ONE PAGE SUMMARY

IT & Software › System Design › Capacity Planning

Introduction to Capacity Planning

Mastering Capacity Planning for Optimal System Performance

DEFINITION

> Capacity planning is the process of determining the necessary resources to meet future demands. It ensures that systems can handle expected workloads without performance issues.

KEY CONCEPTS

Demand Forecasting

Estimating future resource needs based on historical data and trends.

Resource Allocation

Distributing available resources effectively to meet demand without over-provisioning.

Scalability

The ability of a system to grow and manage increased workload by adding resources.

Performance Metrics

Quantitative measures used to assess system performance and capacity utilization.

Bottleneck Identification

Finding and addressing points in the system that limit performance or capacity.

EXAMPLES

→ A web server needs to handle 1,000 concurrent users without slowing down.

→ A database must store 10TB of data while maintaining fast query response times.

→ An application is designed to scale from 100 to 10,000 users seamlessly.

MEMORY TIPS

★ Think 'D-R-P-B' for Demand, Resources, Performance, Bottlenecks.

★ Visualize a water tank: capacity planning is about ensuring it doesn't overflow.

★ Use the acronym 'SCALE' to remember Scalability, Capacity, Allocation, Load, and Efficiency.

COMMON MISTAKES

✗ Neglecting to update forecasts based on changing user behavior.

✗ Overestimating capacity needs, leading to wasted resources.

✗ Failing to consider peak usage times in planning.

QUICK RECAP

Capacity planning is essential for ensuring that systems can meet future demands efficiently. By forecasting demand and allocating resources effectively, you can prevent performance bottlenecks and ensure scalability.