Case Study – ELB, ASG and Route 53

Problem Statement:

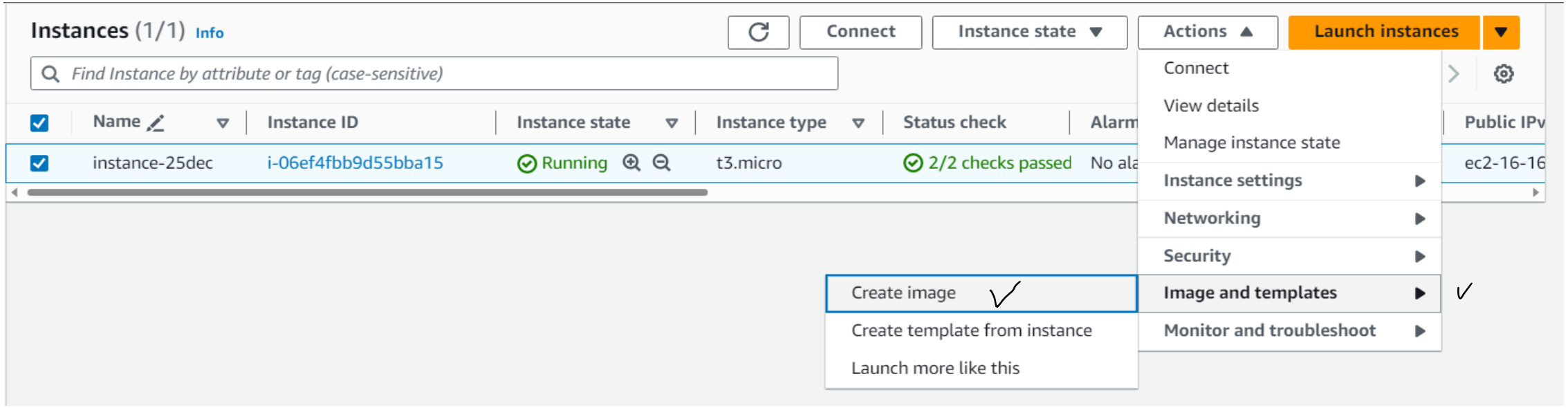
You work for XYZ Corporation that uses on premise solutions and a limited number of systems. With the increase in requests in their application, the load also increases. So, to handle the load the corporation has to buy more systems almost on a regular basis. Realizing the need to cut down the expenses on systems, they decided to move their infrastructure to AWS.

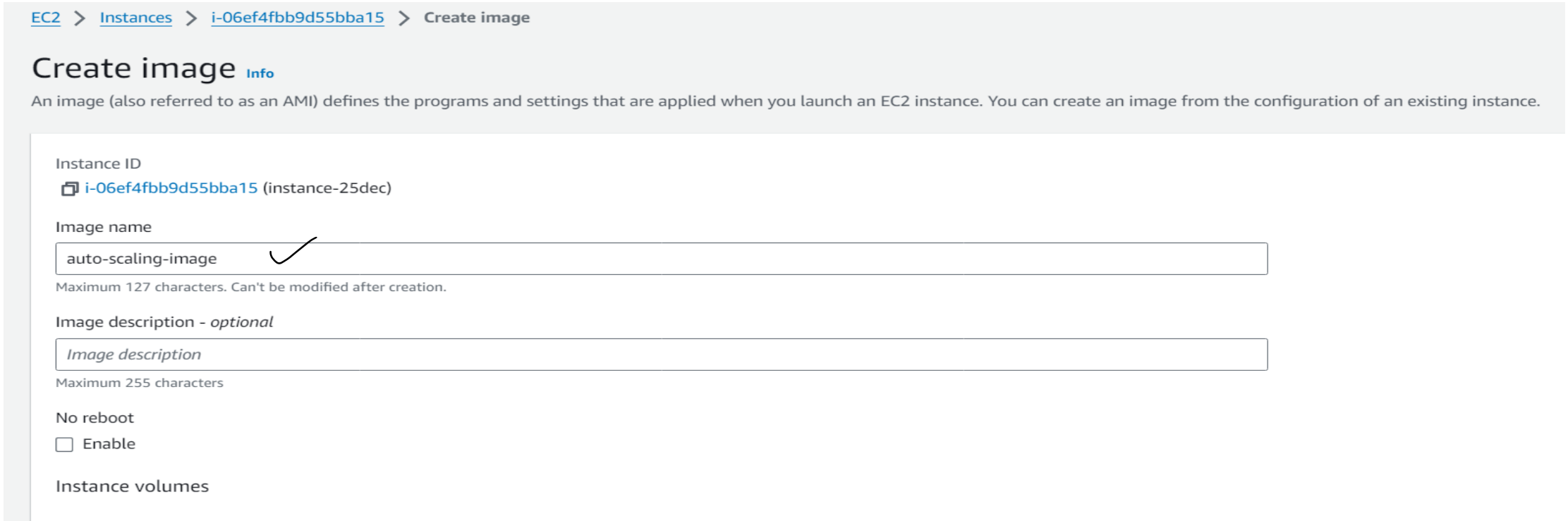
Tasks To Be Performed:

1. Manage the scaling requirements of the company by:
    a. Deploying multiple compute resources on the cloud as soon as the load increases and the CPU utilization exceeds 80%
    b. Removing the resources when the CPU utilization goes under 60%

2. Create a load balancer to distribute the load between compute resources.
3. Route the traffic to the company's domain.

Launch an instance with all traffic allowed- create an image.





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		Enable	Enable

Add volume

During the image creation process, Amazon EC2 creates a snapshot of each of the above volumes.

Tags - *optional*

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

- (•) Tag image and snapshots together
  Tag the image and the snapshots with the same tag.
- ( ) Tag image and snapshots separately
  Tag the image and the snapshots with different tags.

No tags associated with the resource.

Add new tag

You can add up to 50 more tags.

Cancel | Create image

Application and OS Images (Amazon Machine Image) 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

- ( ) Don't include in launch template
- (•) Owned by me
- ( ) Shared with me

Browse more AMIs

Including AMIs from AWS, Marketplace and the Community

Amazon Machine Image (AMI)

auto-scaling-image
ami-0419e5feb0ccfb714
2023-12-25T14:52:12.000Z Virtualization: hvm ENA enabled: true Root device type: ebs

Description

-

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*

my-autoscaling-template

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

Template version description

A prod webserver for MyApp

Max 255 chars

Auto Scaling guidance Info

Select this if you intend to use this template with EC2 Auto Scaling

- [ ] Provide guidance to help me set up a template that I can use with EC2 Auto Scaling

▶ Template tags

▶ Source template

Summary

Software Image (AMI)
auto-scaling-image
ami-0419e5feb0ccfb714

Virtual server type (instance type)
-

Firewall (security group)
-

Storage (volumes)
1 volume(s) - 8 GiB

Free tier: In your first year includes 750 hours of t2.micro (or t3.micro in the Regions in which t2.micro is unavailable) instance usage on free tier AMIs per month, 30 GiB of EBS storage, 2 million IOs, 1 GB of snapshots, and 100 GB of bandwidth to the internet.

Cancel | Create launch template

Choose the same security group – key pair – instance type of image- create launch template.

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.

cs-as-r53

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.

my-autoscaling-template

Create a launch template

Version

Default (1)

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-080f6a5ebfa0d9944
172.31.0.0/16 Default

Create a VPC

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

Select Availability Zones and subnets

eu-north-1a | subnet-04fef07ed760f4ba5
172.31.16.0/20 Default

eu-north-1b | subnet-06dfdd82fb120796a
172.31.32.0/20 Default

eu-north-1c | subnet-06d7f09cca266494f
172.31.0.0/20 Default

Create a subnet

Cancel | Skip to review | Previous | Next

Launch template
Choose a launch template that contains the instance-level settings, such as the Amazon Machine Image (AMI), instance type, key pair, and security groups.

my-autoscaling-template

Create a launch template

Version

Default (1)

Create a launch template version

Description -	Launch template my-autoscaling-template lt-0499aabc3f03006f1	Instance type t3.micro
AMI ID ami-0419e5feb0ccfb714	Security groups -	Request Spot Instances No
Key pair name aws_capstone_project1	Security group IDs sg-0ea6c1cd698da37c8	

Additional details

Storage (volumes) -	Date created Mon Dec 25 2023 20:33:09 GMT+0530 (India Standard Time)

Cancel | Next

Configure advanced options - *optional* Info

Integrate your Auto Scaling group with other services to distribute network traffic across multiple servers using a load balancer or to establish service-to-service communications using VPC Lattice. You can also set options that give you more control over 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 a new load balancer

Define a new load balancer to create for attachment to this Auto Scaling group.

Load balancer type
Choose from the load balancer types offered below. Type selection cannot be changed after the load balancer is created. If you need a different type of load balancer than those offered here, visit the Load Balancing console.

- ◉ Application Load Balancer
  HTTP, HTTPS
- ○ Network Load Balancer
  TCP, UDP, TLS

Load balancer name
Name cannot be changed after the load balancer is created.
cs-as-r53-1

Load balancer scheme
Scheme cannot be changed after the load balancer is created.
- ○ Internal
- ● Internet-facing

Network mapping
Your new load balancer will be created using the same VPC and Availability Zone selections as your Auto Scaling group. You can select different subnets and add subnets from additional Availability Zones.

VPC
vpc-080f6a5ebfa0d9944

Availability Zones and subnets
You must select a single subnet for each Availability Zone enabled. Only public subnets are available for selection to support DNS resolution.

eu-north-1b	subnet-06dfdd82fb120796a
eu-north-1c	subnet-06d7f09cca266494f
eu-north-1a	subnet-04fef07ed760f4ba5

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.
- ☐ 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.
60 seconds

Additional settings

Monitoring | Info
- ☐ Enable group metrics collection within CloudWatch

Default instance warmup | Info
The amount of time that CloudWatch metrics for new instances do not contribute to the group's aggregated instance metrics, as their usage data is not reliable yet.
- ☐ Enable default instance warmup

Cancel | Skip to review | Previous | Next

Listeners and routing
If you require secure listeners, or multiple listeners, you can configure them from the Load Balancing console after your load balancer is created.

Protocol	Port	Default routing (forward to)
HTTP	80	Create a target group

New target group name
An instance target group with default settings will be created.
cs-as-r53-1

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

Add tag
50 remaining

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
  VPC Lattice will not manage your Auto Scaling group's network access and connectivity with other services.
- ○ 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

Configure group size and scaling - *optional* Info
Define your group's desired capacity and scaling limits. You can optionally add automatic scaling to adjust the size of your group.

Group size Info
Set the initial size of the Auto Scaling group. After creating the group, you can change its size to meet demand, either manually or by using automatic scaling.

Desired capacity type
Choose the unit of measurement for the desired capacity value. vCPUs and Memory(GiB) are only supported for mixed instances groups configured with a set of instance attributes.
Units (number of instances)

Desired capacity
Specify your group size.
1

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

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

Min desired capacity	Max desired capacity
1	2
Equal or less than	Equal or greater than

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.

Instance maintenance policy - *new* Info

Control your Auto Scaling group's availability 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.

Control availability and cost during replacement events
An instance maintenance policy determines how much availability your application has when EC2 Auto Scaling replaces instances. It also establishes guardrails that limit the amount of capacity that can be added or removed when replacing instances.

Choose a replacement behavior depending on your availability requirements

Mixed behavior
- (●) **No policy**
  For rebalancing events, new instances will launch before terminating others. For all other events, instances terminate and launch at the same time.

Prioritize availability
- ( ) **Launch before terminating**
  Launch new instances and wait for them to be ready before terminating others. This allows you to go above your desired capacity by a given percentage and may temporarily increase costs.

Control costs
- ( ) **Terminate and launch**
  Terminate and launch instances at the same time. This allows you to go below your desired capacity by a given percentage and may temporarily reduce availability.

Flexible
- ( ) **Custom behavior**
  Set custom values for the minimum and maximum amount of available capacity. This gives you greater flexibility in setting how far below and over your desired capacity EC2 Auto Scaling goes when replacing instances.

Instance scale-in protection

Scale-in protection prevents newly launched instances from being terminated by scaling activities. Make sure to remove scale-in protection for the group or individual instances when instances are ready to be terminated.

- [ ] Enable instance scale-in protection

Cancel | Skip to review | Previous | **Next**

EC2 > Target groups

Target groups (1) Info

Actions ▼ | Create target group

Filter target groups

1

Name	ARN	Port	Protocol	Target type	Load balancer
cs-as-r53-1	arn:aws:elasticloadbalanci...	80	HTTP	Instance	cs-as-r53-1

EC2 > Load balancers

Load balancers (1)

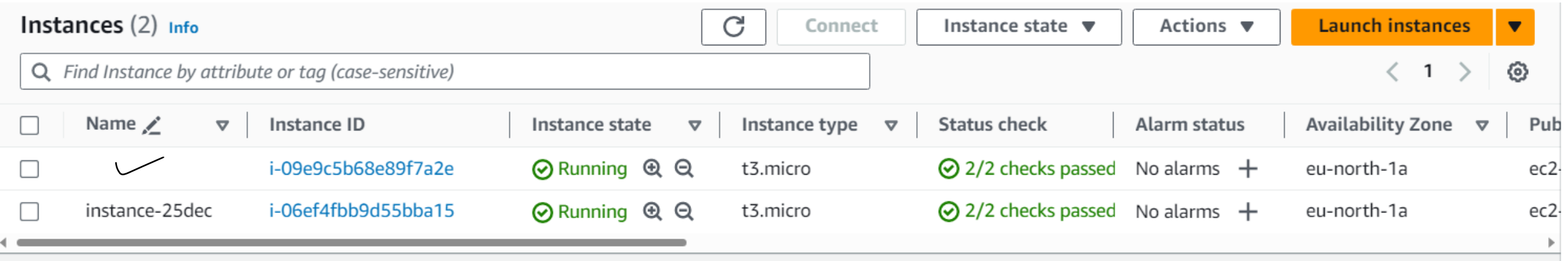
Actions ▼ | Create load balancer ▼

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

Filter load balancers

1

Name	DNS name	State	VPC ID	Availability Zones	Type
cs-as-r53-1	cs-as-r53-1-1417308736.e...	Active	vpc-080f6a5ebfa0d9944	3 Availability Zones	application



There is two methods of directing the traffic . 1- create auto scaling group 2- via load balancer and target group divert the traffic to Router 53.

- Create the target group- using the target group create load balancer

-Create Image of the instance- create the launch template using image – using template create auto scaling group – within the auto scaling group we can choose to auto create or choose our own created load balancer and target group to handle the traffic.

To increase the stress give commands

sudo apt-get install stress

stress --cpu 2 --timeout 60

For Route 53 handle the traffic follow the steps

1-Create the target group and Application Load balancer or we can choose the auto created one.

2- Getting a Domain name from Free Nom and Go daddy.

3- Create hosted Zone in Route53 – Name as Domain we get from above platform- Public hosted zone- create hosted zone.

4- Create Record – open Alias- Route traffic to (Alias to Application load balancer)- Select region where load balancer is available – select the load balancer- create record.

5- Copy every name server in Rote53 and paste them into the Domain providers nameservers.

6- Now copy the Domain name and browse – we can able to see the running website.