

Resource Allocation & Scheduling in Cloud

Rajinder Kumar, Research Scholar, Department of Computer Science & Engineering, Monad University, Hapur, Uttar Pradesh (India)

Abstract

Cloud computing provides on-demand access to computing resources such as virtual machines, storage, networks, and processing power through shared and distributed infrastructures. Efficient resource allocation and scheduling are essential for improving system performance, reducing operational costs, and ensuring effective utilization of cloud resources. Resource allocation involves assigning available resources to users or applications according to their computational requirements and service-level objectives, while scheduling determines the appropriate order and placement of tasks for execution. This paper presents an overview of major resource allocation and scheduling techniques used in cloud computing environments. It examines traditional approaches such as First Come First Serve, Round Robin, and priority-based scheduling, along with more advanced techniques based on optimization, heuristics, metaheuristics, and machine learning. Important performance parameters, including response time, execution time, throughput, resource utilization, energy consumption, load balancing, and cost, are also discussed. The paper further highlights challenges associated with heterogeneous resources, dynamic workloads, scalability, virtualization, task dependencies, and quality-of-service requirements. Effective allocation and scheduling mechanisms can help cloud service providers balance resource demand while maintaining reliability and meeting users' performance expectations. The survey emphasizes the need for adaptive and intelligent scheduling approaches capable of responding to changing workloads and resource availability. Overall, efficient resource allocation and scheduling remain fundamental to achieving scalable, economical, energy-efficient, and high-performing cloud computing systems.

Keywords: Cloud computing, resource allocation, task scheduling, virtual machines, load balancing, resource utilization, optimization, energy efficiency, quality of service, cloud infrastructure.