ONE PAGE SUMMARY

IT & Software › System Design › Scalability

Introduction to Scalability

Mastering scalability is key to successful system design.

DEFINITION

> Scalability is the ability of a system to handle increased load without compromising performance. It ensures that applications can grow and adapt to changing demands.

KEY CONCEPTS

Vertical Scaling

Adding more power to an existing machine, like upgrading CPU or RAM.

Horizontal Scaling

Adding more machines to a system to distribute the load evenly.

Load Balancing

Distributing incoming traffic across multiple servers to optimize resource use.

Elasticity

The ability to automatically adjust resources based on current demand.

Bottlenecks

Points in a system that limit performance and scalability, often requiring optimization.

EXAMPLES

→ A website using multiple servers to handle high traffic during sales.

→ A cloud service that automatically adds resources during peak usage times.

→ A database that can grow by adding more nodes to manage increased queries.

MEMORY TIPS

★ Think of vertical scaling as 'upgrading' and horizontal scaling as 'adding more friends'.

★ Remember 'elasticity' like a rubber band that stretches to meet demand.

★ Visualize bottlenecks as traffic jams that slow down the entire system.

COMMON MISTAKES

✗ Confusing vertical scaling with horizontal scaling.

✗ Neglecting to monitor performance after scaling.

✗ Assuming all applications scale the same way.

QUICK RECAP

Scalability is essential for systems to grow with demand. Understand the differences between vertical and horizontal scaling, and how to manage load effectively. Always monitor for bottlenecks to maintain optimal performance.