Agentic AI for Autonomous Cloud Resource Management

Amarnath J L

*Corresponding Author: [1] Department of Computer Science and Engineering, Nagarjuna College of Engineering and Technology, Visvesvaraya Technological University Belagavi, India,

amarnath.jl@ncetmail.com

ABSTRACT

Cloud computing made the deployment of computing resources by enabling on demand access to infrastructure which are scalable, platforms, and software services. However, modern cloud environments are increasingly challenged by dynamic workloads, heterogeneous infrastructures, energy consumption, service-level agreement (SLA) violations, and operational complexity. Conventional cloud resource management techniques rely heavily on predefined rules, threshold-based policies, or human intervention, making them inadequate for rapidly changing distributed environments. Recent advances in the field of Agentic Artificial Intelligence (Agentic AI) gives an opportunity to develop autonomous systems capable of perception, reasoning, planning, tool utilization, collaboration, and continuous learning without constant human supervision.

This paper proposes an **Agentic AI-based Autonomous Cloud Resource Management Framework (AACRM)** that employs multiple intelligent software agents to automate cloud infrastructure management. The proposed architecture integrates LLMs (Large Language Models), reinforcement learning, cloud monitoring systems, and orchestration platforms such as Kubernetes and OpenStack to create a self-managing cloud ecosystem. Independent AI agents perform workload analysis, resource prediction, virtual machine allocation, container scheduling, anomaly detection, energy optimization, security monitoring, and SLA compliance. The agents collaborate through a centralized coordination framework while maintaining decentralized decision-making capabilities to improve scalability and resilience.

A hybrid decision-making mechanism combining reinforcement learning with predictive machine learning models enables proactive resource allocation before workload spikes occur. Experimental evaluation is designed using CloudSim Plus and Kubernetes-based simulation environments, where the proposed framework is compared against traditional static scheduling, heuristic optimization, and conventional auto-scaling mechanisms. Evaluation metrics include resource utilization, response time, throughput, energy consumption, SLA violation rate, operational cost, and scalability.

Expected results indicate that Agentic AI significantly improves infrastructure efficiency by reducing SLA violations, minimizing energy consumption, increasing server utilization, accelerating provisioning decisions, and lowering operational expenses compared with existing cloud management approaches. The framework also demonstrates enhanced adaptability under unpredictable workloads through continuous autonomous learning and collaborative multi-agent decision-making.

The proposed research contributes a comprehensive architecture, mathematical model, intelligent scheduling algorithm, and evaluation methodology for next-generation autonomous cloud management systems. The framework provides a foundation for self-driving cloud infrastructures capable of supporting Industry 5.0, edge computing, Internet of Things (IoT), smart cities, healthcare, financial services, and large-scale enterprise applications.

KEYWORDS: Agentic AI; Autonomous Cloud Computing; Cloud Resource Management; Multi-Agent Systems; Large Language Models; Reinforcement Learning; Kubernetes; OpenStack; CloudSim Plus; Auto-Scaling; Resource Scheduling; Intelligent Cloud Infrastructure.