WHAT IS AWS ELASTIC LOAD BALANCING?

AWS Elastic Load Balancing allows users to route incoming traffic between multiple EC2 instances, elastic compute cloud servers, containers and IP addresses as appropriate. The distribution of traffic/workloads within a single or between multiple Availability Zones takes place automatically, allowing users to scale the compute capabilities while maintaining high availability as the application usage demand grows.

When the load balancer receives a request from an end-user to access an application, it routes the traffic based on the health of the target instances. It continuously monitors the health status and user requests are always directed to healthy instances. In case an unhealthy instance is detected, the load balancer automatically routes all traffic to other healthy targets as available. The requests are routed back to the server when it returns to a healthy state.

With the Elastic Load balancing capability, users configure the protocol and port of a Listener, a process that identifies connection requests between clients and the load balancer, as well as the load balancer and the instance targets. The Listener follows the predefined rules and policies to route the traffic between the client and the backend instances. These rules are evaluated based on configurable priority, conditions and actions as described on the AWS resource here.

Types of Elastic Load Balancing

The ELB service started in 2009 with the software-based load balancing capability to route traffic, conduct health checks of instances and respond accordingly to maximize availability and fault tolerance of AWS-hosted apps. That load balancing service is now the Classic Load Balancer (CLB), and AWS has since added two new enhanced services to the load balancing portfolio: Application Load Balancer (ALB) and Network Load Balancer (NLB).

Classic Load Balancer

The Classic Load Balancer is primarily developed to deliver balancing services for the EC2 instance network at the Level 4 of the OSI model. Most web applications use the TCP/IP protocol at the Level 4 Transport layer, while also using UDP protocol in some cases. However, AWS load balancing services currently do not support UDP. The Classic Load Balancer uses the information from the protocols and port numbers from incoming request to route the traffic to appropriate AWS EC2 instances hosting the Web application. The process is similar to traditional traffic routing for load balancing purposes using physical devices, except that the CLB performs this task efficiently and automatically within a virtual environment.

Most of the Classic Load Balancing features are offered both with the Application and Network Load Balancers. These include:

- High availability distribution, including automated scaling in response to changing traffic capacity requirements.
- Health checks of the target EC2 instances. The unhealthy targets are automatically off-loaded until the server regains optimal performance capability.
- Security groups for additional networking and security management, including the ability to create internal CLB.
- Centralized management of SSL certificate including encryption of target EC2 servers, controlling ciphers and protocols and offloading SSL decryption.
- Sticky Sessions to route users to same target instances using cookies. This service is not available for Application Load Balancer.
- Operational monitoring via CloudWatch metrics reporting to trigger actions based on events associated with the network performance and traffic requirements.
- Support for both IPv4 and IPv6 for applications relying on TCP and HTTP/HTTPS respectively.
- Log recording of load balancer requests for diagnostic and analyses.

Application Load Balancer

The ALB allows load balancing for HTTP/HTTPS traffic at Layer 7 of the OSI model and can route the traffic to modern application architectures that include containers, IP addresses, EC2 servers, Lambda functions and microservices. The ALB offers the following feature enhancements over the CLB:

- Path-based routing based on the information specified in the HTTP header. The Listener can be configured to route different URL content to appropriate application service.
- Host-Based routing allows the configuration of the Listener to route multiple URL domains from the same ALB.
- Redirect requests between different URLs, return custom HTTP response or route traffic based on IP addresses.
- Support for AWS Lambda functions and containerized applications to maximize cluster resource utilization.
- Dynamic auto-scaling of independent application services attached to specific target instances. Each application service can also be monitored independently.
- Security capabilities such as user authentication before routing the traffic.
- Log data includes additional information. Overall performance of ALB is also better than that of

the CLB.

Network Load Balancer

Load balancing with the NLB works at Layer 4, the Transport Layer. Network Load Balancer is capable of handling millions of routing requests per second between clients and target systems using IP addresses, TCP and port numbers. Although Network Load Balancer doesn't offer Path and Host based routing, and Sticky Sessions capabilities of the Application Load Balancer, it does provide the following advantages over the AWS Classic Load Balancer:

- A robust load balancing solution for volatile workloads.
- Traffic routing to multiple apps on a single target instance or group.
- Traffic routing to target groups outside the Virtual Private Cloud using IP addresses.
- Supports both static and elastic IP addresses.
- Manageable security using Tag-based Identity and Access Management (AWS IAM) permissions.
- Efficient use of clusters with support for Elastic Container Service (AWS ECS).
- Supports health checks for each service independently, within containers or target EC2 instances. The NLB also balances services that scale dynamically in response to varying demand as part of the AWS Auto Scaling capability.

The choice between Classic, Application and Network Load Balancer largely depends on the infrastructure environment, costs, security, and how the traffic must be handled between end-users and target groups. For most of the general use cases where the traffic is handled using IP addresses and TCP ports, CLB may be an appropriate option, especially when the route mapping between the two end-points is direct. For environments handling complex rules for traffic routing at the application level, the ALB would be an appropriate option. For workloads that require extreme performance and routing via static and elastic IP address, or when the source IP addresses must be preserved, the AWS Network Load Balancer would be the most appropriate option.