

AI-DRIVEN ADAPTIVE RESOURCE ALLOCATION AND TASK SCHEDULING FOR PERFORMANCE OPTIMIZATION IN CLOUD COMPUTING ENVIRONMENTS

Yashodhan Akhare¹, Dr. Dharendra Kumar Tripathi²

Research Scholar, Mansarovar Global University, Sehore (Bhopal)¹

Research Supervisor, Mansarovar Global University, Sehore (Bhopal)²

ABSTRACT

Efficient resource allocation and task scheduling are fundamental challenges in cloud computing because of dynamic workloads, heterogeneous virtualized resources and competing requirements for performance, cost, energy efficiency and Quality of Service (QoS). Traditional static and heuristic scheduling techniques frequently fail to adapt effectively to rapidly changing cloud environments. Artificial Intelligence (AI), particularly Machine Learning (ML), Deep Learning (DL) and Deep Reinforcement Learning (DRL), offers an opportunity to develop adaptive scheduling systems capable of learning from workload patterns and continuously optimizing resource allocation decisions. The present study develops a conceptual AI-driven adaptive resource allocation framework for cloud-based systems. The proposed framework combines workload monitoring, state representation, predictive analytics and reinforcement-learning-based scheduling to dynamically allocate computing resources according to workload conditions. The optimization objective incorporates execution time, resource utilization, energy consumption, cost and Service Level Agreement (SLA) violations. Recent studies demonstrate increasing effectiveness of DRL and hybrid AI models in multi-objective cloud-resource optimization. The paper identifies the limitations of conventional approaches, reviews recent developments and proposes an integrated architecture suitable for experimental validation using CloudSim/CloudSim Plus and real-world cloud workload traces. The proposed approach contributes to the broader development of an AI-driven analysis and optimization framework for cloud-based systems.

Keywords: *Cloud computing¹, Artificial intelligence², Resource allocation³, Task scheduling⁴, Deep reinforcement learning⁵, Optimization⁶, QoS⁷, Cloud resource management⁸*

1. INTRODUCTION

Cloud computing has transformed the delivery of computing infrastructure by providing computing, storage and networking resources on demand. Organizations increasingly depend on cloud environments because of their scalability, flexibility and pay-as-you-use model. However, effective management of cloud infrastructure is becoming more difficult because modern cloud environments handle highly heterogeneous and unpredictable workloads. Resource allocation determines how computing resources such as CPU, memory, storage and network bandwidth are distributed among applications and users. Task scheduling determines when and where individual computational tasks are executed. Poor resource allocation can lead to:

- Excessive execution time;
- Underutilization of virtual machines;
- Increased energy consumption;
- Unnecessary operational cost;
- SLA violations; and
- Deterioration in user experience.

Traditional scheduling techniques such as First Come First Serve, Round Robin and static priority scheduling are relatively simple but often fail under highly dynamic workload conditions. Metaheuristic methods such as Genetic Algorithms, Particle Swarm Optimization, Ant Colony Optimization and Cuckoo Search have improved optimization capabilities but can introduce high search cost and may not continuously adapt to changing cloud conditions. Artificial Intelligence provides a different approach. AI-based systems can learn historical workload patterns, predict future resource demands and dynamically modify allocation decisions. Deep Reinforcement Learning is particularly promising because scheduling can be formulated as a sequential decision problem in which an intelligent agent learns an optimal scheduling policy by interacting with the cloud environment. A 2024 review in Artificial Intelligence Review highlighted the growing role of DRL in cloud resource scheduling and identified its ability to address scheduling situations in which conventional heuristic and metaheuristic approaches struggle with complex state spaces. More recently, a 2025 comprehensive review concluded that AI-driven job scheduling is becoming increasingly important because dynamic and heterogeneous clouds require adaptive scheduling mechanisms rather than static optimization. The present paper therefore proposes an AI-driven adaptive framework for intelligent resource allocation and scheduling in cloud-based systems.

2. OBJECTIVES OF THE STUDY

The study aims to:

1. Examine limitations of traditional cloud-resource allocation methods;
2. Investigate AI and reinforcement-learning approaches for cloud scheduling;
3. Develop an adaptive AI-driven resource-allocation framework;
4. Optimize CPU, memory and virtual-machine utilization;
5. Minimize task execution time and makespan;
6. Reduce cloud operating cost and energy consumption;
7. Minimize SLA violations; and
8. Establish a framework suitable for experimental validation.

3. RESEARCH PROBLEM

Resource allocation in cloud systems represents a multi-objective optimization problem.

For a set of tasks:

$$T = T_1, T_2, \dots, T_n \quad T = \{T_1, T_2, \dots, T_n\}$$

and cloud resources:

$$R = R_1, R_2, \dots, R_m \quad R = \{R_1, R_2, \dots, R_m\}$$

the objective is to determine an assignment:

$$T_i \rightarrow R_j$$

that optimizes several competing performance measures.

The optimization function may be expressed as:

$$\text{Minimize } F = w_1(MS) + w_2(C) + w_3(E) + w_4(SLA) - w_5(U) \quad \text{Minimize } F = w_1(MS) + w_2(C) + w_3(E) + w_4(SLA) - w_5(U)$$

where:

- MSMS = Makespan

- CC = Cloud cost
- EE = Energy consumption
- SLASLA = SLA violations
- UU = Resource utilization
- $w_1 \dots w_5$ = optimization weights.

The challenge is that resource demand and availability continuously change, making static optimization insufficient.

4. REVIEW OF RECENT LITERATURE

Zhou et al. (2024) Zhou, Tian, Buyya and colleagues conducted a comprehensive review of Deep Reinforcement Learning methods for resource scheduling in cloud computing. They demonstrated that DRL is particularly suitable for complex scheduling environments because it integrates representation learning with sequential decision-making. They also highlighted challenges including large state spaces, reward-function design and adaptability to realistic cloud conditions.

Neelakantan et al. (2024) Neelakantan and colleagues proposed a Caching-based Deep Q-Network framework for resource allocation in IoT edge-cloud computing. The method dynamically allocates computational and storage resources across edge and cloud infrastructure, demonstrating the growing usefulness of DQN techniques for distributed resource management.

Security-Aware DRL Resource Allocation (2024) Research on serverless multi-cloud edge computing introduced DRL-based allocation mechanisms that jointly address computation-resource allocation and security requirements. This highlights the possibility of incorporating security constraints directly into AI optimization objectives rather than treating resource management and security separately.

Sanjalawe et al. (2025) A comprehensive review of AI-driven job scheduling emphasized that increasing cloud heterogeneity and dynamic workloads have created a need for intelligent scheduling systems capable of continuously adapting scheduling decisions. AI-based approaches were identified as important alternatives to traditional scheduling mechanisms.

AI and ML-Based Cloud Allocation (2025) An IEEE conference study investigated optimization of cloud-resource allocation through AI and ML techniques. The work emphasized predictive analytics, adaptive allocation and real-time decision-making as mechanisms for improving resource utilization, reducing cost and enhancing system performance.

Comparative ML Resource Allocation Review (2025) A 2025 comparative review evaluated ten resource-allocation algorithms across DRL, neural-network, conventional ML and multi-agent categories. The review reported that hybrid AI architectures frequently achieve stronger multi-objective performance than single-algorithm approaches, particularly for makespan, cost and energy optimization.

Table 1: Recent Research on AI-Based Resource Allocation

Study	Approach	Major focus
Zhou et al. (2024)	DRL review	Intelligent resource scheduling
Neelakantan et al. (2024)	Deep Q-Network	Edge-cloud resource allocation
Multi-cloud study (2024)	Security-aware DRL	Secure computation allocation
Sanjalawe et al. (2025)	AI scheduling review	Adaptive cloud job scheduling
IEEE study (2025)	AI/ML	Predictive resource allocation
Comparative review (2025)	Hybrid AI/ML	Multi-objective resource optimization

5. RESEARCH GAP

Existing literature demonstrates the potential of AI for cloud-resource allocation, but several gaps remain. First, many methods optimize only one or two metrics such as makespan or resource utilization. Second, several approaches remain reactive and allocate resources after workload variations have already occurred. Third, models developed for specific workload conditions may not generalize across heterogeneous cloud environments. Fourth, energy consumption, cost and SLA violations are not always optimized simultaneously.

Finally, limited attention has been given to integrated frameworks combining:

workload prediction + adaptive scheduling + multi-objective optimization + continuous feedback

The present study addresses this gap through an integrated AI-driven adaptive optimization architecture.

6. PROPOSED AI-DRIVEN FRAMEWORK

The proposed framework consists of five major layers.

Layer 1: Cloud Monitoring

Real-time information is continuously collected for:

- CPU utilization;
- Memory usage;
- Network usage;

- VM availability;
- Task arrival rate;
- Execution time; and
- Energy consumption.

Layer 2: Data pre-processing

Cloud monitoring data are:

- Cleaned;
- Normalized;
- Transformed; and
- Converted into state vectors.

Layer 3: Workload Analysis

Machine-learning techniques analyse current and historical workload behaviour.

The system identifies:

- Workload intensity;
- Recurring patterns;
- Resource bottlenecks; and
- Anomalous behaviour.

Layer 4: DRL Scheduling Agent

The reinforcement-learning agent observes cloud states and determines the most suitable task-resource allocation.

State

$St = [\text{CPU, Memory, Bandwidth, Queue, VM Load}]$

Action

Possible actions include:

- Allocate task to VM;

- Migrate task;
- Activate additional VM;
- Consolidate workloads; or
- Release idle resource.

Reward

A reward function may be expressed as:

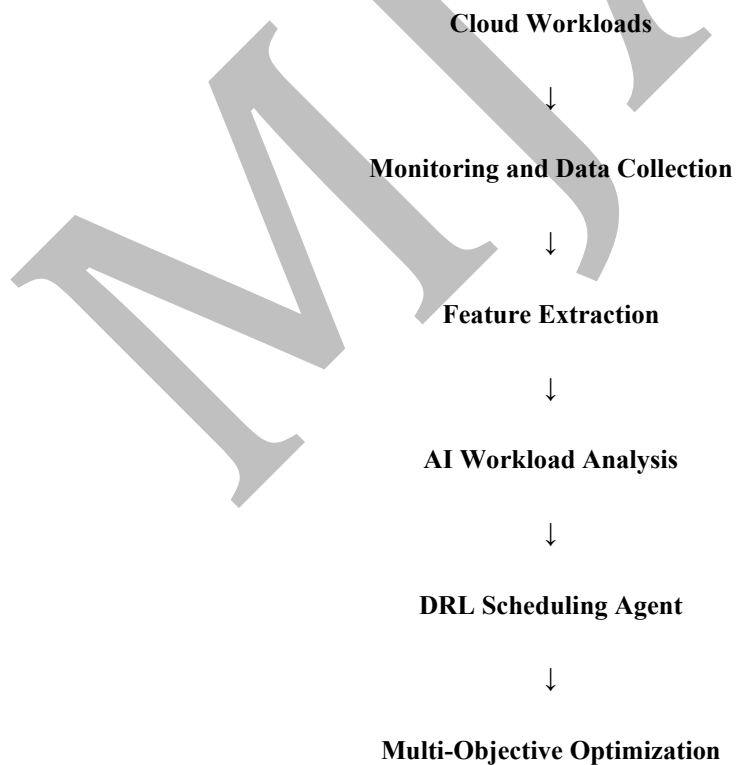
$$R_t = \alpha U - \beta MS - \gamma C - \delta E - \theta SLA, \quad R_t = \alpha U - \beta MS - \gamma C - \delta E - \theta SLA$$

where higher utilization increases reward, while makespan, cost, energy use and SLA violations generate penalties.

Layer 5: Feedback and Learning

Observed outcomes are fed back to the learning agent so that allocation decisions progressively improve.

7. PROPOSED ARCHITECTURE





Resource Allocation



Performance Measurement



Feedback to AI Agent

This creates a continuous adaptive optimization loop.

8. PROPOSED EXPERIMENTAL METHODOLOGY

Simulation Platform

The framework may be implemented using:

- CloudSim;
- CloudSim Plus; or
- IfogSim where edge-cloud environments are included.

Workloads

Experiments may use:

- Synthetic workloads;
- Google cluster traces;
- Alibaba cluster traces; or
- Other publicly available cloud workload datasets.

Algorithms for Comparison

The proposed model should be compared against:

1. First Come First Serve;
2. Round Robin;

3. Min-Min;
4. Max-Min;
5. Particle Swarm Optimization;
6. Genetic Algorithm;
7. Q-Learning;
8. Deep Q-Network; and
9. proposed adaptive DRL model.

9. PERFORMANCE METRICS

Makespan

Makespan represents the total time required to complete the entire workload. It is calculated as:

$$\text{Makespan} = \max_i(\text{Completion Time}_i)$$

A lower makespan value indicates faster overall workload execution.

Resource Utilization

Resource utilization measures the proportion of available computational resources that are actually used. It is calculated as:

$$\text{Resource Utilization} = \frac{\text{Resource Used}}{\text{Resource Available}} \times 100$$

A higher resource utilization value indicates more effective use of the available resources.

SLA Violation Rate

SLA violation rate represents the percentage of tasks that fail to meet the requirements specified in the Service Level Agreement (SLA). It is calculated as:

$$\text{SLA Violation Rate} = \frac{\text{Number of Violated Tasks}}{\text{Total Number of Tasks}} \times 100$$

A lower SLA violation rate indicates better compliance with the defined service-level requirements.

Energy Consumption

Total energy consumed by active physical and virtual resources during execution.

Cost

Total cost represents the overall expenditure incurred for executing the workload, including computing, storage, and network-related expenses. It is calculated as:

$$\text{Total Cost} = \text{Compute Cost} + \text{Storage Cost} + \text{Network Cost}$$

10. EXPECTED RESULTS

Based on recent literature, the proposed AI framework is expected to achieve:

- Higher resource utilization;
- Reduced makespan;
- Lower task waiting time;
- Fewer SLA violations;
- Improved load balancing;
- Reduced energy consumption; and
- Lower operational cost.

The primary advantage is expected to arise from the framework's ability to learn and adapt rather than rely upon fixed scheduling rules.

11. CONCLUSION

Cloud-resource allocation has become increasingly complex because of dynamic workloads, heterogeneous resources and competing performance objectives. Traditional scheduling techniques remain useful for relatively simple environments but lack the adaptability required for modern large-scale cloud systems. The present paper proposes an AI-driven adaptive framework using intelligent workload analysis and Deep Reinforcement Learning. The framework continuously observes cloud conditions, determines adaptive resource-allocation actions and improves future decisions through feedback. Its multi-objective reward mechanism simultaneously considers resource utilization, makespan, cost, energy consumption and SLA violations. The framework provides a strong component of the broader Ph.D. research theme “AI Driven Analysis & Optimization Framework for Cloud Based Systems” and can subsequently be experimentally evaluated using CloudSim Plus and real-world workload traces.

12. REFERENCES

- [1] Zhou, G., Tian, W., Buyya, R., et al. (2024). Deep reinforcement learning-based methods for resource scheduling in cloud computing: A review and future directions. *Artificial Intelligence Review*, 57, 124.
- [2] Neelakantan, P., Gangappa, M., Rajasekar, M., Kumar, T. S., & Reddy, G. S. (2024). Resource allocation for content distribution in IoT edge cloud computing environments using deep reinforcement learning. *Journal of High Speed Networks*, 30(3).
- [3] Sanjalawe, Y., Al-E'mari, S., Fraihat, S., et al. (2025). AI-driven job scheduling in cloud computing: A comprehensive review. *Artificial Intelligence Review*, 58, 197.
- [4] Nandagopal, M., Manavalan, T., Kumar, K. P., et al. (2025). Enhancing energy efficiency in cloud computing through task scheduling with hybrid cuckoo search and transformer models. *Discover Computing*, 28, 199.
- [5] Manhary, F. N., Mohamed, M. H., & Farouk, M. (2025). A scalable machine learning strategy for resource allocation in database. *Scientific Reports*, 15, 30567.
- [6] Chen, Y., Feng, E., & Ling, Z. (2024). Secure resource allocation optimization in cloud computing using deep reinforcement learning. *Journal of Advanced Computing Systems*.