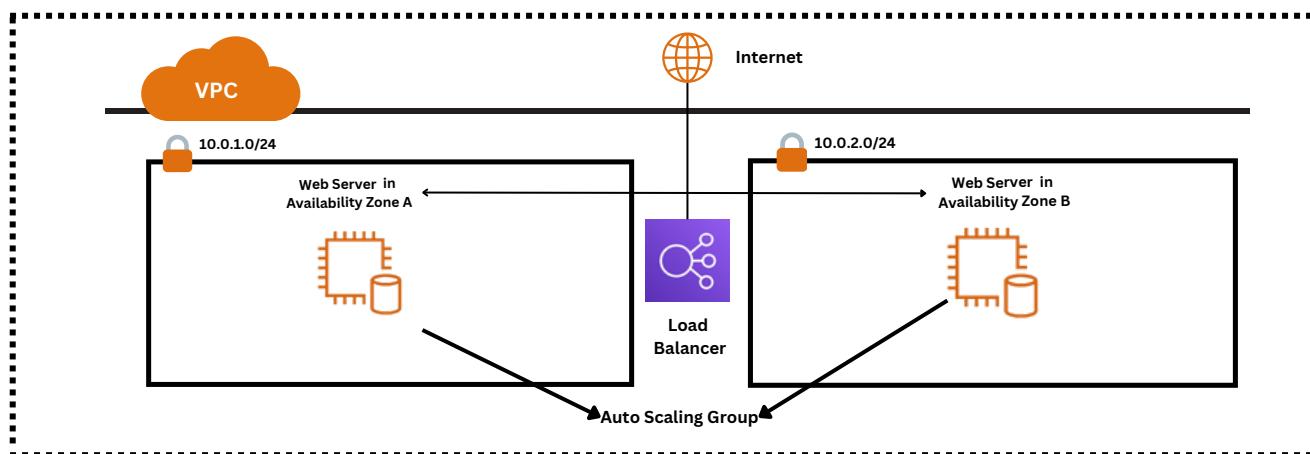


Elastic Load Balancing & Auto Scaling Groups Section



Scalability & High Availability

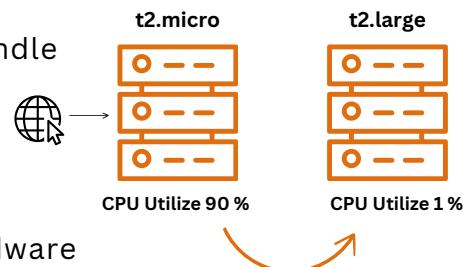
Scalability means a system can handle more work when needed.

- **Vertical Scalability** – Upgrading a single server with more RAM and a faster processor (like upgrading your laptop to a more powerful one)
- **Horizontal Scalability** – Adding more servers to distribute the workload (like using multiple laptops to process tasks together).

Vertical Scalability

Vertical Scalability means making a single server more powerful by upgrading its hardware.

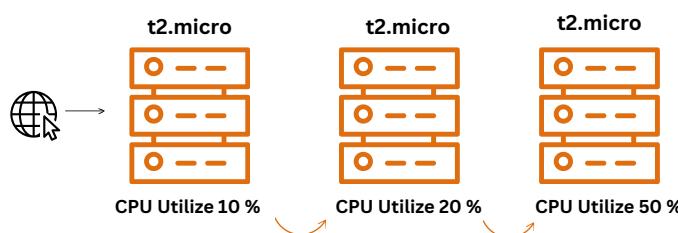
- **Example:** Your application runs on a **t2.micro server**. To handle more load, you upgrade it to a **t2.large server**.
- This type of scaling is common for systems that aren't distributed, like databases.
- **Limitation:** There's a maximum upgrade limit based on hardware capacity.



Horizontal Scalability

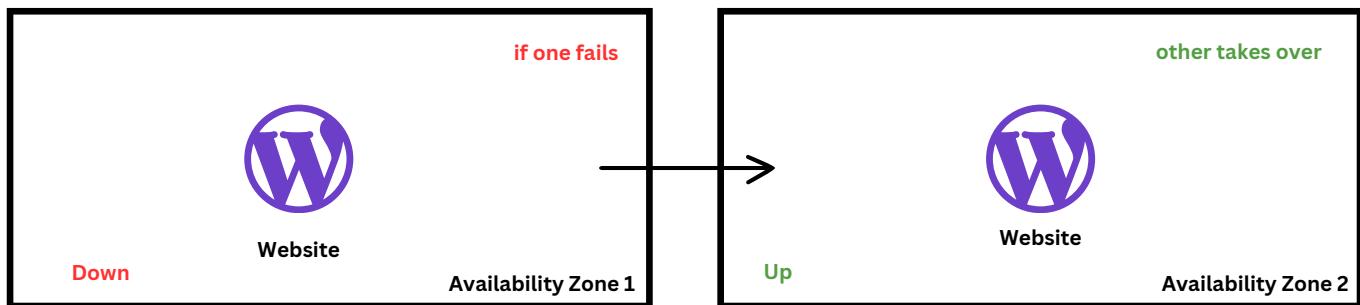
Horizontal Scalability means adding more servers to handle more traffic or workload

- **Example:** Instead of upgrading a single server with more RAM, you add multiple servers to share the work.
- This method is common for web applications and modern systems that need to handle many users.
- Cloud services like Amazon EC2 make it easy to add more servers when needed.





High Availability



Running your application on two or more servers in different locations

High Availability

High Availability means keeping your system running even if one part fails.

- It usually works with horizontal scaling (multiple servers).
- **Example:** Running your application on two or more servers in different locations, so if one fails, the other takes over.
- The goal is to prevent downtime, even if a data center has a problem or stops working.

High Availability & Scalability for EC2 (Simple Explanation)

Vertical Scaling: Make a single server bigger (scale up or down).

- **Example:** Start with a small server (**like t2.nano with 0.5GB of RAM**) and upgrade to a much larger server as your needs grow.

Horizontal Scaling: Add more servers to handle more traffic (scale out or in).

- **Example:** Use Auto Scaling Groups to automatically add or remove servers.
- Use a Load Balancer to distribute the work evenly across all servers.

High Availability: Run your servers in multiple locations (Availability Zones) to ensure the system keeps running, even if one location fails.

- **Example:** Set up your **Auto Scaling Group** and **Load Balancer** to work across **multiple Availability Zones** for better reliability and uptime.

Scalability vs Elasticity vs Agility

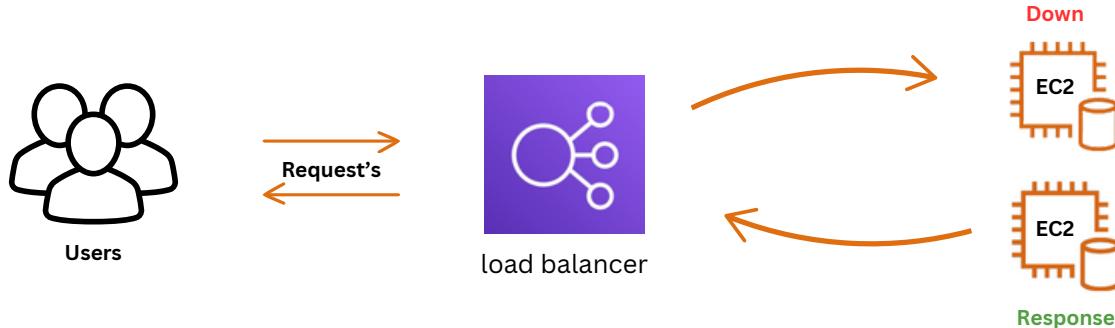
- **Scalability:** The ability to handle more work by either upgrading your existing system (making it faster or stronger) or adding more devices (like getting more cash registers in a store).
- **Elasticity:** Once your system is scalable, elasticity means it can automatically change based on demand. For example, if there are more customers (like during sales), the system can add more cash registers, and when it's quieter, it can remove them. This helps save costs because you only pay for what you need.
- **Agility:** Agility is the ability to quickly get new resources when needed. For example, in the cloud, you can add new computers for your team within minutes instead of waiting days or weeks. (This is more about speed, not directly about scaling).



load balancing

Load balancers are like traffic managers that direct website visitors to different servers (computers) so no single server gets too busy..

- **For example,** imagine a popular restaurant with several chefs in the kitchen. The load balancer is like the host who guides customers to different chefs based on who is least busy, ensuring that no chef gets overwhelmed



- **Distribute traffic:** It sends website visitors to different servers so no server gets too crowded.
- **Single address:** It gives users one easy link (like a website URL) to reach your app, no matter which server is actually handling it.
- **Handle server issues:** If one server stops working, the load balancer sends traffic to other working servers.
- **Health checks:** It regularly checks if the servers are working and stops sending traffic to broken ones.
- **Secure websites:** It keeps websites safe by managing secure connections (HTTPS).
- **Stay available:** The system keeps working even if one server or location has problems by using backup servers in different places.

Elastic Load Balancer (ELB)

- **Managed service:** An ELB is a service that AWS manages for you, so you don't need to worry about running it yourself
- **Reliable:** AWS ensures the ELB is always working and takes care of updates and maintenance.
- **High availability:** AWS makes sure the ELB is always available and ready to use.
- **Easy setup:** AWS provides a simple setup with only a few settings to adjust.

Types of ELBs AWS

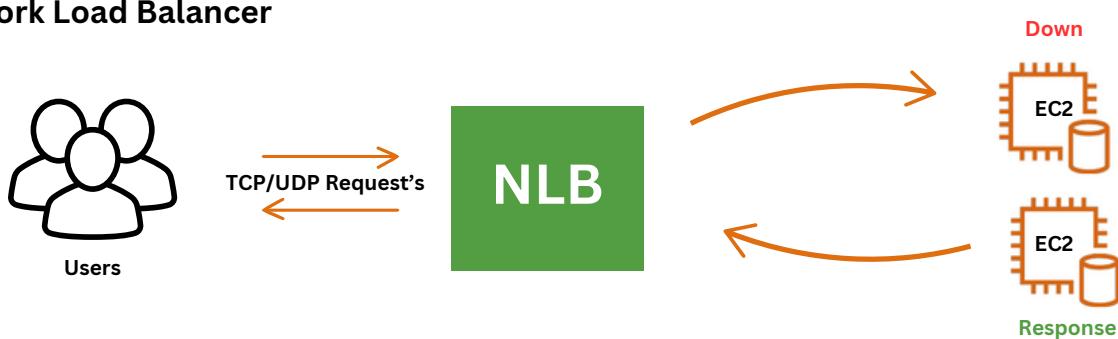
- **Application Load Balancer (Layer 7):** Works with HTTP/HTTPS traffic.
- **Network Load Balancer (Layer 4):** High-performance option that works with TCP traffic.
- **Gateway Load Balancer (Layer 3):** For special types of traffic routing.
- **Classic Load Balancer (Layer 4 & 7):** Retired in 2023, used to handle both types of traffic.

Application Load Balancer



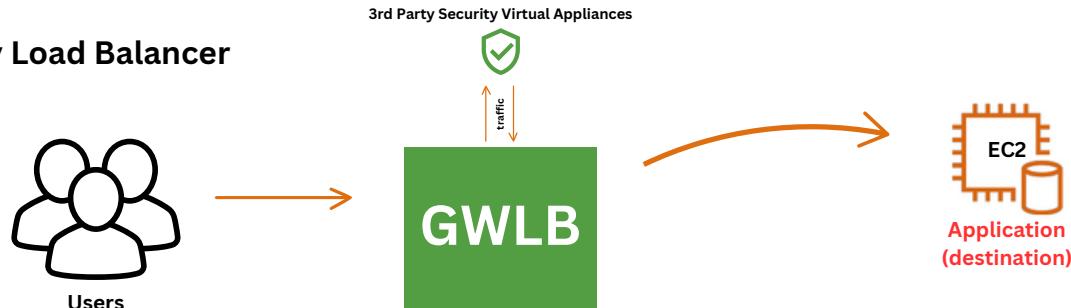
- **Supports HTTP, HTTPS, and gRPC:** The load balancer works with common web protocols like HTTP, HTTPS (for secure websites), and gRPC (used for high-performance communication between apps).
- **HTTP Routing:** It can direct traffic based on specific rules, such as URL paths or domain names, to the correct server.
- **Static DNS:** It provides a fixed web address (URL) that doesn't change, so users can always reach your application

Network Load Balancer



- **Supports TCP and UDP:** The load balancer works with low-level network protocols like TCP and UDP, which are commonly used for real-time applications like gaming or video streaming.
- **High Performance:** It can handle millions of requests per second, making it suitable for very fast, high-traffic applications.
- **Static IP:** It provides a fixed IP address through Elastic IP, so your app always has the same address, even if the underlying infrastructure changes.

Gateway Load Balancer



- **Routes traffic to firewalls:** It can send internet traffic to security systems (firewalls) that you control on EC2 instances to protect your app.
- **Intrusion detection:** It helps detect any unwanted or harmful activities trying to access your system, keeping your app secure.

Application Load Balancer ALB Hands On



Step 1 : Delete All Previous Configuration

- Delete Instance
- Delete Group Policy

The screenshot shows the AWS EC2 Instances page. The sidebar on the left includes sections for AMI Catalog, Elastic Block Store, Network & Security, Load Balancing, and Auto Scaling. The main content area displays a table titled "Instances (0/2) Info" with columns for Name, Instance ID, Instance state, Instance type, Status check, Alarm status, Availability Zone, and Public IPv4 DNS. A message at the top states "No instances" and "You do not have any instances in this region". A "Launch instances" button is located at the bottom of the table.

Create Two Instance

The screenshot shows the AWS EC2 Instances page with two instances listed: "My First Instance" and "My Second Instance". Both instances are in a "Running" state, t2.micro instance type, and us-west-2b availability zone. The Public IPv4 DNS for the first instance is ec2-54-217-144-167.us... and for the second instance is ec2-54-203-12-116.us.... The sidebar on the left shows the EC2 dashboard and various instance-related options.

Create Load Balancer

The screenshot shows the AWS Network Interfaces page. The sidebar on the left has sections for Load Balancing, Target Groups, Trust Stores (with a "New" link), and Auto Scaling. A red arrow points from the "Load Balancing" section to the "Load Balancers" link, which is highlighted in blue. The main content area is currently empty.

Access Web Page Through Instance Live IP Address



Hello Friend, This is Server One ip-172-31-78.us-west-2.compute.internal

Server Second



Hello Friend, This is Server One ip-172-31-28-110.us-west-2.compute.internal

Create Load Balancer

Compare and select load balancer type

A complete feature-by-feature comparison along with detailed highlights is also available. [Learn more](#)

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.
Create
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.
Create
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.
Create

▶ Classic Load Balancer - previous generation

Create Load Balancer

Create Application Load Balancer [Info](#)

The Application Load Balancer distributes incoming HTTP and HTTPS traffic across multiple targets such as Amazon EC2 instances, microservices, and containers, based on request attributes. When the load balancer receives a connection request, it evaluates the listener rules in priority order to determine which rule to apply, and if applicable, it selects a target from the target group for the rule action.

▶ How Application Load Balancers work

Basic configuration

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

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
Serves internet-facing traffic.
Has public IP addresses.
DNS name is publicly resolvable.
Requires a public endpoint.

Internal
Serves internal traffic.
Has private IP addresses.
DNS name is publicly nonresolvable.
Compatible with the IPv4 and Dualstack IP address types.

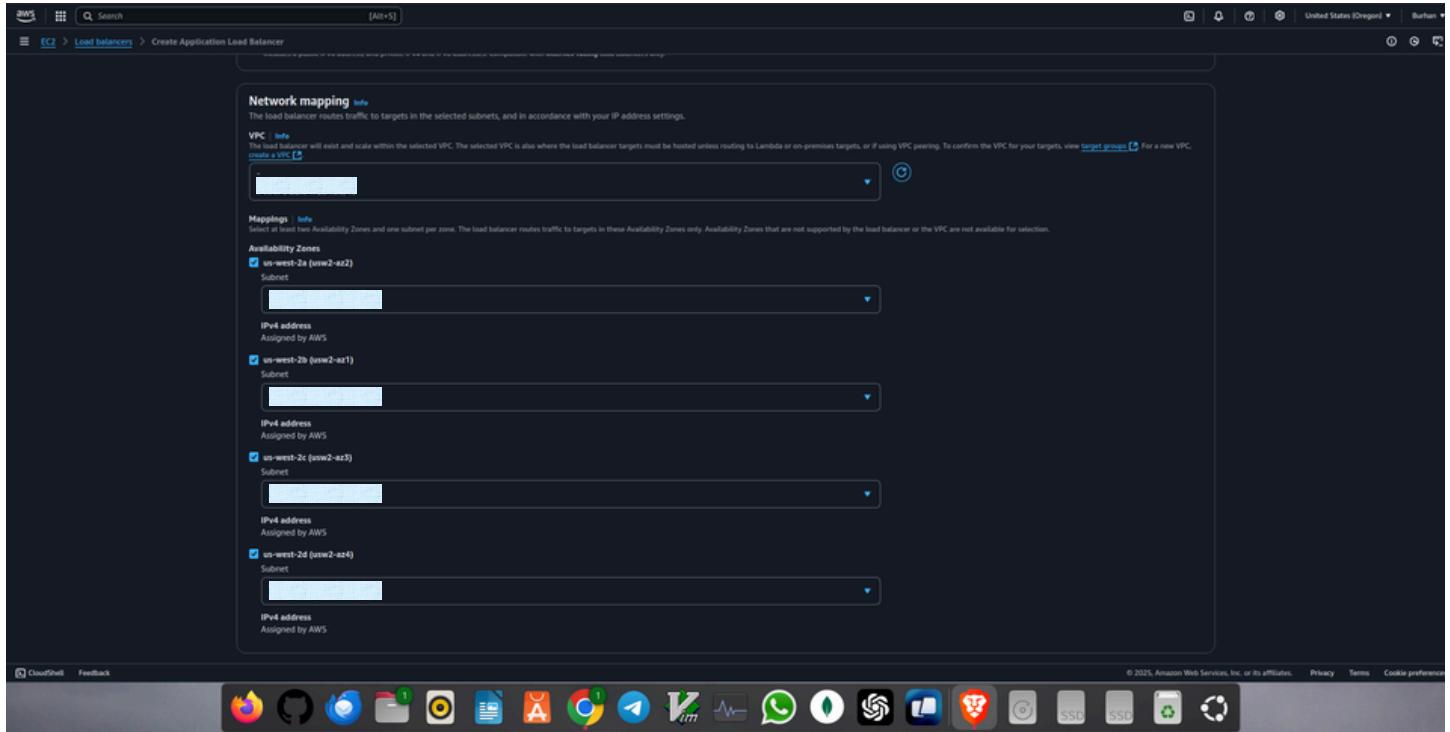
Load balancer IP address type [Info](#)
Select the front-end IP address type to assign to the load balancer. The VPC and subnets mapped to this load balancer must include the selected IP address types. Public IPv4 addresses have an additional cost.

IPv4
Includes only IPv4 addresses.

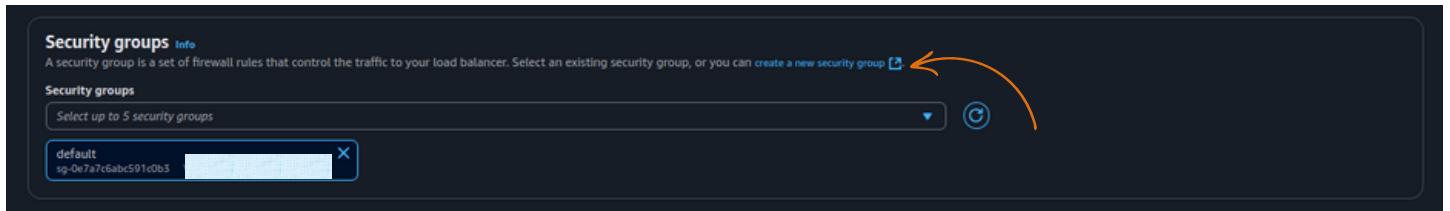
Dualstack
Includes IPv4 and IPv6 addresses.

Dualstack without public IPv4
Includes a public IPv6 address, and private IPv4 and IPv6 addresses. Compatible with internet-facing load balancers only.

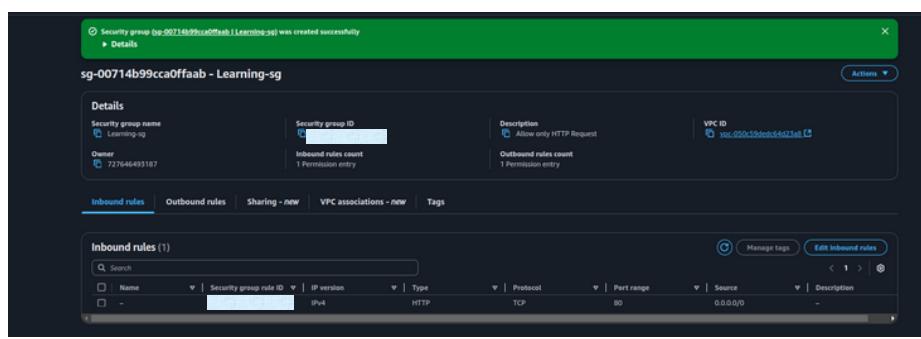
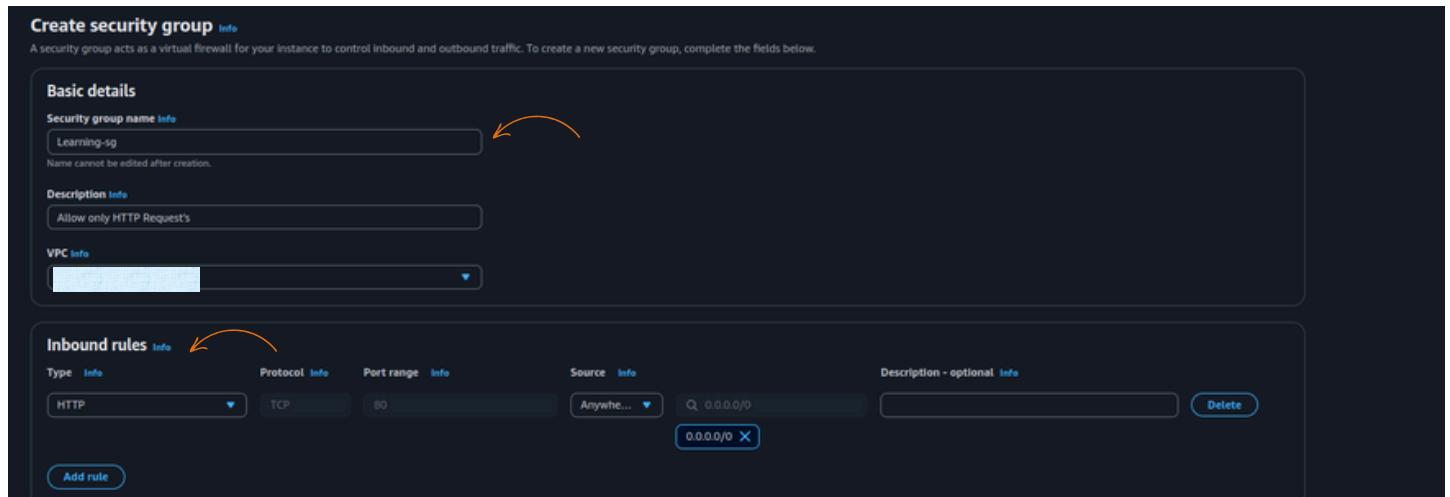
Network Mapping Tick All Availability Zones



Create a new security group



Create Security Group



Create security group

Select your Security Group : Learning-sg

The screenshot shows the 'Security groups' section of the AWS Management Console. A search bar at the top contains the text 'Learning-sg'. Below it, a list box displays 'Learning-sg' with a blue border. A yellow arrow points from the text 'Select your Security Group' in the heading to this list box.

Listeners and routing : Create Target Group

The screenshot shows the 'Listeners and routing' section. Under 'Listener HTTP:80', the 'Default action' dropdown is set to 'Forward to Select a target group'. A yellow arrow points from the text 'Create Target Group' in the heading to this dropdown.

The screenshot shows the 'Specify group details' step. It includes a navigation bar with 'Step 1 Specify group details' (selected) and 'Step 2 Register targets'. The 'Basic configuration' section contains a 'Choose a target type' dropdown where 'Instances' is selected. A yellow arrow points from the text 'Create Target Group' in the heading to this dropdown. The 'Target group name' field is filled with 'Learning-tg'. The 'Protocol : Port' section shows 'HTTP' selected with port '80'.

The screenshot shows the 'Register targets' step. It lists a single target group 'vpc-050c59fed64c7fa8'. The 'Protocol version' section has 'HTTP1' selected. A yellow arrow points from the text 'Create Target Group' in the heading to the 'Protocol version' dropdown. The 'Health checks' section shows 'HTTP' as the health check protocol and a path '/'. A yellow arrow points from the text 'Create Target Group' in the heading to the 'Health check protocol' dropdown.

Next

Register Targets

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)								
Instance ID	Name	State	Security groups	Zone	Private IPv4 address	Subnet ID	Launch time	
I-059b93e51858e2957	My First Instance	Running	launch-wizard-1	us-west-2b	[REDACTED]	[REDACTED]	February 1, 2025, 22:31 (UTC+05:30)	
I-0f5231cedaa39d676	My Second Instance	Running	launch-wizard-1	us-west-2b	[REDACTED]	[REDACTED]	February 1, 2025, 22:31 (UTC+05:30)	

Check Both Instance id > click on 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)								
Instance ID	Name	State	Security groups	Zone	Private IPv4 address	Subnet ID	Launch time	
I-059b93e51858e2957	My First Instance	Running	launch-wizard-1	us-west-2b	[REDACTED]	[REDACTED]	February 1, 2025, 22:31 (UTC+05:30)	
I-0f5231cedaa39d676	My Second Instance	Running	launch-wizard-1	us-west-2b	[REDACTED]	[REDACTED]	February 1, 2025, 22:31 (UTC+05:30)	

0 selected

Ports for the selected instances

Ports for routing traffic to the selected instances.

80
1-65535 (separate multiple ports with commas)

Include as pending below

2 selections are now pending below. Include more or register targets when ready.

Review targets

Targets (2)

Show only pending

Instance ID	Name	Port	State	Security groups	Zone	Private IPv4 address	Subnet ID	Launch time
I-059b93e51858e2957	My First Instance	80	Running	launch-wizard-1	us-west-2b	[REDACTED]	[REDACTED]	February 1, 2025, 22:31 (UTC+05:30)
I-0f5231cedaa39d676	My Second Instance	80	Running	launch-wizard-1	us-west-2b	[REDACTED]	[REDACTED]	February 1, 2025, 22:31 (UTC+05:30)

2 pending

Create

Remove all pending

Cancel **Previous** **Create target group**

Target group: Learning-tg. Created

EC2 > Target groups > Learning-tg

Successfully created the target group: Learning-tg. Anomaly detection is automatically applied to all registered targets. Results can be viewed in the Targets tab.

Learning-tg

Details

Target type	Protocol : Port	Protocol version
Instance	HTTP:80	HTTP1
IP address type	Load balancer	VPC
IPv4	None associated	[REDACTED]

2 Total targets

0 Healthy	0 Unhealthy	2 Unused	0 Initial	0 Draining
0 Anomalous				

Distribution of targets by Availability Zone (AZ)

Select values in this table to see corresponding filters applied to the Registered targets table below.

Targets **Monitoring** **Health checks** **Attributes** **Tags**

Registered targets (2)

Target groups route requests to individual registered targets using the protocol and port number specified. Health checks are performed on all registered targets according to the target group's health check settings. Anomaly detection is automatically applied to HTTP/HTTPS target groups with at least 3 healthy targets.

Instance ID	Name	Port	Zone	Health status	Health status details	Administrative state	Override details	Launch time	Anomaly
I-059b93e51858e2957	My First Instance	80	us-west-2b [REDACTED]	Unused	Target group is not co...	-	-	February ...	Normal
I-0f5231cedaa39d676	My Second Inst...	80	us-west-2b [REDACTED]	Unused	Target group is not co...	-	-	February ...	Normal

Add Target Group Here

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

Protocol	Port
HTTP	: 80 1-65535

Default action | **Info**
Forward to Learning-tg
Target type: Instance, IPv4

HTTP ▾

Create target group Info

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.

Add listener

Load balancer tags - optional
Consider adding tags to your load balancer. Tags enable you to categorize your AWS resources so you can more easily manage them. The 'Key' is required, but 'Value' is optional. For example, you can have Key = production-webserver, or Key = webserver, and Value = production.

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.

Create load balancer

Copy DNS & Paste on Browser

EC2 < **Load balancers**

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	Date created
LearningALB	[REDACTED]	Active		4 Availability Zones	application	February 1, 2025, 1:45:12 PM

Response 1

Not secure

Hello Friend, This is Server One ip-172-31-23-78.us-west-2.compute.internal

Response 2

Not secure

Hello Friend, This is Server One ip-172-31-28-110.us-west-2.compute.internal

Stop One Instance

Successfully Initiated stopping of i-039b93e51858e2957

Instances (1/2) Info

Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS
My First Instance	i-039b93e51858e2957	Stopping	t2.micro	2/2 checks passed	View alarms +	us-west-2b	
My Second Instance	i-0f5231cedaa39d676	Running	t2.micro	2/2 checks passed	View alarms +	us-west-2b	

Till Website Running on Secondary Zone

Not secure learningalb-2132019160.us-west-2.elb.amazonaws.com

Hello Friend, This is Server One ip-172-31-28-110.us-west-2.compute.internal

Registered targets (2) Info

Anomaly mitigation: Not applicable

Deregister Register targets

Target groups route requests to individual registered targets using the protocol and port number specified. Health checks are performed on all registered targets according to the target group's health check settings. Anomaly detection is automatically applied to HTTP/HTTPS target groups with at least 3 healthy targets.

Instance ID	Name	Port	Zone	Health status	Health status details	Admini...	Overrid...	Launch ...	Anomaly detection...
i-039b93e51858e2957	My First Instance	80	us-west-2b (us...)	Unused	Target is in the stoppe...	No override...	No overrid...	February 1...	Normal
i-0f5231cedaa39d676	My Second Inst...	80	us-west-2b (us...)	Healthy	-	No override...	No overrid...	February 1...	Normal

Auto Scaling Group (ASG)

An Auto Scaling Group (ASG) is a feature in AWS that helps manage the number of EC2 instances (virtual servers) running for your application.

- Sometimes, your website or app gets a lot of traffic, and sometimes it doesn't.
- In AWS, you can easily add or remove servers based on the load.

What does an Auto Scaling Group do ?

- Adds more servers (Scale Out) when traffic increases.
- Removes extra servers (Scale In) when traffic decreases.
- Keeps a minimum and maximum number of servers running at all times.
- Automatically connects new servers to a Load Balancer (which distributes traffic).
- Replaces unhealthy servers to keep the application running smoothly.
- Saves costs by ensuring you only use the number of servers you actually need.

In short, ASG helps your application handle traffic changes automatically while saving money!

Auto Scaling Group Hands On

Remove All old Configuration

The screenshot shows the AWS CloudFront console. The left sidebar has a navigation menu with the following items:

- AMI Catalog
- Elastic Block Store** (selected)
 - Volumes
 - Snapshots
 - Lifecycle Manager
- Network & Security** (selected)
 - Security Groups
 - Elastic IPs
 - Placement Groups
 - Key Pairs
 - Network Interfaces
- Load Balancing** (selected)
 - Load Balancers
 - Target Groups
 - Trust Stores New
- Auto Scaling** (selected)
 - Auto Scaling Groups
- Settings

The main content area is titled "Instances" and shows a table with the following columns:

Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public
No instances You do not have any instances in this region							
Launch instances							

At the bottom of the page, there are links for CloudShell, Feedback, and a footer with copyright information: © 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences.

Create ASG Auto Scaling Groups

AMIs Catalog

Elastic Block Store

- Volumes
- Snapshots
- Lifecycle Manager

Network & Security

- Security Groups
- Elastic IPs
- Placement Groups
- Key Pairs
- Network Interfaces

Load Balancing

- Load Balancers
- Target Groups
- Trust Stores New

Auto Scaling

- Auto Scaling Groups**
- Settings

CloudShell Feedback

Search [Alt+S]

United States (Oregon) Burhan

Create Auto Scaling group

Get started with EC2 Auto Scaling by creating an Auto Scaling group.

Create Auto Scaling group

Pricing

Amazon EC2 Auto Scaling features have no additional fees beyond the service fees for Amazon EC2, CloudWatch (for scaling policies), and the other AWS resources that you use. Visit the pricing page of each service to learn more.

Getting started

What is Amazon EC2 Auto Scaling?

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

Give name ASG : aws-learning

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 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.

aws-learning

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

Create a launch template

Create

Cancel Next

Create a launch template

This template is begin used to tell to the Auto Scaling Group How to Create EC2 within it

multiple versions.

Launch template name and description

Launch template name - required

aws-learning-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

Create ASG Auto Scaling Groups

▼ 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

Amazon Linux | macOS | Ubuntu | Windows | Red Hat | SUSE Linux | Debian

Amazon Machine Image (AMI)

Ubuntu Server 24.04 LTS (HVM), SSD Volume Type
ami-00c257e12d6828491 (64-bit (x86)) / ami-0acefc55c3a331fa8 (64-bit (Arm))
Virtualization: hvm ENA enabled: true Root device type: ebs

Free tier eligible ▾

Description

Ubuntu Server 24.04 LTS (HVM), EBS General Purpose (SSD) Volume Type. Support available from Canonical (<http://www.ubuntu.com/cloud/services>).

Canonical, Ubuntu, 24.04, amd64 noble image

Browse more AMIs
Including AMIs from AWS, Marketplace and the Community

Instance type

▼ Instance type [Info](#) | [Get advice](#) Advanced

Instance type

t2.micro Free tier eligible

Family: t2 1 vCPU 1 GiB Memory Current generation: true
On-Demand Linux base pricing: 0.0116 USD per Hour On-Demand SUSE base pricing: 0.0116 USD per Hour
On-Demand Windows base pricing: 0.0162 USD per Hour On-Demand RHEL base pricing: 0.026 USD per Hour
On-Demand Ubuntu Pro base pricing: 0.0134 USD per Hour

All generations

Compare instance types

Additional costs apply for AMIs with pre-installed software

Network settings

▼ Network settings [Info](#)

Subnet | [Info](#)

Don't include in launch template

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 groups | [Info](#)

Select security groups

Advanced network configuration

Compare security group rules



Advance > User data

User data - optional | Info
Upload a file with your user data or enter it in the field.

```
#!/bin/bash
# Update system
sudo apt update -y && sudo apt upgrade -y

# Install Apache, MySQL, PHP, and required extensions
sudo apt install -y apache2 mysql-server php php-mysql libapache2-mod-php php-cli unzip wget

# Start Apache and MySQL services
sudo systemctl start apache2
sudo systemctl enable apache2
sudo systemctl start mysql
sudo systemctl enable mysql
sudo chmod -R 755 /var/www/html
sudo systemctl restart apache2
```

User data has already been base64 encoded

```
#!/bin/bash
# Update system
sudo apt update -y && sudo apt upgrade -y

# Install Apache, MySQL, PHP, and required extensions
sudo apt install -y apache2 mysql-server php php-mysql libapache2-mod-php php-cli unzip wget

# Start Apache and MySQL services
sudo systemctl start apache2
sudo systemctl enable apache2
sudo systemctl start mysql
sudo systemctl enable mysql
sudo chmod -R 755 /var/www/html
sudo systemctl restart apache2
```

Refresh Launch Template

Name

Auto Scaling group name
Enter a name to identify the group.

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.
    
[Create a launch template](#)

Click On Next

Instance type requirements [Info](#)

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

Launch template	Version
<input type="text" value="awslearningtemplate1"/> 	Default
lt-0c35cf21361fb0ac1	
Instance type	
t2.micro	

Select Multi Availability

Zone

click on **Next**



Advance > User data

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

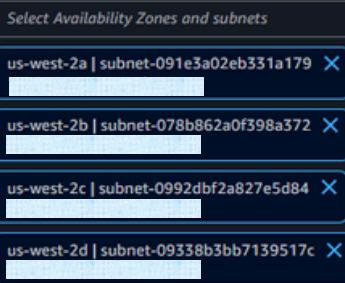
Choose the VPC that defines the virtual network for your Auto Scaling group.



[Create a VPC](#)

Availability Zones and subnets

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



Integrate with other services - *optional* Info

Use a load balancer to distribute network traffic across multiple servers. Enable service-to-service communications with VPC Lattice. Shift resources away from impaired Availability Zones with zonal shift. You can also customize health check replacements and monitoring.

Load balancing Info

Use the options below to attach your Auto Scaling group to an existing load balancer, or to a new load balancer that you define.

No load balancer
Traffic to your Auto Scaling group will not be fronted by a load balancer.

Attach to an existing load balancer
Choose from your existing load balancers.

Attach to a new load balancer
Quickly create a basic load balancer to attach to your Auto Scaling group.



Attach to an existing load balancer

Select the load balancers that you want to attach to your Auto Scaling group.

Choose from your load balancer target groups
This option allows you to attach Application, Network, or Gateway Load Balancers.

Choose from Classic Load Balancers

Existing load balancer target groups

Only instance target groups that belong to the same VPC as your Auto Scaling group are available for selection.

[Select target groups](#)



target-newLearning | HTTP
Application Load Balancer: LearningALB



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 Recommended

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.

EC2 Auto Scaling will start to detect and act on health checks performed by Elastic Load Balancing. To avoid unexpected terminations, first verify the settings of these health checks in the [Load Balancer console](#).



[Turn on VPC Lattice health checks](#)

[Click On Next](#)

Desired capacity



Desired capacity type
Choose the unit of measurement for the desired capacity value. vCPUs and Memory(GB) 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 Equal or less than desired capacity

Max desired capacity 2 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.

Review

Click On Next

Review Info

Step 1: Choose launch template

Group details

Auto Scaling group name
aws-learning

Launch template

Launch template	Version	Description
awslearningtemplate1	Default	

Step 2: Choose instance launch options

Network

VPC
vpc- [REDACTED]

Availability Zones and subnets

Availability Zone	Subnet	Subnet CIDR range
us-west-2a	[REDACTED]	[REDACTED]
us-west-2b	[REDACTED]	[REDACTED]
us-west-2c	[REDACTED]	[REDACTED]
us-west-2d	[REDACTED]	[REDACTED]

Availability Zone distribution
Balanced best effort

Desired capacity

Auto Scaling groups (1) [Info](#)

Search your Auto Scaling groups

Name	Launch template/configuration	Instances	Status	Desired capacity	Min	Max	Availability Zones
aws-learning	awslearningtemplate1 Version Default	0	Updating capacity...	2	1	2	us-west-2a, us-west-2b, us-west-2c, us-west-2d

Instance management

Auto Scaling groups > [aws-learning](#)

Desired capacity: 2 | Scaling limits (Min - Max): 1 - 2 | Desired capacity type: Units (number of instances) | Status: -

Date created: Tue Feb 04 2025 11:14:34 GMT+0530 (India Standard Time)

Details | Integrations - new | Automatic scaling | **Instance management** | Instance refresh | Activity | Monitoring

Instances (2)

Instance ID	Lifecycle	Instance type	Weighted capacity	Launch template...	Availability Zone	Health status	Protected from
i-01aa3e691879daa55	InService	t2.micro	-	awslearningtemplate1	us-west-2c	Healthy	
i-0a72b9dfc3f48206f	InService	t2.micro	-	awslearningtemplate1	us-west-2b	Healthy	

Tow Instance are running this instance are created by ASG : auto scaling group

Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS	Public IPv4 ...	Elastic IP
	i-01aa3e691879daa55	Running	t2.micro	2/2 checks passed	View alarms	us-west-2c			
	i-0a72b9dfc3f48206f	Running	t2.micro	2/2 checks passed	View alarms	us-west-2b			

Target group

EC2 > Target groups > [target-newLearning](#)

target-newLearning

Details

Target type	Protocol	Protocol version
Instance	HTTP: 80	HTTP1
IP address type	Load balancer	
IPv4	None associated	
2 Total targets	2 Healthy	0 Unhealthy
	0 Anomalous	

Distribution of targets by Availability Zone (AZ)

Select values in this table to see corresponding filters applied to the Registered targets table below.

Targets | Monitoring | Health checks | Attributes | Tags

Registered targets (2) [Info](#)

Target groups route requests to individual registered targets using the protocol and port number specified. Health checks are performed on all registered automatically applied to HTTP/HTTPS target groups with at least 3 healthy targets.

Instance ID	Name	Port	Zone	Health status	Health status details
i-01aa3e691879daa55		80	us-west-2c (us...)	Healthy	-
i-0a72b9dfc3f48206f		80	us-west-2b (us...)	Healthy	-

Access From Load Balancer

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
LearningALB	[DNS Name]	Active	[VPC ID]	4 Availability Zones	application

Paste on Browser

The page shows the Apache2 Default Page with the Ubuntu logo and the text "It works!". Below it, there is a configuration overview section and a file structure listing for /etc/apache2.

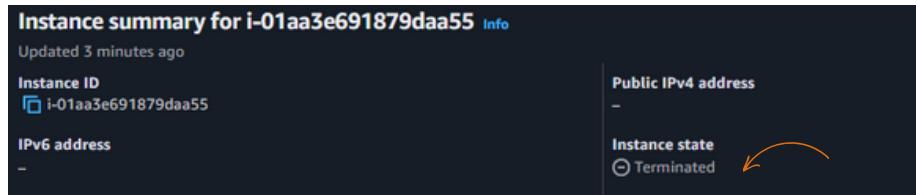
```
/etc/apache2/
  |-- apache2.conf
  |  `-- ports.conf
  `-- mods-enabled
```

Webpage access form both instance we can Terminated 1 running instance

The screenshot shows the AWS Auto Scaling groups console for the 'aws-learning' group. It displays two instances: one in us-west-2c and one in us-west-2b, both marked as healthy. A green arrow points to the instance in us-west-2c. A modal window titled 'Terminate (delete) instance?' is open over the instance details, asking if the user wants to terminate the selected instance.

Desired capacity

We have set a minimum of 2 and a desired capacity in the Auto Scaling Group. If one instance gets terminated, the Auto Scaling Group automatically creates a new instance



Power of ASG

	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Pub
□	i-01aa3e691879daa55	i-01aa3e691879daa55	Terminated	t2.micro	X	-	us-west-2c	-
□	i-05c15e2c9243b9f50	i-05c15e2c9243b9f50	Running	t2.micro	2/2 checks passed	View alarms +	us-west-2c	ec2
□	i-0a72b9dfc3f48206f	i-0a72b9dfc3f48206f	Running	t2.micro	2/2 checks passed	View alarms +	us-west-2b	ec2

Activity history (4)					
Status	Description	Cause	Start time	End time	
Successful	Launching a new EC2 instance: i-05c15e2c9243b9f50	At 2025-02-04T06:04:33Z an instance was launched in response to an unhealthy instance needing to be replaced.	2025 February 04, 11:34:35 AM +05:30	2025 February 04, 11:34:42 AM +05:30	
Connection draining in progress	Terminating EC2 instance: i-01aa3e691879daa55 - Waiting For ELB Connection Draining.	At 2025-02-04T06:04:33Z an instance was taken out of service in response to an EC2 health check indicating it has been terminated or stopped.	2025 February 04, 11:34:33 AM +05:30		
Successful	Launching a new EC2 instance: i-01aa3e691879daa55	At 2025-02-04T05:44:34Z a user request created an AutoScalingGroup changing the desired capacity from 0 to 2. At 2025-02-04T05:44:46Z an instance was started in response to a difference between desired and actual capacity, increasing the capacity from 0 to 2.	2025 February 04, 11:14:48 AM +05:30	2025 February 04, 11:14:55 AM +05:30	
Successful	Launching a new EC2 instance: i-0a72b9dfc3f48206f	At 2025-02-04T05:44:34Z a user request created an AutoScalingGroup changing the desired capacity from 0 to 2. At 2025-02-04T05:44:46Z an instance was started in response to a difference between desired and actual capacity, increasing the capacity from 0 to 2.	2025 February 04, 11:14:48 AM +05:30	2025 February 04, 11:14:55 AM +05:30	

Auto Scaling Group (ASG) in AWS automatically detects unhealthy instances and replaces them with new ones

Clean Up All Configuration

If you manually delete an instance that is part of an Auto Scaling Group (ASG), the ASG will automatically launch a new instance to maintain the desired capacity

Click on Auto Scaling Group

Auto Scaling groups (1) Info

Name	Launch template/configuration	Instances	Status	Desired capacity	Min	Max	Availability Zones
<input type="checkbox"/> aws-learning	awslearningtemplate1 Version Default	2	-	2	1	2	us-west-2a, us-west-2b, us-west-2c, us-west-2d

Auto Scaling groups (1/1) Info

Name	Launch template/configuration	Instances	Status	Desired capacity	Min	Max	Availability Zones
<input checked="" type="checkbox"/> aws-learning	awslearningtemplate1 Version Default	2	-	2	1	2	us-w...

Delete Auto Scaling group

Auto Scaling group contains running instances
Deleting these Auto Scaling groups will terminate all instances in each group.
This action cannot be undone.

Are you sure you want to delete this Auto Scaling group?

- aws-learning

Deleting the Auto Scaling group will terminate all instances in the group. This action cannot be undone.

To confirm deletion, type *delete* in the field.

Cancel **Delete**

Delete Load Balancer

Load balancers (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
<input type="checkbox"/> LearningALB	[REDACTED]	Active	vpc-050c59dedc64d23a8	4 Availability Zones

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
<input checked="" type="checkbox"/> LearningALB	[REDACTED]	Active	vpc-050c59dedc64d23a8	4 Availability Zones

Load balancer: LearningALB

Details Listeners and rules Network mapping Resource map - new

Details

Load balancer type: [REDACTED]

Status: [REDACTED]

Delete load balancer

Delete load balancer LearningALB permanently? This action can't be undone.

⚠ Proceeding with this action deletes the load balancer and its listeners. Target groups associated to this load balancer will become available for association to another load balancer and their registered targets remain unaffected.

To avoid accidental deletion we ask you to provide additional written consent.

Type *confirm* to agree.

Cancel **Delete**

Should i delete my target group?
it not cost your any money and the TG is going to empty because we are delete Auto Scaling Group

When Your ASG is Deleted the your EC2 Instances auto deleted

<input type="radio"/> Terminated	<input type="radio"/> <input type="radio"/>	t2.micro
<input type="radio"/> Terminated	<input type="radio"/> <input type="radio"/>	t2.micro
<input type="radio"/> Terminated	<input type="radio"/> <input type="radio"/>	t2.micro