Create and Configure the Auto Scaling Group in EC2

Auto Scaling is an Amazon Web Service it allows instances to scale when traffic or CPU load increases.

Auto-scaling is a service that monitors all instances that are configured into the Auto Scaling group and ensures that loads are balanced in all instances.

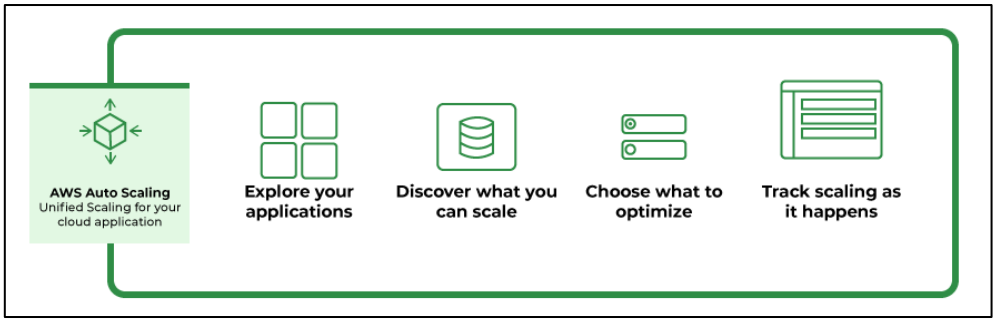
Depending on the load scaling group, increase the instance according to the configuration. When we created the auto-scaling group, we configured the Desired capacity, Minimum capacity, maximum capacity, and CPU utilization. If CPU utilization increases by 60% in all instances, one more instance is created, and if CPU utilization decreases by 30% in all instances, one instance is terminated. These are totally up to us; what is our requirement. If any Instance fails due to any reason, then the Scaling group maintains the Desired capacity and starts another instance.

The auto-scaling group follows Horizontal Scaling. This service is very important for us nowadays because we do not need to create new instances manually and do not require manual monitoring.

Benefits of Auto Scaling

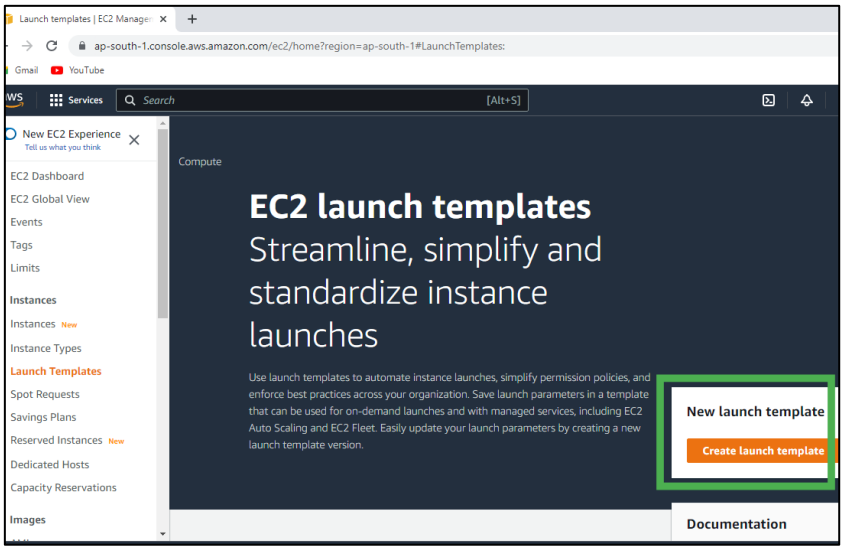
Dynamical scaling: AWS auto-scaling service doesn't required any type of manual intervention it will automatically scale the application down and up by depending up on the incoming traffic.
Pay For You Use: Because of auto scaling the resource will be utilised in the optimised way where the demand is low the resource utilisation will be low and the demand will high the resource utilisation will increase so the AWS is going to charge you only for the amount of resources you really used.
Automatic Performance Maintenance: AWS autoscaling maintains the optimal application performance with considering the workloads it will ensures that the application is running to desired level which will decrease the latency and also the capacity will be increased by based on your application

How AWS Auto Scaling Works?
AWS autoscaling will scale the application based on the load of application. Instead of scaling manually AWS auto scaling will scale the application automatically when the incoming traffic is high it will scale up the application and when the traffic is low it will scale down the application.

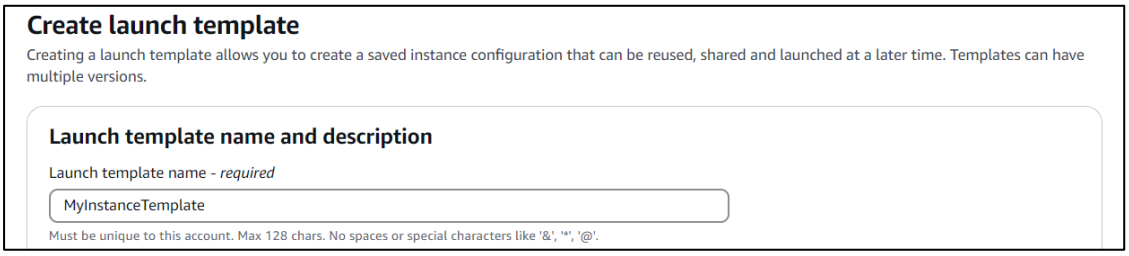


Steps To create Auto Scaling Launch Template

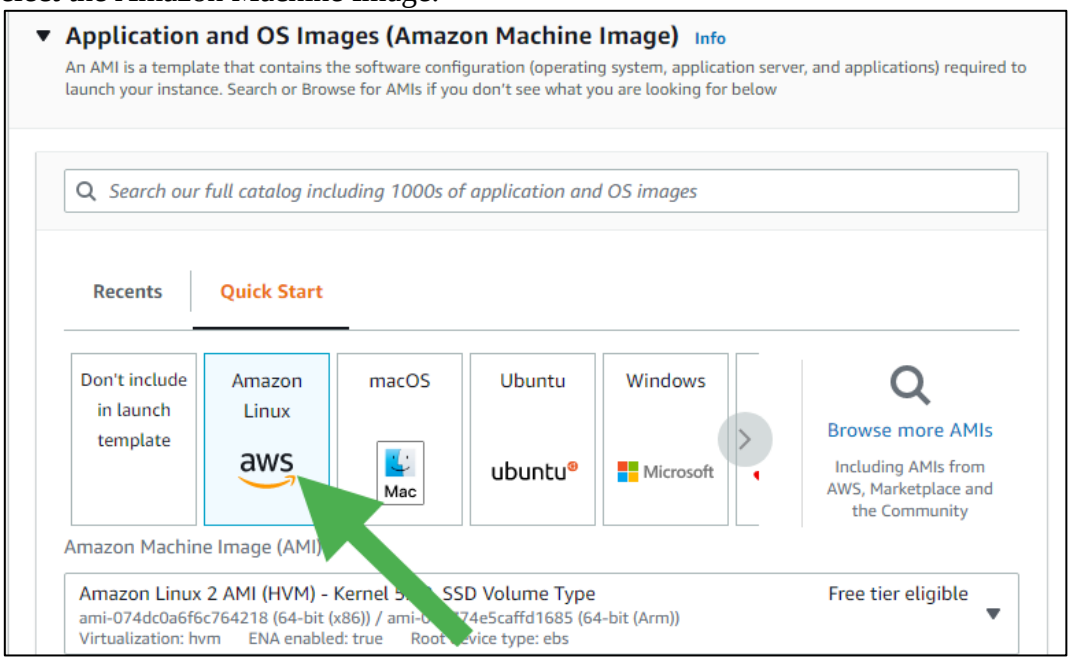
- Login to your AWS Account.
- Click on the EC2(Elastic Cloud Computing) in the homepage search bar.
- Scroll Down and click on the **Launch Templates** and click on the **Create launch template**



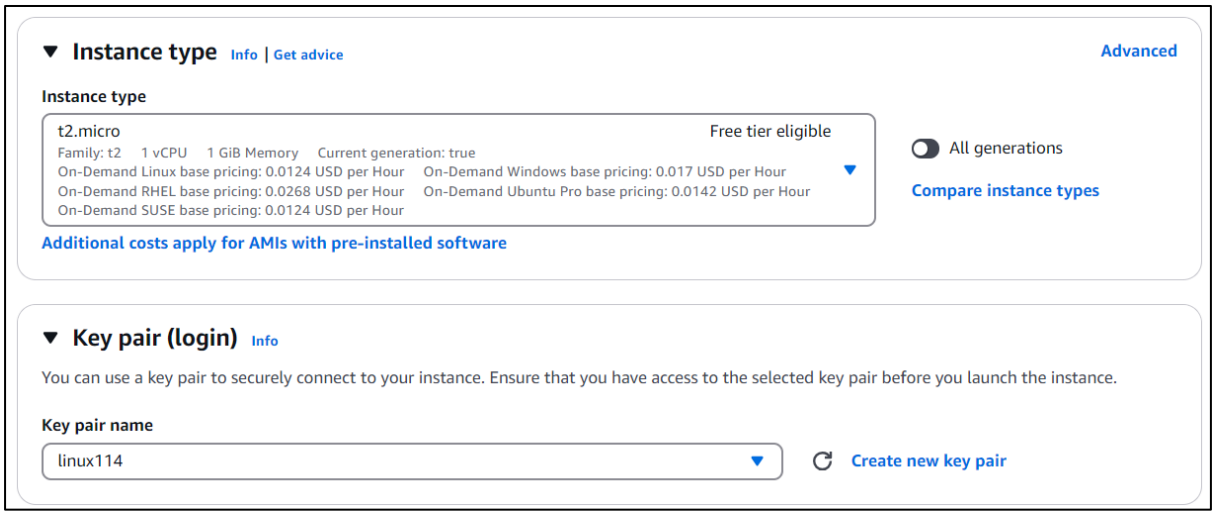
- Type the Template name.



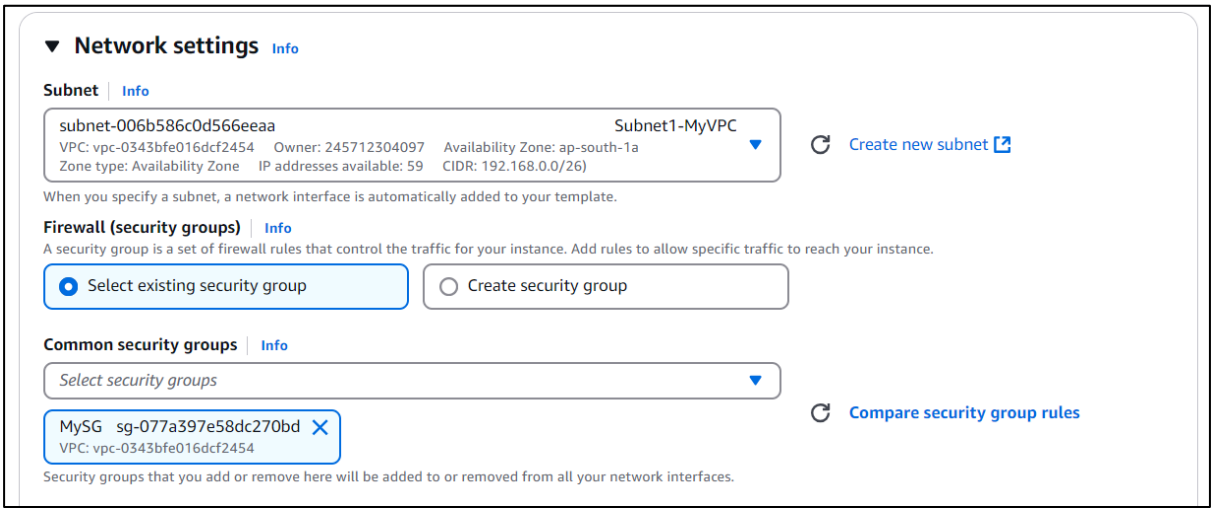
- Select the Amazon Machine Image.



- Select the Instance Type and Key pair.

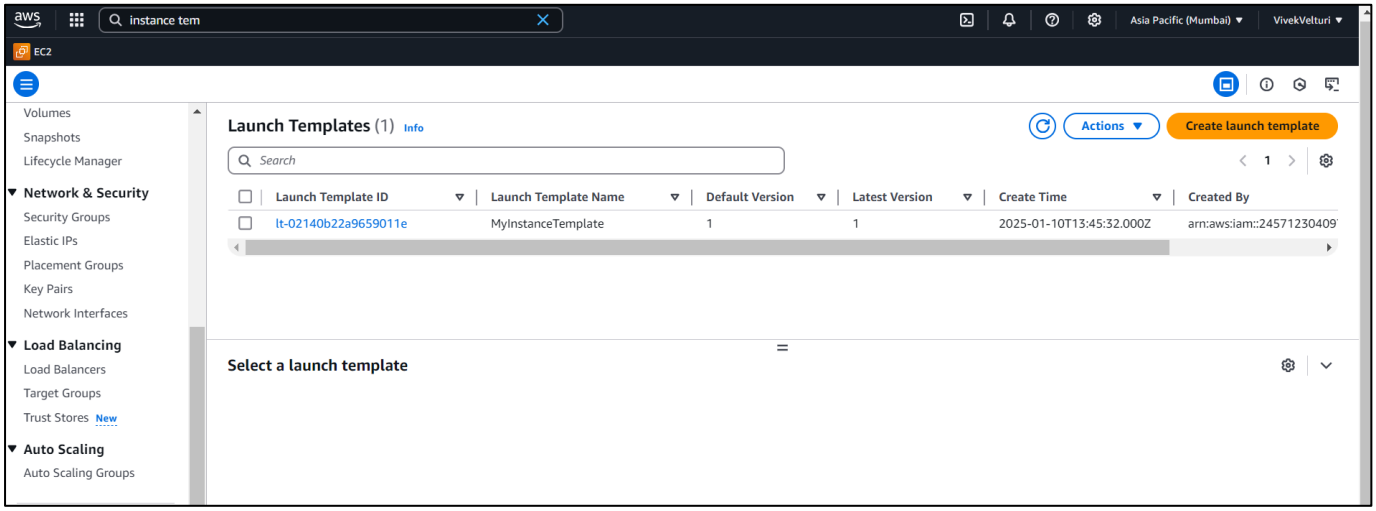


- In Network Settings, Provide Your Subnet or Create a new one & Select the Security Group or Create the new one.

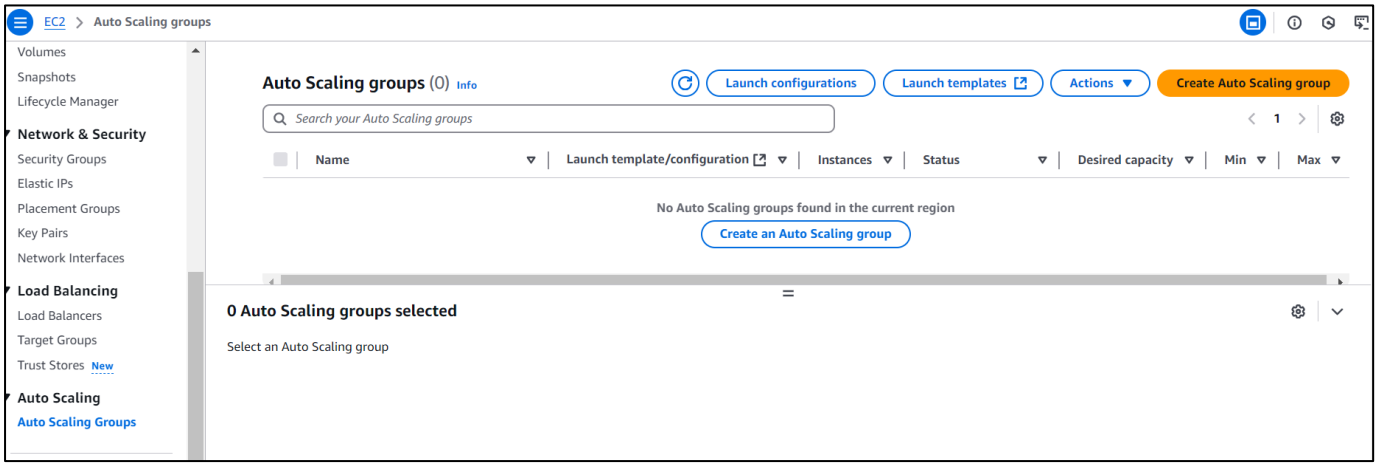


- In Advanced Details, provide your User Data it you have any and click Create launch template on the right-hand side.

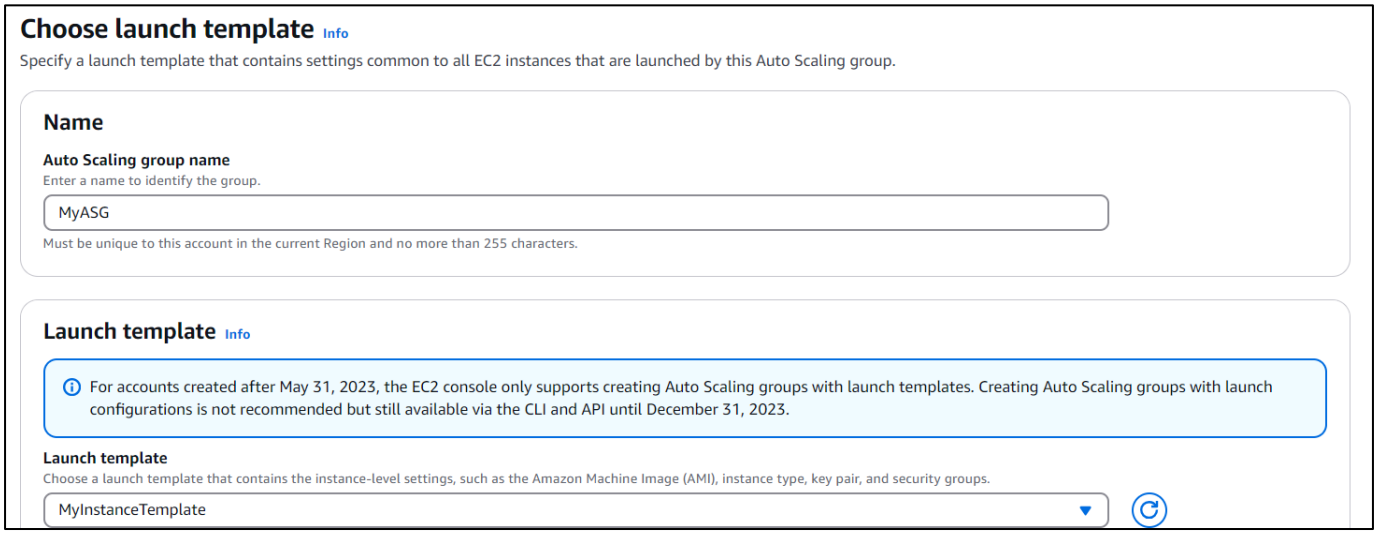
- Your Instance template has been created, Now, scroll down and click on the Auto Scaling Groups.



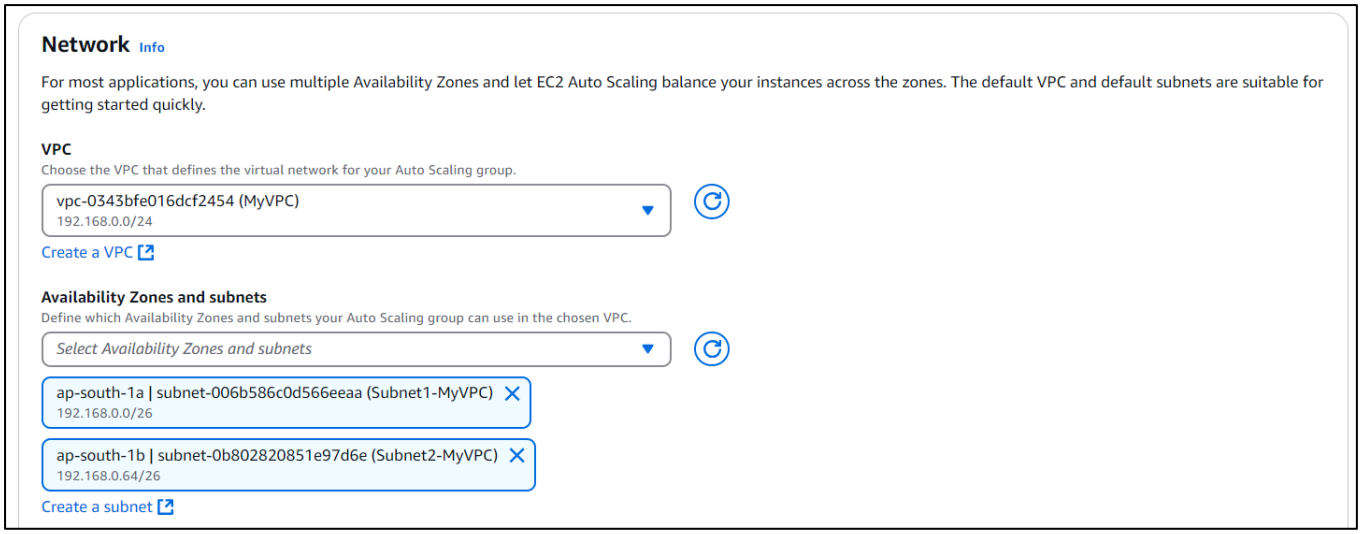
Create An Auto Scaling Group Using a Launch Template



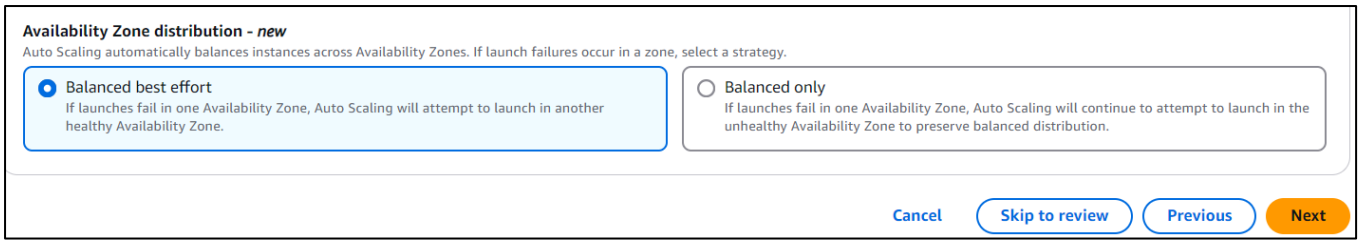
- Type the Auto Scaling group name and select the launch template we have created.



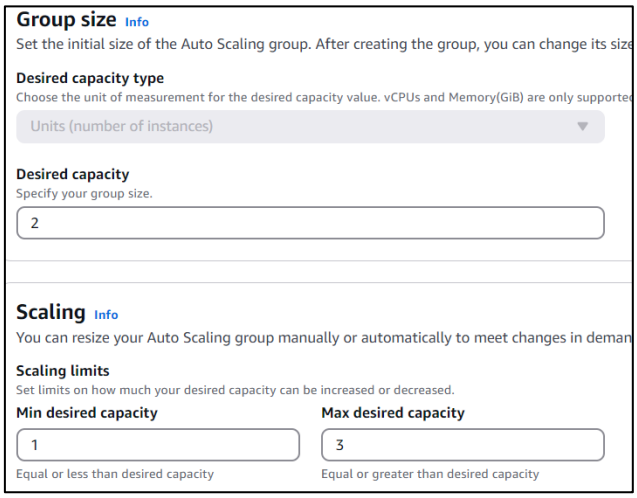
- Click Next
- Select the VPC or go with the default VPC and also select the Availability zone.



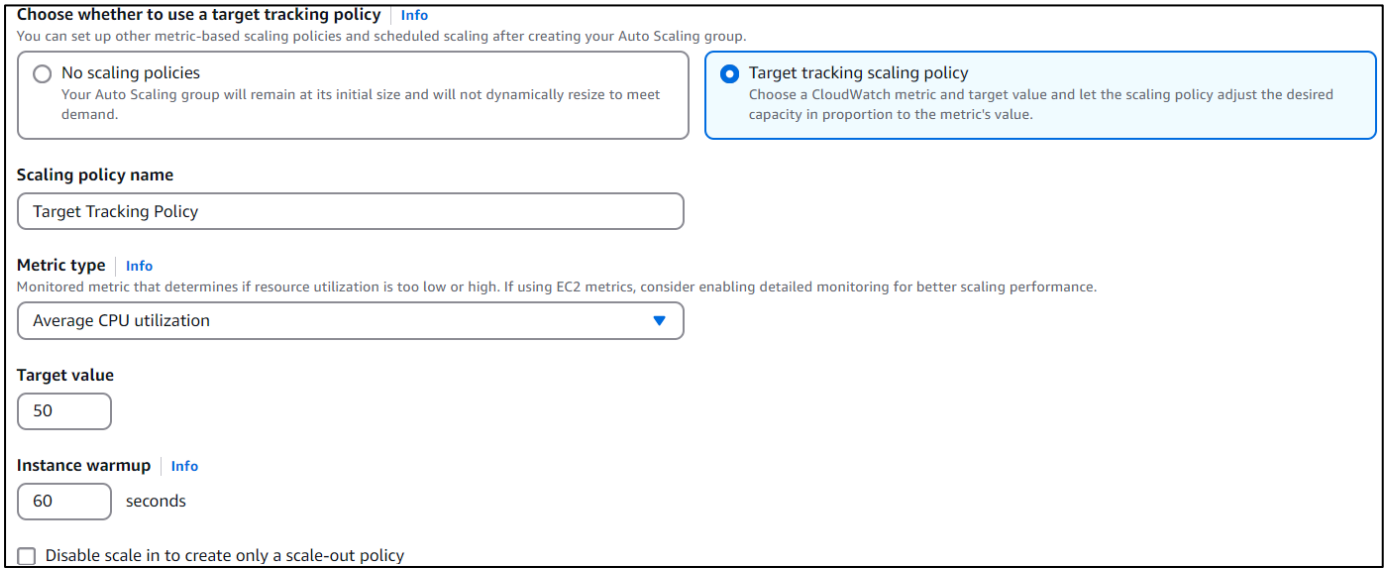
- Select Availability Zone distribution to Balance best effort as default and click Next.



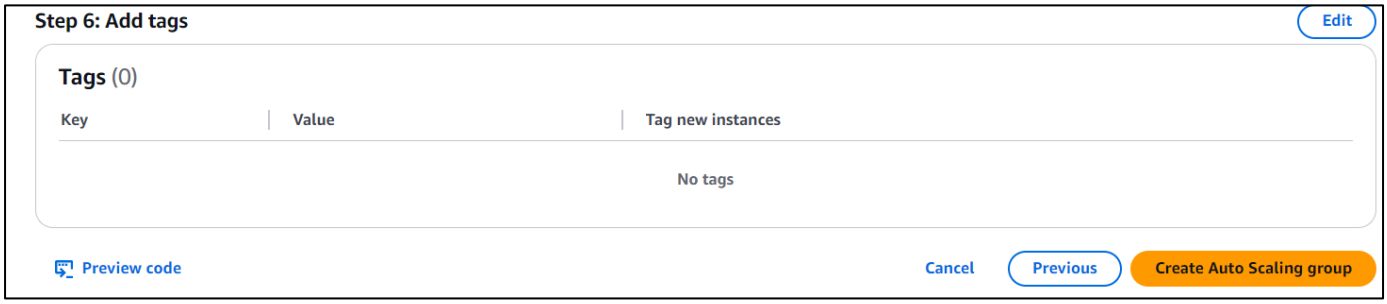
- In Integrate with other services, leave the settings as default and click Next or edit as per you requirement if your project demands for the use of load balancer, health check, etc.
- Configure the Group size and Scaling policies.
    - Select as per your requirement:
    - Desired: 2
    - Minimum: 1
    - Maximum: 3



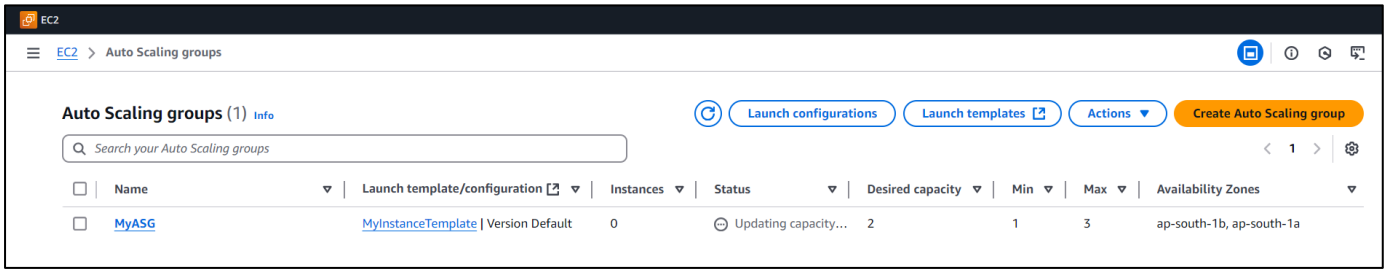
- Select the Target tracking scaling policy and provide the values as per your requirement, in my scenario I have assigned the Metric type to be Average CPU utilization, and the Target value to be 50 while the instance warmup time to be 60 seconds.



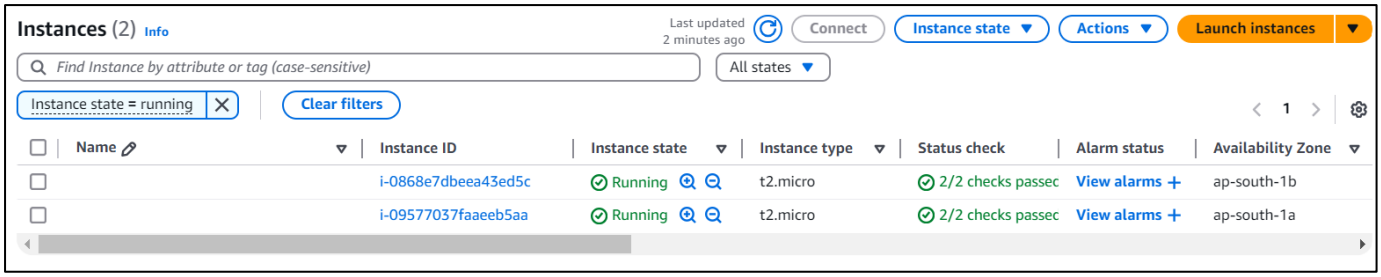
- Click on the Create Auto Scaling Group.



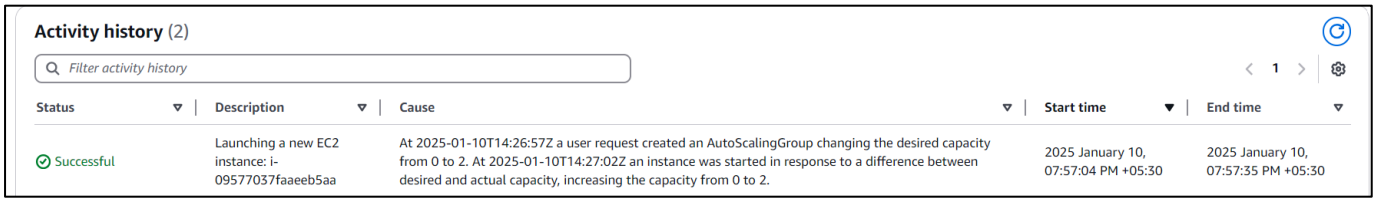
- Now you can see the Auto Scaling is creating and it is also creating the desired state of the EC2 Instance

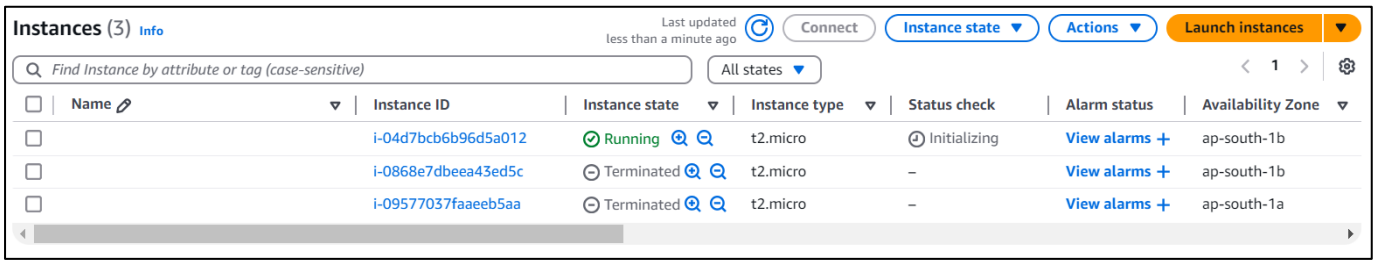


- We selected the Desired state equal to 2 and we can see observe 2 Instances running in the EC2 panel.

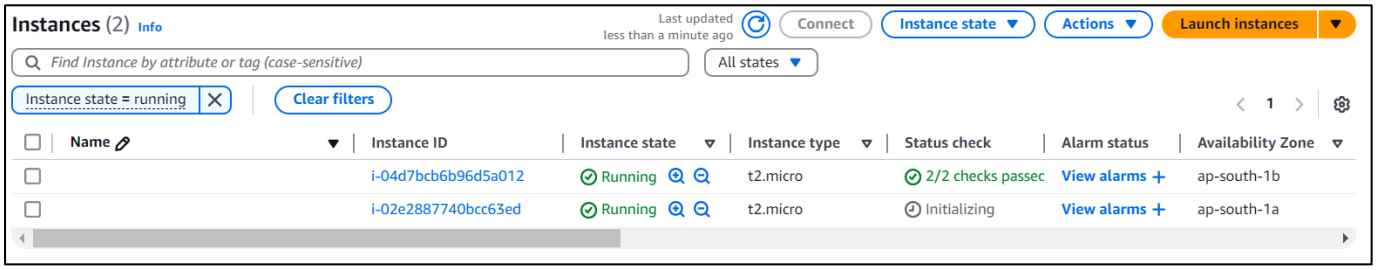


- Now if we manually Terminate the Instance created, The ASG which is monitoring will again start launching instances as per our requirement given, in our case we have requested for a minimum capacity of 1 and desirable capacity of 2 and maximum capacity of 3.
- Upon termination of the instances manually, if we look into the activity of our ASG created we can observe that a new instance has been launched.

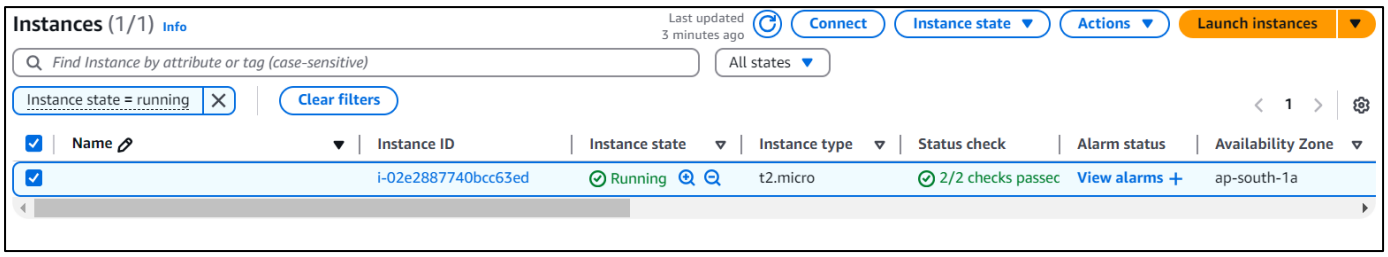




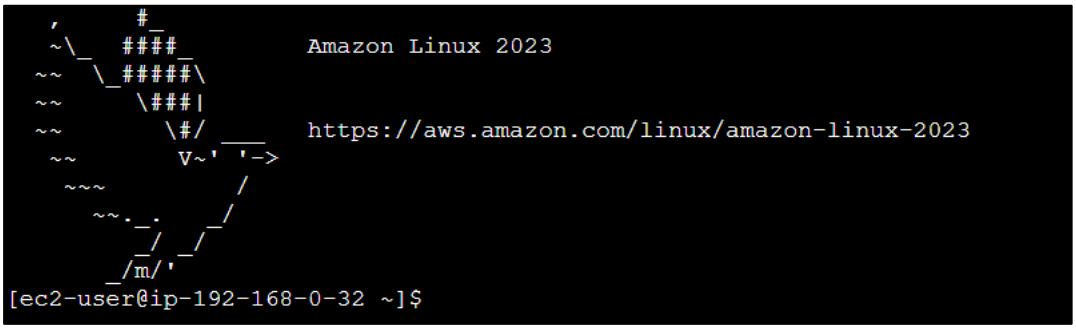
- After a while the launch of a second instance by our ASG has also been observed



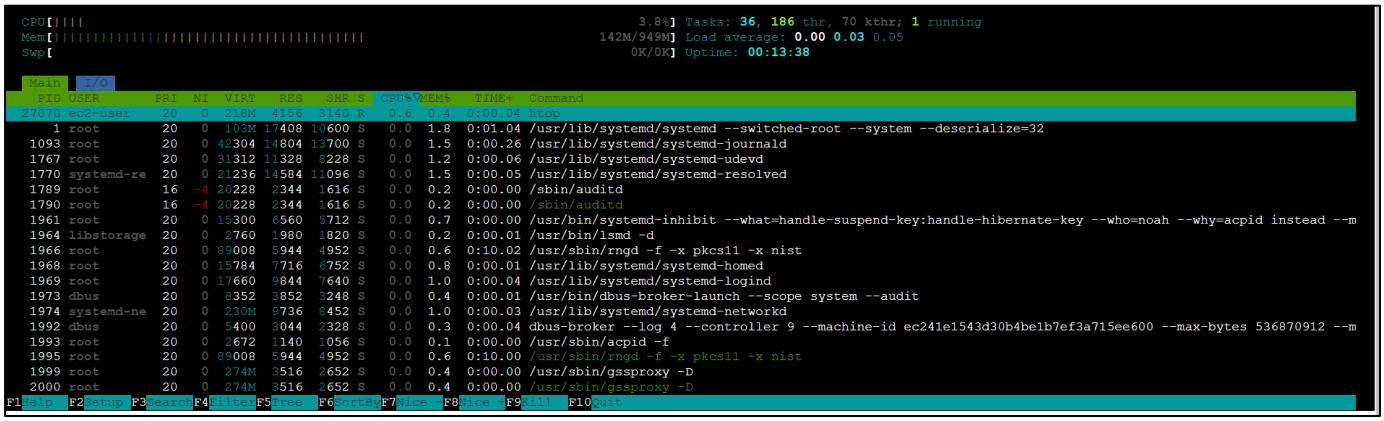
- **I have knowingly given 2 subnets, one private and one public, the instance which was launched with the private subnet was shutdown automatically as it was not able to connect to internet due to lack of an IGW not being assigned to it. While the instance launched with public subnet was observed to be healthy and running.**



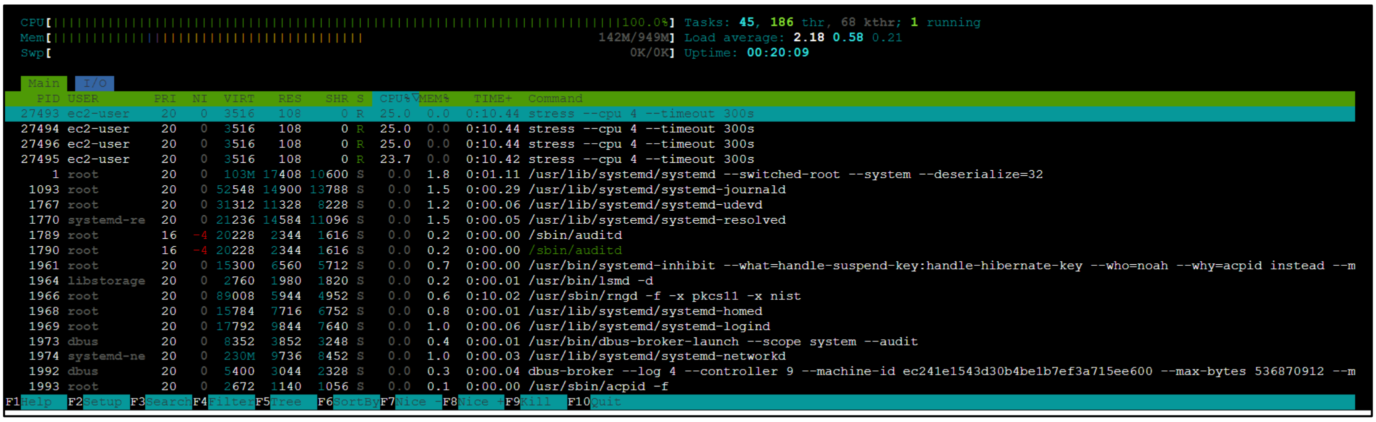
- Lets now connect to this EC2 Instance.

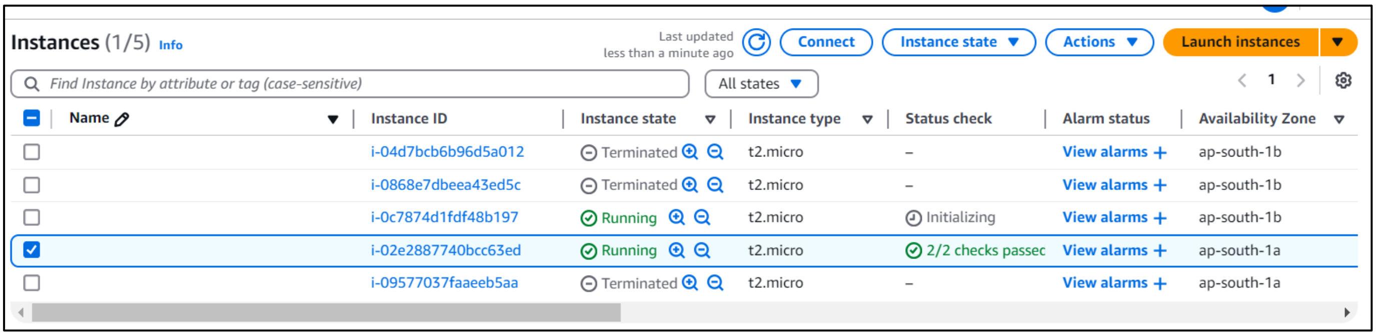


- Execute the following commands to monitor the status of the cpu:
  Sudo yum install htop -y
  htop



- We can observe the cpu utilization percent to be 0.6%, we have already assigned in our target tacking scaling policy to create another instance if the cpu utilization goes over 50%, now lets execute the following commands to incorporate stress into the instance and observe if our ASG launches another instance to handle the load we have manually exerted onto our system:
  sudo yum install stress -y
  stress –cpu 4 –timeout 300s
- Upon executing the above command we can observe that the usage capacity of the cpu has increased to 98.7%, it should now trigger the ASG to launch another instance to handle the load.



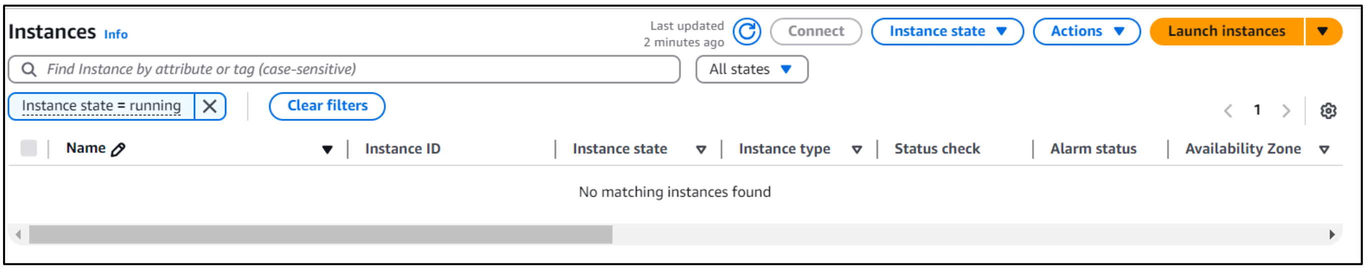


- A new instance has been created to balance out the load we have applied on the old instance.

Delete Your Resources

Delete your ASG:

- In your AWS Management Console, head to **Amazon EC2 and Auto Scaling Groups**
- Select Actions on the right-hand side, Click **Delete**.
- Type delete in the pop up window and Click **Delete.**
- The Instances launched by the ASG will also be deleted automatically as we can go to the EC2 Dashboard and observe if any instances are still Running.



Delete your instance template:

- In your AWS Management Console, head to **Amazon EC2 and Launch Templates.**
- Select Actions on the right-hand side, Click **Delete template.**

- Type delete in the pop up window and Click **Delete.**