Monitoring and Alerting with CloudWatch

1.1 Creating an AMI for auto scaling

- create an AMI from the existing Web Server 1 Instance
- In the left navigation pane, locate the Instances section, and choose Instances.
- From the **Actions** dropdown list, choose **Image and templates** > **Create image**, and then configure the following.
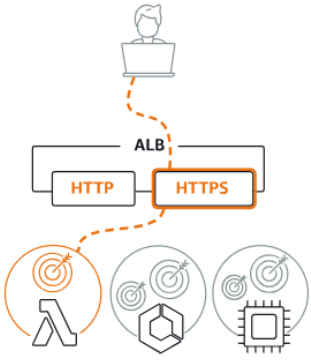

2 Creating a load Balancer

- In the left navigation pane, locate the Load Balancing section, and choose Load Balancers

- In the Load balancer types section, for Application Load Balancer, choose Create.
- Select Application load balancer

Compare and select load balancer type

A complete feature-by-feature comparison along with detailed highlights is also available. Learn more ↗
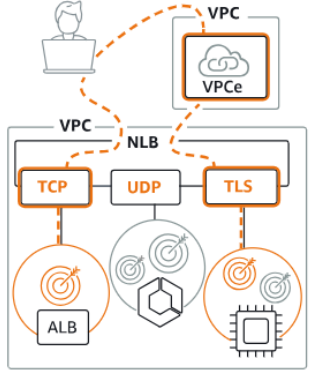
Load balancer types

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*

Close

- Map network to the VPC you create with private and public subnets and select two public subnets
- Select at least two Availability Zones and a subnet for each zone. A load balancer node will be placed in each selected zone and will automatically scale in response to traffic. The load balancer routes traffic to targets in the selected Availability Zones only.

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 ↗.

Lab VPC
vpc-0e44cca92a56b2d3d5
IPv4 VPC CIDR: 10.0.0.0/16

IP pools - *new* | Info

You can optionally choose to configure an IPAM pool as the preferred source for your load balancers IP addresses. Create or view Pools in Amazon VPC IP Address Manager console ↗.

☐ Use IPAM pool for public IPv4 addresses

The IPAM pool you choose will be the preferred source of public IPv4 addresses. If the pool is depleted IPv4 addresses will be assigned by AWS.

Availability Zones and subnets | Info

Select at least two Availability Zones and a subnet for each zone. A load balancer node will be placed in each selected zone and will automatically scale in response to traffic. The load balancer routes traffic to targets in the selected Availability Zones only.

☑ **us-west-2a (usw2-az1)**

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-0e44b3809b0e64d02 | Public Subnet 1 |
| IPv4 subnet CIDR: 10.0.0.0/24 | |

☑ **us-west-2b (usw2-az2)**

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-01e2bcae9a475fe1d | Public Subnet 2 |
| IPv4 subnet CIDR: 10.0.2.0/24 | |

- Select security group which you have to create and enable port 80 for http.

- At Listeners and routing section and click on create target group to create a target group.

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	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.

Add listener

Create target group

Step 2
Register targets

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

lab-target-group

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

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

| TCP ▼ | 80 |
| | 1-65535 |

VPC

Select the VPC with the Application Load Balancer that you want to include in the target group.

Lab VPC

- Add instance to the target

	Instance ID	Name	State	Security groups	Zone
☑	i-00f322f216b4b64e4	Web Server 1	⊘ Running	Web Security Group	us-west-2b

1 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

Review targets

Targets (0)   Remove all pending

Filter targets   ◯ Show only pending

Instance ID ▼	Name ▼	Port ▼	State ▼	Security groups ▼	Zone ▼	Private IPv4 address	Subnet ID ▼	Launch time ▲

- Add the target group in Listing and Routing

✅ **Successfully created load balancer: LabELB**
It might take a few minutes for your load balancer to fully set up and route traffic. Targets will also take a few minutes to complete the registration process and pass initial health checks.

ⓘ **Application Load Balancers now support public IPv4 IP Address Management (IPAM)**
You can get started with this feature by configuring **IP pools** in the **Network mapping** section.

[Edit IP pools]

LabELB

[Actions ▼]

▼ Details

Load balancer type	**Status**	**VPC**	**Load balancer IP address type**
Application	⊙ Provisioning	vpc-0e4cca92a56b2d3d5 ↗	IPv4
Scheme	**Hosted zone**	**Availability Zones**	**Date created**
Internet-facing	Z1H1FL5HABSF5	subnet-0e44b3809b0e64d02 ↗ us-west-2a (usw2-az1)	April 28, 2025, 21:29 (UTC+05:30)
		subnet-01e2bcae9a475fe1d ↗ us-west-2b (usw2-az2)	

Load balancer ARN	**DNS name** Info
📋 arn:aws:elasticloadbalancing:us-west-2:662892777925:loadbalancer/app/LabELB/607ca1df0087208e	📋 LabELB-344476822.us-west-2.elb.amazonaws.com (A Record)

- We have successfully created load Balancer.

3. Creating a launch template

- create a launch template for your Auto Scaling group. A launch template is a template that an Auto Scaling group uses to launch EC2 instances. When you create a launch template, you specify information for the instances, such as the AMI, instance type, key pair, security group, and disks

- In the left navigation pane, locate the **Instances** section, and choose **Launch Templates** and create launch template

Create launch template

Creating a launch template allows you to create a saved instance configuration that can be reused, shared and launched at a later time. Templates can have multiple versions.

Launch template name and description

Launch template name - *required*

`lab-app-launch-template`

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

Template version description

`A web server for the load test app`

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

▶ 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.

▼ Summary

Software Image (AMI)
Lab AMI for Web Server
ami-0c1105ef7bb23591f

Virtual server type (instance type)
t3.micro

Firewall (security group)
Web Security Group

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 GB of snapshots, and 100 GB of bandwidth to the internet.

Cancel   [Create launch template]

- For Auto Scaling guidance, choose Provide guidance to help me set up a template that I can use with EC2 Auto Scaling.
- In the **Application and OS Images (Amazon Machine Image) - required** section, choose the **My AMIs** tab. Notice that **Web Server AMI** and choose that AMI.
- In the Key pair (login) section, confirm that you create a key pair to log to instances in autoscaling group.

- Click on launch template

4. Creating an Auto Scaling group

- Choose lab-app-launch-template, and then from Actions dropdown list, choose Create Auto Scaling group

- On the Choose instance launch options page, in the Network section, configure the VPC and SUBNETS

- In the Integrate with other services – optional page choose existing load balancer which we have already created before.

- On the Configure group size and scaling policies – optional page

- In the **Group size –** *optional* section, enter the following values: **Desired capacity**:2, **Minimum capacity**: 2 **Maximum capacity**: 4
- In the Scaling policies – optional section, configure the following options:
- Choose **Target tracking scaling policy**.
- For **Metric type**, choose **Average CPU utilization**.
- Change the **Target value** to 50
- This change tells Auto Scaling to maintain an average CPU utilization across all instances of 50 percent. Auto Scaling automatically adds or removes capacity as required to keep the metric at or close to the specified target value. It adjusts to fluctuations in the metric due to a fluctuating load pattern.

- Add Notification: Send notifications to SNS topics whenever Amazon EC2 Auto Scaling launches or terminates the EC2 instances in your Auto Scaling group.

- Create a topic

ⓘ **New Feature**
Amazon SNS now supports High Throughput FIFO topics. Learn more ↗

Create topic

Details

Type | Info
Topic type cannot be modified after topic is created

○ FIFO (first-in, first-out)
  • Strictly-preserved message ordering
  • Exactly-once message delivery
  • Subscription protocols: SQS

● Standard
  • Best-effort message ordering
  • At-least once message delivery
  • Subscription protocols: SQS, Lambda, Data Firehose, HTTP, SMS, email, mobile application endpoints

Name

```
MyTopic
```

Maximum 256 characters. Can include alphanumeric characters, hyphens (-) and underscores (_).

Display name - *optional* | Info
To use this topic with SMS subscriptions, enter a display name. Only the first 10 characters are displayed in an SMS message.

```
My Topic
```

Maximum 100 characters.

- Review everything and choose create auto scaling group.

⊘ Lab Auto Scaling Group, 1 Scaling policy created successfully

Auto Scaling groups (1) Info

Last updated less than a minute ago ↻ **Launch configurations** **Launch templates** ↗ **Actions** ▼

```
Search your Auto Scaling groups
```

☐	Name ▽	Launch template/configuration ↗ ▽	Instances ▽	Status ▽	Desired capacity ▽	Min ▽	Max ▽	Availability Zones
☐	**Lab Auto Scaling Group**	lab-app-launch-template	Version Defaul	0	⊖ Updating capacity...	2	2	4	us-west-2b

---

- In the left navigation pane, locate the Instances section, and choose Instances.
- You should see two new instances named **Lab Instance**. These instances were launched by auto scaling.

Instances (1/3) Info

Last updated 1 minute ago ↻ **Connect** **Instance state** ▼ **Actions** ▼ **Launch instances** ▼

```
Find Instance by attribute or tag (case-sensitive)
```

All states ▼   ‹ 1 › ⚙

☐	Name ✎ ▽	Instance ID	Instance state ▽	Instance type ▽	Status check	Alarm status	Availability Zone ▽	Public IPv4 DNS ▽	Publi
☑	Web Server 1	i-00f322f216b4b64e4	⊘ Running ⊕ ⊖	t3.micro	⊘ 3/3 checks passec	View alarms +	us-west-2b	–	34.2?
☐	Lab Instance	i-0292e65c2a73825d4	⊘ Running ⊕ ⊖	t3.micro	⏱ Initializing	View alarms +	us-west-2b	–	–
☐	Lab Instance	i-07994dc6e699aec01	⊘ Running ⊕ ⊖	t3.micro	⏱ Initializing	View alarms +	us-west-2b	–	–

═

i-00f322f216b4b64e4 (Web Server 1)   ⚙ ⌄

- Verifying that load balancing is working

- in the Load Balancing section, choose Target Groups.
- Choose lab-target-group which we have created.
- In the Registered targets section, two Lab Instance targets should be listed for this target group.
- Wait until the Health status of both instances changes to healthy.

5. Verifying that load balancing is working

- You should see two new instances named Lab Instance. These instances were launched by auto scaling.

- Open a new web browser tab, paste the DNS name that you copied before, and press Enter.

- If we refresh website we can see change in availability zone

6: Testing auto scaling

- In the AWS Management Console, in the search bar, enter and choose CloudWatch
- In the left navigation pane, in the **Alarms** section, choose **All alarms**.
- Two alarms are displayed. The Auto Scaling group automatically created these two alarms. These alarms automatically keep the average CPU load close to 50 percent while also staying within the limitation of having 2–4 instances.
- Choose the alarm that has **AlarmHigh** in its name. This alarm should have a **State** of OK. The *OK* state indicates that the alarm has not been initiated. It is the alarm for **CPU Utilization > 50**, which adds instances when the average CPU utilization is high. The chart should show very low levels of CPU at the moment.
- You should see the **AlarmHigh** chart indicating an increasing CPU percentage. Once it crosses the 50 percent line for more than 3 minutes, it initiates auto scaling to add additional instances.

- More than two instances named **Lab Instance** should now be running. Auto scaling created the new instances in response to the alarm.