

Introduction to Load Balancing

Load Balancing: Distributing Workloads for Optimal Performance

DEFINITION

Load balancing is the process of distributing network traffic across multiple servers to ensure no single server becomes overwhelmed. This enhances performance, reliability, and availability of applications.

KEY CONCEPTS

Types of Load Balancers

There are hardware and software load balancers, each serving to distribute traffic differently.

Sticky Sessions

Also known as session persistence, this keeps a user's requests on the same server for the duration of their session.

Scalability

Load balancing allows systems to scale easily by adding or removing servers as needed.

Round Robin

This method distributes requests sequentially to each server in a rotating manner.

Health Checks

Load balancers perform regular checks to ensure servers are operational and can handle requests.

EXAMPLES

- A website uses load balancing to manage traffic spikes during sales events.
- An online gaming platform distributes player connections across multiple servers.
- A cloud service provider balances workloads among virtual machines for optimal resource use.

MEMORY TIPS

- ★ Think of a traffic cop directing cars (requests) to different roads (servers).
- ★ Remember 'Round Robin' as a game where everyone gets a turn.
- ★ Use 'Sticky Sessions' to recall that some users stick with the same server.

COMMON MISTAKES

- ✗ Confusing load balancing with failover systems, which are different.
- ✗ Neglecting to configure health checks, leading to server overload.
- ✗ Assuming all load balancers are the same without understanding their types.

QUICK RECAP

Load balancing is essential for optimizing server performance by distributing traffic. Key methods include round robin and sticky sessions, while health checks ensure reliability. Understanding these concepts is crucial for effective network management.