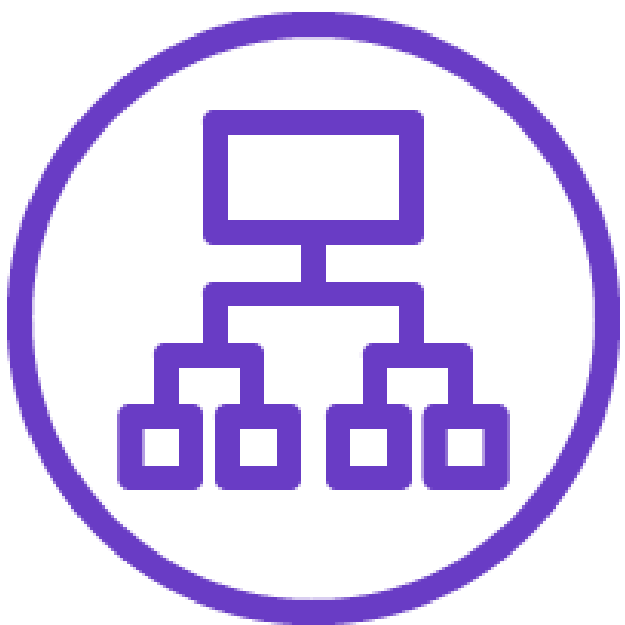




45°

Lab - AWS re/Start

Escalado y balanceo de carga de una arquitectura



Amazon EC2 Auto Scaling





Interactuando con Arquitecturas

Nota. *ELB distribuye automáticamente el tráfico entrante de las aplicaciones entre varias instancias de EC2. ELB proporciona la capacidad de balanceo de carga necesaria para enrutar el tráfico de las aplicaciones y ayudarlo a conseguir *tolerancia a errores* en sus aplicaciones.*

*Auto Scaling le ayuda a mantener la *disponibilidad* de las aplicaciones y le ofrece la posibilidad de *reducir o aumentar automáticamente la capacidad (cant de instancias)* de Amazon EC2 en función de las condiciones que defina. Puede utilizar el escalado automático para asegurarse de que está ejecutando el número deseado de instancias EC2. También puede aumentar automáticamente el número de instancias durante picos de demanda para mantener el rendimiento y puede reducir la capacidad durante periodos de inactividad para reducir costes. El escalado automático es adecuado para aplicaciones que tienen *patrones de demanda estables* o que experimentan una *variabilidad de uso horaria, diaria o semanal*.*

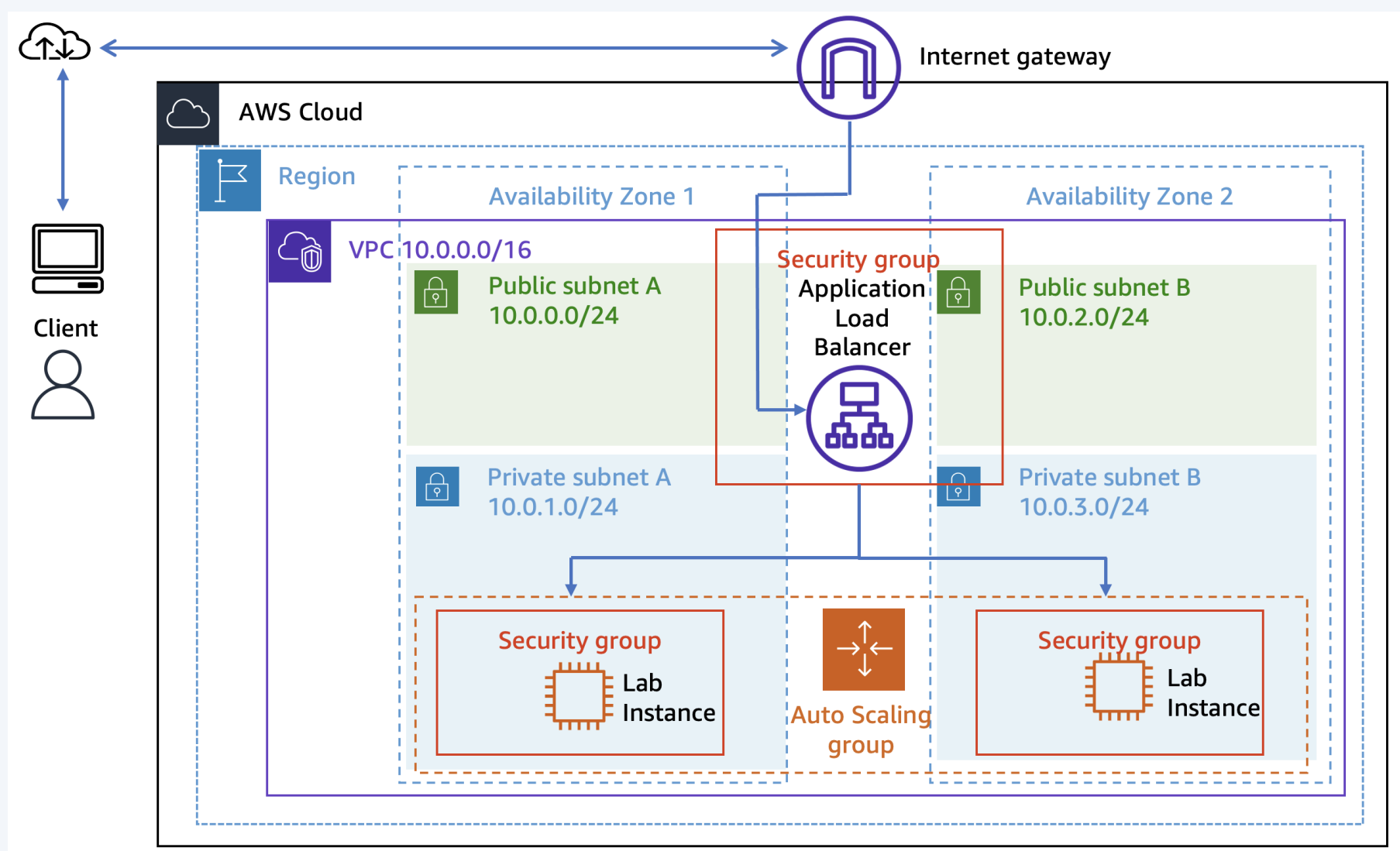
Tarea 01



Los objetivos son:

- Crear una AML a partir de una instancia EC2.
- Crear un equilibrador de carga.
- Cree una plantilla de lanzamiento y un grupo de Auto Scaling.
- Configure un grupo de Auto Scaling para escalar nuevas instancias dentro de subredes privadas.
- Utilice las alarmas de Amazon CloudWatch para monitorear el desempeño de su infraestructura.

Y esta es la arquitectura final a desarrollar:



Tarea 01



Empezaremos creando una AMI, que es una imagen de una instancia personalizada, en este caso de *Web Server 1*. Y a partir de esta imagen, podremos crear un plantilla de lanzamiento, que nos permita ejecutar nuevas instancias con las configuraciones, aplicaciones, permisos de *Web Server 1*

EC2 > Instances > i-0179e0931b0e1e276 > Create image

Create image [Info](#)

An image (also referred to as an AMI) defines the programs and settings that are applied when you launch an EC2 instance. You can create an image from the configuration of an existing instance.

Instance ID
i-0179e0931b0e1e276 (Web Server 1)

Image name

Maximum 127 characters. Can't be modified after creation.

Image description - optional

Maximum 255 characters

No reboot
☐ Enable

Instance volumes

Storage type	Device	Snapshot	Size	Volume type	IOPS	Throughput	Delete on termination	Encrypted
EBS	/dev/...	Create new snapshot fr...	8	EBS General Purpose S...	100		<input checked="" type="checkbox"/> Enable	<input type="checkbox"/> Enable

[Add volume](#)

Luego, procedemos a crear el balanceador de carga

EC2 > Load balancers > Create Application 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
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.

IP address type [Info](#)
Select the type of IP addresses that your subnets use.

☒ IPv4
Recommended for internal load balancers.

☐ Dualstack
Includes IPv4 and IPv6 addresses.

Tarea 01



Continuando con la configuración del ALB:

Network mapping [Info](#)

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

VPC [Info](#)

Select the virtual private cloud (VPC) for your targets or you can [create a new VPC](#). Only VPCs with an internet gateway are enabled for selection. The selected VPC can't be changed after the load balancer is created. To confirm the VPC for your targets, view your [target groups](#).

Lab VPC
vpc-04d9a2ddd9090d596
IPv4: 10.0.0.0/16

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.

☒ **us-west-2a (usw2-az2)**

Subnet

subnet-0a880e1dd71aa7622Public Subnet 1

IPv4 address

Assigned by AWS

☒ **us-west-2b (usw2-az1)**

Subnet

subnet-0b4644e0a78713e35Public Subnet 2

IPv4 address

Assigned by AWS

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

Web Security Group
sg-0b6a99f54fdb57f1dVPC: vpc-04d9a2ddd9090d596

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

ProtocolPortDefault action

HTTP:80Forward tolab-target-groupTarget type: Instance, IPv4

1-65535

Info

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

[EC2](#) > [Target groups](#) > Create target group

Step 1
Specify group details

Step 2
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 (1)

Filter instances

Instance ID	Name	State	Security groups	Zone	Private IPv4 address
i-0179e0931b0e1e276	Web Server 1	Running	Web Security Group	us-west-2b	10.0.2.121

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

Summary

Review and confirm your configurations. [Estimate cost](#)

Basic configuration [Edit](#)

LabELB

- Internet-facing
- IPv4

Security groups [Edit](#)

- Web Security Group
[sg-0b6a99f54fdb57f1d](#)

Network mapping [Edit](#)

VPC: [vpc-04d9a2ddd9090d596](#)
Lab VPC

- us-west-2a
[subnet-0a880e1dd71aa7622](#)
Public Subnet 1
- us-west-2b
[subnet-0b4644e0a78713e35](#)
Public Subnet 2

Listeners and routing [Edit](#)

- HTTP:80 defaults to
[lab-target-group](#)

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.

Luego, creamos nuestra plantilla de lanzamiento para nuestro ASG (Auto Scaling Group). Es decir que el ASG, crea nueva instancias a partir de la AMI que asignemos

Tarea 01



Acerca de la configuración de nuestra plantilla:

[EC2](#) > [Launch templates](#) > 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*

LabELB-1242855894.us-west-2.elb.amazonaws.com

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

▼ Key pair (login) [Info](#)

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

Don't include in launch template ▼

↻ [Create new key pair](#)

▼ Network settings [Info](#)

Subnet [Info](#)

Don't include in launch template ▼

↻ [Create new subnet](#) [↗](#)

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 ▼

Web Security Group sg-0b6a99f54fdb57f1d ✕
VPC: vpc-04d9a2ddd9090d596

↻ [Compare security group rules](#)

► Advanced network configuration

▼ 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

My AMIs

Quick Start

☒ Owned by me

☐ Shared with me



[Browse more AMIs](#)

Including AMIs from AWS, Marketplace and the Community

Amazon Machine Image (AMI)

Web Server AMI
ami-0ec3ff70dab36fc02
2024-03-05T21:21:56.000Z Virtualization: hvm ENA enabled: true Root device type: ebs

Description

Lab AMI for Web Server

Architecture

x86_64

AMI ID

ami-0ec3ff70dab36fc02

▼ Summary

Software Image (AMI)

Lab AMI for Web Server
ami-0ec3ff70dab36fc02

Virtual server type (instance type)

t3.micro

Firewall (security group)

Web Security Group

Storage (volumes)

1 volume(s) - 8 GiB

Cancel

Create launch template

Tarea 01



Ahora, creamos nuestro Grupo de Auto Escalado:

Launch Templates (1/1) Info

Q Search

Launch Template ID	Launch Template Name	Default Version	Latest Version	Create Time	Created By
lt-0959d87b201f6ea92	LabELB-1242855894.us-west-...	1	1	2024-03-05T21:44:38.000Z	arn:aws:sts::33971287059

Actions

Create launch template

Launch instance from template

Modify template (Create new version)

Delete template

Delete template version

Set default version

Manage tags

Create Spot Fleet

Create Auto Scaling group

View details

- 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

Instance type requirements Info

Override launch template

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
LabELB-1242855894.us-west-2.elb.amazonaws.com	Default	A web server for the load test app
lt-0959d87b201f6ea92		

Instance type
t3.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-04d9a2ddd9090d596 (Lab VPC)

10.0.0.0/16

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

us-west-2a | subnet-0512fa5324c312c85 (Private Subnet 1)

10.0.1.0/24

us-west-2b | subnet-0fb1e678694efb7e1 (Private Subnet 2)

10.0.3.0/24

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

lab-target-group | HTTP

Application Load Balancer: LabELB

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

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

Health check grace period Info

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

300

seconds

<

SWIPE

Tarea 01



Y un poco más de configuración del ASG:

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

2

Equal or less than desired capacity

Max desired capacity

4

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.

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

50

Es momento de revisar el funcionamiento del ELB. En primer lugar vemos que el ASG funciona bien, creo el número mínimo deseado de dos instancias. Y en el target group, estas dos nuevas instancias son a donde apunta el ELB

Instances (3) [Info](#)

Find Instance by attribute or tag (case-sensitive)

Any state

Refresh

Connect

Instance state

Actions

Launch instances

<input type="checkbox"/>	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS	Public IPv4 ...	Elastic IP
<input type="checkbox"/>	Lab Instance	i-0d5d7de2cbd059fce	Running	t3.micro	2/2 checks passed	View alarms	us-west-2a	-	-	-
<input type="checkbox"/>	Web Server 1	i-0179e0931b0e1e276	Running	t3.micro	2/2 checks passed	View alarms	us-west-2b	-	35.160.182.189	-
<input type="checkbox"/>	Lab Instance	i-0a9654606059d5ab1	Running	t3.micro	2/2 checks passed	View alarms	us-west-2b	-	-	-

Targets

Monitoring

Health checks

Attributes

Tags

Registered targets (2) [Info](#)

Anomaly mitigation: Not applicable

Refresh

Deregister

Register targets

Filter targets

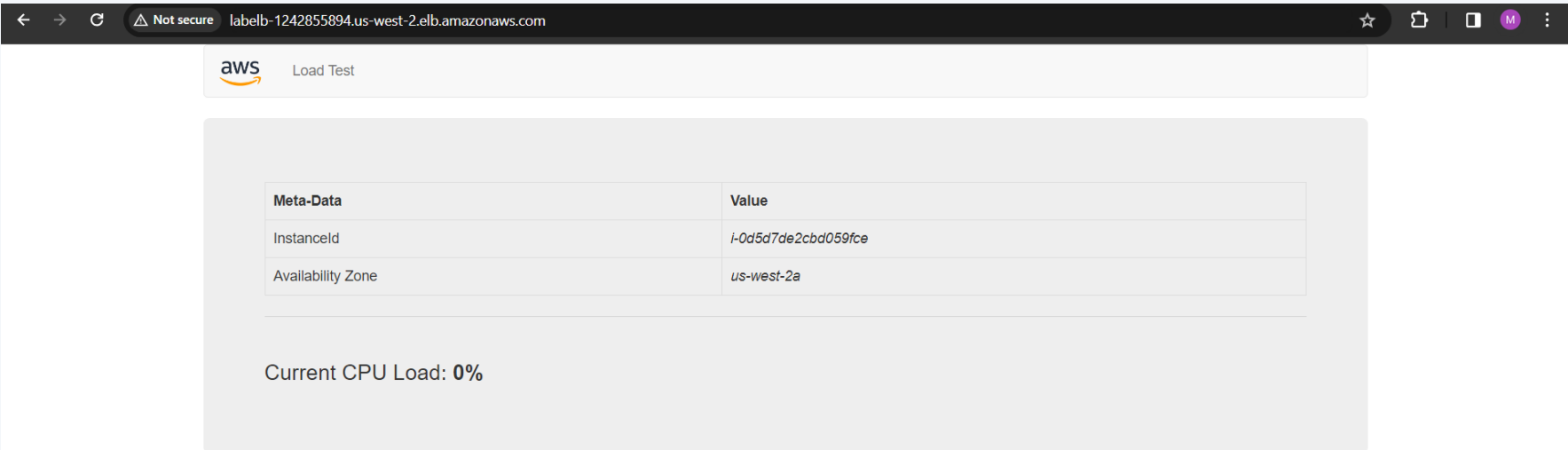
1

<input type="checkbox"/>	Instance ID	Name	Port	Zone	Health status	Health status details	Launch time	Anomaly detection...
<input type="checkbox"/>	i-0d5d7de2cbd059fce	Lab Instance	80	us-west-2a	Healthy	-	March 5, 2024, 16:59 (UTC-05:00)	Normal
<input type="checkbox"/>	i-0a9654606059d5ab1	Lab Instance	80	us-west-2b	Healthy	-	March 5, 2024, 16:59 (UTC-05:00)	Normal

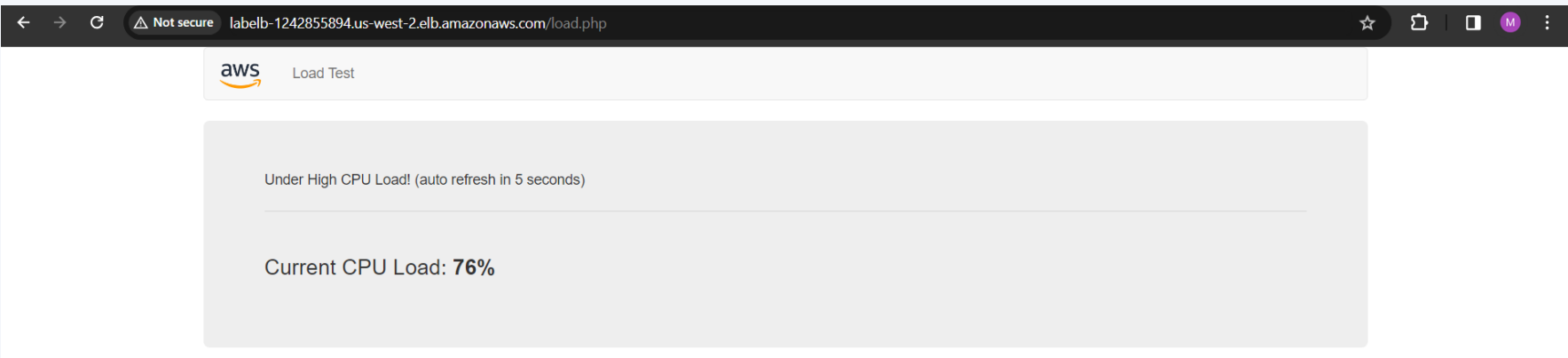
Tarea 01



Efectivamente, el ELB distribuye correctamente el tráfico



Ahora, pondremos a prueba nuestro ASG, pues testaremos un comportamiento mayor a 50% de CPU Utilization:



CloudWatch

> Alarms

Alarms (1/2)

Search

Alarm state: Any

▼

Alarm type: Any

▼

Actions status: Any

▼

< 1 >

Hide Auto Scaling alarms

Clear selection

Create composite alarm

Actions ▼

Create alarm

<div></div>	Name ▼	State ▼	Last state update ▼	Conditions	Actions ▼
<div><div><div></div></div></div>	TargetTracking-Lab Auto Scaling Group-AlarmHigh-24441282-aeae-45b1-9983-644bef2abde8	<div><div></div><div>In alarm</div></div>	2024-03-05 22:14:28	CPUUtilization > 50 for 3 datapoints within 3 minutes	<div><div></div><div>Actions enabled</div></div>
<div><div><div></div></div></div>	TargetTracking-Lab Auto Scaling Group-AlarmLow-c6d4132f-a2f2-41aa-aabf-8fc7ce5d3fa8	<div><div></div><div>OK</div></div>	2024-03-05 22:11:23	CPUUtilization < 35 for 15 datapoints within 15 minutes	<div><div></div><div>Actions enabled</div></div>

Activity history (3)

Filter activity history

< 1 >

Status ▼	Description ▼	Cause ▼	Start time ▼	End time ▼
<div><div></div><div>Waiting for instance war mup</div></div>	Launching a new EC2 instance: i-0c5e2407066283545	At 2024-03-05T22:14:28Z a monitor alarm TargetTracking-Lab Auto Scaling Group-AlarmHigh-24441282-aeae-45b1-9983-644bef2abde8 in state ALARM triggered policy Target Tracking Policy changing the desired capacity from 2 to 3. At 2024-03-05T22:14:39Z an instance was started in response to a difference between desired and actual capacity, increasing the capacity from 2 to 3.	2024 March 05, 05:14:41 PM -05:00	
<div><div></div><div>Successful</div></div>	Launching a new EC2 instance: i-0a9654606059d5ab1	At 2024-03-05T21:59:12Z a user request created an AutoScalingGroup changing the desired capacity from 0 to 2. At 2024-03-05T21:59:26Z an instance was started in response to a difference between desired and actual capacity, increasing the capacity from 0 to 2.	2024 March 05, 04:59:28 PM -05:00	2024 March 05, 04:59:35 PM -05:00
<div><div></div><div>Successful</div></div>	Launching a new EC2 instance: i-0d5d7de2cbd059fce	At 2024-03-05T21:59:12Z a user request created an AutoScalingGroup changing the desired capacity from 0 to 2. At 2024-03-05T21:59:26Z an instance was started in response to a difference between desired and actual capacity, increasing the capacity from 0 to 2.	2024 March 05, 04:59:28 PM -05:00	2024 March 05, 04:59:35 PM -05:00

Finalmente, eliminamos la instancia *Web Server 1*

Successfully terminated i-0179e0931b0e1e276										
Instances (1/4) Find Instance by attribute or tag (case-sensitive) Any state Connect Instance state Actions Launch instances										
<input type="text" value="Find Instance by attribute or tag (case-sensitive)"/> Any state Instance state = running Clear filters < 1 >										
<input type="checkbox"/>	Lab Instance	i-0d5d7de2cbd059fce	Running	t3.micro	2/2 checks passed	View alarms	us-west-2a	-	-	-
<input type="checkbox"/>	Lab Instance	i-0c5e2407066283545	Running	t3.micro	2/2 checks passed	View alarms	us-west-2b	-	-	-
<input type="checkbox"/>	Lab Instance	i-0a9654606059d5ab1	Running	t3.micro	2/2 checks passed	View alarms	us-west-2b	-	-	-

