

ABSTRACT

Resource allocation in cloud computing is a critical aspect that directly impacts the efficiency and performance of cloud services. The advent of reinforcement learning (RL) has provided innovative approaches to optimize resource allocation dynamically. This report explores the integration of reinforcement learning techniques in cloud resource allocation, highlighting existing systems, motivations for using RL, and a proposed system designed to enhance resource management. The study aims to address the challenges associated with traditional static allocation methods by employing RL algorithms to adaptively manage resources based on real-time demands and conditions. The rapid evolution of cloud computing has transformed the landscape of IT resource management, necessitating advanced strategies for effective resource allocation. As organizations increasingly rely on cloud services for their computational needs, the challenge of efficiently distributing resources to meet dynamic user demands has become paramount.

Traditional resource allocation methods often fall short in addressing the complexities and variability of workloads, leading to issues such as underutilization, overprovisioning, and increased operational costs. In this context, reinforcement learning emerges as a promising approach to optimize resource allocation in cloud environments. The transformative potential of reinforcement learning in revolutionizing resource allocation practices within cloud computing. By leveraging RL techniques, organizations can achieve greater efficiency and adaptability in their resource management strategies, ultimately leading to enhanced service delivery and customer satisfaction. Future research directions include exploring multi-agent reinforcement learning frameworks for collaborative resource management across distributed cloud environments

Keywords: Cloud computing, Reinforcement Learning, Dynamic workloads, Cloud environment, Markovnikov process