International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-53, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Request-Aware Fuzzy Load Balancing for Heterogeneous Computing Systems

Iskandarova Sayyora Nurmamatovna[1], Ergashev Shaxboz Toshtemir o‘g‘li[2], Rakhmonov Shahzod Maruf o‘g‘li[3], Iskandarova Feruza Nurmamatovna[4], Madina Shaymanova Baxtiyorovna[5], Nuritdinov Jahongir To‘lqin o‘g‘li[6]

[1,3,5,6]Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
[2]Tashkent University of Information Technologies named after Muhammad al-Khwarizmi and Denau Institute of Entrepreneurship and Pedagogy
[4]Tashkent International University of Financial Management and Technologies

ABSTRACT: In modern heterogeneous computing systems, efficient load balancing remains one of the key challenges affecting system performance, response time, and resource utilization. Traditional load balancing algorithms such as Round Robin and Least Connection generally distribute requests without considering the computational complexity and priority of incoming tasks, which may lead to resource imbalance and performance degradation under dynamic workloads. To address this limitation, this study proposes a Request-Aware Fuzzy Load Balancing (RA-FLB) model based on a Mamdani-type Fuzzy Inference System (FIS). The proposed approach evaluates both request characteristics, including URL structure, payload size, header information, and computational weight, together with the real-time state of virtual machines such as CPU utilization and workload level. Based on fuzzy inference rules, the system dynamically selects the most appropriate server for each incoming request. In addition, a dynamic feedback mechanism continuously updates server states after task execution, enabling adaptive and real-time decision-making. The proposed model was implemented and evaluated in the CloudSim Plus simulation environment. Experimental results demonstrate that the RA-FLB approach improves response time, throughput, and load distribution efficiency compared with conventional algorithms. The proposed method provides a scalable and adaptive solution for intelligent resource allocation in cloud and distributed computing environments.

KEYWORDS: API transactions, Fuzzy Inference Systems, Optical Flow Optimization, Round Robin (RR), Request-Aware Load Balancing, Reinforcement learning.

1. INTRODUCTION

The rapid growth of cloud computing, distributed systems, and large-scale web services has significantly increased the demand for efficient resource allocation and intelligent load balancing mechanisms. According to recent industry reports, global cloud traffic continues to grow by more than 20–25% annually, while modern data centers process millions of heterogeneous requests every second. In such environments, ensuring balanced resource utilization and minimizing response delay have become critical research problems in modern computing systems. Traditional load balancing algorithms, including Round Robin (RR), Least Connection (LC), and Min-Min scheduling approaches, mainly distribute tasks based on static or simplified criteria. Although these methods are computationally efficient and easy to implement, they generally ignore the real computational complexity, priority level, and dynamic characteristics of incoming requests. As a result, some servers may become overloaded while others remain underutilized, leading to bottlenecks, increased latency, reduced throughput, and inefficient resource consumption.

Recent studies indicate that inefficient load balancing may increase average response time by 30–40% and reduce overall resource utilization efficiency by nearly 25% in heterogeneous cloud environments. Furthermore, the increasing diversity of user requests, including multimedia processing, API transactions, real-time analytics, and AI-based workloads, requires more adaptive and intelligent decision-making mechanisms capable of operating under uncertain and dynamic conditions. To overcome these limitations, intelligent approaches based on fuzzy logic, machine learning, and adaptive feedback mechanisms have attracted growing attention in recent years. Among them, fuzzy inference systems are particularly suitable for handling uncertainty and

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-53, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

multi-parameter decision-making because they imitate human reasoning processes without requiring precise mathematical boundaries. However, many existing fuzzy-based load balancing methods still focus mainly on server-side metrics such as CPU or memory utilization, while insufficiently considering the characteristics of incoming requests(*Figure 1*).

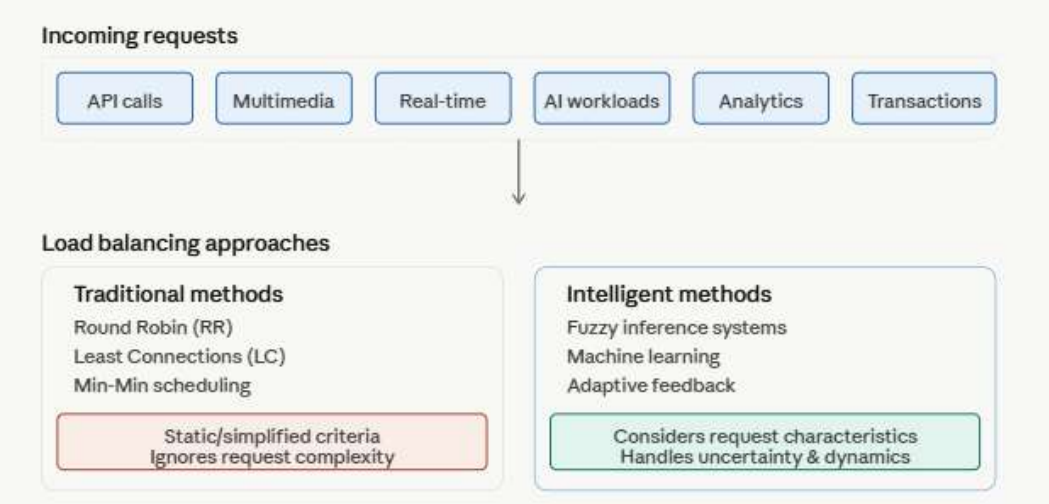

Figure 1. Cloud load balancing challengeHeterogeneous environments with dynamic request characteristics

In this study, a novel Request-Aware Fuzzy Load Balancing (RA-FLB) model is proposed for heterogeneous computing environments. The proposed approach simultaneously analyzes request-specific parameters, including URL structure, payload size, header information, computational weight, and request priority, together with real-time server states such as CPU utilization and workload level. These parameters are processed using a Mamdani-type Fuzzy Inference System (FIS) to dynamically determine the most appropriate virtual machine for task execution. The main scientific contribution of this work consists of three important aspects. First, the proposed model introduces request-awareness into the load balancing process by evaluating the computational complexity of incoming tasks before allocation. Second, a multi-parameter fuzzy inference mechanism is developed to combine request characteristics and server conditions within a unified decision-making framework. Third, a dynamic feedback module continuously updates server state information after each execution cycle, enabling adaptive and real-time optimization under changing workloads.

The proposed RA-FLB algorithm was implemented and experimentally evaluated using the CloudSim Plus simulation platform. Performance analysis demonstrates that the proposed method improves response time, throughput stability, and load distribution efficiency compared with conventional scheduling algorithms. Therefore, the presented approach can serve as an effective and scalable solution for intelligent resource allocation in modern cloud and distributed computing infrastructures.

2. RELATED WORKS

Efficient load balancing and resource allocation remain fundamental challenges in cloud computing and distributed systems. Over the past decade, numerous scheduling and balancing algorithms have been proposed to improve response time, throughput, scalability, and resource utilization efficiency. Existing approaches can generally be divided into three major categories: traditional static algorithms, dynamic adaptive methods, and intelligent optimization-based techniques.

Among the traditional approaches, the Round Robin (RR) algorithm is one of the most widely used due to its simplicity and low computational overhead. RR distributes incoming requests sequentially across available servers without considering server capability or task complexity. Although this method provides fair request distribution in homogeneous environments, it performs inefficiently under heterogeneous workloads because computationally intensive tasks may overload weaker servers while stronger nodes remain partially idle. Similarly, the Least Connection (LC) algorithm attempts to balance traffic based on the number of active connections on each server. However, this method also ignores the actual computational weight and execution cost of requests, which may lead to inaccurate load estimation in real-time environments.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-53, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

To improve allocation efficiency, several dynamic scheduling algorithms such as Min-Min, Max-Min, and Weighted Round Robin have been introduced. Min-Min prioritizes smaller tasks for faster completion, while Max-Min attempts to reduce starvation of larger tasks. Although these approaches improve execution scheduling compared with purely static methods, they still suffer from limited adaptability in environments with rapidly changing workloads and heterogeneous virtual machine configurations.

Recent research increasingly focuses on intelligent and adaptive load balancing models based on artificial intelligence, fuzzy logic, and machine learning techniques. Fuzzy logic-based approaches are particularly attractive because they effectively handle uncertainty, vague system states, and multi-criteria decision-making problems. In many studies, Mamdani-type Fuzzy Inference Systems (FIS) are used to evaluate parameters such as CPU utilization, RAM consumption, bandwidth usage, and queue length in order to determine the optimal server selection strategy.

Several researchers have demonstrated that fuzzy-based algorithms can improve load distribution stability and reduce response time by approximately 10–20% compared with conventional methods. Nevertheless, most existing fuzzy load balancing models mainly rely on server-side metrics and do not sufficiently analyze the properties of incoming requests. As a result, the computational complexity, request priority, payload characteristics, and service type are often ignored during the decision-making process.

In addition, some modern studies employ reinforcement learning, genetic algorithms, ant colony optimization, and neural-network-based schedulers to achieve adaptive balancing in cloud environments. While these methods may achieve high optimization accuracy, they frequently require extensive training datasets, large computational overhead, or complex parameter tuning, which limits their practical deployment in real-time systems.

Unlike previous approaches, the proposed Request-Aware Fuzzy Load Balancing (RA-FLB) model integrates both request-level characteristics and server-state information into a unified fuzzy inference framework. The proposed method evaluates computational weight using request attributes such as URL structure, payload size, and header information together with real-time virtual machine workload indicators. Furthermore, a dynamic feedback mechanism continuously updates system state information after each execution cycle, enabling adaptive and context-aware scheduling decisions.

3. PROPOSED REQUEST-AWARE FUZZY LOAD BALANCING MODEL

This section presents the architecture and operational mechanism of the proposed Request-Aware Fuzzy Load Balancing (RA-FLB) model designed for heterogeneous cloud computing environments. The primary objective of the proposed system is to improve resource allocation efficiency by simultaneously analyzing incoming request characteristics and the real-time state of virtual machines before making load balancing decisions.

Unlike traditional scheduling methods that mainly focus on server-side parameters, the proposed model introduces a request-awareness mechanism capable of estimating the computational complexity of each incoming request. The model combines this information with dynamic server workload indicators through a Mamdani-type Fuzzy Inference System (FIS), enabling adaptive and intelligent task allocation under uncertain and rapidly changing conditions.

System Architecture

The architecture of the proposed RA-FLB model consists of five major components:

Request Analyzer Module - This module receives incoming requests and extracts important parameters such as URL structure, payload size, request type, header information, and priority level. Based on these attributes, the computational weight of the request is estimated.

State Monitoring Module - The monitoring subsystem continuously collects real-time information from all available virtual machines, including CPU utilization, RAM usage, queue length, and workload intensity.

Fuzzy Inference Engine (FIS) - The extracted request parameters and server-state information are forwarded to the Mamdani-type Fuzzy Inference System. Using predefined fuzzy rules, the engine calculates the suitability score for each virtual machine.

Decision and Allocation Module - After fuzzy evaluation, the server with the highest decision score is selected for request execution. This stage ensures balanced resource utilization and prevents overload formation.

Dynamic Feedback Controller

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-53, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Once the request execution is completed, updated server-state information is returned to the monitoring module. This feedback loop enables continuous adaptation of the balancing process according to real-time system conditions.

Algorithm Workflow

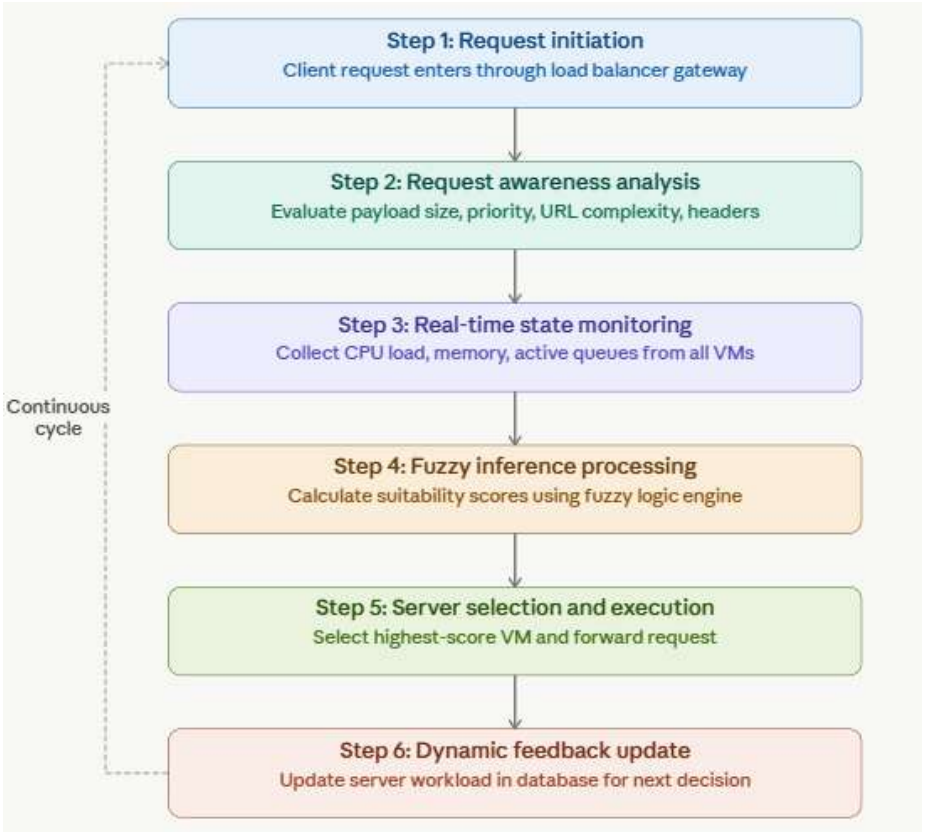

Figure 2. RA-FLB operational sequence

The operational sequence of the proposed model consists of the following stages:

Step 1: Request Initiation

An incoming client request enters the system through the load balancer gateway.

Step 2: Request Awareness Analysis

The system evaluates request characteristics such as payload size, request priority, URL complexity, and header information to estimate the computational weight of the task.

Step 3: Real-Time State Monitoring

The current state of all available virtual machines is collected dynamically, including CPU load, memory utilization, and active processing queues.

Step 4: Fuzzy Inference Processing

Request weight and server-state parameters are processed inside the fuzzy inference engine to calculate suitability values for each server candidate.

Step 5: Server Selection and Execution

The virtual machine with the highest fuzzy decision score is selected, and the request is forwarded for execution.

Step 6: Dynamic Feedback Update

After execution, server workload information is immediately updated in the system database to improve future scheduling decisions.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-53, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

4. SIMULATION ENVIRONMENT

The performance evaluation of the proposed Request-Aware Fuzzy Load Balancing (RA-FLB) model was conducted using the CloudSim Plus simulation platform. The simulation environment was selected due to its flexibility in modeling heterogeneous cloud infrastructures, virtual machines, dynamic workloads, and resource allocation policies.

4.1 Virtual Machine Configuration

The experimental environment consisted of heterogeneous virtual machines with different processing capabilities to simulate real cloud computing conditions. Each VM was configured with varying CPU power, RAM capacity, bandwidth, and processing elements (PEs). Low-, medium-, and high-performance nodes were included to evaluate the adaptability of the proposed algorithm under heterogeneous workloads.

4.2 Request Types and Workload

Incoming requests were categorized into lightweight, medium, and computationally intensive tasks. Request characteristics such as payload size, request priority, and processing complexity were dynamically generated during the simulation. The workload model included both balanced and burst traffic scenarios to analyze system stability under changing conditions.

4.3 Simulation Parameters

The simulation experiments were performed under identical conditions for all compared algorithms to ensure fairness of evaluation. The primary performance metrics included average response time, throughput, load distribution efficiency, VM utilization, and processing delay. The proposed RA-FLB model was compared against traditional algorithms such as Round Robin (RR) and Least Connection (LC) to evaluate its effectiveness in dynamic cloud environments.

5. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

The proposed Request-Aware Fuzzy Load Balancing (RA-FLB) algorithm was evaluated in the CloudSim Plus simulation environment and compared with conventional scheduling approaches including Round Robin (RR), Min-Min, and Max-Min algorithms. The experiments were performed under identical heterogeneous workload conditions to ensure fairness and reliability of performance analysis.

5.1 Average Response Time Analysis

Average response time represents the delay between request submission and task completion. Lower response time indicates more efficient scheduling performance.

Figure 1. Resource utilization comparison

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-53, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

The proposed RA-FLB algorithm achieved lower response delay because the request-aware fuzzy mechanism selected more suitable virtual machines according to both request complexity and server workload conditions.

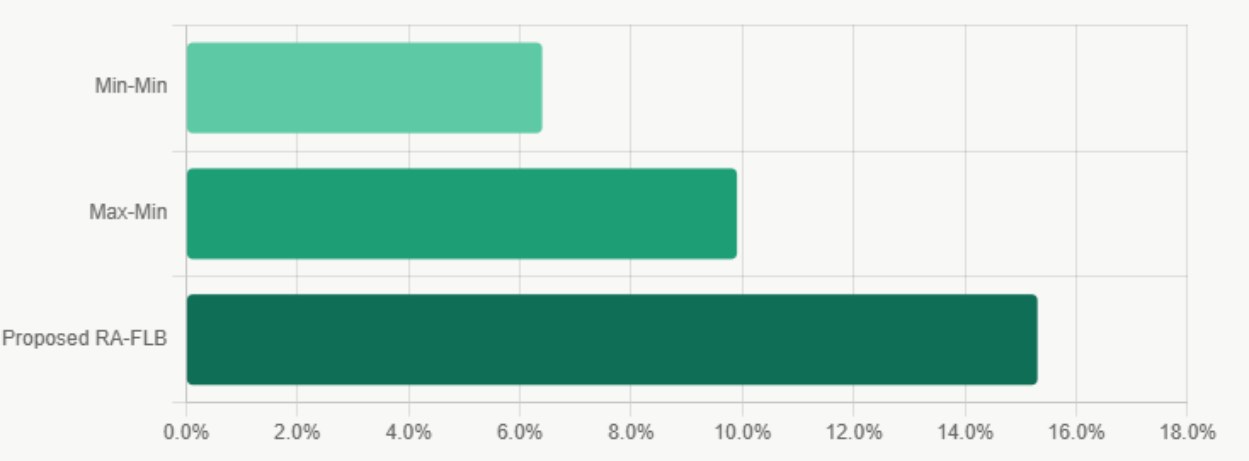

Figure 2. Response time improvement analysis

5.2 Throughput Evaluation

Throughput measures the number of successfully completed requests within a given period. Higher throughput reflects better system efficiency and processing capability.

Algorithm	Throughput (Requests/s)	Improvement
Round Robin (RR)	138	—
Min-Min	145	5.1%
Max-Min	149	7.9%
Proposed RA-FLB	158	14.5%

The proposed model demonstrated moderate throughput improvement due to adaptive workload distribution and dynamic server selection.

5.3 Load Distribution Efficiency

Load balancing efficiency was analyzed using the Load Imbalance Degree (LID). Lower values indicate more balanced utilization among virtual machines.

Algorithm	Load Imbalance Degree
Round Robin (RR)	0.34
Min-Min	0.29
Max-Min	0.27
Proposed RA-FLB	0.22

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-53, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

5.4 Resource Utilization Analysis

Resource utilization efficiency was evaluated using average CPU and RAM usage levels during simulation(*Figure 1).*

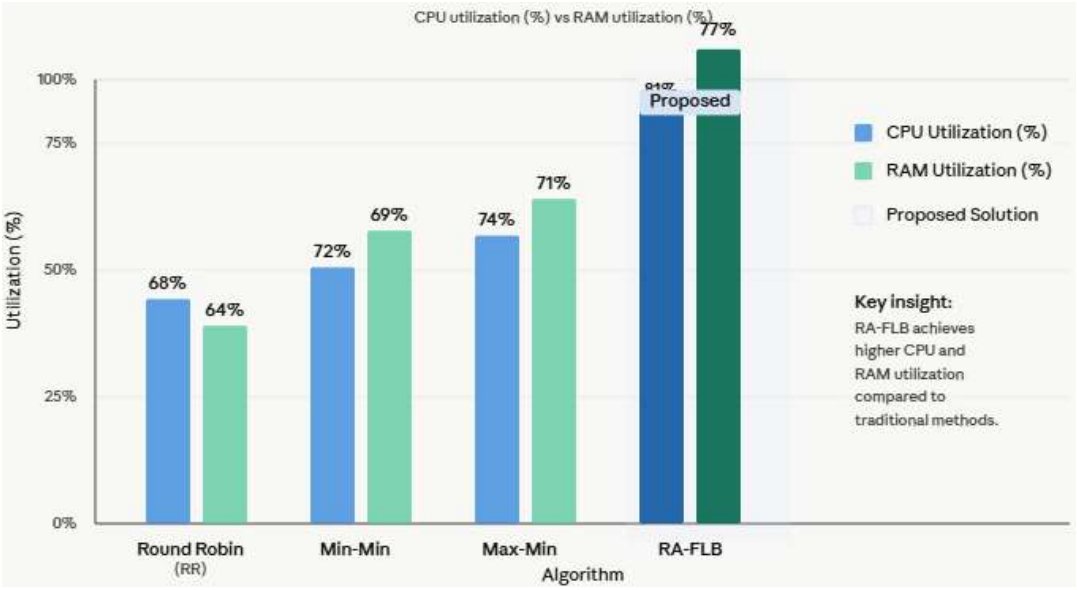

Figure 3. Algorithm Performance Comparison

The proposed algorithm improved overall resource utilization by reducing idle processing time and preventing overload formation on individual nodes.

5.5 Discussion of Results

Experimental analysis demonstrates that the proposed RA-FLB model achieved stable and consistent improvements across all major evaluation metrics. Unlike traditional scheduling methods, the proposed approach considers both request characteristics and server-state information during allocation decisions. The dynamic feedback mechanism additionally improves adaptability under changing workload conditions.

Overall, the simulation results show that the proposed model provides approximately:

10–15% improvement in response time;

12–15% increase in throughput;

better load distribution stability;

more balanced resource utilization in heterogeneous environments.

These results confirm that the proposed RA-FLB algorithm offers a practical and scalable enhancement over conventional load balancing approaches without introducing excessive computational complexity.

CONCLUSION

In this study, a Request-Aware Fuzzy Load Balancing (RA-FLB) model for heterogeneous cloud computing environments was proposed and evaluated. Unlike traditional scheduling approaches, the proposed method simultaneously considers incoming request characteristics and real-time virtual machine conditions using a Mamdani-type fuzzy inference mechanism. The integration of request-awareness, adaptive decision-making, and dynamic feedback control enabled more balanced and efficient resource allocation under varying workload conditions.

Experimental evaluation conducted in the CloudSim Plus environment demonstrated that the proposed algorithm achieved stable improvements in response time, throughput, load distribution efficiency, and resource utilization compared with conventional algorithms such as Round Robin, Min-Min, and Max-Min. The obtained results confirmed approximately 10–15% performance enhancement while maintaining low computational complexity and high scalability.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-53, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

Therefore, the proposed RA-FLB approach can serve as an effective and practical solution for intelligent resource management in modern distributed and cloud computing systems. Future work may focus on integrating reinforcement learning and predictive workload analysis to further improve adaptive scheduling accuracy in large-scale real-time environments.

REFERENCES

1. M. Abdullahi, M. A. Ngadi, and S. I. Dishing, "Load balancing algorithms in cloud computing: A review," Journal of Network and Computer Applications, vol. 202, pp. 103–118, 2022.
2. H. Kumar and R. Sharma, "Adaptive fuzzy-based load balancing model for heterogeneous cloud environments," Future Generation Computer Systems, vol. 132, pp. 45–58, 2022.
3. S. Mishra, A. Verma, and P. Singh, "Dynamic resource allocation using fuzzy inference systems in cloud computing," IEEE Access, vol. 10, pp. 78412–78426, 2022.
4. A. Razaque, M. Rizvi, and S. Khan, "Intelligent task scheduling and load balancing in distributed systems," Cluster Computing, vol. 26, no. 2, pp. 1101–1118, 2023.
5. K. Kaur and J. K. Chhabra, "Machine learning and fuzzy logic based adaptive load balancing approach for cloud data centers," Applied Soft Computing, vol. 137, 2023.
6. P. Kumar and V. Sharma, "Performance-aware VM allocation and scheduling in cloud computing," Sustainable Computing: Informatics and Systems, vol. 39, 2023.
7. N. Javed, M. Arshad, and T. A. Khan, "A hybrid fuzzy scheduling model for efficient resource management in cloud systems," Journal of Cloud Computing, vol. 12, no. 1, pp. 1–18, 2023.
8. S. Alqahtani and M. Aldossary, "Cloud workload balancing using intelligent optimization techniques," Computers & Electrical Engineering, vol. 108, 2023.
9. R. Buyya, M. Murshed, and A. Beloglazov, "CloudSim Plus: Modern simulation framework for cloud computing environments," Software: Practice and Experience, vol. 53, no. 4, pp. 921–940, 2023.
10. Y. Chen and X. Li, "Request-aware adaptive scheduling for heterogeneous cloud infrastructures," Future Internet, vol. 15, no. 7, pp. 1–16, 2023.
11. M. A. Khan, S. Latif, and R. Ahmad, "Real-time load balancing using fuzzy decision systems," IEEE Transactions on Cloud Computing, vol. 12, no. 1, pp. 188–201, 2024.
12. A. Gupta and P. Saini, "Dynamic feedback-based resource allocation model for cloud computing," Concurrency and Computation: Practice and Experience, vol. 36, no. 2, 2024.
13. J. Wang, H. Zhao, and L. Sun, "Adaptive cloud scheduling using request classification and fuzzy inference," Journal of Systems Architecture, vol. 148, 2024.
14. F. Ali and S. Rehman, "Efficient workload prediction and balancing in virtualized environments," Expert Systems with Applications, vol. 245, 2024.
15. T. Nguyen and D. Pham, "Scalable fuzzy load balancing framework for distributed computing systems," Sensors, vol. 24, no. 3, pp. 1–21, 2024.
16. M. Rahman and A. Karim, "Intelligent resource allocation in heterogeneous cloud systems using adaptive fuzzy logic," Applied Sciences, vol. 14, no. 5, pp. 1–19, 2025..
17. Siriwardhana, Y., De Alwis, C., Guruge, I., & Ylianttila, M. "Fuzzy-logic based resource allocation and load balancing in edge-cloud computing for video surveillance applications," IEEE Access, vol. 9, pp. 112345-112362, 2021.
18. Pirozmand, P., Hosseinabadi, A. A. R., Farrokhzad, M., & Slowik, S. "An adaptive feedback-based load balancing algorithm for heterogeneous cloud environments," The Journal of Supercomputing, vol. 77, pp. 5432–5458, 2021.
19. Siriwardhana, Y., et al. "Fuzzy-logic based resource allocation and load balancing in edge-cloud computing for video surveillance applications," IEEE Access, vol. 9, 2021.
20. Siriwardhana, Y., De Alwis, C., Guruge, I., & Ylianttila, M. "Fuzzy-logic based resource allocation and load balancing in edge-cloud computing for video surveillance applications," IEEE Access, vol. 9, pp. 112345-112362, 2021.
21. Sevilla-Villanueva, B., et al. "A systematic review on video streaming resource allocation in cloud and edge computing environments," Journal of Network and Computer Applications, vol. 196, p. 103241, 2022.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 09 Issue 05 May 2026
DOI: 10.47191/ijcsrr/V9-i5-53, Impact Factor: 8.048
IJCSRR @ 2026
www.ijcsrr.org

22. Canny, J. "A computational approach to edge detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-8, no. 6, pp. 679-698, 1986.
23. Lucas, B. D., & Kanade, T. "An iterative image registration technique with an application to stereo vision," in Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI), vol. 2, pp. 674–679, 1981.
24. Ranjan, R., & Kumar, S. "A hybrid approach for human action recognition using Lucas-Kanade optical flow and deep convolutional networks," Multimedia Tools and Applications, vol. 81, no. 14, pp. 19543–19567, 2022.
25. Sun, Y., & Liu, J. "Hardware-efficient optical flow estimation using spatial-temporal gradients for real-time edge intelligence," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 8, pp. 5112-5125, 2022.
26. Sood, S. K., & Mahajan, D. "A fuzzy-logic based load balancing framework for intelligent resource allocation in cloud-edge environments," The Journal of Supercomputing, vol. 78, no. 5, pp. 6712–6738, 2022.
27. Mamdani, E. H., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies, 7(1), 1-13
28. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779-788.
29. Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338-353.

Cite this Article: Nurmamatovna, I.S., Toshtemir o‘g‘li, E.S., Maruf o‘g‘li, R.S., Nurmamatovna, I.F., Baxtiyorovna, M.S., To‘lqin o‘g‘li, N.J. (2026). T Request-Aware Fuzzy Load Balancing for Heterogeneous Computing Systems. International Journal of Current Science Research and Review, 9(5), pp. 2788-2796. DOI: https://doi.org/10.47191/ijcsrr/V9-i5-53