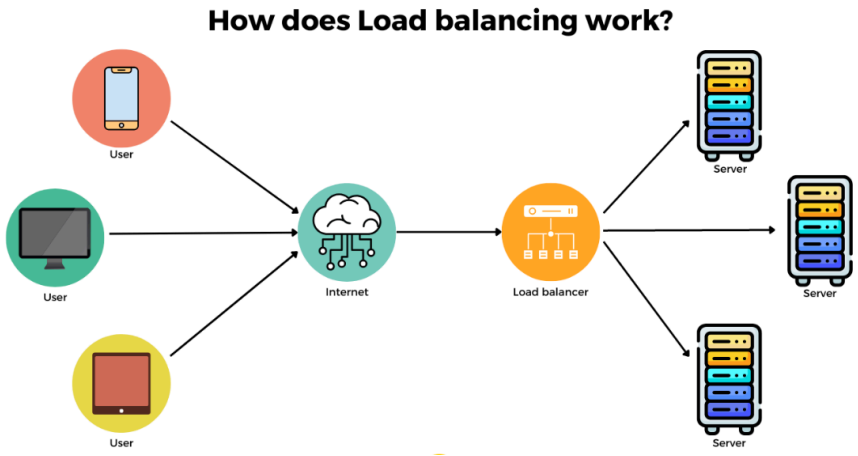
Load Balancing

Load balancing is the process of distributing incoming network traffic across multiple servers to optimize resource usage, maximize throughput, and improve application availability and responsiveness.

Here are some Load Balancing examples:

- There is application load balancer which distributes one single application over the servers; there is another which distributes only between the server cluster; another directs the traffic from multiple paths to a single destination.
- Other load balancing solutions are very advanced. They can shape the traffic and act as intelligent traffic switches, do different health checks on the content, applications, and servers, add extra security on the network and protect it from malicious software and improve availability.



How does it work?

Load balancing is achieved and managed with a tool or application that is called a load balancer. Despite the form of the load balancer (hardware or software), its main goal is to spread the network traffic among different servers and prevent overloading.

1. Round Robin

- **How it works:** Requests are distributed sequentially to each server in turn.
- **Use case:** Simple, evenly distributed workloads.

2. Least Connections

- **How it works:** Routes each new request to the server with the fewest active connections.
- **Use case:** Good when request duration varies significantly.

3. Least Response Time

- **How it works:** Sends traffic to the server with the fastest average response time and fewest active connections.
- **Use case:** Real-time systems needing low-latency responses.