AWS Load Balancers

Load balancing is the method of distributing network traffic equally across a pool of resources that support an application. Modern applications must process millions of users simultaneously and return the correct text, videos, images, and other data to each user in a fast and reliable manner.

To handle such high volumes of traffic, most applications have many resource servers with duplicate data between them.

A load balancer is a device that sits between the user and the server group and acts as an invisible facilitator, ensuring that all resource servers are used equally.

sticky sessions:

Sticky sessions — also known as session persistence — is the method that makes it possible for the load balancer to identify requests coming from the same client and to always send those requests to the same server. In sticky sessions, all user information is stored on the server side, and this method is commonly used in stateful services. This functionality is primarily available for HTTP load balancers.

Benefits:

Load balancing directs and controls internet traffic between the application servers and their visitors or clients. As a result, it improves an application's availability, scalability, security, and performance.

Application availability

Server failure or maintenance can increase application downtime, making your application unavailable to visitors. Load balancers increase the fault tolerance of your systems by automatically detecting server problems and redirecting client traffic to available servers. You can use load balancing to make these tasks easier:

- Run application server maintenance or upgrades without application downtime
- Provide automatic disaster recovery to backup sites

- Perform health checks and prevent issues that can cause downtime

Application scalability

You can use load balancers to direct network traffic intelligently among multiple servers. Your applications can handle thousands of client requests because load balancing does the following:

- Prevents traffic bottlenecks at any one server
- Predicts application traffic so that you can add or remove different servers, if needed
- Adds redundancy to your system so that you can scale with confidence

Application security

Load balancers come with built-in security features to add another layer of security to your internet applications. They are a useful tool to deal with distributed denial of service attacks, in which attackers flood an application server with millions of concurrent requests that cause server failure. Load balancers can also do the following:

- Monitor traffic and block malicious content
- Automatically redirect attack traffic to multiple backend servers to minimize impact
- Route traffic through a group of network firewalls for additional security

Application performance

Load balancers improve application performance by increasing response time and reducing network latency. They perform several critical tasks such as the following:

- Distribute the load evenly between servers to improve application performance
- Redirect client requests to a geographically closer server to reduce latency
- Ensure the reliability and performance of physical and virtual computing resources

Note: In Host-based routing, incoming traffic is routed on the basis of the domain name or host name given in the Host Header.

AWS Load Balancer Types

There are four AWS load balancer types supported:

- AWS Classic Load Balancer
- AWS Network Load Balancer (NLB)
- AWS Application Load Balancer (ALB)
- AWS Gateway Load Balancer (GLB)

Classic Load Balancers

Classic Load Balancers distribute upcoming traffic to different EC2 instances in multiple Availability Zones. During this process, there is a chance of the fault tolerance of your application. These Load Balancers detect healthy and unhealthy instance and direct the traffic towards only healthy ones.

It also helps in a way such that without disrupting the flow of requests to your application you can add or remove instances from your load balancers as your need changes.

AWS ELB can calculate the majority of workloads automatically. Protocol and port which a person configures are used to detect the connection requests from clients it also forwards requests to one or more registered instances.

The number of instances can be modified. Health checks can be monitored so that the load balancer only sends requests to the healthy instances.

- Note: This balancer does not support host-based routing and results in low efficiency of resources.

Application Load Balancer

- It is responsible for performing the tasks in the application layer of the OSI model and is the advanced form of load balancing.
- It is used when there are HTTP and HTTPS traffic routing.
- It supports host-based and path-based routing.

Network Load Balancer

- It performs the task at layer 4 of the connection level in the OSI model.
- Its primary purpose is to route TCP traffic.
- It can manage a massive amount of traffic and is suitable for managing low latencies.

Connection Draining (Deregistration Delay)

- By default, if a registered EC2 instance with the ELB is deregistered or becomes unhealthy, the load balancer immediately closes the connection
- Connection draining can help the load balancer to complete the in-flight requests made while keeping the existing connections open, and preventing any new requests from being sent to the instances that are de-registering or unhealthy.