What is Load Balancing?

Load balancing is the process of distributing incoming network traffic across multiple servers (or resources) so that no single server is overwhelmed.

- It improves **performance** (by handling more requests).
- Ensures **high availability** (if one server fails, traffic shifts to others).
- Provides **scalability** (easy to add/remove servers).

Think of it like a traffic policeman at a busy junction — directing vehicles (requests) so that no single road (server) gets jammed.

Common Load Balancing Strategies

1. **Round Robin**
    - Requests are distributed **sequentially** across servers.
    - Example: If there are 3 servers, requests go → Server 1 → Server 2 → Server 3 → back to Server 1.
    - **Best for**: When all servers have roughly equal capacity.
2. Limitation: Doesn't consider server load (one might get overloaded if requests are heavy).

---

2. **Least Connections**
    - Requests go to the server with the **fewest active connections**.
    - Helps when some requests are "heavier" and take longer.
    - **Best for**: When request processing times vary a lot.
3. Example:

- Server A has 2 active requests, Server B has 5.
- Next request goes to **Server A**.

---

3. **Random**
    - Requests are sent to a **random server**.
    - Simple to implement, avoids patterns.
    - **Best for**: Large clusters where randomness statistically balances load over time.
4. Limitation: Not efficient if only a few servers are available (may overload one by chance).