SYMBIOSIS INSTITUTE OF TECHNOLOGY, PUNE

Symbiosis International (Deemed University)

(Established under section 3 of the UGC Act, 1956)

Re-accredited by NAAC with 'A' grade (3.58/4) | Awarded Category – I by UGC

Founder: Prof. Dr. S. B. Mujumdar, M. Sc., Ph. D. (Awarded Padma Bhushan and Padma Shri by President of India)

Assignment No. 07

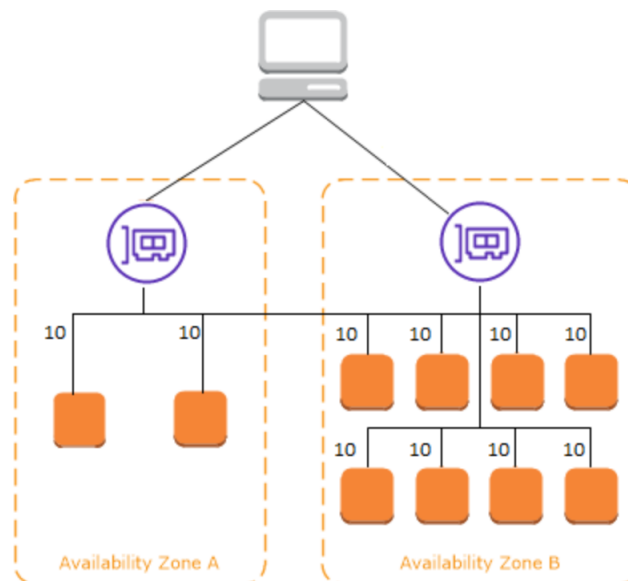|	
Subject:	Cloud Computing Tools and Techniques Lab
TCODE:	TE7949
Name of Student	**Pema Rinchen**
PRN No.	**22070122142**
Branch	CSE, Batch (2022-26)
Academic Year & Semester	2025-26
Date of Performance	**27/02/2025**
Title of Assignment:	To Implement Auto Scaling on a Cloud Platform like AWS/GCP/Microsoft Aure/ Open source cloud. Deploy multiple technology/multiple applications to see load distribution variation.
Objectives:	To Implement Auto Scaling on a Cloud Platform like AWS/GCP/Microsoft Aure/ Open source cloud. Deploy multiple technology/multiple applications to see load distribution variation.

Theory: *(Write about the assignment briefly etc.)*	Elastic Load Balancing automatically distributes your incoming traffic across multiple targets, such as EC2 instances, containers, and IP addresses, in one or more Availability Zones. It monitors the health of its registered targets, and routes traffic only to the healthy targets. Elastic Load Balancing scales your load balancer capacity automatically in response to changes in incoming traffic. A load balancer accepts incoming traffic from clients and routes requests to its registered targets (such as EC2 instances) in one or more Availability Zones. The load balancer also monitors the health of its registered targets and ensures that it routes traffic only to healthy targets. When the load balancer detects an unhealthy target, it stops routing traffic to that target. It then resumes routing traffic to that target when it detects that the target is healthy again. You configure your load balancer to accept incoming traffic by specifying one or more *listeners*. A listener is a process that checks for connection requests. It is configured with a protocol and port number for connections from clients to the load balancer. Likewise, it is configured with a protocol and port number for connections from the load balancer to the targets. **Availability Zones and load balancer nodes** When you enable an Availability Zone for your load balancer, Elastic Load Balancing creates a load balancer node in the Availability Zone. If you register targets in an Availability Zone but do not enable the Availability Zone, these registered targets do not receive traffic. Your load balancer is most effective when you ensure that each enabled Availability Zone has at least one registered target. We recommend enabling multiple Availability Zones for all load balancers. With an Application Load Balancer however, it is a requirement that you enable at least two or more Availability Zones. This configuration helps ensure that the load balancer can continue to

|	
	route traffic. If one Availability Zone becomes unavailable or has no healthy targets, the load balancer can route traffic to the healthy targets in another Availability Zone. **Load balancer benefits:** A load balancer distributes workloads across multiple compute resources, such as virtual servers. Using a load balancer increases the availability and fault tolerance of your applications. You can add and remove compute resources from your load balancer as your needs change, without disrupting the overall flow of requests to your applications. You can configure health checks, which monitor the health of the compute resources, so that the load balancer sends requests only to the healthy ones. You can also ofload the work of encryption and decryption to your load balancer so that your compute resources can focus on their main work.
Output with Steps (*Add text on work done with screenshots , code etc.)*	
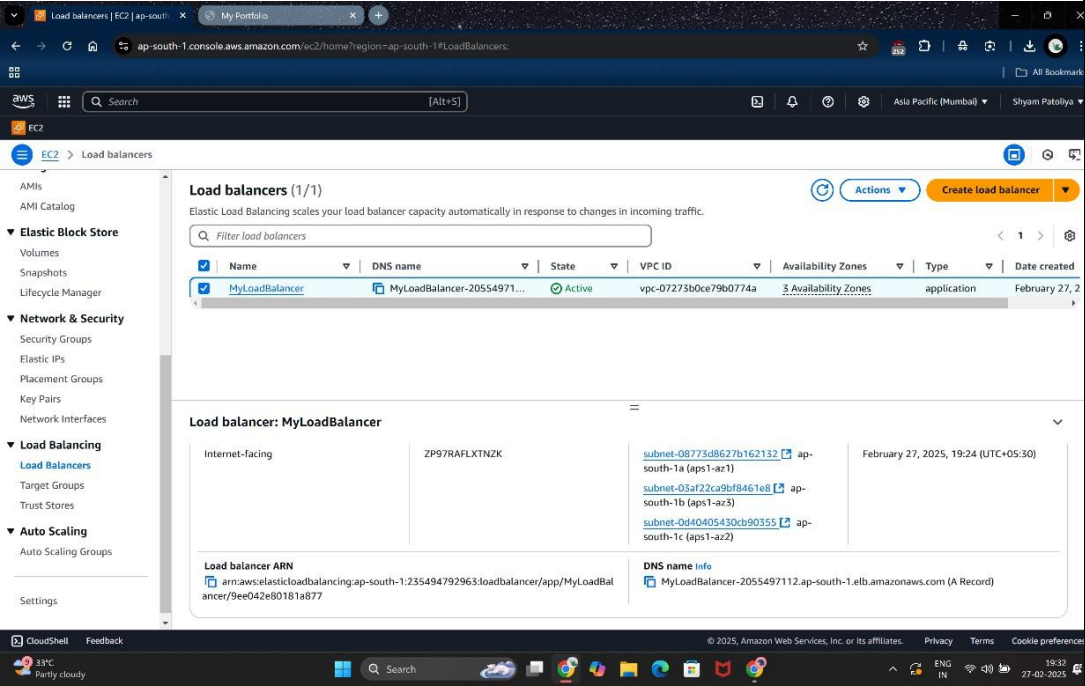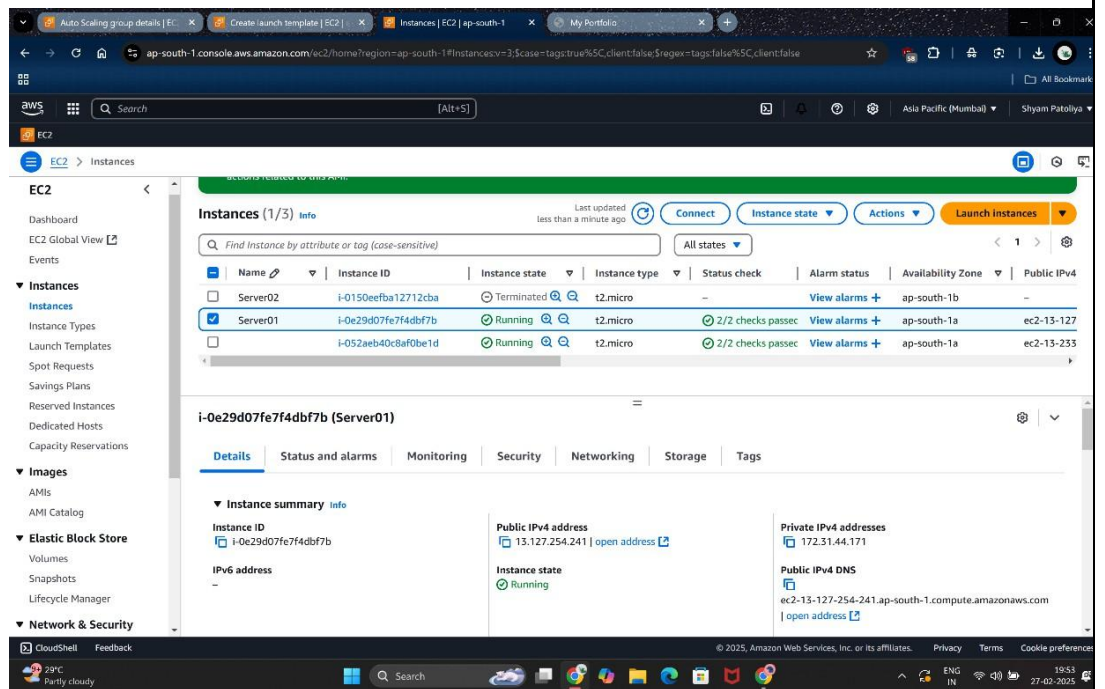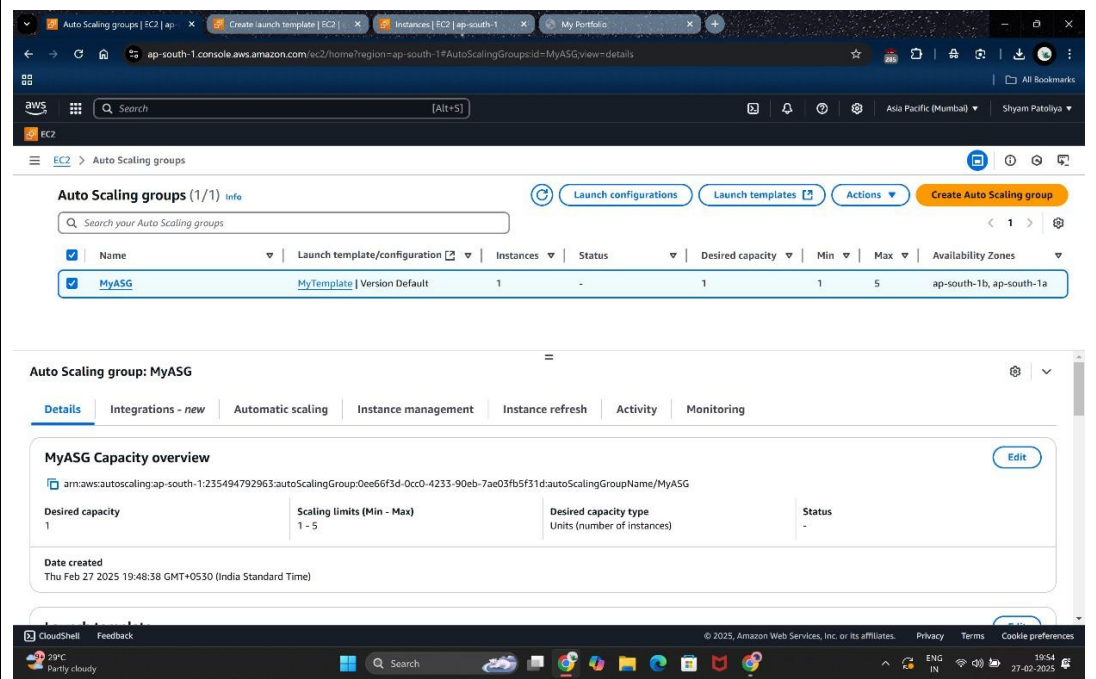
Conclusion *(Write your overall hands-on learning experience using Excel etc.)*	Implementing auto scaling on a cloud platform like AWS is a powerful strategy for managing application performance and resource utilization effectively. By leveraging Auto Scaling groups, you can dynamically adjust the number of instances based on real-time demand, ensuring that your applications remain responsive during peak loads while minimizing costs during low-traffic periods. Deploying multiple applications across different technologies allows you to gain valuable insights into load distribution and performance characteristics. This approach not only helps in understanding how various applications respond under different conditions but also aids in identifying potential bottlenecks and optimizing resource allocation.

Subject Faculty Name: Prof. Aditi Sharma