Auto Scaling and Elastic Load Balancing

- Auto Scaling and Elastic Load Balancing are features of AWS that you can use separately or together for elasticity and high availability.

AWS Auto Scaling

- Application scaling to optimize performance and costs
- Amazon EC2 Auto Scaling automates the process of launching (scaling out) and terminating (scaling in) Amazon EC2 instances based on the traffic demand for your application.
- Amazon EC2 Auto Scaling provides elasticity and scalability.

Amazon Elastic Load Balancing (ELB)

- Elastic Load Balancing (ELB) automatically distributes incoming application traffic across multiple targets and virtual appliances in one or more Availability Zones (AZs).
- ELB features high availability, automatic scaling, and robust security necessary to make your applications fault tolerant.

There are four types of Elastic Load Balancer (ELB) on AWS:

- Application Load Balancer (ALB) – layer 7 load balancer that routes connections based on the content of the request.
- Network Load Balancer (NLB) – layer 4 load balancer that routes connections based on IP protocol data.
- Classic Load Balancer (CLB) – this is the oldest of the three and provides basic load balancing at both layer 4 and layer 7 (not on the exam anymore).
- Gateway Load Balancer (GLB) – distributes connections to virtual appliances and scales them up or down (not on the exam).

Application Load Balancer (ALB)

ALB is best suited for load balancing of HTTP and HTTPS traffic and provides advanced request routing targeted at the delivery of modern application architectures, including microservices and containers.

Network Load Balancer (NLB)

NLB is best suited for load balancing of TCP traffic where extreme performance is required.

Network Load Balancer is also optimized to handle sudden and volatile traffic patterns.