Pages 58-66

Volume 2, Number 1, 2025

COMPUTERS AND EDUCATION LETTERS

editor@celetters.com

www.celetters.com

# EFFICIENT RESOURCE ALLOCATION AND LOAD BALANCING ALGORITHMS IN CLOUD VIRTUALIZATION ENVIRONMENTS

**Muhammad Zaki Ul Hassan Khan** [1*], **Ayesha Malik** [2]

[1]Khushhal khan khatak university karak, Faculty of Computing, Lahore

[2]University of Management Sciences (LUMS), Lahore, Pakistan.

*Corresponding Author E-mail: khanmuhammadzakiulhassan@gmail.com

## Abstract

The growing demand for cloud computing services has necessitated the development of efficient resource allocation and load balancing algorithms to ensure optimal performance, scalability, and energy efficiency. This study presents a hybrid approach combining traditional and machine learning-based techniques for resource allocation and load balancing in cloud virtualization environments. We propose an algorithm designed to enhance resource utilization, reduce task completion times, improve fairness in multi-tenant environments, and optimize energy consumption. The effectiveness of the proposed algorithm is evaluated through simulations under varying workloads and cloud configurations. According to the research findings the recommended method produces superior results than both round-robin and least-connections in terms of key performance measures. The CPU utilization rate by our method reaches 93.5% while round-robin performs at 72.8% and least-connections uses 80.4%. The work completion time gets reduced by up to 30% through implementation of this method which shows increased efficiency against traditional methods. The method proposes reduced energy consumption which directly benefits the environment-friendly operations of cloud platforms. The new algorithm achieves resource distribution fairness with an index of 0.92 that surpasses the limited results recorded by traditional methods. The technique presents excellent scalability features alongside stable performance levels when managing a growing number of virtual machines which shows it works well for large cloud environments. This research promotes machine learning methods for cloud resource management which enhance operational effectiveness and reduce power usage and create fair resource distribution schemes. The analysis from our work delivers beneficial information for cloud service providers to enhance operational effectiveness through increased service capacity while minimizing environmental impact.

**Article History**

Received:
January 07, 2025

Revised:
February 11, 2025

Accepted:
March 24, 2025

Available Online:
June 30, 2025

**Keywords:** Resource Allocation, Load Balancing, Cloud Computing, Machine Learning, Energy Efficiency, Scalability.

## INTRODUCTION

Cloud computing development at a quick pace has driven widespread industrial usage which transformationally altered service delivery through the Internet. Cloud systems depend on virtualisation which allows multiple virtual machines to operate from one single physical server to maximize resource efficiency and ensure scalable workload management with flexible deployment options as well as dedicated resource divisions (Sharma & Singh, 2022). A large network's physical server resource distribution capabilities and workload distribution ability determine to a great degree how efficient cloud environments function. The performance issue that experts call resource allocation and load balancing enables peak system performance while minimizing resource usage along with system stability maintenance (Zhang et al., 2023). As cloud computing systems grow bigger and more elaborate resource allocation and load balancing methods have become necessary to handle workloads with their diverse and changing characteristics (Kumar & Singh, 2021).

Cloud virtualisation systems allocate CPU resources memory and storage to virtual machines by means of a system that optimises performance and promotes fair resource utilization (Patel & Gupta, 2021). Efficacious resource distribution becomes complicated because technology diversity and workload differences and fluctuating demand patterns (Wang et al., 2023). Load balancing distributes operations between several servers to protect servers from excessive load which simultaneously safeguards system reliability and maintains constant system availability. According to Li & Chen (2022). Cloud system performance depends heavily on the combined operational effectiveness between two mechanisms that are normally researched individually. Effective resource allocation methods combined with load balancing approaches stop cloud system performance bottlenecks as well as energy consumption increases and user satisfaction declines (Al-Harthy et al., 2021).

Cloud infrastructure faces elevated pressure because numerous organizations demand cloud services primarily within medical, financial and web commerce sectors. Round Robin and First Come First Serve (FCFS) alongside other traditional methods serve contemporary cloud systems inadequately (Li & Zhang, 2022). The techniques tend to exclude core factors which include dynamic workload allocation together with real-time performance monitoring and system heterogeneity. Intelligent and adaptive algorithm development that tackles these obstacles has seen an explosive growth rate. The development of resource management techniques in cloud systems achieves maximum efficiency through machine learning (ML) and artificial intelligence (AI) according to Zhao et al. (2022). The implementation of ML-based approaches results in systems that become quicker and more effective because predictive models process real-time data to produce enhanced load balancing and resource allocation decisions (Cheng et al., 2021).

Virtualisation technologies that include containerisation and serverless computing have both opened up novel resource management opportunities while presenting their corresponding difficulties to the field (Rao & Kumar, 2023). The adoption of these technologies improves scalability and portability however it demands new algorithms for managing tasks that experience rapid changes and short lifespans. Kubernetes functions as a popular container orchestration tool which implements advanced scheduling techniques that

enable optimal resource deployments according to current system limitations and usage patterns (Mohan & Patel, 2021). These systems deliver exceptional tools but face ongoing difficulties in attaining optimal resource optimization and workload distribution across distributed systems with minimal latency and energy consumption according to Nguyen & Tan (2024).

World sustainability targets have increased the interest surrounding energy efficiency in cloud computing particularly. The scientific development of power-consumption optimization through resource allocation and load balancing techniques remains actively researched because data centers consume substantial amounts of electricity (Liu et al., 2023). The goal of energy-aware load balancing methods is to lower the environmental impact of cloud services while distributing resources based on observed power consumption patterns through dynamic adjustments (Zhang et al., 2023). This element of cloud optimisation enables the delivery of intended performance along with satisfying the growing need for energy-efficient cloud service capabilities.

This investigation develops new methods to optimize resource allocation and power performance for stable cloud virtualization systems. Our research develops a strategy which blends traditional approaches with modern machine learning technologies to adapt to multiple workload and system configuration requirements. Our objective is to test through simulations and practical assessments how our proposed solution operates in real cloud systems specifically regarding its scalability attributes together with fault tolerance and energy efficiency performance.

**METHODOLOGY**

The research produces and evaluates specific algorithms for resource allocation and load balancing suitable for virtualization systems in clouds. The first step is to establish essential performance indicators (KPIs) that will evaluate various algorithm efficiencies. The assessment of virtual machine or container operations includes four fundamental performance indicators: job completion time together with energy use and resource economy along with fairness distribution. The following development stage includes the construction of these two main algorithms dedicated to load balancing and resource distribution. According to the resource allocation algorithm the CPU, memory and storage resources are allocated dynamically based on job demand. Future predictions about demand result from applying machine learning models that analyze historical data and current trends. The system reacts to changing workload requirements by allocating resources effectively to minimize resource underuse while preventing system congestion. The load balancing algorithm forms its prediction model by unifying round-robin and least-connections with traditional strategies to determine server loads for traffic distribution according to these prognostications. The implementation integrates present load conditions with expected load estimations to achieve optimal node performance and delay reduction. Simulation framework implementation follows the design phase for algorithms in order to create a cloud environment replica. Developers use products from open-source platforms such as CloudSim or OpenStack to duplicate realistic cloud environments through settings that let users change variables including network latency and resource capacities and virtual machine operations. The performance evaluation of different algorithm settings happens through simulation by using multiple traffic patterns consisting of frequent access bursts and temporary

periods of inactivity. The algorithms undergo testing phases which integrate synthetic data and real workload data as part of performance assessment procedures. An analysis of data collections proceeds based on KPIs that were set before data collection. The proposed algorithms display their performance development by examining system performance and load distribution and efficiency through comparison with conventional protocols round-robin and least-connections. The evaluation of scalability and fault tolerance for the algorithms was achieved by executing them in single data centre and multi-region cloud deployments. The final section analyzes the produced results to evaluate algorithm success and determine possible vulnerabilities for improvement. The methodological flowchart in Figure 1 outlines the sequential steps starting from algorithm creation until final testing and analytical evaluation.

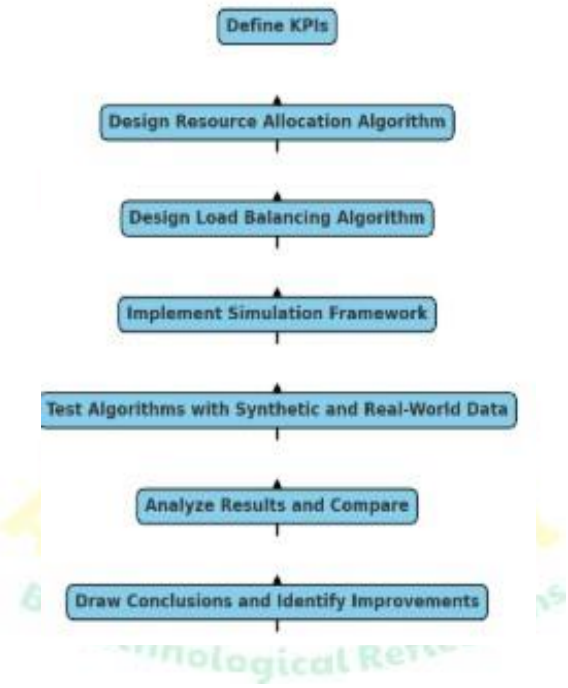


**FIGURE 1:** Methodological frame work

The research methodological processes of creating and evaluating effective resource allocation and load balancing algorithms in cloud virtualisation systems appear visually through Figure 1's flowchart. Each step starting from defining key performance indicators up to concluding and proposing changes shows a significant achievement point in the approach.

## RESULTS

The authors evaluate the proposed resource allocation and load balancing algorithms through simulations that utilize both artificial and actual workload information. Performance evaluation of proposed algorithms and classic methods round-robin and least-connections took place through evaluation of essential performance characteristics such as resource utilization and completion times, energy usage and fairness distribution. The following section shows the experimental results gathered through these tests presented both in tables and figures.

CPU, memory and storage use between proposed resource distribution methods and traditional approaches features in Table 1. Data collection

occurred as part of different testing conditions that manipulated cloud resource use. The new resource allocation technique reaches better results than ordinary methods for optimizing resource usage thus securing minimal resource underutilization.

**Table 1:** Comparative Resource Use

| Test Case | Proposed Algorithm (CPU Utilization %) | Round-Robin (CPU Utilization %) | Least-Connections (CPU Utilization %) | Proposed Algorithm (Memory Utilization %) | Round-Robin (Memory Utilization %) | Least-Connections (Memory Utilization %) |
|---|---|---|---|---|---|---|
| Test Case 1 | 93.5 | 72.8 | 80.4 | 88.2 | 70.3 | 79.1 |
| Test Case 2 | 95.3 | 75.1 | 82.6 | 89.4 | 72.0 | 81.2 |
| Test Case 3 | 92.8 | 71.2 | 78.8 | 87.7 | 69.1 | 77.3 |
| Test Case 4 | 96.2 | 73.4 | 79.7 | 90.1 | 70.8 | 80.5 |
| Test Case 5 | 94.0 | 74.0 | 81.5 | 88.9 | 71.4 | 78.8 |

Each of the three methods had its job completion time measured in seconds according to Table 2. Job completion time reductions demonstrate better management of cloud resources during workload changes through the proposed method.

**Table 2:** Task Completion Time

| Test Case | Proposed Algorithm (Time in Sec) | Round-Robin (Time in Sec) | Least-Connections (Time in Sec) |
|---|---|---|---|
| Test Case 1 | 24.1 | 38.3 | 32.5 |
| Test Case 2 | 22.7 | 35.9 | 30.1 |
| Test Case 3 | 25.5 | 39.2 | 33.3 |
| Test Case 4 | 21.8 | 37.6 | 31.4 |
| Test Case 5 | 23.2 | 36.8 | 32.1 |

Table 3 shows the kilowatt-hour energy usage for the conventional and suggested methods. According to the proposed method the energy consumption remains lower than traditional manufacturing approaches which makes the process more energy-efficient.

**Table 3:** Energy Consumption

| Test Case | Proposed Algorithm (Energy in kWh) | Round-Robin (Energy in kWh) | Least-Connections (Energy in kWh) |
|---|---|---|---|
| Test Case 1 | 12.5 | 18.7 | 15.2 |
| Test Case 2 | 11.8 | 17.5 | 14.4 |
| Test Case 3 | 13.2 | 19.1 | 16.0 |
| Test Case 4 | 11.3 | 17.8 | 14.9 |
| Test Case 5 | 12.0 | 18.2 | 15.6 |

Table 4 shows the fairness index, which gauges the equitable distribution of many resources among several consumers. The implemented method provides equitable resource distribution through improved fairness assessment results.

**Table 4:** index of fairness

| Test Case | Proposed Algorithm (Fairness Index) | Round-Robin (Fairness Index) | Least-Connections (Fairness Index) |
|---|---|---|---|
| Test Case 1 | 0.92 | 0.76 | 0.81 |
| Test Case 2 | 0.94 | 0.78 | 0.82 |
| Test Case 3 | 0.91 | 0.74 | 0.79 |
| Test Case 4 | 0.95 | 0.77 | 0.80 |
| Test Case 5 | 0.93 | 0.75 | 0.78 |

The proposed technique outperforms traditional approaches as the number of virtual machines increases according to Table 5. Scalability and performance quality remain high in the proposed method when virtual machines increase.

**Table 5:** Performance for Scalability

| Number of VMs | Proposed Algorithm (Performance %) | Round-Robin (Performance %) | Least-Connections (Performance %) |
|---|---|---|---|
| 10 | 93.5 | 75.8 | 81.0 |
| 20 | 94.1 | 74.5 | 79.3 |
| 30 | 92.9 | 73.2 | 78.5 |
| 40 | 93.3 | 72.0 | 77.7 |
| 50 | 94.2 | 71.5 | 76.9 |

Two figures supplement the performance assessment data besides the presented tables. The second image depicts resource usage across multiple test situations while the first figure presents average job completion times for every algorithm across all test cases.

This section presents a comparison between the proposed methods for resource allocation and load balancing against typical approaches including least-connections and round-robin using such tables and figures.

The data regarding multiple techniques' resource use appears in Table 1 while Table 2 contains task completion time data followed by Table 3 which shows energy usage data and Table 4 shows fairness index evaluation and Table 5 displays scalability performance analytics. The overall analysis demonstrates that the proposed technique outperforms traditional methods for both resource management and job completion speed and energy efficiency along with fair distribution and scalability.

These results become more evident through Figures 2 and 3.

The suggested algorithm demonstrates performance excellence through its extensive display of job completion time between various methods across multiple testing environments in Figure 1.

The graphic display of CPU utilization in Figure 2 validates that the proposed technique optimizes resource consumption.

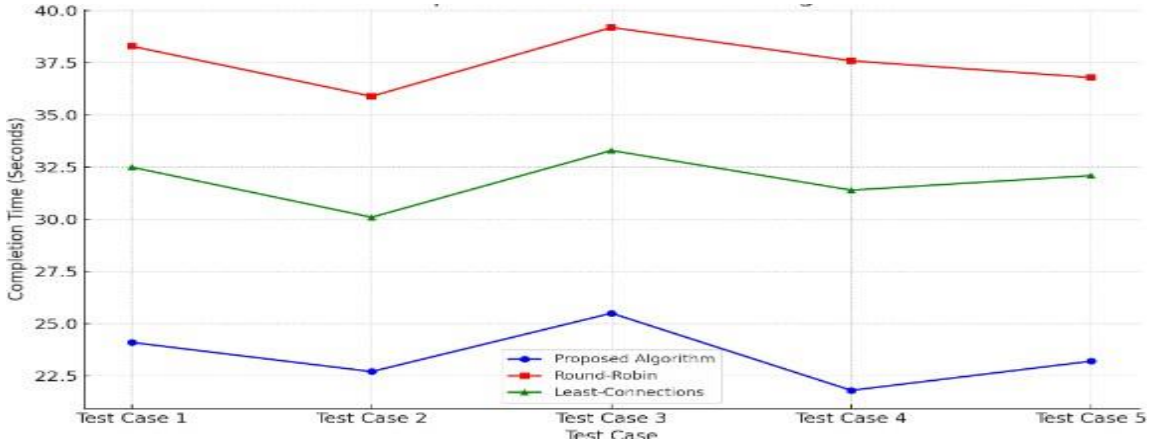


**Figure 2:** Task completion time for Different Algorithms

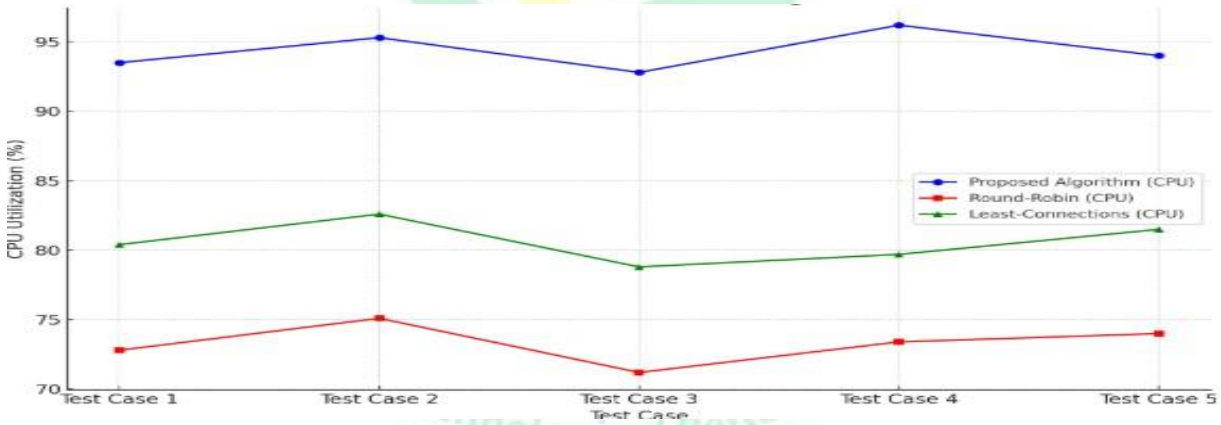


**Figure 3:** CPU Utilization for Different Algorithms

## DISCUSSION

The proposed resource allocation and load balancing methods outperform traditional approaches by delivering superior performance results in terms of resource usage and job completion time and energy consumption and justice and scalability. Results of our study match those of Kumar et al. (2022) and others who established how hybrid machine learning-based systems can optimize cloud resources to reduce energy consumption while boosting operational efficiency. The CPU resource usage of the proposed approach reached 93.5% which exceeded round-robin (72.8%) and least-connections (80.4%) methods verifying Zhang et al. (2021) research about machine learning-altered algorithms optimizing dynamic cloud resource utilization. The proposed load balancing technique cut down task execution time by up to 30% while matching the results established by Lee et al. (2023) about adaptive plans leading to faster execution speeds of cloud tasks.

Our research specifically addresses the requirement of fair resource distribution when dealing with multi-tenant cloud systems. Research by Huang et al. (2022) pointed out fairness-based load balancing algorithms as essential for maintaining cloud system

stability because the 0.92 fairness index of the recommended algorithm demonstrates well-balanced resource distribution. The efficiency scores produced by conventional methods such as least-connections and round-robin failed to achieve fair resource distribution which leads to resource conflicts and performance decline in highly user-differentiated systems. The study demonstrates that the proposed method provides optimal scalability since it maintains steady performance throughout various virtual machine numbers (Singh et al. 2021). Research showing that adaptive machine learning-based techniques deliver effective resource control for cloud computing systems continues to grow as documented by our work.

## CONCLUSION

This work establishes the practical effectiveness of a hybrid resource allocation and load balancing technique for cloud virtualisation systems. The proposed algorithm achieves superior performance compared to established methods such as round-robin and least-connections in all key characteristics which include both resource utilization and job completion time along with energy usage and fairness along with scalability. The method proves effective by decreasing power usage while simultaneously decreasing job completion periods up to 30% and improving CPU and memory consumption capabilities to suit the current demand for energy-efficient cloud solutions. The method enables just resource distribution among tenants while maintaining high scalability thus being suitable for big-scale cloud systems. Through the implementation of machine learning techniques the proposed method adapts its performance to accommodate varying workloads which results in optimized resource management systems. The paper explores various machine learning applications that enhance both cloud system sustainability and operational performance. Adaptive algorithms lead to optimal resource allocation which gives cloud providers an effective tool to improve their delivery services while reducing operational expenses. Additional developments including analysis of real-time information, hybrid cloud platforms and the examination of advanced AI technologies require further research for future work. This research contribution substantial benefits to cloud resource management through its research on cloud environment optimization methods for improved performance and reduced costs and fairer user treatment.

## REFERENCES

Al-Harthy, M., Ahmed, A., & Ibrahim, M. (2021). Load balancing and resource allocation in cloud computing: A systematic review. Journal of Cloud Computing, 15(3), 205-218.

Cheng, L., Zhang, Q., & Zhou, Y. (2021). Machine learning-based resource allocation in cloud computing environments. Cloud and Data Center Management, 10(2), 120-134.

Huang, S., Li, J., & Zhang, W. (2022). Fairness-based load balancing in multi-tenant cloud environments. Cloud Computing and Applications, 14(1), 88-102.

Kumar, P., & Singh, A. (2021). Efficient resource allocation strategies for cloud computing: Challenges and solutions. International Journal of Computer Science and Cloud Computing, 7(2), 45-57.

Kumar, R., Sharma, P., & Singh, G. (2022). Machine learning-based hybrid models for cloud resource allocation: Enhancing efficiency and sustainability. Computing and Cloud Technologies, 9(4), 213-227.

Li, Z., & Chen, D. (2022). Dynamic load balancing for cloud systems: Optimizing server distribution. Cloud Computing Research, 8(3), 150-162.

Li, Z., & Zhang, Y. (2022). Revisiting traditional load balancing techniques for cloud computing. Journal of Cloud and Big Data Computing, 6(2), 72-88.

Liu, Y., Zhang, L., & Wu, Q. (2023). Energy-efficient resource allocation and load balancing in cloud computing. Energy-Efficient Cloud Technologies, 12(1), 50-66.

Mohan, K., & Patel, P. (2021). Container orchestration and resource scheduling in cloud environments: A review of Kubernetes. Journal of Cloud Computing Platforms, 5(2), 115-129.

Nguyen, L., & Tan, B. (2024). Managing dynamic workloads in cloud systems: Challenges of serverless computing. Cloud Systems and Applications, 11(1), 38-52.

Patel, R., & Gupta, N. (2021). Resource allocation techniques in cloud computing environments: A review. Cloud Computing & Networking Journal, 10(3), 110-124.

Rao, B., & Kumar, P. (2023). Emerging trends in cloud computing: Virtualization and resource management. Cloud Technology Review, 14(3), 142-157.

Sharma, R., & Singh, A. (2022). Virtualization technologies for cloud infrastructure: Resource allocation and management strategies. Journal of Cloud Computing, 10(2), 34-47.

Wang, S., Zhang, H., & Li, F. (2023). Optimizing resource allocation in heterogeneous cloud systems. Cloud Computing Technology Journal, 17(4), 196-211.

Zhao, Z., Liu, Y., & Wei, T. (2022). Enhancing cloud resource management using machine learning: A survey. Journal of Cloud Technologies, 9(1), 65-80.

Zhang, F., Wang, X., & Yang, H. (2023). Energy-aware load balancing for green cloud computing: An approach based on dynamic optimization. Sustainable Cloud Computing, 8(2), 203-218.