

Scalable Web Application using NLB + EC2 Auto Scaling

1. Create Launch Template (EC2)

Screenshot of the AWS CloudShell interface showing the creation of a Launch Template named "myec2t1".

Launch template name and description:
Launch template name: myec2t1
Template version description: A prod webserver for MyApp

Auto Scaling guidance:
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:

Launch template contents:
Specify the details of your launch template below. Leaving a field blank will result in the field not being included in the launch template.

Application and OS Images (Amazon Machine Image):
An AMI contains the operating system, application server, and applications for your instance. If you don't see a suitable AMI below, use the search field or choose [Browse more AMIs](#).

Key pair (login):
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: mydebian13

Network settings:
Subnet: Don't include in launch template
Availability Zone: Don't include in launch template
Firewall (security group): Select existing security group (myec2Securitygroup)
Security groups: myec2Securitygroup sg-00sec2b2e28f558b23
Advanced network configuration

Metadata response hop limit:
Allow tags in metadata:
User data - optional:
Upload a file with your user data or enter it in the field.
Choose file: #!/bin/bash
yum update -y
amazon-linux-extras install nginx1 -y
systemctl start nginx
systemctl enable nginx
echo "Severed by \$hostname via NLB" > /usr/share/nginx/html/index.html

Summary:
Software Image (AMI): Amazon Linux 2023 kernel-6.1 A...
Virtual server type (instance type): t2.micro
Firewall (security group): myec2Securitygroup
Storage (volumes): 1 volume(s) - 8 GiB

Free tier: In your first year of opening an AWS account, you get 750 hours per month of t2.micro instance usage (or t3.micro where t2.micro isn't available) when used with free tier AMIs, 750 hours per month of public IPv4 address usage, 30 GiB of EBS storage, 2 million I/Os, 1 GiB of snapshots, and 100 GiB of bandwidth to the internet. Data transfer charges are not included as part of the free tier allowance.

Create launch template

The screenshot shows the AWS EC2 console with the 'Launch Templates' section selected. A single launch template named 'myec2lt' is listed. The details pane shows the launch template ID, name, version, creation time, and owner. The 'Launch template version details' section shows the default version and its creation date.

2. Create Target Group (Critical for NLB)

The screenshot shows the AWS EC2 console with the 'Target Groups' section selected. A new target group is being created, as indicated by the 'Create target group' button. The 'Create target group' wizard is open, showing Step 1: Create target group. It lists three options: 'Create target group' (selected), 'Step 2 - recommended', and 'Review and create'. The 'Create target group' step details the target type as 'Instances' (selected) and provides options for ALB, NLB, and GWLB. Other target types like 'IP address' and 'Lambda function' are also shown. The 'Target group name' field is set to 'mylbtgt'. The 'Protocol' is set to 'TCP' and the port is '80'. The URL at the bottom is <https://ap-southeast-1.console.aws.amazon.com/console/home?region=ap-southeast-1&serviceId=EC2&resourceId=TargetGroups&resourceType=TargetGroup&resourceName=Create%20target%20group>.

Screenshot of the AWS CloudFront console showing the creation of a new target group named "mynlbtg".

Target group details:

- VPC:** [vpc-05073fe61b42c061b](#) (selected)
- Health check protocol:** TCP
- Health check port:** 80 (Traffic port)
- Healthy threshold:** 5
- Unhealthy threshold:** 2

Target group summary:

Total targets	Healthy	Unhealthy	Unused	Initial	Draining
0	0	0	0	0	0

Registered targets (0): No registered targets.

3. Create Network Load Balancer

Screenshot of the AWS CloudFront console comparing three types of load balancers:

- Application Load Balancer (ALB):** Handles HTTP and HTTPS traffic at the application level, providing advanced routing and visibility features.
- Network Load Balancer (NLB):** Handles TCP and UDP traffic at the request level, supporting TLS offloading and centralized certificate deployment.
- Gateway Load Balancer (GWLB):** Handles traffic for third-party virtual appliances, supporting GENEVE and handling millions of requests per second.

Classic Load Balancer - previous generation: A link to the previous generation of load balancers.

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

- Internet-facing**
 - Serves internal-facing traffic.
 - Has public IP addresses.
 - DNS name resolves to public IPs.
 - Requires a public subnet.
- Internal**
 - Serves internal traffic.
 - Has private IP addresses.
 - DNS name resolves to private IPs.

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.

- IPv4** Includes only IPv4 addresses.
- Dualstack** Includes IPv4 and IPv6 addresses.

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

VPC
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 on-premises targets or if using VPC peering. To confirm the VPC for your targets, view [target groups](#).

vpc-050796f61b42c061b (default) [Create VPC](#)

Availability Zones and subnets

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

VPC
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 on-premises targets or if using VPC peering. To confirm the VPC for your targets, view [target groups](#).

vpc-050796f61b42c061b (default) [Create VPC](#)

Availability Zones and subnets
Select one or more Availability Zones and corresponding subnets. Enabling multiple Availability Zones increases the fault tolerance of your applications. The load balancer routes traffic to targets in the selected Availability Zones only. Availability Zones that are not supported by the load balancer or the VPC are not available for selection.

ap-southeast-1a (apse1-a2z)
Subnet
Only CIDR blocks corresponding to the load balancer IP address type are used. At least 8 available IP addresses are required for your load balancer to scale efficiently.
subnet-04d4f478a07c66d07
IPv4 subnet CIDR: 172.31.16.0/20

IPv4 address
The front-end IPv4 address of the load balancer in the selected Availability Zone.
 Assigned by AWS Use an Elastic IP address

ap-southeast-1b (apse1-a2z)
Subnet
Only CIDR blocks corresponding to the load balancer IP address type are used. At least 8 available IP addresses are required for your load balancer to scale efficiently.
subnet-06d80066a0edc0d47
IPv4 subnet CIDR: 172.31.32.0/20

IPv4 address
The front-end IPv4 address of the load balancer in the selected Availability Zone.
 Assigned by AWS Use an Elastic IP address

ap-southeast-1c (apse1-a2z)

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 group - recommended
AWS Lambda and API Gateway listeners can only be enabled at creation by including at least one security group. You can change security groups after creation. The security groups for your load balancer must allow it to communicate with registered targets on both the listener port and the health check port. For PrivateLink Network Load Balancers, security group rules are enforced on **private** traffic; however, you can turn off egress rate evaluation after creation with the load balancer's security tab or using the API.

Select up to 5 security groups [Create security group](#)

mylbsecuritygroup sg-029981115af1x1 VPC: vpc-050796f61b42c061b

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 TCP:80
Protocol: TCP Port: 80 Forward to target group: mylb1 Target group: mylb1 Weight: 1 Percent: 100% [Add target group](#) [Evaluate](#)

Forward to target group Info
Creates a target group and configures weight according to assigned target group weights. [Create target group](#)

Target group mylb1 Target type: Instance, IPv4 Target stickiness: Off [Edit](#)

Weighted routing evaluation - recommended
Your listeners' load balancers won't route to destination requests according to assigned target group weights if your target groups don't follow best practices.

[Evaluate](#)

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

mylb

Listeners

Protocol	Port	Default action	ARN	Security policy	Forwarding rule group	Target group	ARN	Not applicable	ALPN policy	Tags
TCP	80	Forward to target group	arn:aws:elasticloadbalancing:ap-southeast-1:368314293908:loadbalancer/nel/mylb/43f62430756761a	arn:aws:elasticloadbalancing:ap-southeast-1:368314293908:securitygroup/nel/mysg-0298c1133d91c1 - mylbsecuritygroup	1	mylb-target-group	arn:aws:elasticloadbalancing:ap-southeast-1:368314293908:targetgroup/nel/mylb-target-group-11904	Target group stickiness: Off	None	0 tags
TCP	443	Forward to target group	arn:aws:elasticloadbalancing:ap-southeast-1:368314293908:loadbalancer/nel/mylb/43f62430756761a	arn:aws:elasticloadbalancing:ap-southeast-1:368314293908:securitygroup/nel/mysg-0298c1133d91c1 - mylbsecuritygroup	1	mylb-target-group	arn:aws:elasticloadbalancing:ap-southeast-1:368314293908:targetgroup/nel/mylb-target-group-11904	Target group stickiness: Off	None	0 tags

NLB Security group(Inbound rules):-

Edit inbound rules

Inbound rules control the incoming traffic that's allowed to reach the instance.

Security group rule ID	Type	Protocol	Port range	Source	Description - optional
-	Custom TCP	TCP	443	Anywhere... <input type="text" value="0.0.0.0/0"/>	<input type="button" value="Delete"/>
-	Custom TCP	TCP	80	Anywhere... <input type="text" value="0.0.0.0/0"/>	<input type="button" value="Delete"/>
-	Custom TCP	TCP	22	Custom <input type="text" value="Q"/>	<input type="button" value="Delete"/>

Rules with source of 0.0.0.0/0 or /0 allow all IP addresses to access your instance. We recommend setting security group rules to allow access from known IP addresses only.

EC2 Security group(Inbound rules):-

Edit inbound rules

Inbound rules control the incoming traffic that's allowed to reach the instance.

Security group rule ID	Type	Protocol	Port range	Source	Description - optional
sg-06d12024910302c	Custom TCP	TCP	443	Anywhere... <input type="text" value="0.0.0.0/0"/>	<input type="button" value="Delete"/>
sg-0ff1a46a250721932	Custom TCP	TCP	80	Anywhere... <input type="text" value="0.0.0.0/0"/>	<input type="button" value="Delete"/>
sg-0959400c28512ec	SSH	TCP	22	Custom <input type="text" value="Q"/>	<input type="button" value="Delete"/>

Rules with source of 0.0.0.0/0 or /0 allow all IP addresses to access your instance. We recommend setting security group rules to allow access from known IP addresses only.

4. Create Auto Scaling Group

The screenshot shows a multi-step wizard for creating an Auto Scaling group. The steps are as follows:

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

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

Launch template Info
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.
myEc221
Create a launch template L Version L
Version L Create a launch template version L

Description
AMIS
ami-059f07c65e32875a8
Key pair name
myKeyPair13

Security groups
-
Security group IDs
sg-0aa26a39f58621

Instance type
t2.micro

Request Spot Instances
No

Additional details
Storage (volume)
-
Date created
Mon Oct 22 2023 22:16:49 GMT+0530 (India Standard Time)

Cancel **Next**
- Step 3: Choose instance launch options**

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
myEc221 (t2.micro)	Lastest	-

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-050716161942001b
172.17.0.0/16

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 L
ap-southeast-1a | subnet-04047479a0cf7d5d07
172.17.40.0/16 - default
ap-southeast-1a | subnet-06db806cbeeb5a47
172.17.41.0/16 - default

Create a subnet L

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

Balanced best effort
Launch instances in every Availability Zone. Auto Scaling will attempt to launch in another healthy Availability Zone.

Balanced only
Launch instances in one healthy Availability Zone. Auto Scaling will continue to attempt to launch in the unhealthy Availability Zone to preserve balanced distribution.

Override launch template
- Step 4: Integrate with other services - optional**

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.

Select Load balancing option
 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.
 Choose from Classic Load Balancers
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 to attach
 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 L
myHttpLog TCP
Network Load Balancer: myts

VPC Lattice integration options Info
To improve networking capabilities and scalability, integrate your Auto Scaling group with VPC Lattice. VPC Lattice facilitates communications between AWS services and helps you connect and manage your applications across compute services in AWS.

Select VPC Lattice service to attach
 No VPC Lattice service
Leave this setting empty if your Auto Scaling group's network access and connectivity with other services is not required.
 Attach to VPC Lattice service
Incoming requests associated with specified VPC Lattice target groups will be routed to your Auto Scaling group.

Create new VPC Lattice service L

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

AWS CloudWatch Metrics > Auto Scaling groups > Create Auto Scaling group

Step 1
 Choose launch template or configuration
 Choose instance launch options
 Step 3 - optional
 Integrate with other services

Step 2
 Configure group size and scaling
 Step 1 - optional
 Add notifications
 Step 6 - optional
 Add tags
 Review

Configure group size and scaling - optional
 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
 Desired capacity is the minimum number of instances required for the desired capacity value. vCPUs and MemoryGB are only supported for mixed instances groups configured with a set of instance attributes.

Group size info
 Desired capacity
 Enter your group size.
 2

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

Scaling limits
 Enter the amount 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
 An automatic scaling policy uses CloudWatch Metrics and target value for scheduled scaling after creating your Auto Scaling group. It automatically scales up or down based on the CloudWatch Metrics and target value.

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
Info
 Control what Auto Scaling group does 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.

Choose a replacement behavior depending on your availability requirements

Shortest timeout
 For short lived events, like instance refreshes and

Longest running
 Launch before terminating

Current tasks
 Continue launching instances until the group reaches the desired capacity

Flexible
 Auto scales across zones for better availability

AWS CloudWatch Metrics > Auto Scaling groups

Auto Scaling groups (1/1) Info

Name	Launch template/configuration	Instances	Status	Desired capacity	Min	Max	Availability Zones	Creation time
<input checked="" type="checkbox"/> mynlbsg	myc2t1 Version Latest	2	-	2	1	3	2 Availability Zones	Mon Dec 22 2022 23:41:16 GMT+0530 (India Standard Time)

Auto Scaling group: mynlbsg

Details Integrations Automatic scaling Instance management Instance refresh Activity Monitoring Tags - moved

mynlbsg Capacity overview
[Edit](#)

arn:aws:autoscaling:ap-southeast-1:368314293908:autoScalingGroup:a449b65-772a-45d0-a2c0-c01706a2fmac:autoScalingGroupName/mynlbg

Desired capacity 2 **Scaling limits** 1 - 3 **Desired capacity type** Units (number of instances) **Status** -

Date created Mon Dec 22 2022 23:41:16 GMT+0530 (India Standard Time)

Launch template
Info
 AMI ID: ami-05f071c6e52875a8
 Security group IDs: sg-004cb2c2e28f558b23
 Owner: arn:aws:iam::368314293908:root
 Create time: Mon Dec 22 2022 23:16:49 GMT+0530 (India Standard Time)
 Request Spot Instances: No

Version Latest
Description -
 View details in the launch template console

AWS CloudWatch Metrics > instances

Instances (2) Info

Instances (2) Info									
Connect Instance state Actions Launch instances									
Running X Clear filters									
Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS	Public IPv4 IP	Elastic IP
i-0950417826fc2891	ami-05f071c6e52875a8	Running Q	12 micro	Initializing	View alarms +	ap-southeast-1b	ec2-13-212-144-114.ap...	-	-
i-0aa0505a0eac2189d	ami-05f071c6e52875a8	Running Q	12 micro	Initializing	View alarms +	ap-southeast-1a	ec2-13-1-222-65.ap-sout...	3.1.222.65	-

Select an instance