What is Load Balancing?

Load Balancing is the process of distributing incoming network traffic (or tasks/requests) across multiple servers or resources (like application servers, databases, or services) to:

Maximize performance

Increase fault tolerance

Avoid overloading any single node

Ensure high availability and scalability

A load balancer sits between clients and servers and decides where to send each request.

Common Load Balancing Strategies

Here are the most widely used algorithms:

1. Round Robin (Most Common, Simple)

Requests are distributed in order, one by one, to each server.

After reaching the last server, it starts again from the first.

📌 Example:

If you have 3 servers, the requests go like this:

Request 1 → Server A

Request 2 → Server B

Request 3 → Server C

Request 4 → Server A

... and so on

Good when all servers are equal in capacity

Bad if servers have different performance

2. Least Connections

The request is sent to the server with the fewest active connections.

Useful for long-lived sessions (like WebSocket, database connections).

Example:

If:

Server A has 10 active connections

Server B has 4

Server C has 2

Next request → Server C.

Good for dynamic load

Slight overhead in tracking active connections

3. Random

Selects a server at random for each request.

Sometimes combined with weights (e.g., weighted random).

📌 Example:

Request 1 → Server B

Request 2 → Server C

Request 3 → Server A

Random every time.

Very simple

Can result in uneven distribution, especially short-term