
Inspire Eduversity Publications
In Association with
International Academic and Research Foundation
National Articles on Mixed Methodology
Souvenir 2025
Organised by Payanam

INSPIRE EDUVERSITY PUBLICATIONS

In Association with

International Academic & Research Foundation

National Articles on Mixed Methodology Souvenir 2025

Published on 10th December 2025

Organised by Payanam

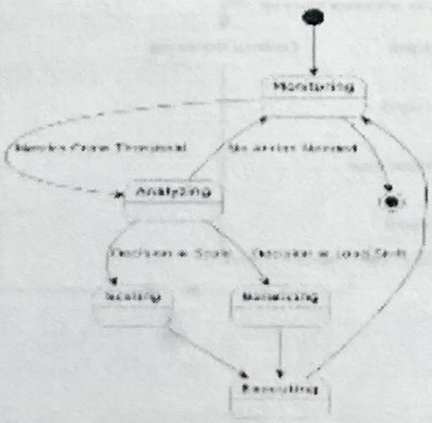
ENHANCING LOAD BALANCING AND SCALING ALGORITHMS IN CLOUD COMPUTING USING AGENT-BASED AI SYSTEMS

DANIEL RAJA SINGH,
Research Scholar,
Department of Advanced Computing and Analytics, VISTAS, Chennai.
Email: daniel.mca.09@gmail.com

Dr. R. DURGA,
Professor,
Department of Advanced Computing and Analytics, VISTAS, Chennai.
Email: durga.scs@vistas.ac.in

1. Introduction

Cloud computing has revolutionized the way applications and services are delivered by enabling dynamic resource allocation. Load balancing and auto-scaling are two key components that ensure optimal resource utilization and service availability. However, traditional methods can struggle with real-time decision-making in complex and dynamic environments. This paper explores how Agent-Based Artificial Intelligence (Agent AI) can improve the effectiveness of load balancing and scaling algorithms in cloud computing.



2. AI-Driven Resource Management in Cloud Computing

Load balancing in cloud computing distributes network traffic across multiple servers to prevent overloading, using methods like Round Robin and Least Connections. Auto-scaling dynamically adjusts server capacity based on demand, employing reactive or predictive strategies. Agent AI involves autonomous software entities that perceive environments, make decisions, and act to achieve objectives. These agents can be reactive, deliberative, or hybrid, enabling intelligent automation. By integrating AI-driven agents, cloud systems enhance efficiency through real-time decision-making and adaptive scaling. This synergy optimizes performance while maintaining stability under varying workloads.

3. Architecture of Agent-Based Load Balancing and Scaling

A typical architecture includes a Monitoring Agent that tracks performance metrics like CPU usage and latency. The Decision Agent applies AI algorithms to determine scaling or workload redistribution. An Execution Agent interfaces with cloud APIs (AWS, Azure, GCP) to implement these actions. Finally, a Learning Agent refines future decisions using reinforcement learning or neural networks.

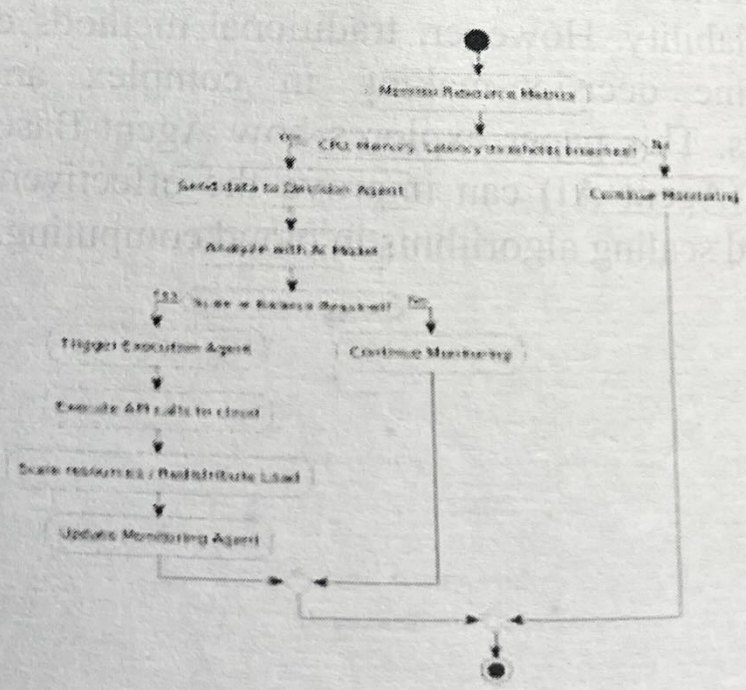


Diagram: Agent AI Framework flow and sequence diagram for Cloud Load Balancing

4. Algorithms and AI Techniques

AI-driven cloud optimization leverages reinforcement learning for policy refinement and swarm intelligence (e.g., Ant Colony) for decentralized coordination. Multi-agent systems (MAS) enable collaborative cluster management, while genetic algorithms evolve near-optimal scaling configurations. Together, these techniques enhance adaptability and efficiency in dynamic cloud environments.

5. Benefits and Challenges

This approach improves resource utilization and reduces costs while enhancing responsiveness for lower latency and better user experience. However, implementation complexity and the need for reliable real-time data pose significant challenges. Potential agent conflicts or coordination issues may also arise, demanding careful management. Balancing these benefits and challenges is key to maintaining system stability and efficiency.

6. Conclusion

Agent AI brings a new level of intelligence and autonomy to load balancing and scaling in cloud computing. By learning from the environment and making smart decisions in real time, these systems can significantly enhance performance and efficiency in large-scale, dynamic environments.

ISBN No. 978-93-47021-37-4