



## Amazon Elastic Load Balancer (ELB)

**Load balancing** is the process of distributing incoming network traffic across multiple servers or resources to ensure no single server becomes overwhelmed. This helps improve the availability and reliability of applications by balancing the load across multiple servers, optimizing resource use, maximizing throughput, minimizing response time, and avoiding overload on any single resource.

### 1. Application Load Balancer (ALB)

#### Definition:

The Application Load Balancer operates at the **application layer (Layer 7)** of the OSI model. It intelligently routes requests based on HTTP/HTTPS protocols, considering various aspects of the request such as URL path, host headers, HTTP method, query string, and more.

#### Based on what – routing happens

- **Round Robin** (default for HTTP/HTTPS): Distributes incoming requests evenly across all available targets in the target group. When weighted routing is required, Route 53 DNS-level routing is preferred.
- **Least Outstanding Requests** (optional setting for HTTP/HTTPS): Routes each new request to the target with the least number of outstanding (pending) requests.

### 2. Network Load Balancer (NLB)

#### Definition:

The Network Load Balancer operates at the **transport layer (Layer 4)** of the OSI model. It routes connections based on IP protocol data, making it capable of handling millions of requests per second with ultra-low latency.

#### Based on what – routing happens

**Flow Hash Routing:** Distributes traffic based on a hash of the source IP address, source port, destination IP address, destination port, and protocol, ensuring that connections from a client are routed to the same target.



## Amazon Elastic Load Balancer (ELB)

### 3. Gateway Load Balancer (GWLB)

#### Definition:

The Gateway Load Balancer operates at **Layer 3 (Network layer)** and combines a transparent network gateway (L3) with load balancing capabilities. It routes traffic through virtual appliances, such as firewalls, IDS/IPS, and other network appliances, deployed on AWS.

#### Based on what – routing happens

**Flow Hash Routing:** Similar to NLB, it routes traffic based on a hash of the packet, which maintains consistent routing paths for individual flows.

### 4. Classic Load Balancer (CLB)


The **Classic Load Balancer (CLB)** is the original Elastic Load Balancer (ELB) offered by AWS and operates at both the **application layer (Layer 7)** and the **transport layer (Layer 4)** of the OSI model. While it's considered a legacy service, it's still available for use in AWS, although AWS recommends using either the **Application Load Balancer (ALB)** or **Network Load Balancer (NLB)** for new applications due to their enhanced features and performance capabilities.



## Amazon Elastic Load Balancer (ELB)

### Intro Parameters

- Regional Service
- Supports Multi AZ
- Spread load across multiple EC2 instances
- Separate public traffic from private traffic
- Health checks allow ELB to know which instances are working properly (done on a port and a route, /health is common)
- **Does not support weighted routing**
- If no targets are associated with the target groups ⇒ **503 Service Unavailable**
- Targets are unreachable (possibly due to NACL or SG rules) ⇒ **504 Timeout Error**
- Using ALB & NLB, instances in peered VPCs can be used as targets using IP addresses.

 **Weighted routing** is a traffic management technique used in DNS (Domain Name System) to distribute network traffic across multiple resources (such as servers, data centers, or cloud regions) based on assigned weights. This allows you to control the proportion of requests that are directed to each resource, effectively distributing traffic according to your desired configuration.

While **Elastic Load Balancing (ELB)** provides powerful mechanisms for distributing traffic and ensuring high availability, it does not support weighted routing. For scenarios where weighted traffic distribution is required, **AWS Route 53** is the appropriate service to use.



## **Amazon Elastic Load Balancer (ELB)**

### **Classic Load Balancer (CLB) - deprecated**

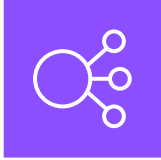
- **Load Balancing to a single application**
- Supports HTTP, HTTPS (layer 7) & TCP (layer 4)
- Health checks are HTTP or TCP based
- Provides a fixed hostname ([xxx.region.elb.amazonaws.com](https://xxx.region.elb.amazonaws.com))



## Amazon Elastic Load Balancer (ELB)

### Application Load Balancer (ALB)

- **Load balancing to multiple applications** (target groups) based on the request parameters
- Operates at Layer 7 (HTTP, HTTPS and WebSocket)
- Provides a fixed hostname ([xxx.region.elb.amazonaws.com](https://xxx.region.elb.amazonaws.com))
- ALB terminates the upstream connection and creates a new downstream connection to the targets
- **Security Groups can be attached to ALBs** to filter requests
- Great for micro services & container-based applications (Docker & ECS)
- Client info is passed in the request headers
  - Client IP ⇒ X-Forwarded-For
  - Client Port ⇒ X-Forwarded-Port
  - Protocol ⇒ X-Forwarded-Proto
- Target Groups
  - Health checks are done at the target group level
  - Target Groups could be
    - EC2 instances - HTTP
    - ECS tasks - HTTP
    - Lambda functions - HTTP request is translated into a JSON event
    - Private IP Addresses
- **Listener Rules** can be configured to route traffic to different target groups based on:
  - Path ([example.com/users](https://example.com/users) & [example.com/posts](https://example.com/posts))
  - Hostname ([one.example.com](https://one.example.com) & [other.example.com](https://other.example.com))
  - Query String ([example.com/users?id=123&order=false](https://example.com/users?id=123&order=false))
  - Request Headers
  - Source IP address



## **Amazon Elastic Load Balancer (ELB)**

- Important Points
  - When the target type is IP, you can specify private IP addresses only.
  - If you specify targets using an instance ID, traffic is routed to instances using the primary private IP address specified in the primary network interface for the instance.
  - If you specify targets using IP addresses, you can route traffic to an instance using any private IP address from one or more network interfaces. This enables multiple applications on an instance to use the same port.



## Amazon Elastic Load Balancer (ELB)

### Network Load Balancer (NLB)

- Operates at Layer 4 (TCP, UDP)
- Can handle millions of request per seconds (extreme performance)
- **Lower latency** ~ 100 ms (vs 400 ms for ALB)
- **1 static public IP per AZ** (vs a static hostname for CLB & ALB)
- **Elastic IP can be assigned to NLB** (helpful for whitelisting specific IP)
- Maintains the same connection (TCP or UDP) from client all the way to the target application
- **No security groups can be attached to NLBs.** Since they operate on layer 4, they cannot see the data available at layer 7. They just forward the incoming traffic to the right target group as if those requests were directly coming from client. So, the **target instances must allow TCP traffic on port 80 from anywhere.**
- Within a target group, NLB can send traffic to
  - **EC2 instances**
    - If you specify targets using an instance ID, traffic is routed to instances using the **primary private IP address**
  - **IP addresses**
    - Used when you want to balance load for a physical server having a static IP.
  - **Application Load Balancer (ALB)**
    - Used when you want a static IP provided by an NLB but also want to use the features provided by ALB at the application layer.



## Amazon Elastic Load Balancer (ELB)

### Gateway Load Balancer (GWLB)

- Operates at layer 3 (Network layer) - IP Protocol
- Used to route requests to a fleet of 3rd party virtual appliances like Firewalls, Intrusion Detection and Prevention Systems (IDPS), etc.
- Performs two functions:
  - **Transparent Network Gateway** (single entry/exit for all traffic)
  - Load Balancer (distributes traffic to virtual appliances)
- Uses GENEVE protocol
- Target groups for GWLB could be
  - EC2 instances
  - IP addresses





## Amazon Elastic Load Balancer (ELB)

### Sticky Sessions (Session Affinity)

- Requests coming from a client is always redirected to the same instance based on a cookie. After the cookie expires, the requests coming from the same user might be redirected to another instance.
- **Only supported by CLB & ALB** because the cookie can be seen at layer 7
- Used to ensure the user doesn't lose his session data, like login or cart info, while navigating between web pages.
- **Stickiness may cause load imbalance**
- Cookies could be
  - **Application-based** (TTL defined by the application)
  - **Load Balancer generated** (TTL defined by the load balancer)
- ELB reserved cookie names (should not be used)
  - AWSALB
  - AWSALBAPP
  - AWSALBTG



## Amazon Elastic Load Balancer (ELB)

### Cross-zone Load Balancing

- Allows ELBs in different AZ containing unbalanced number of instances to distribute the traffic evenly across all instances in all the AZ registered under a load balancer.
- Supported Load Balancers
  - Classic Load Balancer
    - Disabled by default
    - No charges for inter AZ data
  - Application Load Balancer
    - Always on (can't be disabled)
    - No charges for inter AZ data
  - Network Load Balancer
    - Disabled by default
    - Charges for inter AZ data
- **Server Name Indication (SNI)**
  - SNI allows us to load multiple SSL certificates on one Load Balancer to serve multiple websites securely
  - **Only works for ALB & NLB** (CLB only supports one SSL certificate)
  - Newer protocol, not every client supports it yet
  - **Supported in CloudFront** also



## **Amazon Elastic Load Balancer (ELB)**

AWS ELB Scenario based Questions

<https://lisireddy.medium.com/aws-elb-scenario-based-questions-eebda4ac6dd9>