ONE PAGE SUMMARY

IT & Software › Networking › Load Balancing

Introduction to Load Balancing

Load Balancing: Distributing Traffic for Optimal Performance

DEFINITION

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

KEY CONCEPTS

Traffic Distribution

Load balancers evenly distribute incoming traffic to prevent server overload.

High Availability

Load balancing increases system reliability by rerouting traffic if a server fails.

Scalability

It allows the addition of more servers to handle increased traffic without downtime.

Session Persistence

Also known as sticky sessions, it ensures a user's requests are sent to the same server.

Health Checks

Load balancers perform regular checks on servers to ensure they are operational.

EXAMPLES

→ A website using multiple servers to manage user requests during peak hours.

→ An online store balancing traffic during a major sale event.

→ A cloud service distributing workloads across various data centers.

MEMORY TIPS

★ Think of load balancing as a traffic cop directing cars (data) to avoid jams (overload).

★ Remember 'HATS' for High availability, Application distribution, Traffic management, Scalability.

★ Visualize a seesaw balancing weights (servers) to keep it level (performance).

COMMON MISTAKES

✗ Assuming all load balancers are the same without understanding their types.

✗ Neglecting to configure health checks, leading to traffic directed to down servers.

✗ Overlooking session persistence needs for applications requiring user state.

QUICK RECAP

Load balancing is essential for distributing traffic across servers to enhance performance and reliability. Key concepts include traffic distribution, high availability, and health checks. Remember to avoid common mistakes like neglecting health checks and confusing load balancer types.