Max 255 chars

Auto Scaling guidance Info

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

4) After launch template created click next.



Launch template Info

For accounts created after May 31, 2023, the EC2 console only supports creating Auto Scaling groups with launch templates. Creating Auto Scaling groups with launch configurations is not recommended but still available via the CLI and API until December 31, 2023.

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.

my-demo-template

Create a launch template

Version

Default (1)

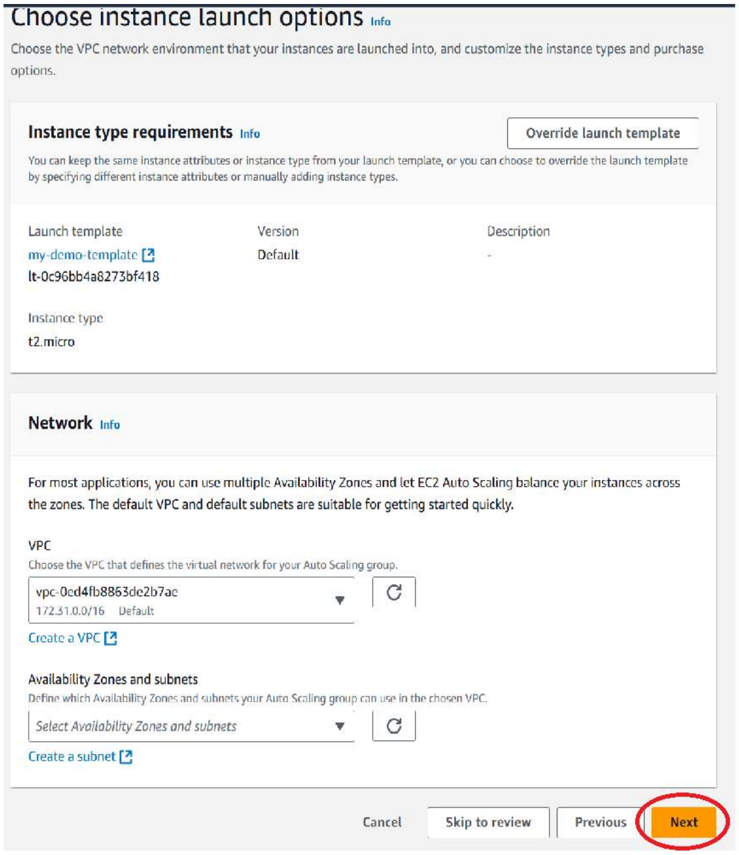
Create a launch template version

Description	Launch template	Instance type
-	my-demo-template lt-0c96bb4a8273bf418	t2.micro
AMI ID	Security groups	Request Spot Instances
ami-04a37924ffe27da53	-	No
Key pair name	Security group IDs	
molly	sg-0f4dda968c7ab5e42	

Additional details

Storage (volumes)	Date created
-	Sat Oct 26 2024 16:35:18 GMT+0530 (India Standard Time)

5) Under network choose the network specifications and click next



Choose instance launch options Info

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

Instance type requirements Info

You can keep the same instance attributes or instance type from your launch template, or you can choose to override the launch template by specifying different instance attributes or manually adding instance types.

Launch template	Version	Description
my-demo-template lt-0c96bb4a8273bf418	Default	-

Instance type
t2.micro

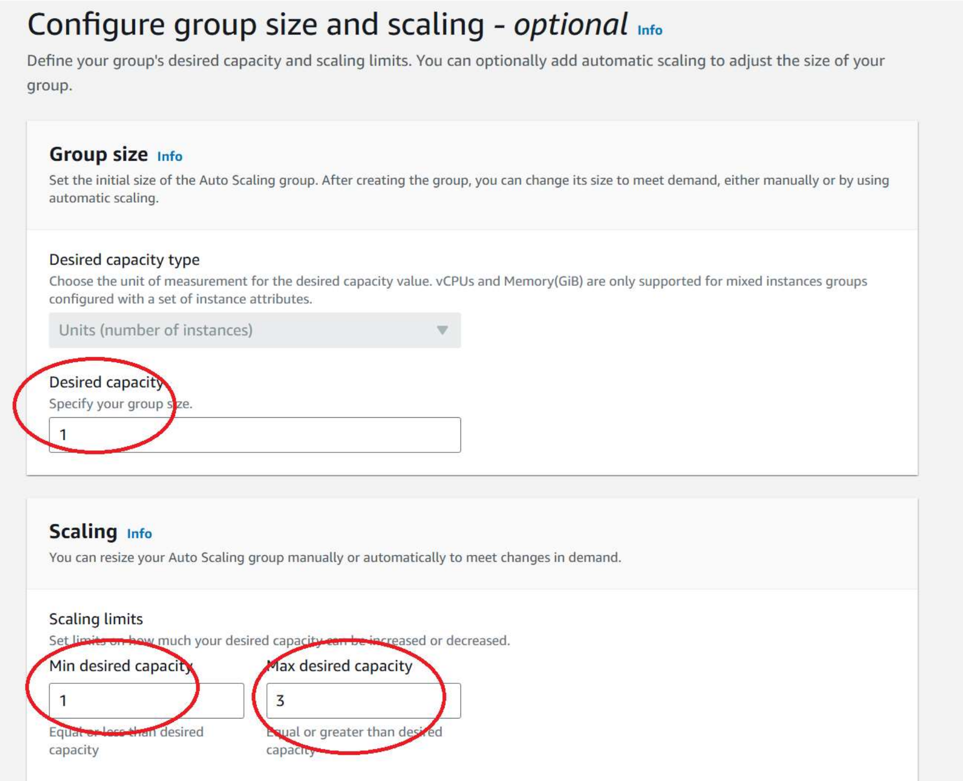
Network Info

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.
vpc-0ed4fb8863de2b7ae
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
Create a subnet

6) Under group size and scaling mention the desired, minimum and maximum capacity.



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

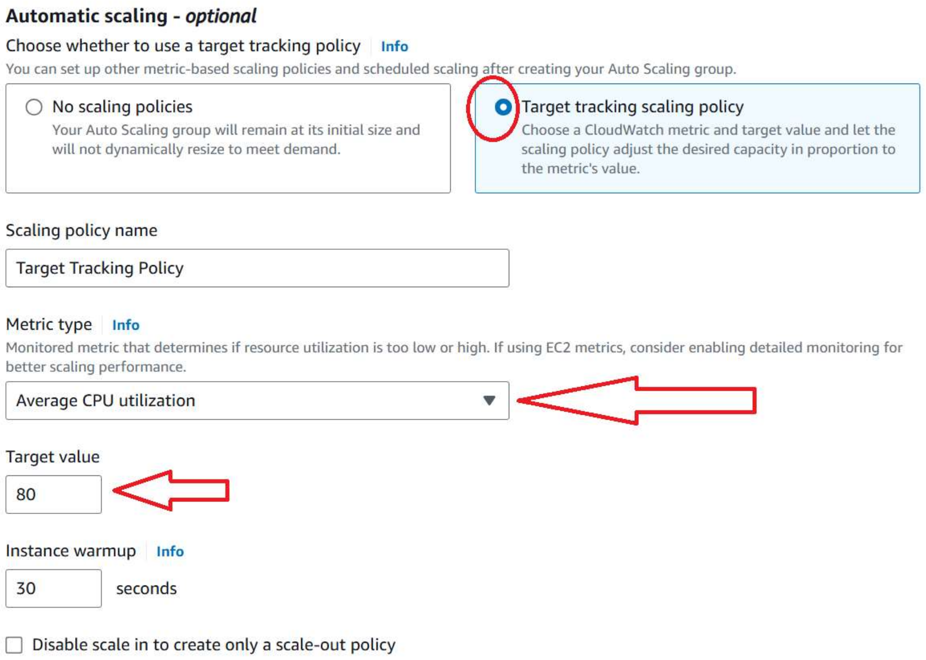
You can resize your Auto Scaling group manually or automatically to meet changes in demand.

Scaling limits

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

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

7) Under automatic scaling choose target tracking scaling policy, choose metric type CPU utilization and give the target value as 80. Review and then create.



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.

Scaling policy name

Target Tracking Policy

Metric type Info

Monitored metric that determines if resource utilization is too low or high. If using EC2 metrics, consider enabling detailed monitoring for better scaling performance.

Average CPU utilization ▼

Target value

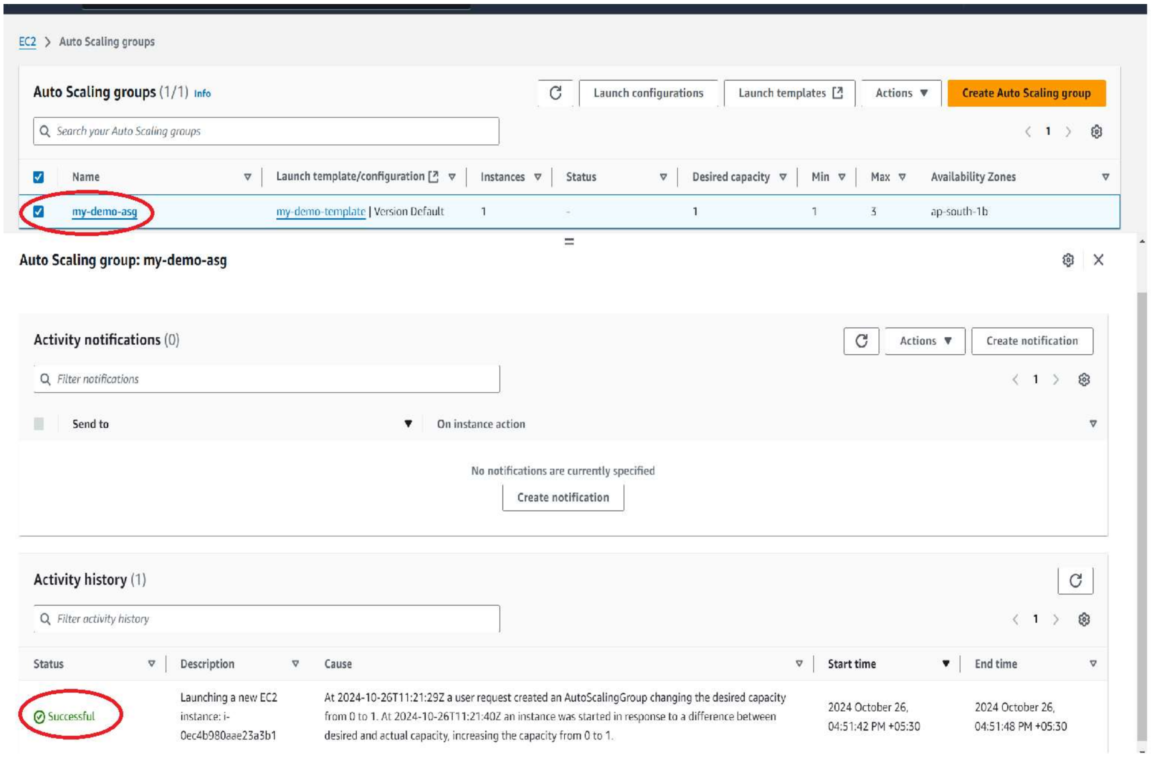
80

Instance warmup Info

30 seconds

□ Disable scale in to create only a scale-out policy

8) We can see that ASG is active and under activity we can see it has automatically created a EC2 instance.


EC2 > Auto Scaling groups
Launch configurations
Launch templates
Actions ▼
Create Auto Scaling group
Search your Auto Scaling groups

Auto Scaling groups (1/1) Info

☑	Name	Launch template/configuration	Instances	Status	Desired capacity	Min	Max	Availability Zones
☑	my-demo-asg	my-demo-template | Version Default	1	-	1	1	3	ap-south-1b

Auto Scaling group: my-demo-asg

Activity notifications (0)

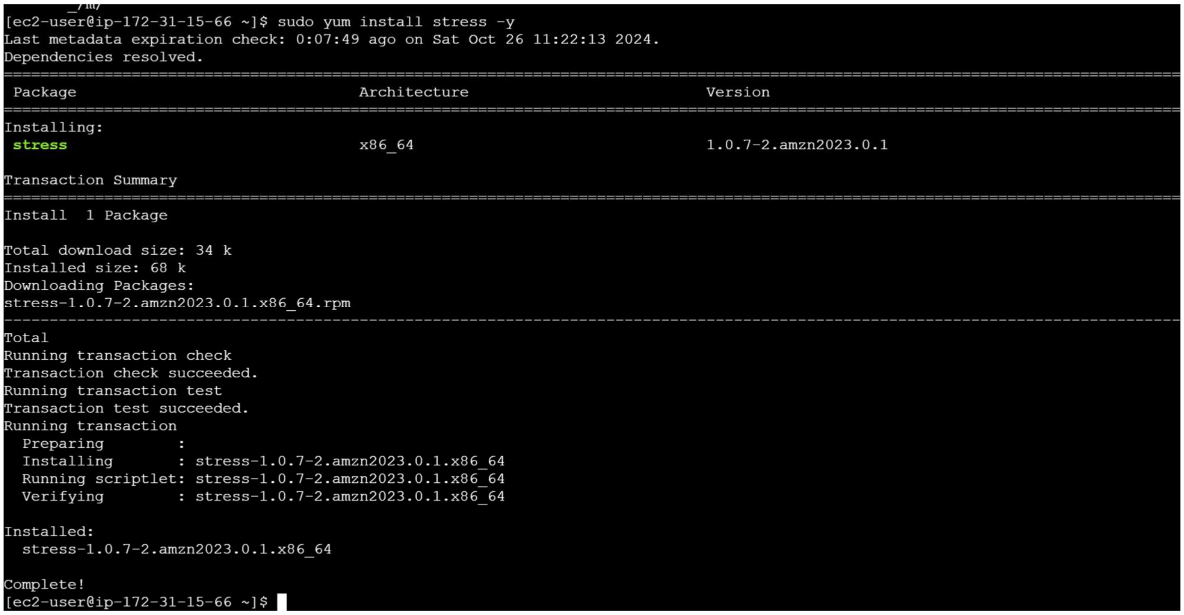
□	Send to	On instance action

No notifications are currently specified

Activity history (1)

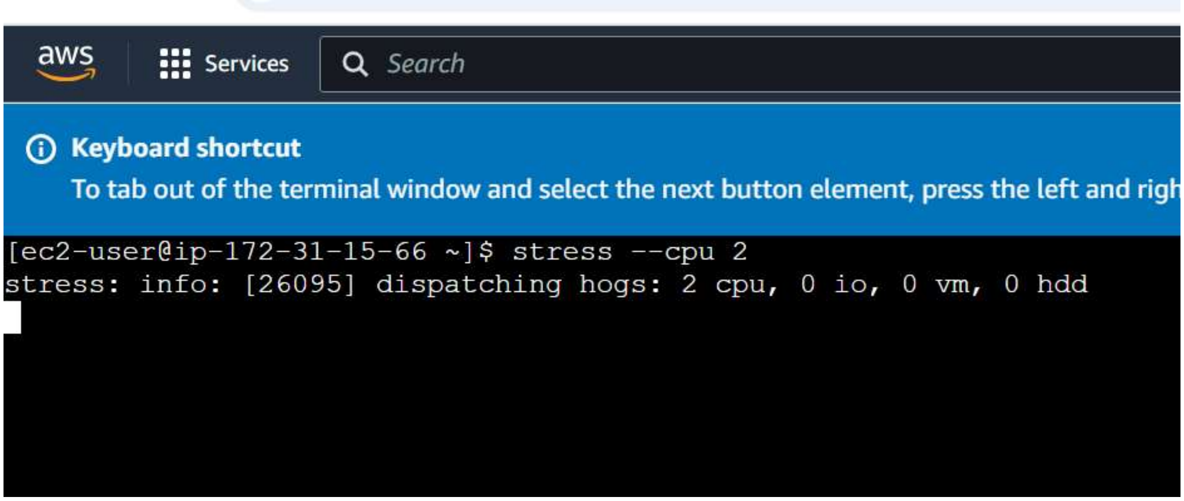
Status	Description	Cause	Start time	End time
✓ Successful	Launching a new EC2 instance: i-0ec4b980aae23a3b1	At 2024-10-26T11:21:29Z a user request created an AutoScalingGroup changing the desired capacity from 0 to 1. At 2024-10-26T11:21:40Z an instance was started in response to a difference between desired and actual capacity, increasing the capacity from 0 to 1.	2024 October 26, 04:51:42 PM +05:30	2024 October 26, 04:51:48 PM +05:30

9) Now to check the scaling we will put stress to our EC2. Connect to the EC2 . Then install stress package using (sudo yum install stress -y)command



```
[ec2-user@ip-172-31-15-66 ~]$ sudo yum install stress -y
Last metadata expiration check: 0:07:49 ago on Sat Oct 26 11:22:13 2024.
Dependencies resolved.
==========================================================================================================================
 Package                              Architecture                          Version
==========================================================================================================================
Installing:
 stress                               x86_64                                1.0.7-2.amzn2023.0.1

Transaction Summary
==========================================================================================================================
Install  1 Package

Total download size: 34 k
Installed size: 68 k
Downloading Packages:
stress-1.0.7-2.amzn2023.0.1.x86_64.rpm
--------------------------------------------------------------------------------------------------------------------------
Total
Running transaction check
Transaction check succeeded.
Running transaction test
Transaction test succeeded.
Running transaction
  Preparing        :
  Installing       : stress-1.0.7-2.amzn2023.0.1.x86_64
  Running scriptlet: stress-1.0.7-2.amzn2023.0.1.x86_64
  Verifying        : stress-1.0.7-2.amzn2023.0.1.x86_64

Installed:
  stress-1.0.7-2.amzn2023.0.1.x86_64

Complete!
[ec2-user@ip-172-31-15-66 ~]$
```

10) Give stress to your instance (stress --cpu 2) command



Keyboard shortcut
To tab out of the terminal window and select the next button element, press the left and righ

```
[ec2-user@ip-172-31-15-66 ~]$ stress --cpu 2
stress: info: [26095] dispatching hogs: 2 cpu, 0 io, 0 vm, 0 hdd
```

11) After some time we can see ASG automatically creates another instance as CPU utilization increases and later the Instance is removed as CPU utilization decreases.

Route the traffic to the company's domain

Solution:

1) First register the domain name that you want your users to use to access your content.

a) how you register a domain name with Amazon Route 53:

You choose a domain name and confirm that it's available, meaning that no one else has registered the domain name that you want.

If the domain name you want is already in use, you can try other names or try changing only the top-level domain, such as .com, to another top-level domain, such as .ninja or .hockey.

b) When you register a domain with Route 53, the service automatically makes itself the DNS service for the domain by Creating a hosted zone that has the same name as your domain.
   And Assigns a set of four name servers to the hosted zone. Gets the name servers from the hosted zone and adds them to the domain.

2) After you register your domain name, Route 53 automatically creates a public hosted zone that has the same name as the domain.

3) To route traffic to your resources, you create records, also known as resource record sets, in your hosted zone. Each record includes information about how you want to route traffic for your domain, such as the following:

   a) Name
      The name of the record corresponds with the domain name (example.com) or subdomain name (www.example.com, retail.example.com) that you want Route 53 to route traffic for.The name of every record in a hosted zone must end with the name of the hosted zone. For example, if the name of the hosted zone is example.com, all record names must end in example.com. The Route 53 console does this for you automatically.

b) Type
   The record type usually determines the type of resource that you want traffic to be routed to. For example, to route traffic to an email server, you specify MX for Type. To route traffic to a web server that has an IPv4 IP address, you specify A for Type.

c) Value
   Value is closely related to Type. If you specify MX for Type, you specify the names of one or more email servers for Value. If you specify A for Type, you specify an IP address in IPv4 format, such as 192.0.2.136.

d) You can also create special Route 53 records, called alias records, that route traffic to Amazon S3 buckets, Amazon CloudFront distributions, and other AWS resources