**Load Balancer in System Design**

A **Load Balancer** is a component that distributes incoming traffic across multiple servers to ensure reliability, efficiency, and scalability.

**Types of Load Balancing**

1️⃣ **Layer 4 Load Balancing (Transport Layer - TCP/UDP Based)**

* Routes traffic based on IP address and port (without inspecting actual data).
* Example: Amazon ELB (Elastic Load Balancer) using TCP-based routing.
* **When to Use:**
  + For handling high volumes of simple TCP/UDP traffic.
  + For applications where session persistence isn’t required (e.g., gaming, VoIP).

2️⃣ **Layer 7 Load Balancing (Application Layer - HTTP/HTTPS Based)**

* Routes traffic based on application content (URL, headers, cookies).
* Example: Nginx, HAProxy, AWS Application Load Balancer.
* **When to Use:**
  + For intelligent routing (e.g., sending /api requests to API servers).
  + When session persistence, SSL termination, or caching is needed.

**Load Balancer Example: Amazon (AWS)**

**1️⃣ Initial Phase (Small Scale)**

* Amazon’s early e-commerce website ran on **a single web server**.
* The server handled user requests, payment processing, and database queries.
* As traffic increased, the server **became overloaded** during peak shopping seasons.

**2️⃣ Performance Bottlenecks**

* The single server **couldn’t handle millions of users**, leading to slow page loads.
* Traffic spikes (e.g., Black Friday sales) caused **server crashes**.
* **No failover mechanism**, meaning downtime affected all users.

**3️⃣ Implementing Load Balancers**

✅ **Horizontal Scaling:**

* Amazon introduced **multiple web servers** behind a **Load Balancer**.
* Incoming requests were distributed among different servers to prevent overload.

✅ **Traffic Routing:**

* Requests for product pages were sent to **web servers**.
* Requests for payments were sent to **secure payment servers**.

✅ **High Availability & Fault Tolerance:**

* If one server **fails**, the Load Balancer redirects traffic to other healthy servers.
* **Auto-Scaling Groups (AWS)** dynamically add or remove servers based on demand.

**4️⃣ Final Scalable System**

* **AWS ELB (Elastic Load Balancer)** dynamically distributes traffic.
* **Region-based Load Balancing** ensures users are connected to the nearest data center.
* **CDN (CloudFront)** caches static content to reduce load on servers.

**Key Takeaways**

* **Use Layer 4 Load Balancers** for TCP/UDP-based applications where raw traffic needs to be distributed quickly.
* **Use Layer 7 Load Balancers** for HTTP/HTTPS traffic when intelligent routing (e.g., API gateways, authentication) is required.
* **Use Load Balancers with Auto-Scaling** to handle fluctuating traffic and ensure reliability.