Auto Scaling with Application Load Balancer (ALB)

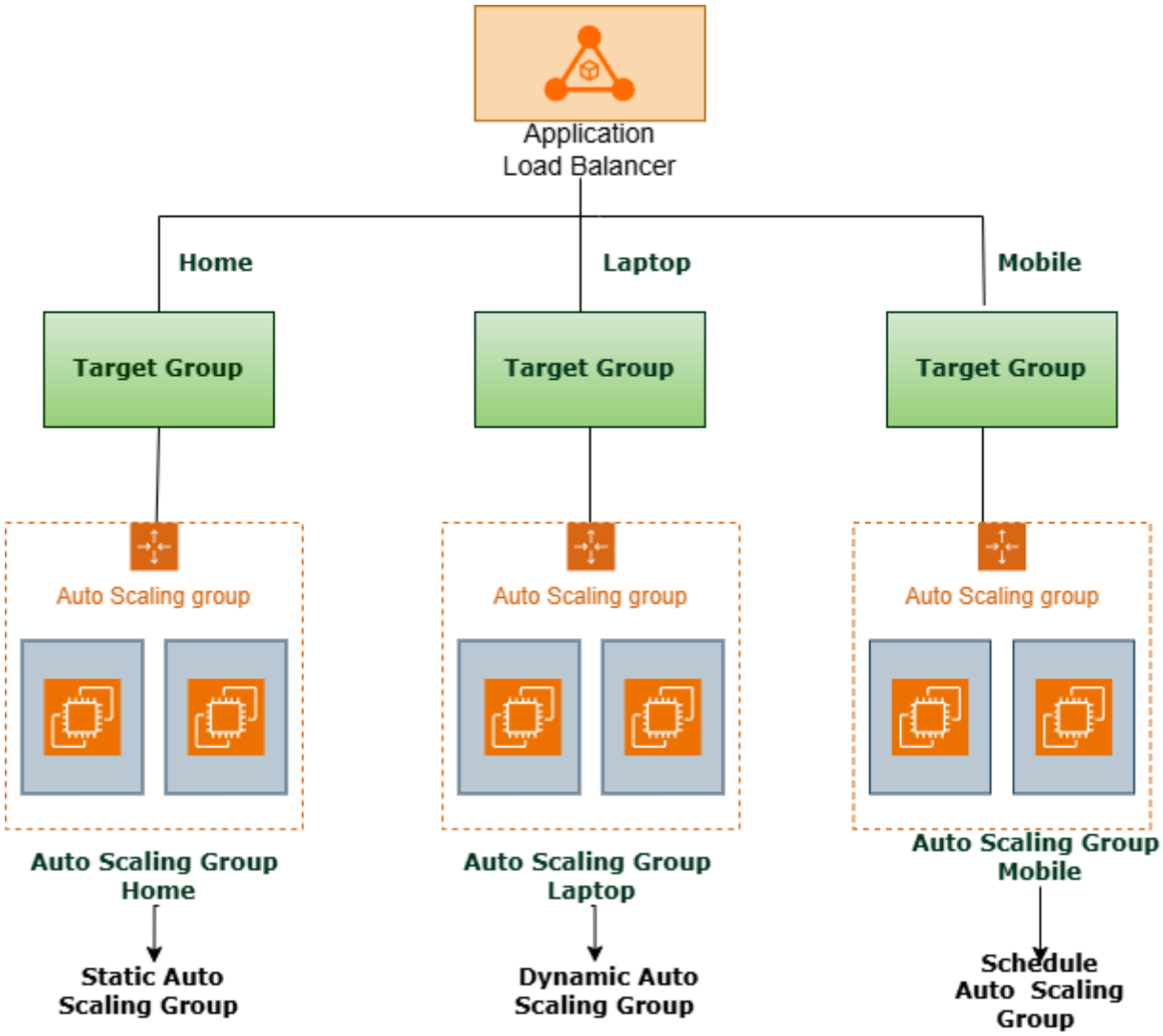
Introduction.

Efficient traffic management and scalability are essential for modern cloud applications. By combining an **Application Load Balancer(ALB)** with multiple **Auto Scaling Groups (ASGs)**, workloads can be distributed intelligently across different user categories **Home, Laptop, and Mobile.** Each group uses a unique scaling strategy (Static, Dynamic, and Scheduled) to ensure consistent performance, high availability, and optimized resource utilization.

Architecture Components.

1. Application Load Balancer (ALB)
    - Distributes incoming traffic across multiple targets (EC2 instances) in different Auto Scaling Groups.
    - Ensures high availability and fault tolerance.
    - Routes requests based on listener rules to the appropriate Target Group (Home, Laptop, or Mobile).
2. Target Groups
    - Logical groups of instances that receive traffic from the ALB.
    - Each target group is associated with an Auto Scaling Group.

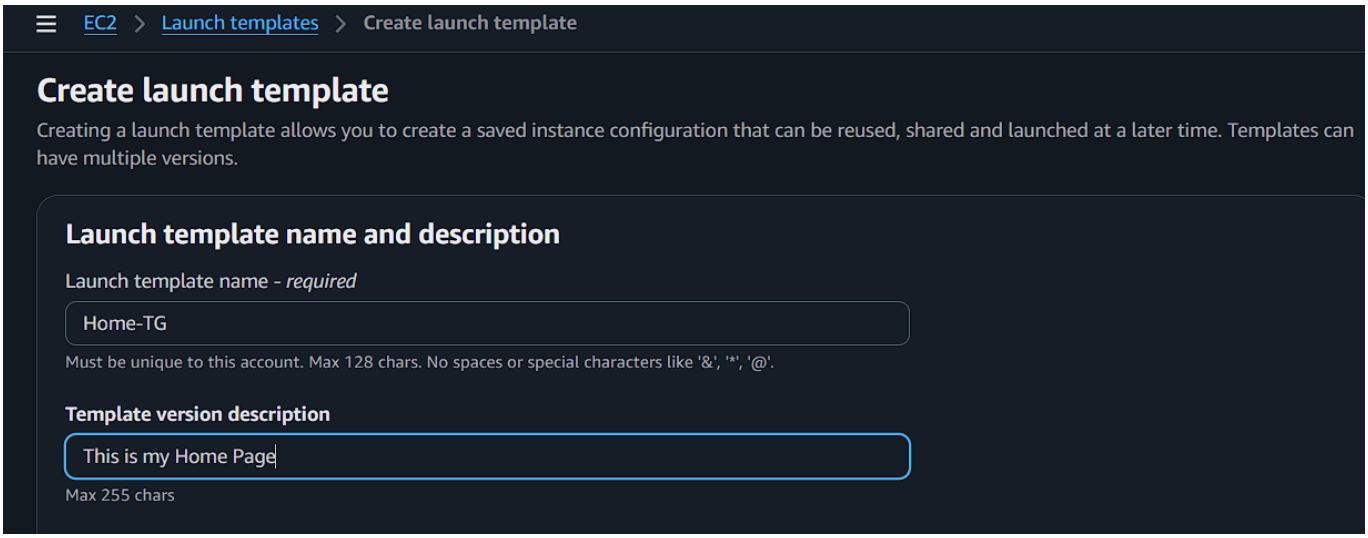
Target Group	Description	Scaling Strategy
Home	Handles traffic from home devices	Static Auto Scaling
Laptop	Manages traffic from laptops	Dynamic Auto Scaling
Mobile	Serves mobile users	Scheduled Auto Scaling

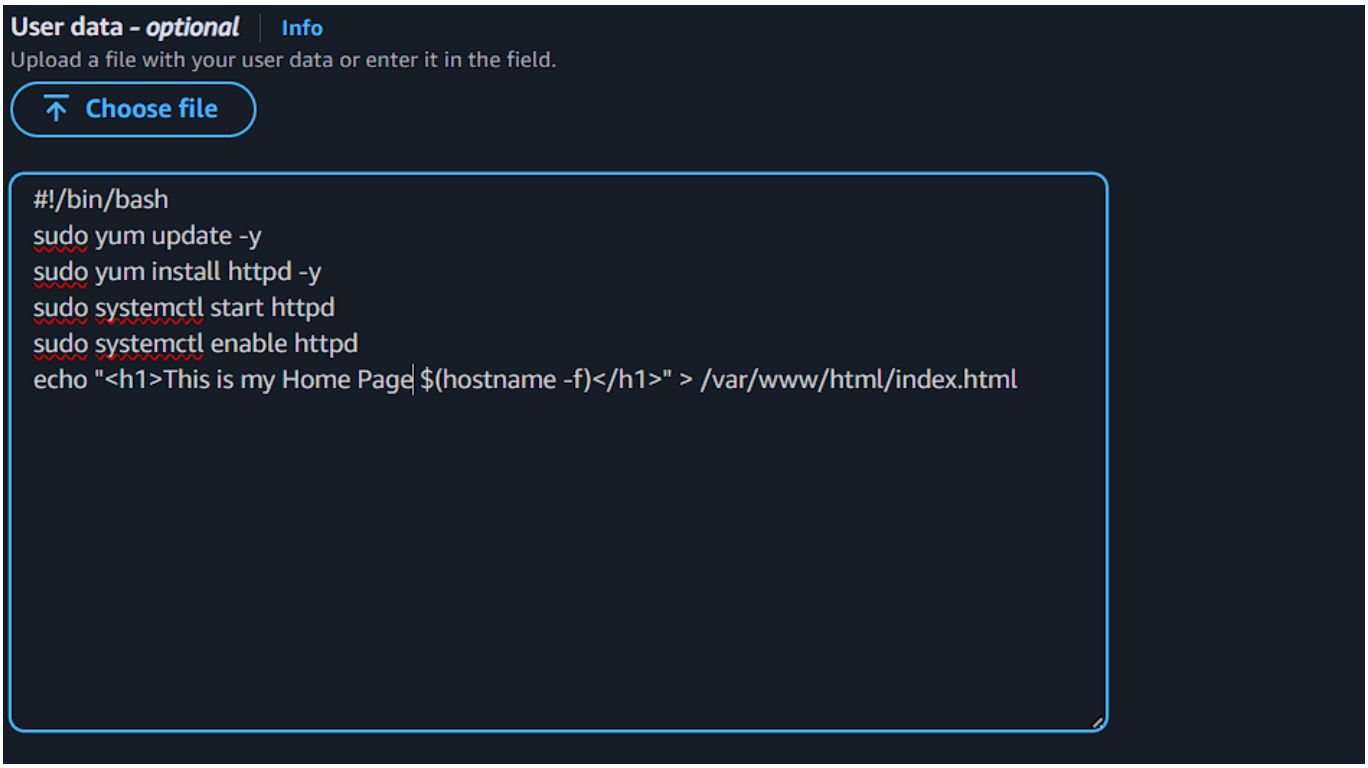


Step-by-Step Deployment Guide.

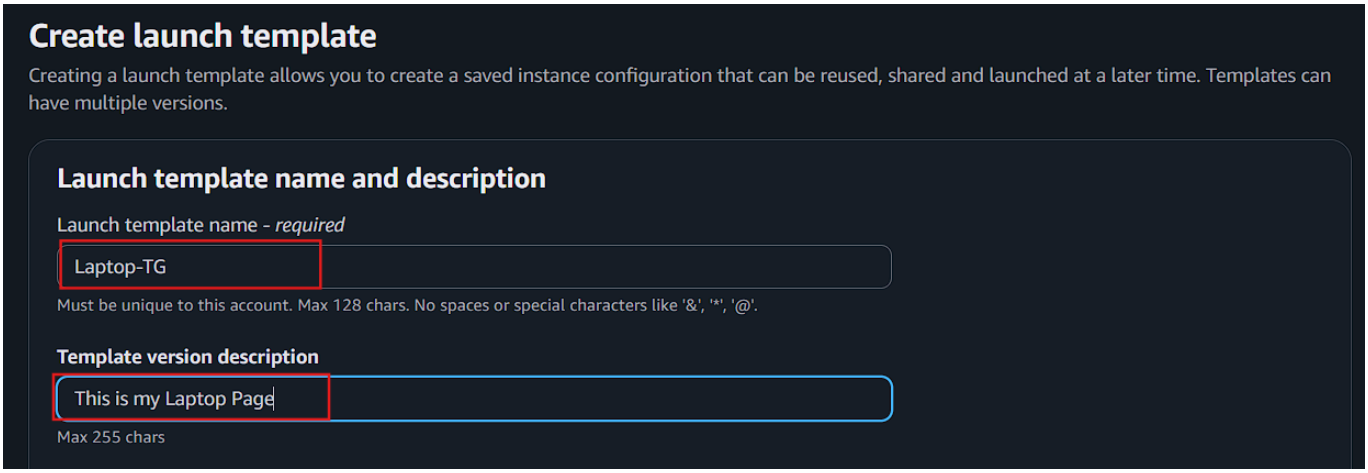
Step 1: Create Launch Template.

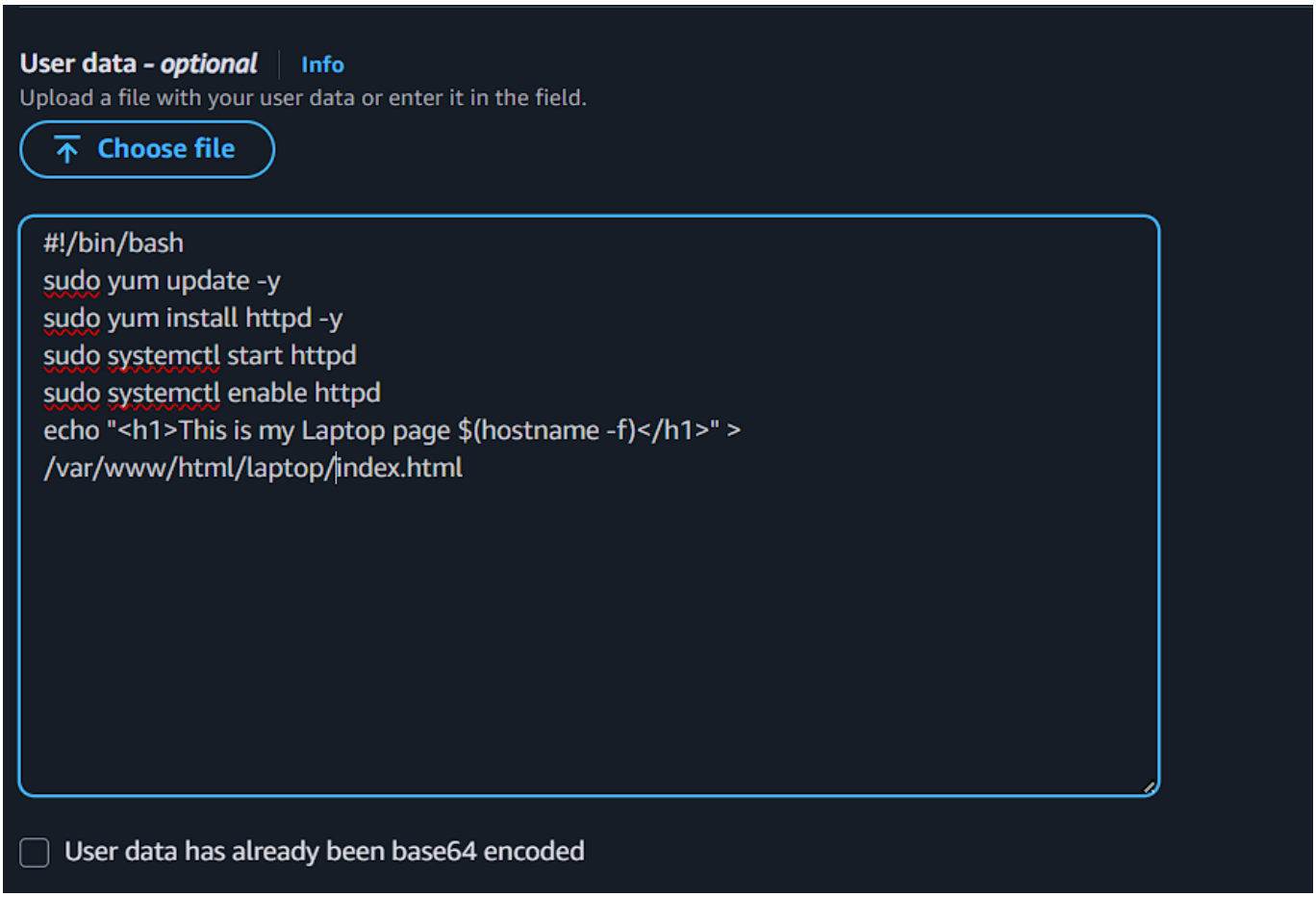
Launch Template (Home-TG).



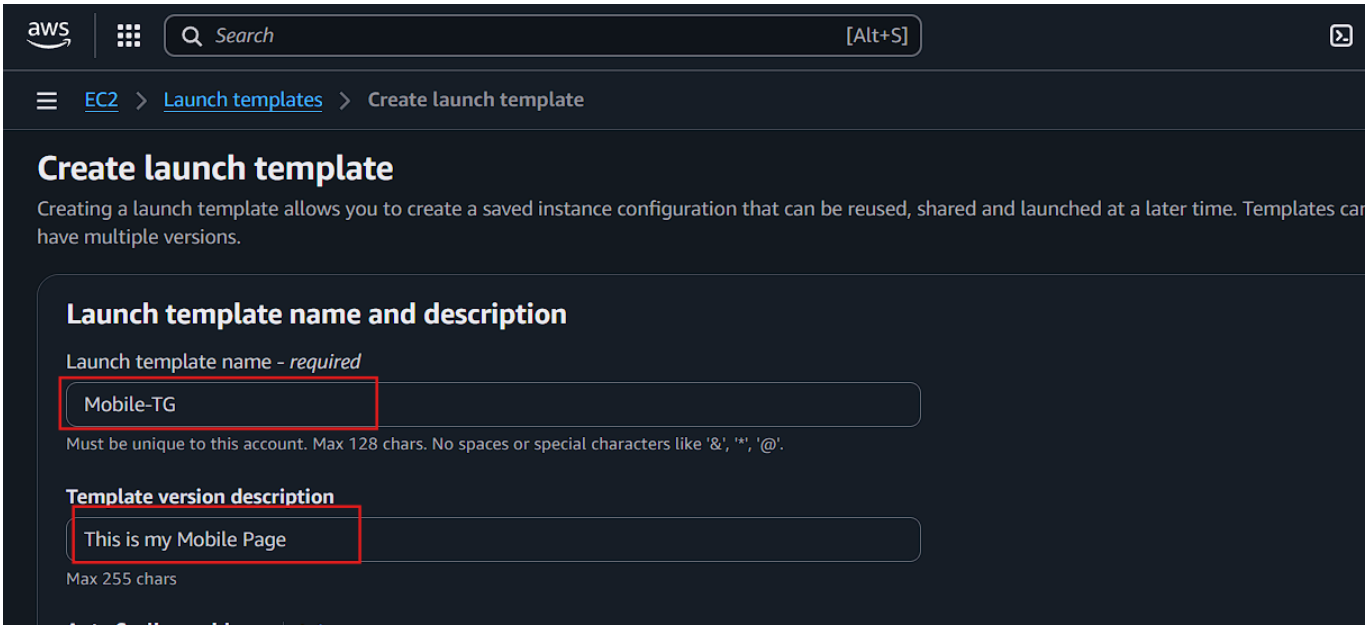


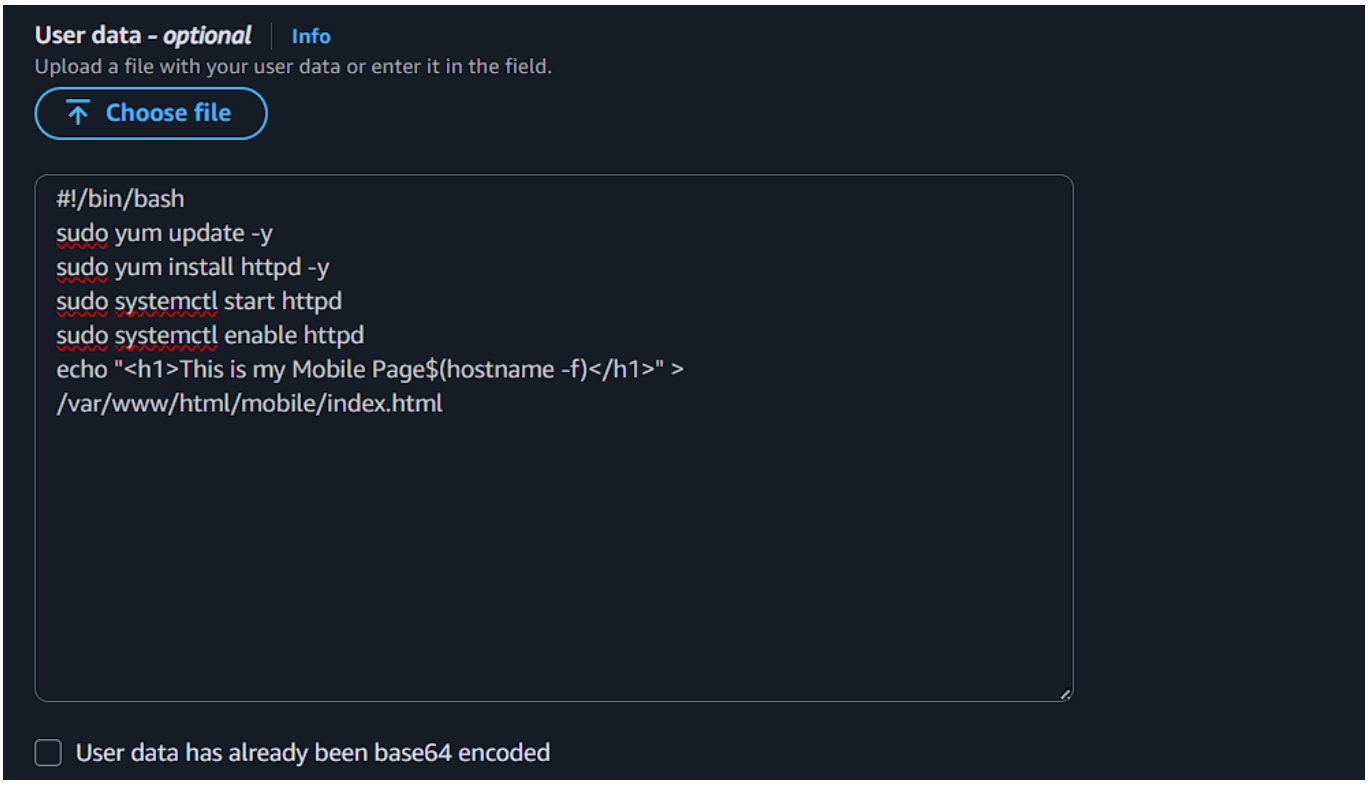
Launch Template (Laptop-TG).

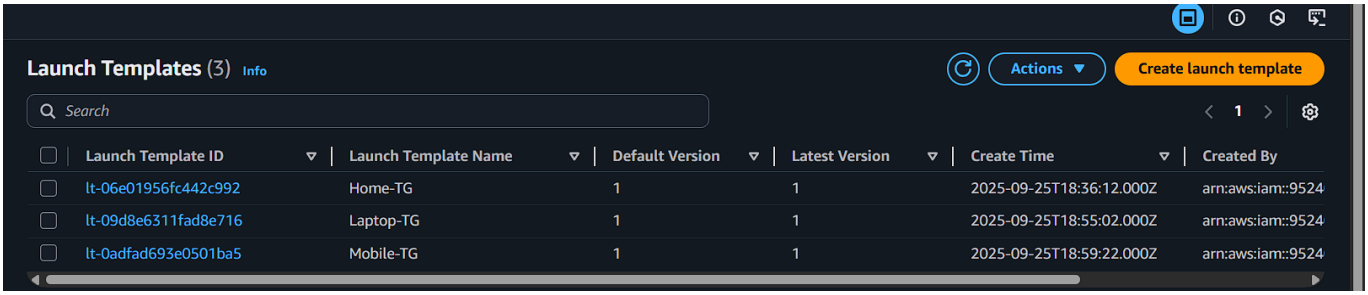




Launch Template(Mobile-TG).

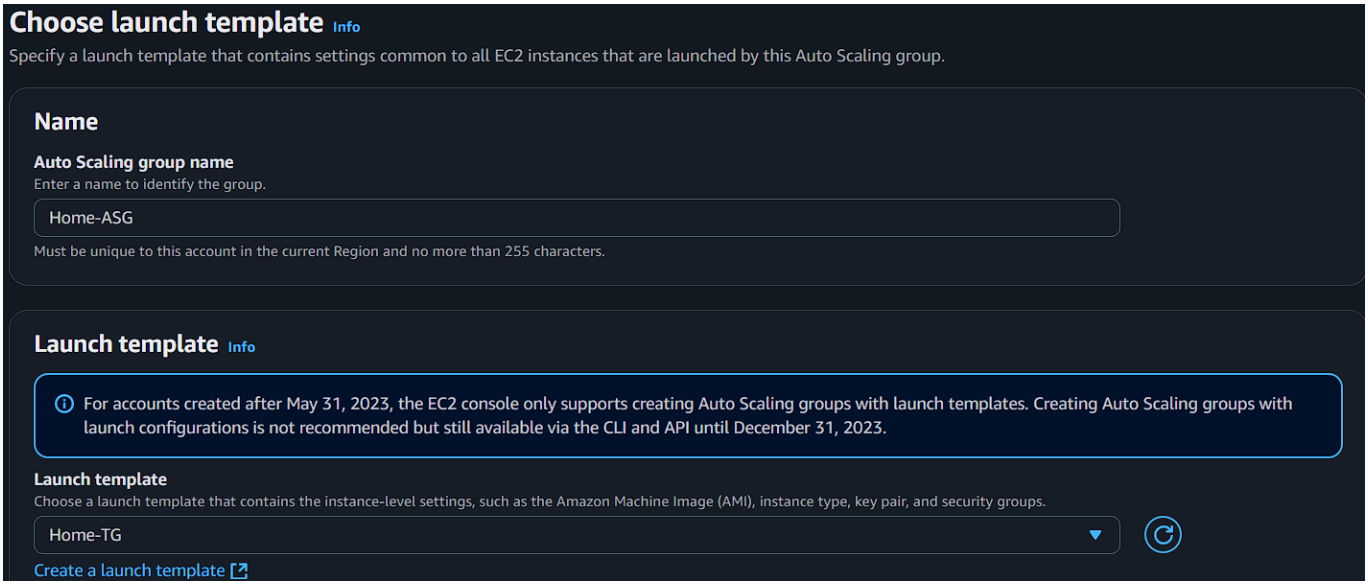


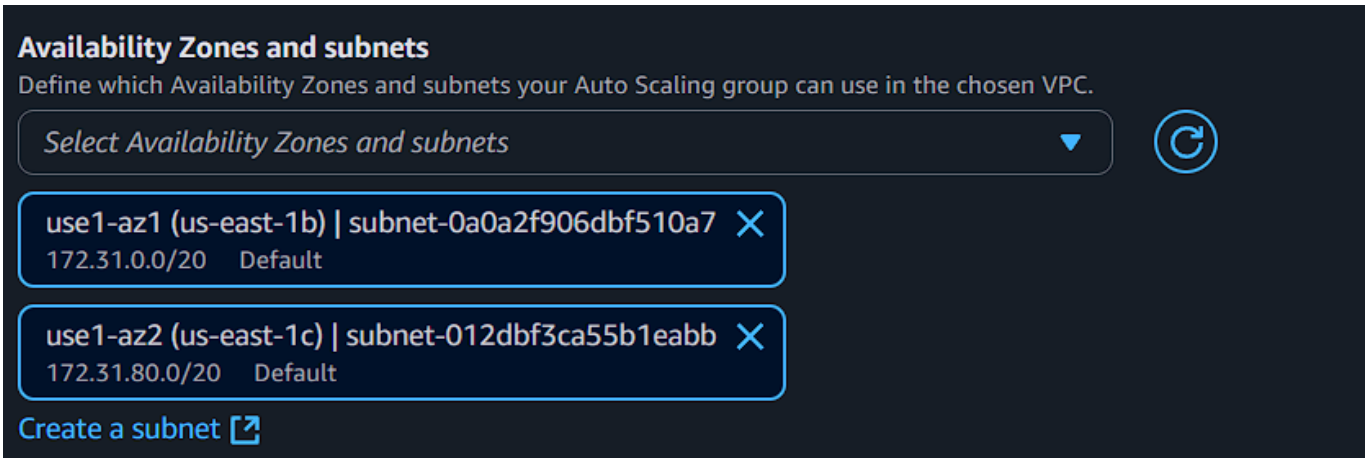


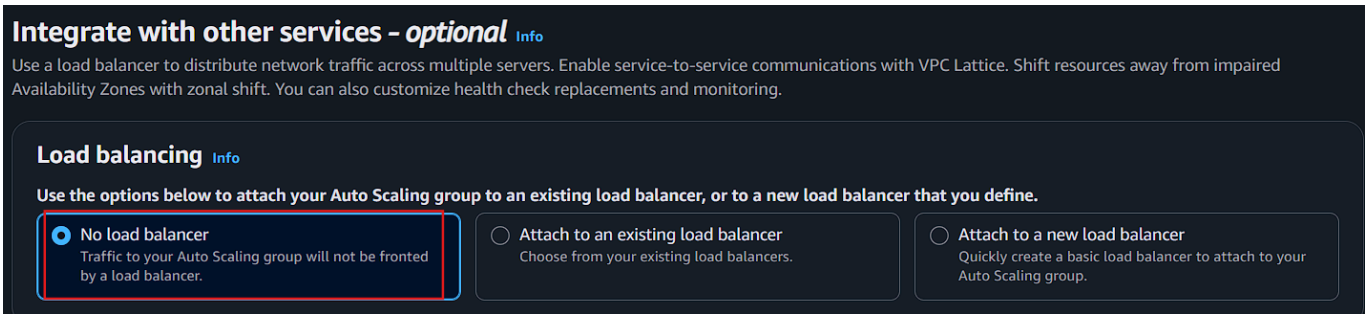


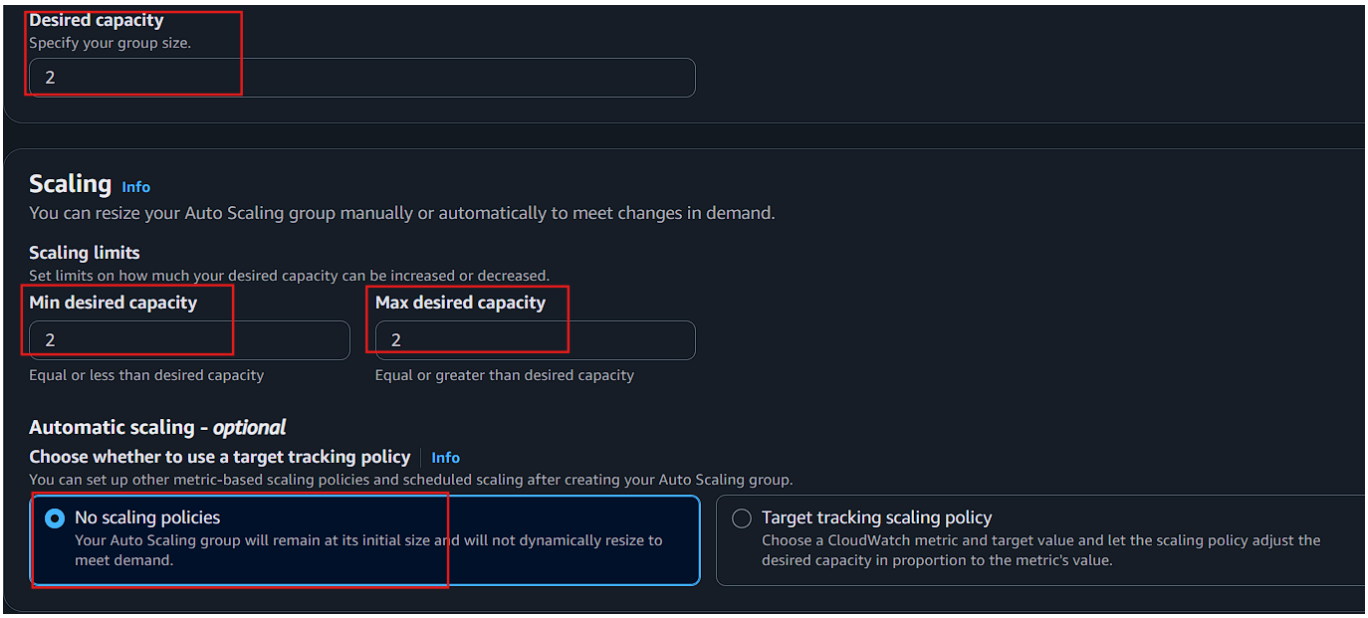
Step 2: Create Auto Scaling Groups (ASGs).

Home ASG: Use Static Scaling for a fixed number of instances.

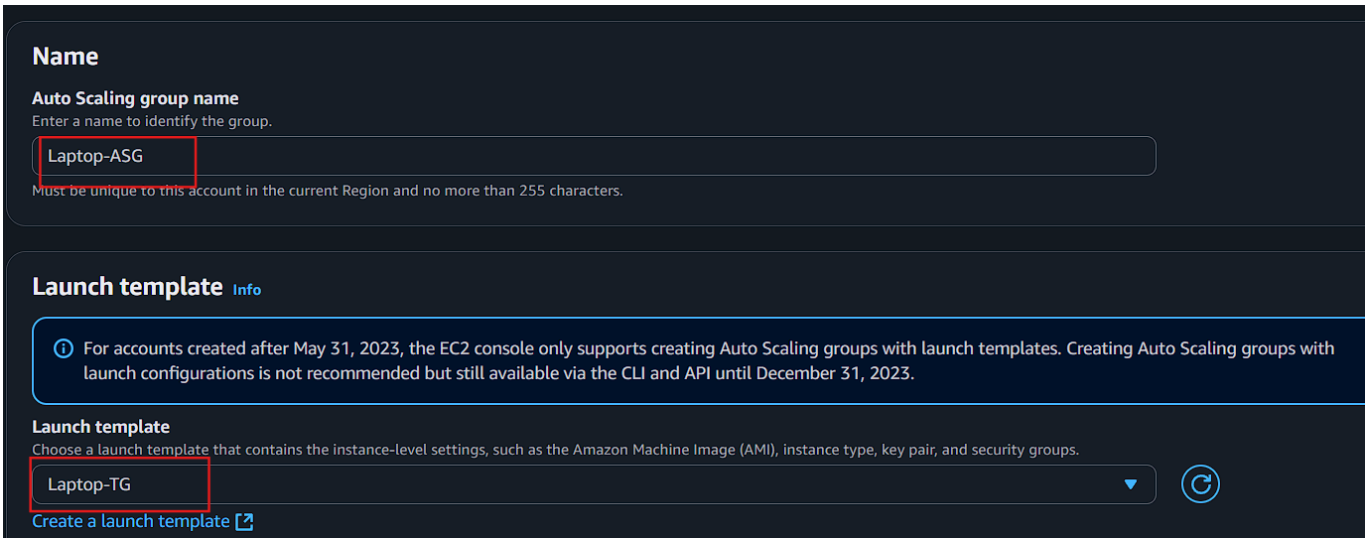


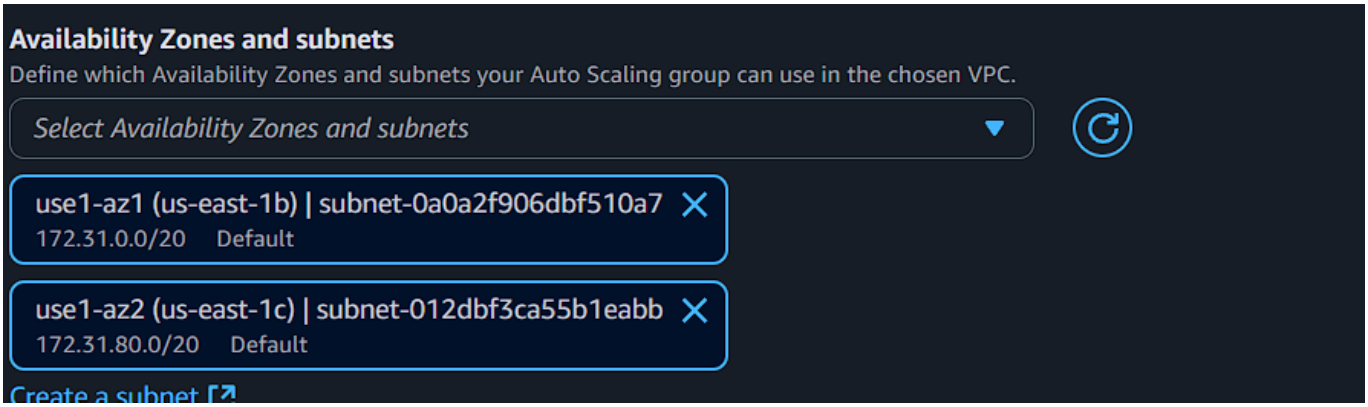


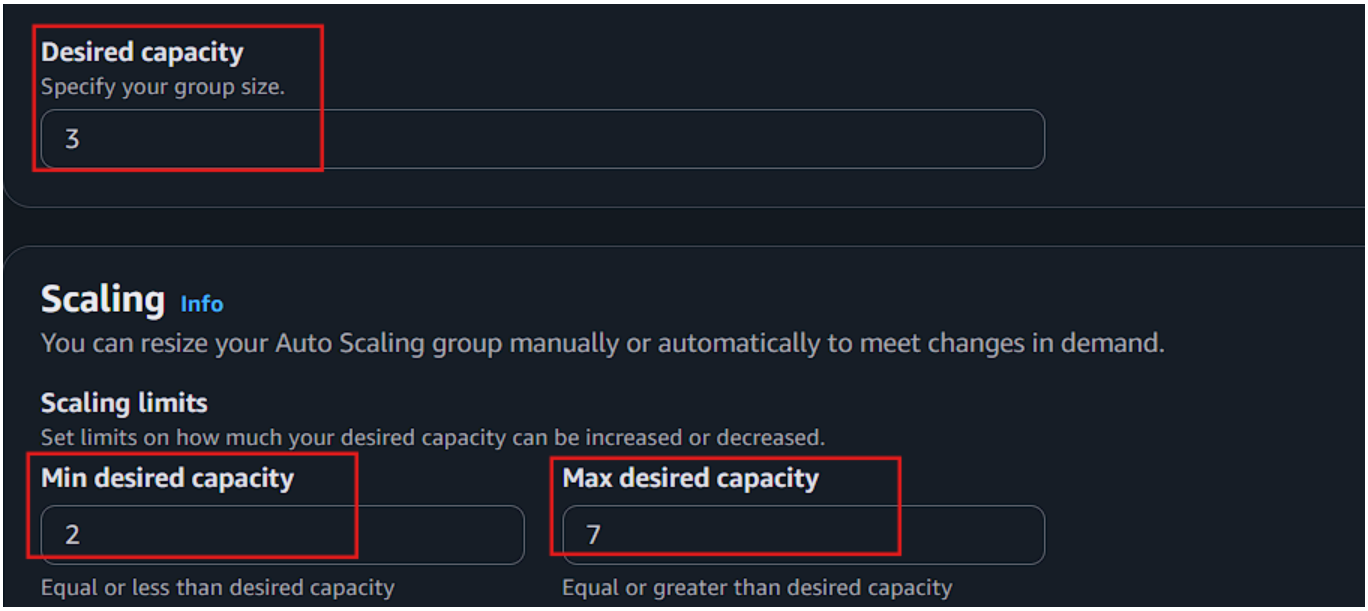


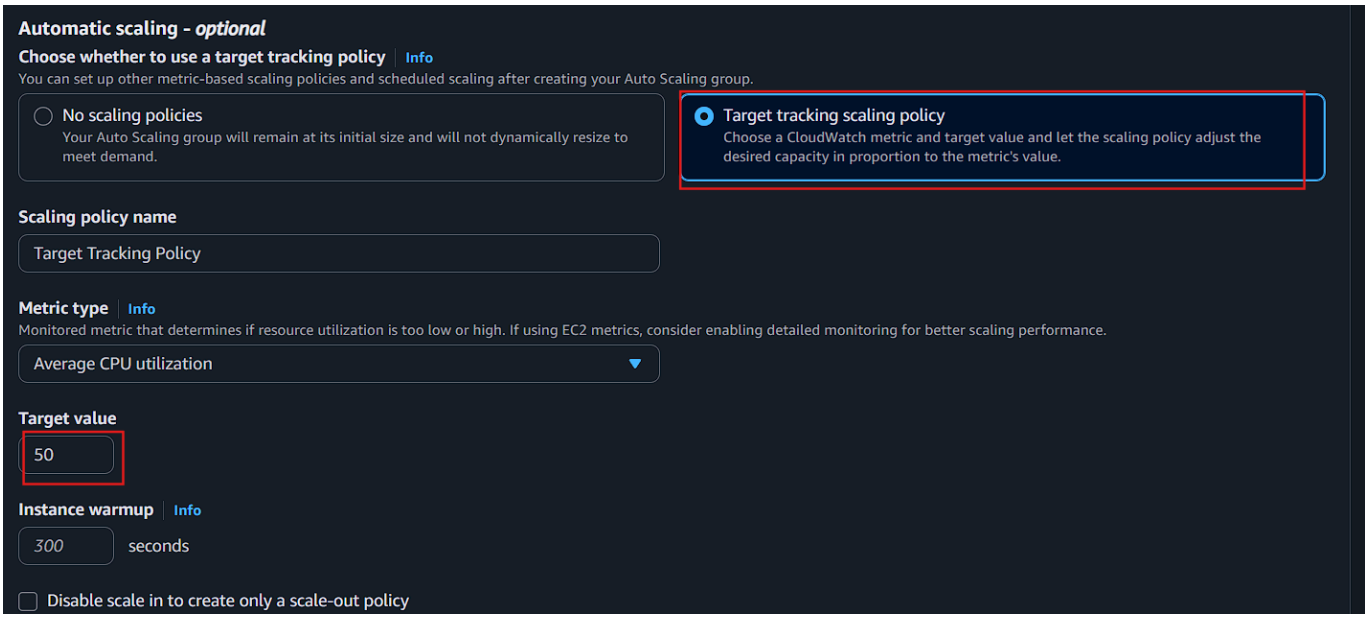


Laptop ASG: Implement Dynamic Scaling based on performance metrics.

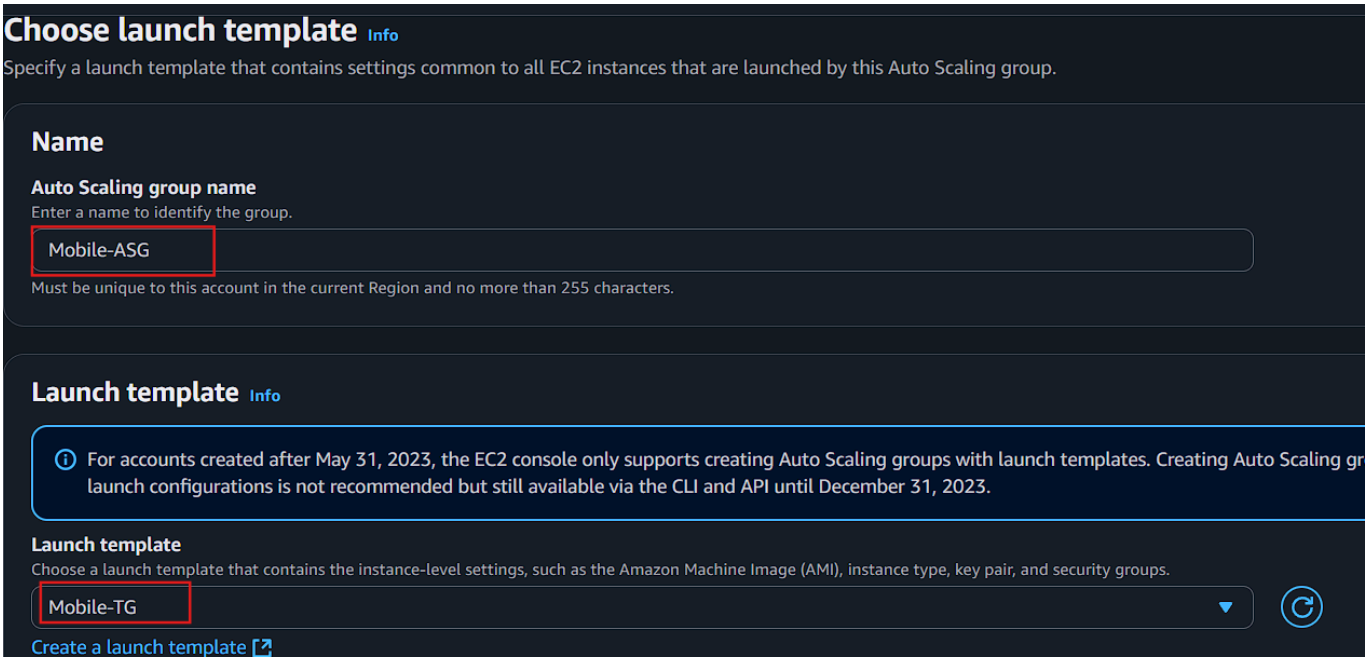

Name
Auto Scaling group name
Enter a name to identify the group.
Laptop-ASG
Must be unique to this account in the current Region and no more than 255 characters.
Launch template Info
For accounts created after May 31, 2023, the EC2 console only supports creating Auto Scaling groups with launch templates. 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.
Laptop-TG
Create a launch template

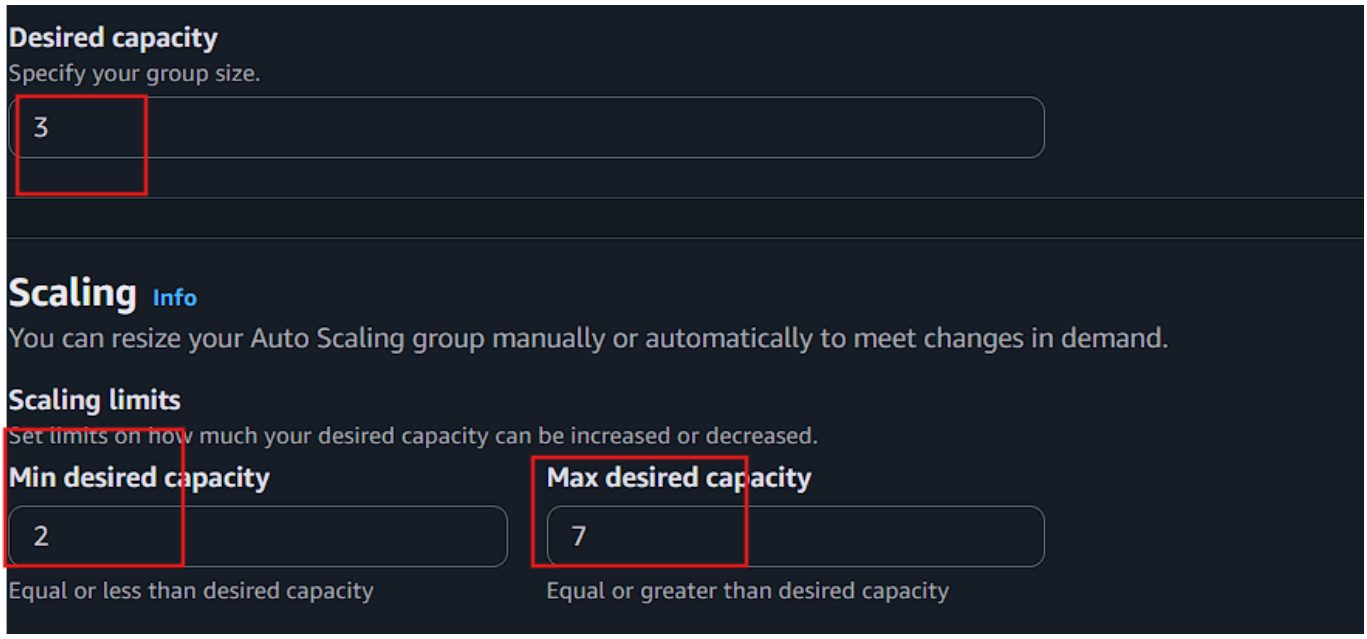

Availability Zones and subnets
Define which Availability Zones and subnets your Auto Scaling group can use in the chosen VPC.
Select Availability Zones and subnets
use1-az1 (us-east-1b) | subnet-0a0a2f906dbf510a7
172.31.0.0/20 Default
use1-az2 (us-east-1c) | subnet-012dbf3ca55b1eabb
172.31.80.0/20 Default
Create a subnet

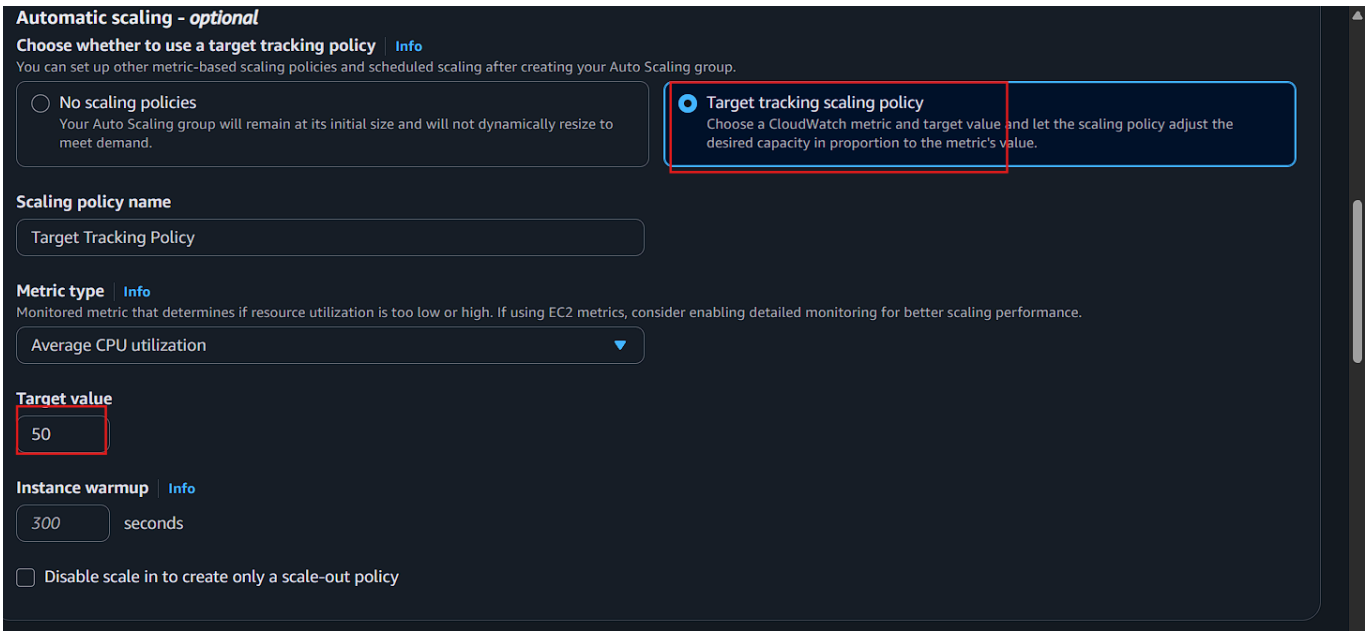

Desired capacity
Specify your group size.
3
Scaling Info
You can resize your Auto Scaling group manually or automatically to meet changes in demand.
Scaling limits
Set limits on how much your desired capacity can be increased or decreased.
Min desired capacity
2
Equal or less than desired capacity
Max desired capacity
7
Equal or greater than desired capacity

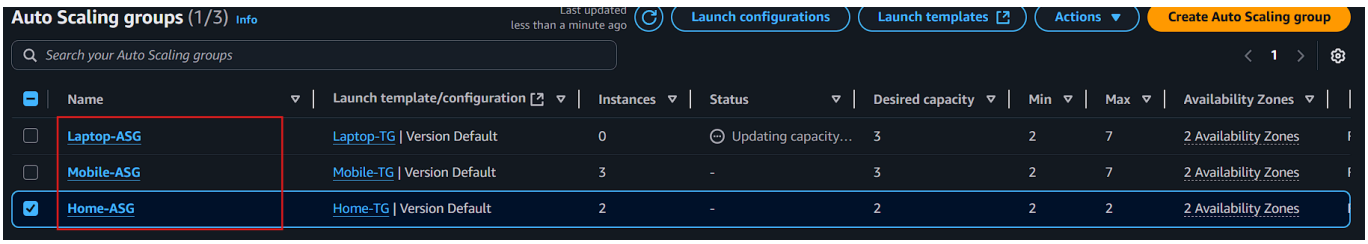


Mobile ASG: Apply Scheduled Scaling for predictable workloads.

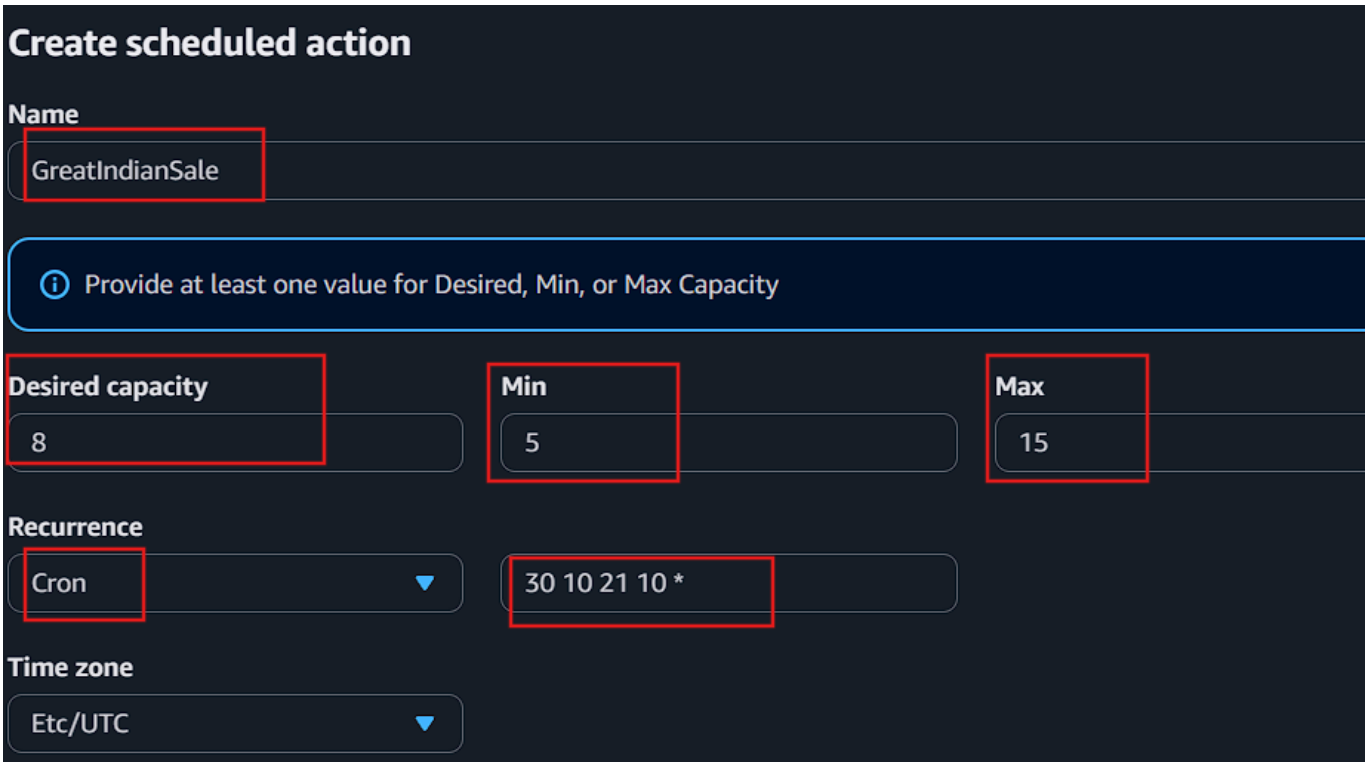


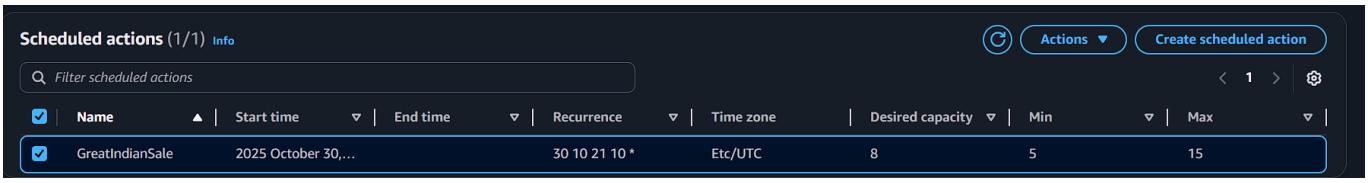


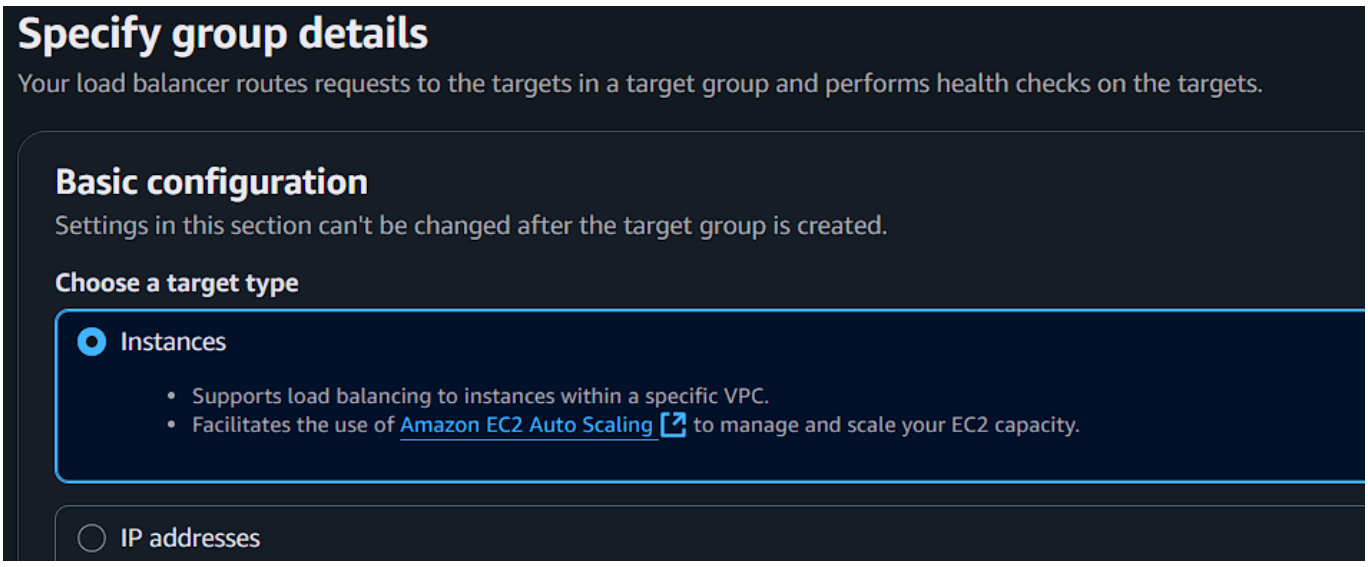




Step 3: Create scheduled action.



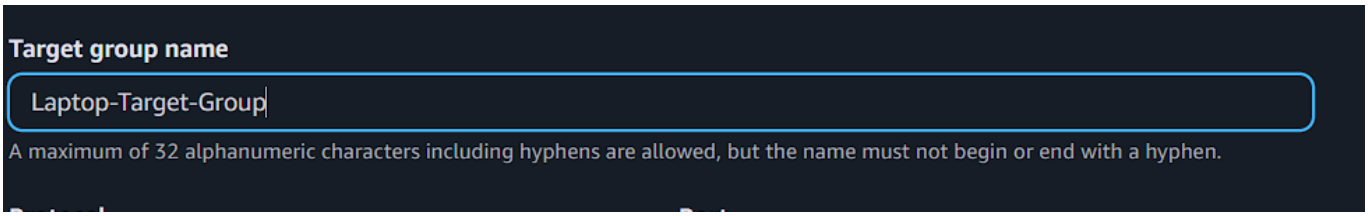


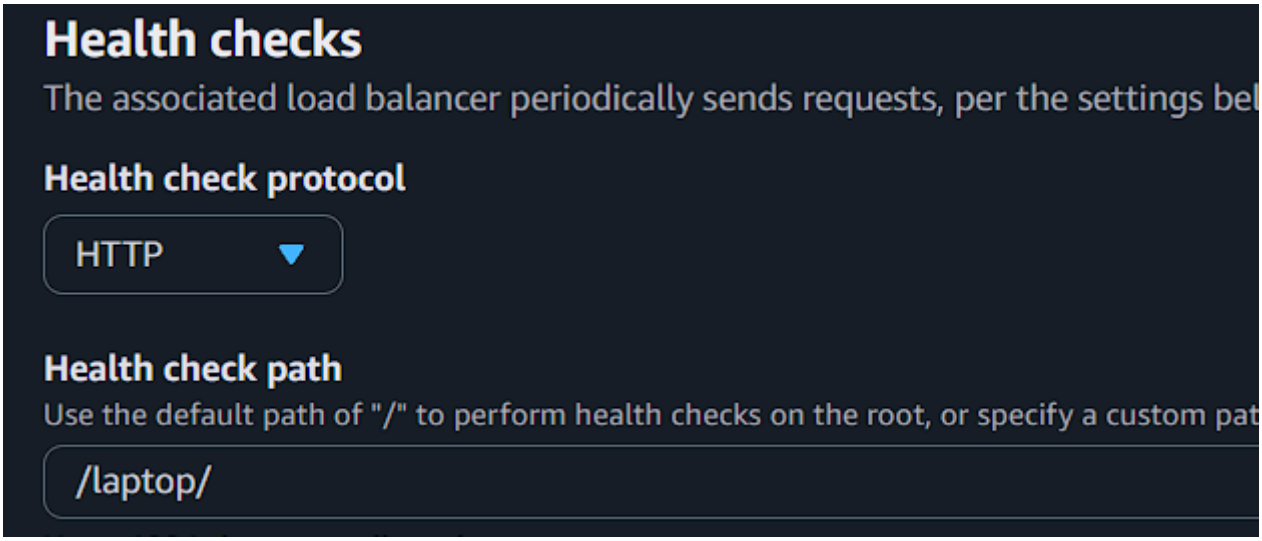


Step 4: Create Target Groups.

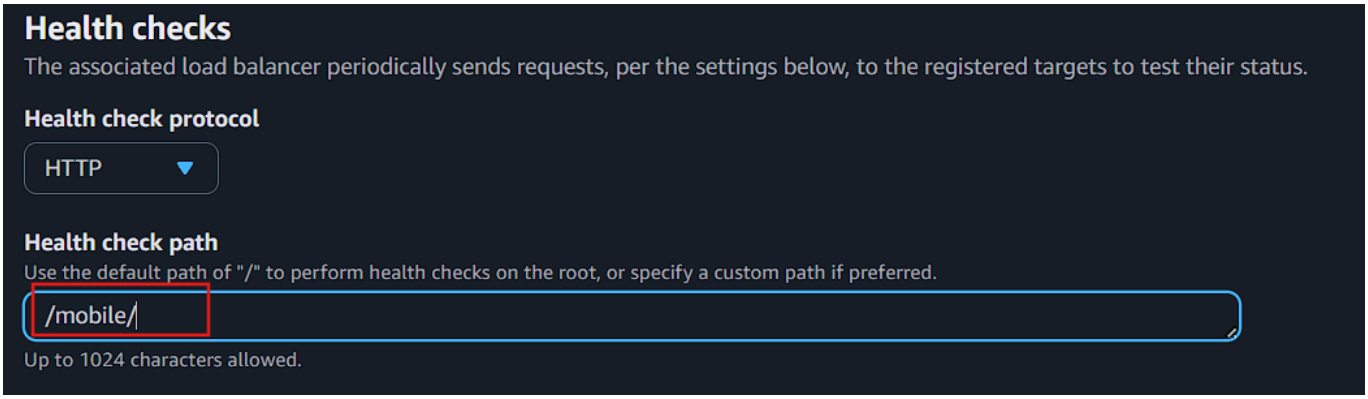
Home-Target-Group

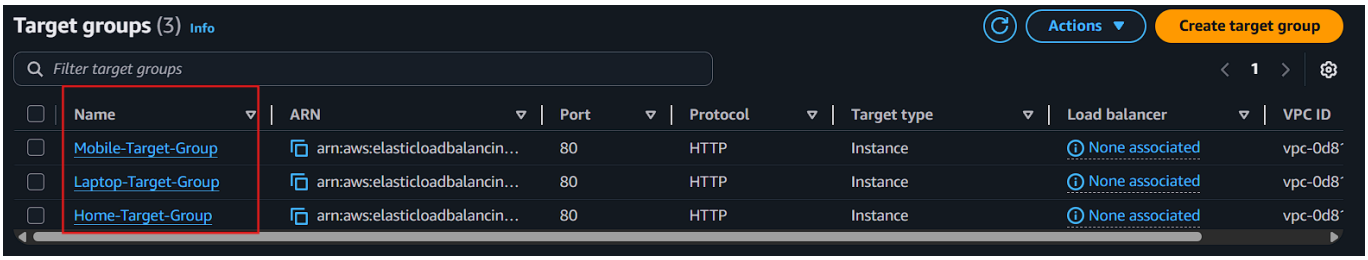
Laptop-Target-Group



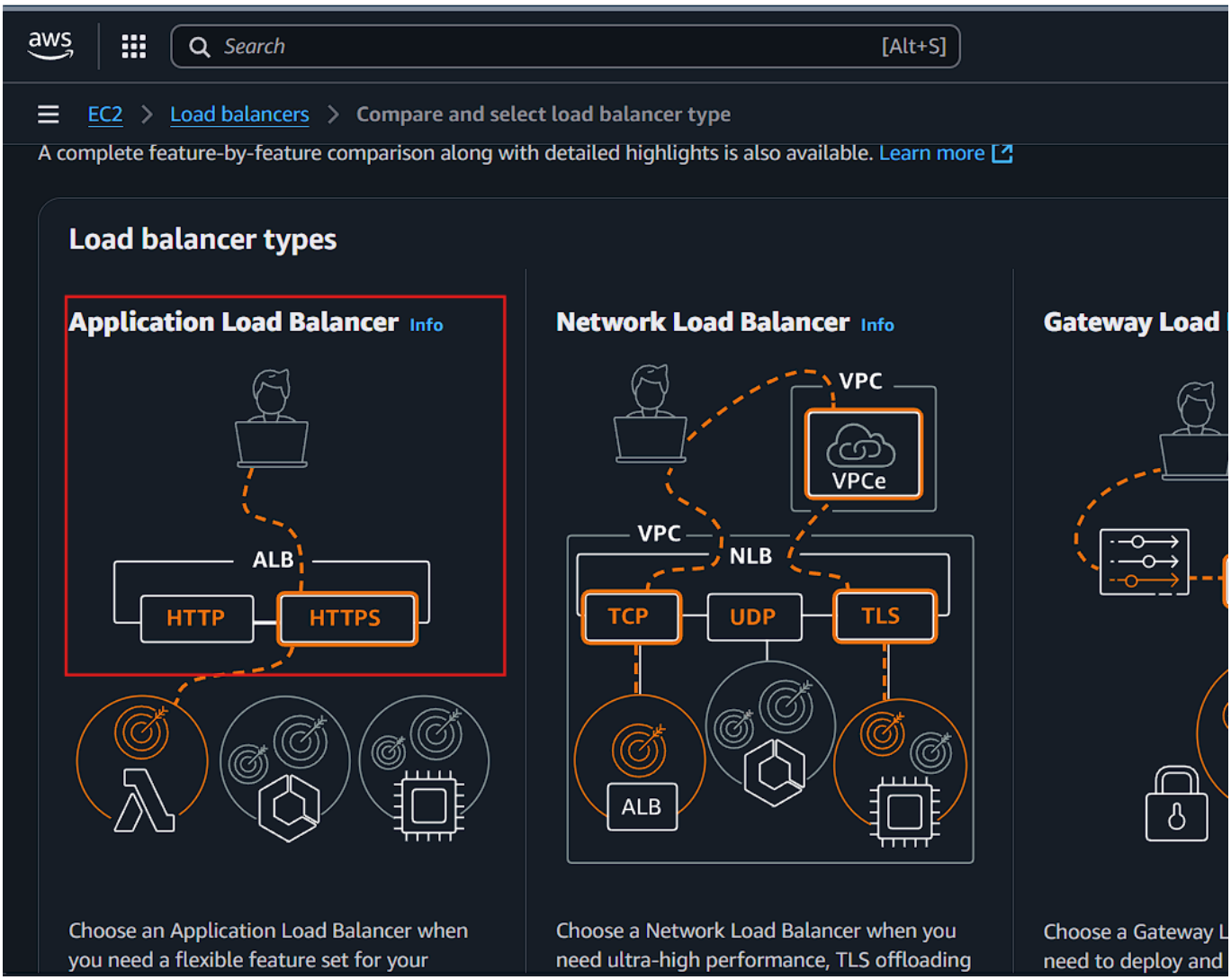


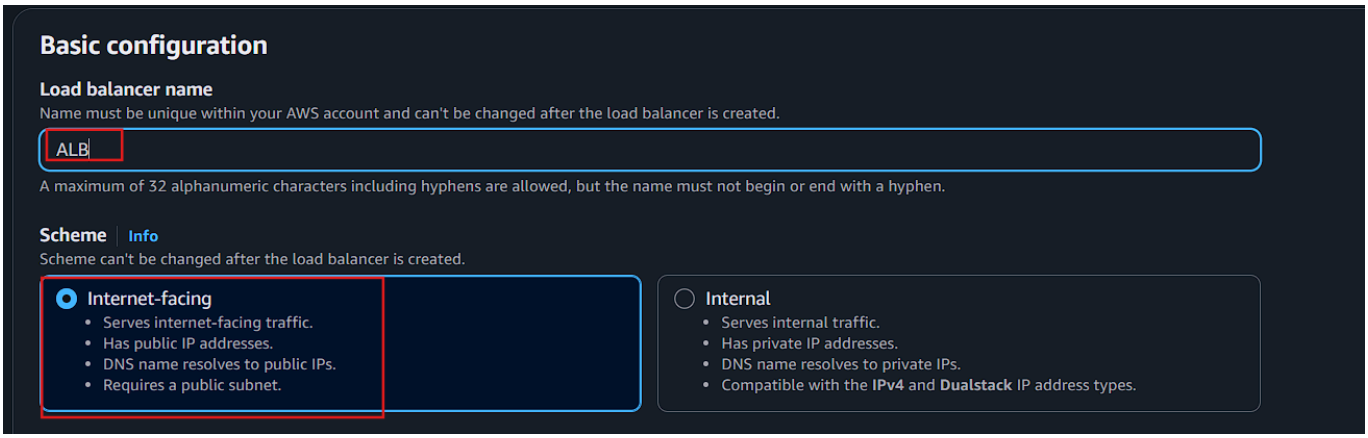
Mobile-Target-Group

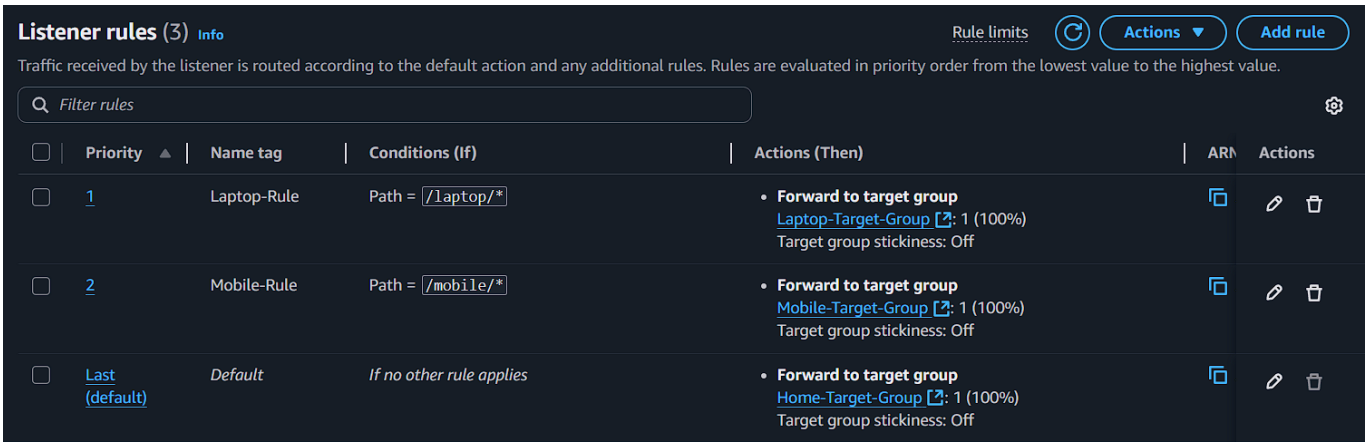


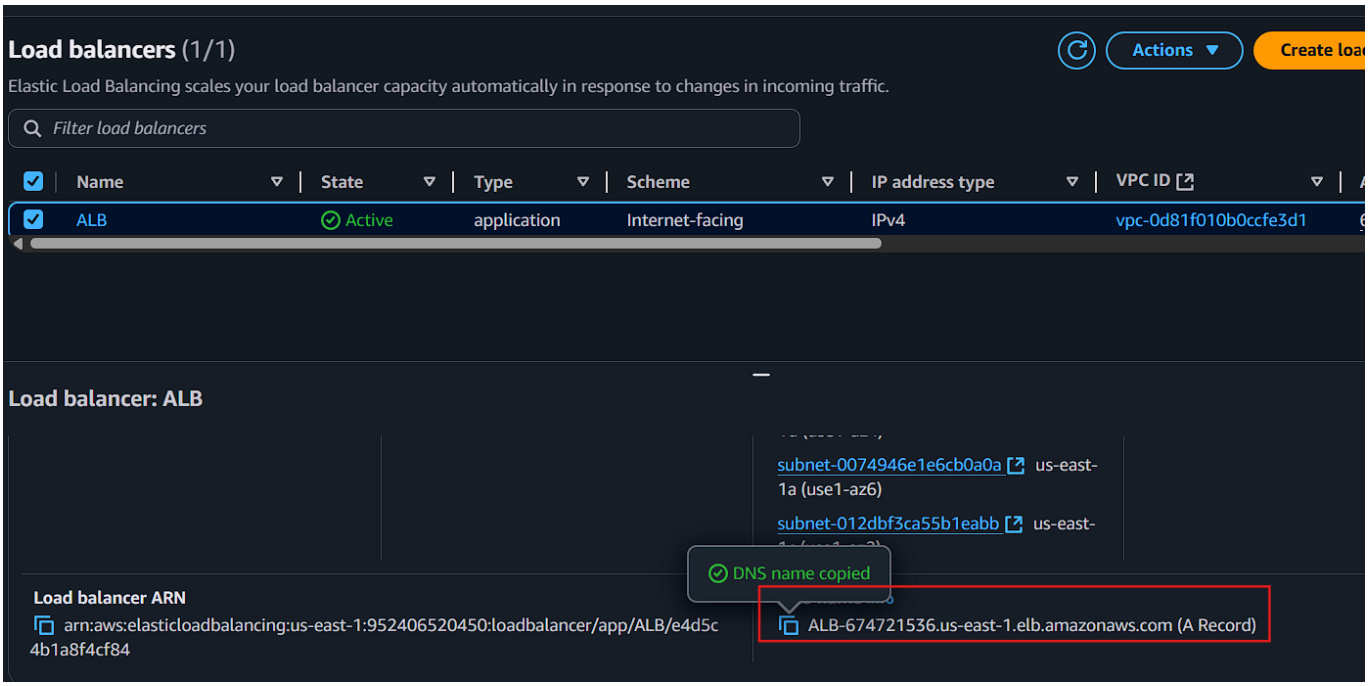


Step 5: Set Up Application Load Balancer (ALB).



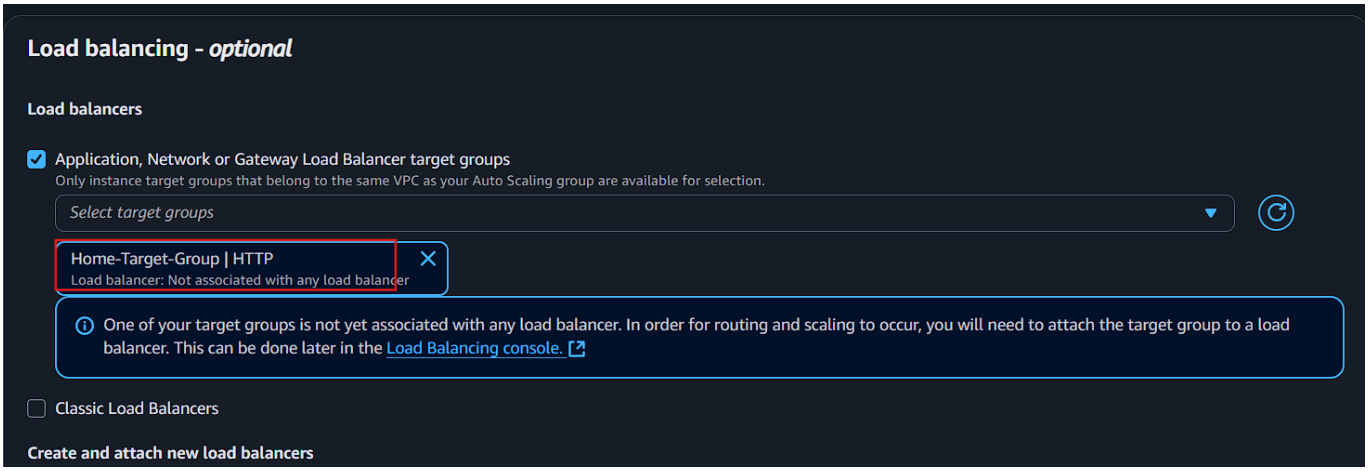


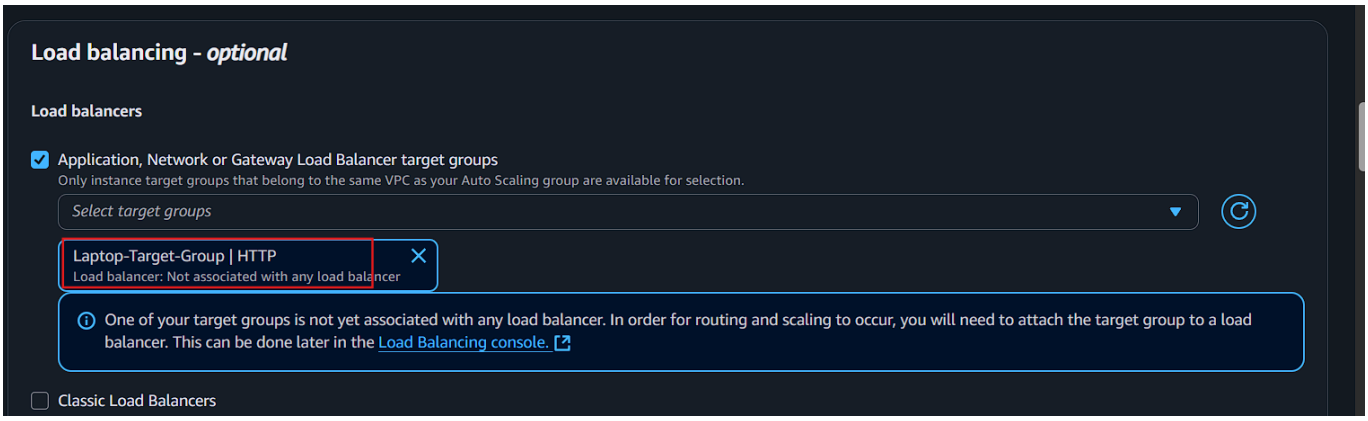


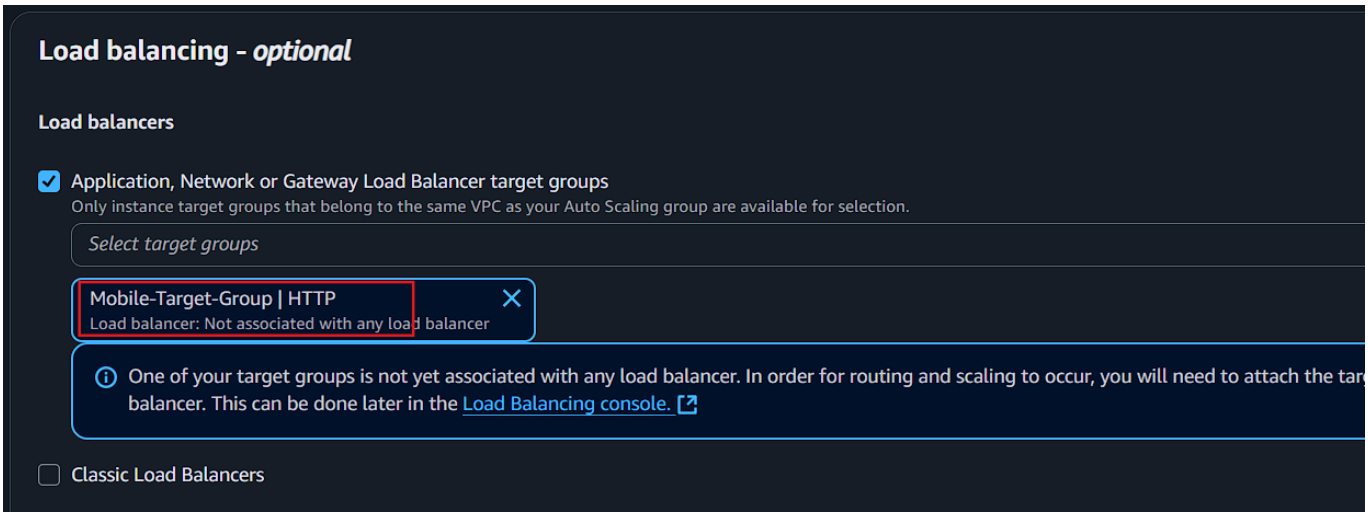


Step 6: Attach ALB to ASGs.

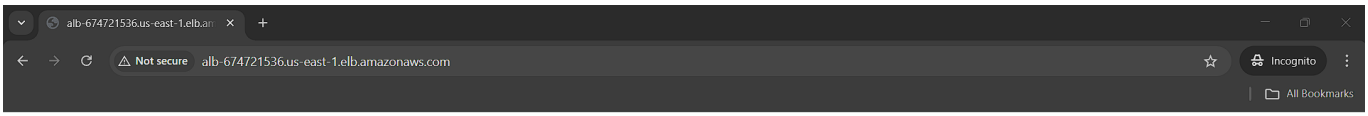
Connect each Auto Scaling Group with its corresponding target group.



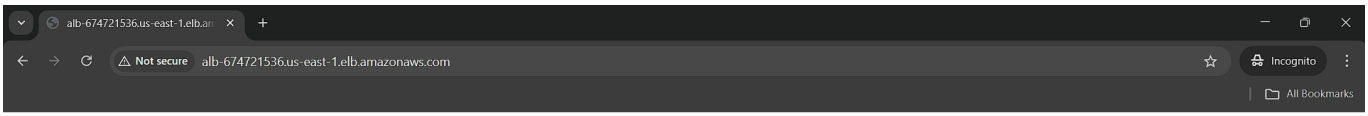




Step 7: Test and Verify.

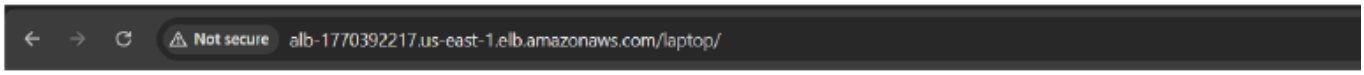


This is my Home Page ip-172-31-2-23.ec2.internal

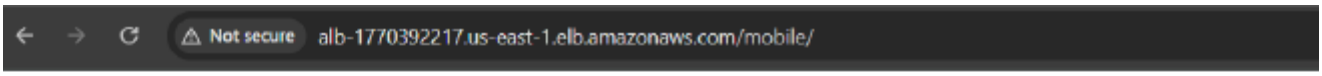


This is my Home Page ip-172-31-81-153.ec2.internal

this is laptop page ip-172-31-27-238.ec2.internal



this is laptop page ip-172-31-24-241.ec2.internal



this is mobile page ip-172-31-20-81.ec2.internal

this is mobile page ip-172-31-22-198.ec2.internal

Conclusion

This setup provides a flexible and robust cloud architecture combining the power of **AWS Auto Scaling** and **Application Load Balancer**. By using different scaling strategies for various workloads, organizations can achieve both performance optimization and cost efficiency across all user categories.