SESSION 7

Elastic Load Balancer(ELB) and Auto Scaling Groups(ASG)

Scalability:

- It means that an application/ system can handle greater loads by adapting.
- There are two kinds:

Vertical Scalability: means changing the size of instance, t2.nano->t2.large, for databases, but there is a limit to hardware. It is scale up/down

Horizontal Scalability: Increasing the number of instances/systems for your applications. It is scale in/out using Auto scaling groups

High Availability

- It means running application/ system in at least 2 availability Zones.
- The goal is to survive a data centre disaster(power outage, earthquake)

OSI(Open Systems Interconnection) Model

It is a 7-layer architecture with each layer having specific functionality to perform. All these 7 layers work collaboratively to transmit the data from one person to another across the globe.

— Human-computer interaction layer, where applications can access the network services

— Ensures that data is in a usable format and is where data encryption occurs

— Maintains connections and is responsible for controlling ports and sessions

— Transmits data using transmission protocols including TCP and UDP

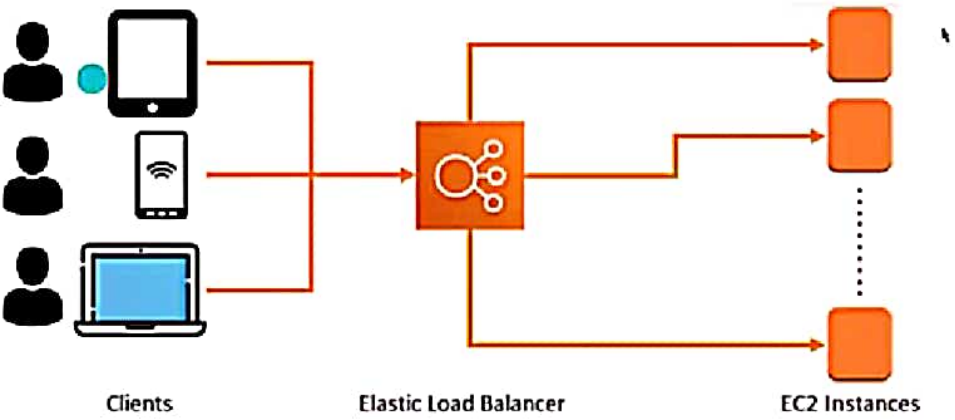
— Decides which physical path the data will take

— Defines the format of data on the network

— Transmits raw bit stream over the physical medium

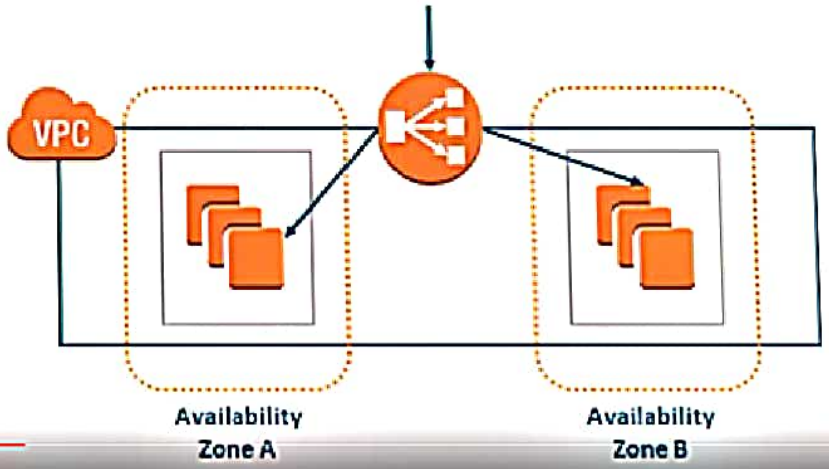
Elastic Load Balancer(ELB)

Elastic load balancer is a service provided by Amazon in which the incoming traffic is efficiently automatically distributed across a group of backend servers in a manner that increases speed and performance. It helps to improve scalability of your application and secures your applications. Load Balancer allows you to configure health checks for the registered targets. In case any of registered target fails the health check, the load balancer will not route traffic to that unhealthy target. Thereby ensuring your application is highly available and fault tolerant.



2. Network Load Balancer:

- It takes routing decisions in the Transport layer (TCP/SSL) of the OSI model (layer 4) it can handle millions of requests per second. Widely used to load balancing the TC traffic and it will also support elastic or static IP.
- It is capable of handling millions of requests per second while maintaing ultra low latency
- Optimised for sudden and volatile traffic patterns while using a single static IP address per AZ.



3. Application Load Balancer

- It is used when decisions are to be made related to HTTP and HTTPS traffic routing (Layer 7). It supports path-based routing and host-based routing. This load balancer works at the Application layer of the OSI Model. The load balancer also supports dynamic host port mapping.

4. Gateway Load Balancer:

Gateway Load Balancers provides you the facility to deploy, scale, and manage virtual appliances like firewall. Gateway Load Balancers combines a transparent network gateway and then distributes the traffic.

Auto Scaling Groups

- It is used to manage Amazon EC2 capacity automatically
- Maintain right number of instances for your application
- operate healthy groups of instance and scale it accordingly to your needs.
- It replace unhealthy instances
- Are cost effective by running at optimal capacity
- If min=1, and no instance is running, it will launch a new instance automatically to meet the requirements.
- If we terminate a running instance and no instance is running, the health status changes from healthy to unhealthy, so it will launch an instance and is back to healthy.
- In real life load on your website and application can change. The goal of ASG is
  Scale in (remove EC2 instances) to match decrease load.
  Scale out(add EC2 instances) to match increase load.
  It automatically registers new instances to a load balancer.

Questions

1.What is the main purpose of High Availability in the Cloud?
2. Which AWS offered Load Balancer should you use to handle hundreds of thousands of connections with low latency?
3.Changing an EC2 Instance Type from a t3a.medium to a t3a.2xlarge is an example of?
4.What can you use to handle quickly and automatically the changing load on your websites and applications by adding compute resources?
5. Which Load Balancer is best suited for HTTP/HTTPS load balancing traffic?

1.Applications running even in disasters 2.NLB 3. Vertical Scaling 4. ASG 5. ALB