ISSN : 2350-0743

International Journal of Recent Advances in Multidisciplinary Research

Vol. 12, Issue 03 pp.10929-10932, March, 2025

REVEIW ARTICLE

ADAPTIVE LOAD BALANCING ALGORITHM FOR OPTIMIZED RESOURCE ALLOCATION IN CLOUD DATA CENTRES

Thirusubramanian Ganesan[1], Mohanarangan Veerapperumal Devarajan[2], Akhil Raj Gaius Yallamelli[3], Vijaykumar Mamidala[4], Rama Krishna Mani Kanta Yalla[5] and Veerandra Kumar R[6],

[1]Cognizant Technology Solutions, Texas, USA; [2]Ernst & Young (EY), Sacramento, USA; [3]Amazon Web Services Inc, Seattle, USA; [4]Conga (Apttus), Broomfield, CO, USA; [5]Amazon Web Services, Seattle, WA, USA; [6]Saveetha Engineering College, Saveetha Nagar, Thandalam, Chennai

ARTICLE INFO

Article History

Received 19th December, 2024
Received in revised form
17th January, 2025
Accepted 26th February, 2025
Published online 28th March, 2025

Keywords:

Adaptive Load Balancing, Resource Allocation, Cloud Computing, Virtual Machine Scheduling, Energy Efficiency, Dynamic Workload Management.

Corresponding author:
Thirusubramanian Ganesan

ABSTRACT

Efficient resource allocation is crucial for cloud data centers performance, scalability, and cost effectiveness. Traditional load balancing systems frequently fail to adjust changing workloads resulting in poor resource use, increased latency and excessive energy consumption. Adaptive Load Balancing Algorithm that optimizes resource allocation in cloud systems is implemented. It combines heuristic and metaheuristic techniques to efficiently distribute workloads among virtual machines resulting in lower response times and balanced computing loads. Centralised Control System monitors real time system data and dynamically reallocates resources in response to workload needs. Performance evaluation shows suggested solution reduces delay by 35 percentage from 120 ms to 85 ms and reduces energy consumption by 30 percent from 200W to 140W. This increases efficiency of resources from 65 to 98 percent. Suggested adaptive technique improves distribution of loads, energy usage and level of service in cloud computing settings. It provides fault tolerance by dynamically shifting jobs during server failures, reducing downtime and increasing system dependability. Comparative research with existing load balancing approaches confirms its advantages in managing large scale dynamic workloads demonstrating suitability for implementation in cloud data centers.

Citation: Thirusubramanian Ganesan, Mohanarangan Veerapperumal Devarajan, Akhil Raj Gaius Yallamelli, Vijaykumar Mamidala, Rama Krishna Mani Kanta Yalla, Veerandra Kumar R. 2025. "Adaptive Load Balancing Algorithm for Optimized Resource Allocation in Cloud Data Centres", *International Journal of Recent Advances in Multidisciplinary Research,* 12,(03), 10929-10932.

INTRODUCTION

In today's dynamic cloud computing world precise workload planning is critical for optimal resource management (Devarajan, 2022). Predicting computing resource consumption is vital for optimal cloud performance and cost effectiveness (Yallamelli, 2019). Cloud computing offers scalable resources for largescale data processing and storage (Valivarthi). Itenables indepth data analysis, longterm storage and combining several services (Narla, 2020). Analyzing prior data flow predict potential harm and illicit conduct (Yalla, 2021). RBM-BiGRU model combines Restricted Boltzmann Machines for hierarchical information illustration and BiGRU for precise load prediction. Spark based Density Peak Clustering helps detect anomalies while Fuzzy C-Means Clustering detects load curve periodicity. Short-term load forecasting is difficult due to load uncertainty (Alagarsundaram *et al*., 2024). Cloud Testing Adoption Assessment Model by utilizing FMCDM and empirical survey approaches assess and successfully apply cloud testing tactics inside framework for maintaining resources (Gattupalli, 2024). Load balancing based on real time data is achieved through Incremental Gaussian Mixture Model (Kadiyala). By dynamically adjusting number of components ensures efficient resource distribution among nodes.

Problem Statement: Balancing effectiveness and security while protecting sensitive data flows remains a challenge (Samudrala, 2020). Handling bigger datasets and more intricate computations in PMDP causes scalability problem (Gollavilli, 2022). Cloud storage customized cache for data intensive scientific computing workflow has drawback in frequent reliance on data reduction techniques causing workflow bottlenecks and inadequately addressing capacity demands of large scale scientific datasets (Narla, 2024).

OBJECTIVE

- Develop adaptive load balancing algorithm that constantly optimises resource allocation in cloud data centres to improve performance and efficiency.

- Establish a sophisticated scheduling mechanism use heuristic and metaheuristic approaches to distribute workloads among virtual machines.
- Analyze the effect of adaptive load balancing on latency, energy consumption and resource usage efficiency.

LITERATURE SURVEY

Ganesan et al. (2024) investigates Distributed Learning (DL) techniques and Network Slicing (NS) for optimizing 6G vehicular environments' resource management based on a DL-as-a-Service (DLaaS) paradigm. The model exploits edge-cloud capabilities and dynamic provisioning strategies for proactive and efficient deployment of DL (Devarajan). The strategy optimizes network flexibility, intelligence, and performance while improving traffic handling and customer experience. Ganesan, Almusawi, et al. (2024) familiarizes improved resource allocation and task scheduling methods usingEnhanced Bat Optimization Algorithm (IBOA) with dynamic weights and Modified Social Group Optimization (MSGO) technique. Simulations protest that the proposed methods outperform existing approaches like MOTSGWO attaining better energy efficiency, response time and resource utilization. Kadiyala (2019) introduces a hybrid DBSCAN, fuzzy C-Means, and ABC-DE optimization clustering model for improved resource utilization and safe data exchange in IoT-based fog computing. The proposed method has higher clustering accuracy (93%), latency (18 ms), and optimized bandwidth use (85%) compared to the existing approach. Model has improved efficient, secure IoT data processing with high compliance (88%) and access control (93%) (Gudivaka, 2021). Cloud computing has transformed IT administration by offering elastic access to storage, applications, and compute power. Allur, (2021) presents a novel AI-based load-balancing technique utilizing edge computing and machine learning is presented to optimize resource utilization in dynamic cloud infrastructures. The suggested method selectively loads workloads on data centres and virtual machines based on proximity, leading to improved scalability, efficiency, and system responsiveness (Narla, 20221). Sareddy and Khan (20250) resolves the sparsity problem in recommendation engine collaborative filtering systems through the application of graph neural networks (GNNs) to generate early user clusters. Employing GitHub as a study case, the model improves individualized recommendations in human resource management with increased accuracy, recall, and F-measure. It successfully overcomes sparsity, assisting project managers in making accurate HR decisions. Kethu et al. (2025) introduces AI-based system that combines Social Determinants of Health (SDOH), Electronic Health Records (EHRs), Multi-Omics Data, and Resource Optimization Models to improve geriatric care. System enhances the management of chronic diseases, maximizes resource utilization, and provides scalable and equitable care with 94% accuracy and a 95% F1 score. By solving systemic inefficiencies, the model revolutionizes elderly healthcare delivery.

ADAPTIVE LOAD BALANCING ALGORITHM FOR OPTIMIZED RESOURCE ALLOCATION: Adaptive Load Balancing Method illustrated in Figure 1 starts with user requests being routed via Load Balancer that distributes jobs depending on real time available resources. Load Monitoring module monitors CPU, memory and bandwidth utilization and sends data to Centralized Control System. CCS examines resource consumption using optimization approaches allowing Decision Module to modify job distribution. Jobs are given to VMs on server or nodes resulting in balanced workloads, lower latency and better Quality of Service.

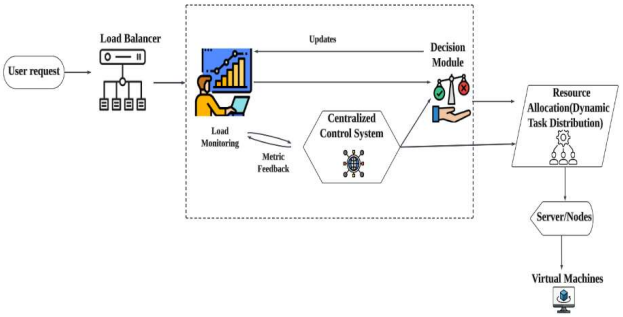

Figure 1. Architecture of adaptive load balancing

Data Collection: Data gathering method uses real time logging from virtual machines, servers and network traffic to monitor critical cloud parameters like resource usage, workload intricacy, cost and performance. CPU, memory, disk I/O and bandwidth metrics are continually monitored to optimize resource allocation.

User Request: When user submits request the Load Balancer allocates workloads according to resource demand and availability. Request is processed and sent to the proper server or VM for processing. Load Monitoring System monitors real time indicators and sends data to CCS to continually optimize resource allocation and ensure QoS.

$$P(k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \quad (1)$$

Where λ average request arrival rate

Load Balancer: Load Balancer distributes incoming requests across many servers to guarantee resource use and prevent overload. Balancing algorithm decides how to assign jobs using various techniques like Round Robin, Least Connection or Weighted Load Distribution.

$$L_i = \frac{W_i}{\sum_{j=1}^{n} W_j} \times T \quad (2)$$

Where W_i weight assigned and T is total number of tasks

Load Monitoring: They regularly monitor each servers current utilization state. Load balancing decisions are based on metrics such as CPU use, memory consumption, network speed and response time.

$$LU = \frac{CPU_{used}+MEM_{used}+BW_{\boldsymbol{used}} \quad (utilized\ resources)}{CPU_{total}+MEM_{total}+BW_{\boldsymbol{total}} \quad (total\ available\ resources)} \quad (3)$$

Centralized Control System: CCS serves as coordinator gathering metric data from Load Monitoring and make work allocation choices using established rules. It employs RL or Optimization Algorithms such as Genetic Algorithms and PSO to optimize resource allocation.

$$C = \sum_{i=1}^{n} (\alpha \times E_i + \beta \times T_i + \gamma \times U_i) \quad (4)$$

Where E_i energy consumption, T_i Task completion time and U_i resource utilization
α, β, γ are weight factors

Decision Module: Decision Module optimizes cloud resource management by processing changes from CCS. It uses system parameters like workload complexity, availability of assets *R*and performance scores to make choices. It assigns additional resources, redistributes workloads or dynamically scales resources to ensure maximum utilization, cost efficiency and QoS. They improve cloud performance by utilizing heuristic approaches ensures load balancing and avoid resource bottlenecks.

$$D = \omega_1 U + \omega_2 R + \omega_3 Q \quad (5)$$

Where $\omega_1, \omega_2, \omega_3$ are priority weights

Resource Allocation: Resource Allocation Module periodically distributes tasks to VMs by selecting best option from option Module.

Minimum Completion Time$MCT(T_j) = \min_{VM_i}(A_i + E_{i,j})$ (6)

Where A_i available processing time and $E_{i,j}$ estimated execution time

Server/Nodes & Virtual Machines: Activities assigned are correctly carried out by VMs when working on actual servers or nodes in cloud architecture.

RESULT AND DISCUSSION

Dataset Description: Cloud Resource Management Dataset (https://www.kaggle.com/datasets/bhagvendersingh/cloud-resource-management-dataset) from Kaggle includes critical parameters for resource allocation, usage, cost and operational improvement in cloud systems. It compares baseline resource pool, workload difficulty, utilization, cost per unit and achievement score before and after reduction. It also monitors enhanced resource allocation, increased utilization and cost efficiency with utilization enhancement quantifying efficiency improvements and allowing for evaluation of cloud load balancing and handling resources solutions.

Performance Analysis of Proposed Work: Existing method has latency 120 ms but Proposed Method obtains 85 ms latency shows increased processing efficiency. This decrease shows that adaptive load balancing technique improves resource allocation resulting in quicker task execution and better cloud performance as displayed in Figure 2.

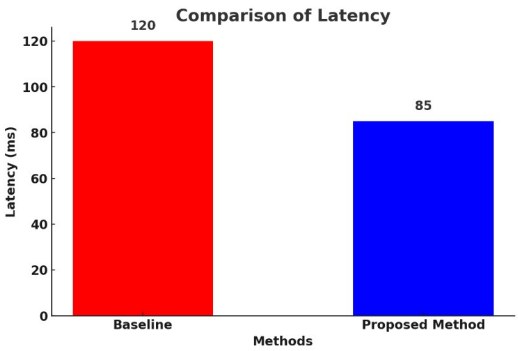

Figure 2. Latency comparison

Energy consumption and resource utilization efficiency are compared between baseline and proposed methods in Figure 3. Proposed method lowers energy usage from 200W to 140W exhibiting increased power efficiency. Resource Utilization Efficiency rises from 65 to 98 percent indicating improved task allocation and effective resource utilization.

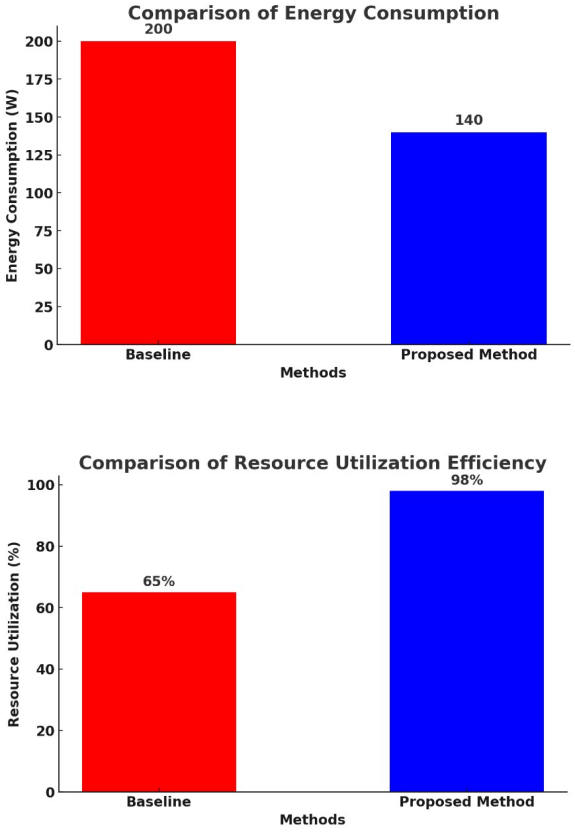

Figure 3. Energy consumption and resource utilization comparison

Table 1. Performance metrics

Metric	Proposed method
Latency (ms)	85
Energy Consumption (W)	140
Resource Utilization (%)	98

CONCLUSION

Adaptive Load Balancing Algorithm optimize resource allocation in cloud data centers addresses major obstacles like poor workload transportation, latency issues and excessive energy consumption. Framework combines real time tracking, heuristic improvement and dynamic task reallocation to improve cloud performance. Result support algorithms efficacy revealing reduction in latency from 120 ms to 85 ms, drop in energy expenditure from 200W to 140W and boost in resource utilization to 98. This implies adaptive load balancing greatly enhance cloud computing performance ensuring higher scalability and dependability. It promotes fault recovery and resilience providing uninterrupted services even under shifting workload situations. Future work includes incorporating AI powered predictive analytics, improving security procedures and expanding the method to edge computing settings. By combining self learning capabilities adapt model to changing cloud topologies enhancing productivity in large scale

distributed computation while retaining energy efficiency and great Quality of Service.

REFERENCES

Devarajan M. V. and C. Solutions, "An improved bp neural network algorithm for forecasting workload in intelligent cloud computing," vol. 10, no. 9726, 2022.

Yallamelli, A. R. G. "Wipro Ltd, Hyderabad, Telangana, India," vol. 7, no. 9726, 2019.

Valivarthi, D. T. "Optimizing cloud computing environments for big data processing," Int. J. Eng., vol. 13, no. 2.

Narla, S. "Transforming smart environments with multi-tier cloud sensing, big data, and 5g technology," vol. 5, 2020.

Yalla, R. K. M. K. "Cloud-Based Attribute-Based Encryption and Big Data for Safeguarding Financial Data," Int. J. Eng. Res. Sci. Technol., vol. 17, no. 4, pp. 23–32, Oct. 2021.

Alagarsundaram, P.S. K. Ramamoorthy, D. Mazumder, V. Malathy, and M. Soni, "A Short-Term Load Forecasting model using Restricted Boltzmann Machines and Bi-directional Gated Recurrent Unit," in 2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON), Aug. 2024, pp. 1–5. doi: 10.1109/NMITCON62075.2024.10699152.

Gattupalli, K. "A Survey on Cloud Adoption for Software Testing: Integrating Empirical Data with Fuzzy Multicriteria Decision-Making," vol. 10, no. 4, 2022.

Kadiyala B. and H. Kaur, "Dynamic load balancing and secure iot data sharing using infinite gaussian mixture models and plonk," vol. 7, no. 2.

Samudrala, V. K. "AI-powered anomaly detection for cross-cloud secure data sharing in multi-cloud healthcare networks," Curr. Sci., 2020.

Gollavilli, V. S. B. H. "PMDP: A Secure Multiparty Computation Framework for Maintaining Multiparty Data Privacy in Cloud Computing," J. Sci. Technol. JST, vol. 7, no. 10, Art. no. 10, Dec. 2022.

Narla, S. "A Blockchain-Based Method for Data Integrity Verification in Multi-Cloud Storage Using Chain-Code and HVT," vol. 12, no. 1, 2024.

Ganesan, T. R. R. Al-Fatlawy, S. Srinath, S. Aluvala, and R. L. Kumar, "Dynamic Resource Allocation-Enabled Distributed Learning as a Service for Vehicular Networks," in 2024 Second International Conference on Data Science and Information System (ICDSIS), Hassan, India: IEEE, May 2024, pp. 1–4. doi: 10.1109/ICDSIS61070.2024.10594602.

Devarajan, M. V. A. R. G. Yallamelli, R. K. M. Kanta Yalla, V. Mamidala, T. Ganesan, and A. Sambas, "Attacks classification and data privacy protection in cloud-edge collaborative computing systems," Int. J. Parallel Emergent Distrib. Syst., vol. 0, no. 0, pp. 1–20, doi: 10.1080/17445760.2024.2417875.

Ganesan, T. M. Almusawi, K. Sudhakar, B. R. Sathishkumar, and K. S. Kumar, "Resource Allocation and Task Scheduling in Cloud Computing Using Improved Bat and Modified Social Group Optimization," in 2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON), Bengaluru, India: IEEE, Aug. 2024, pp. 1–5. doi: 10.1109/NMITCON62075.2024.10699250.

Kadiyala, B. "Integrating dbscan and fuzzy c-means with hybrid abc-de for efficient resource allocation and secured iot data sharing in fog computing," Int. J. HRM Organ. Behav., vol. 7, no. 4, pp. 1–13, 2019.

Gudivaka, R. L. "A Dynamic Four-Phase Data Security Framework for Cloud Computing Utilizing Cryptography and LSB-Based Steganography," Int. J. Eng. Res. Sci. Technol., vol. 17, no. 3, pp. 90–101, Aug. 2021.

Allur, N. S. "Optimizing Cloud Data Center Resource Allocation with a New Load-Balancing Approach," Int. J. Inf. Technol. Comput. Eng., vol. 9, no. 2, pp. 188–201, 2021.

Narla, S. S. Peddi, and D. T. Valivarthi, "Optimizing Predictive Healthcare Modelling in a Cloud Computing Environment Using Histogram-Based Gradient Boosting, MARS, and SoftMax Regression," Int. J. Manag. Res. Bus. Strategy, vol. 11, no. 4, pp. 25–40, Nov. 2021.

Sareddy M. R. and S. Khan, "AI-Driven Human Resource Management: Enhancing Transparency and Security with Machine Learning," J. Artif. Intell. Capsule Netw., vol. 6, no. 4, pp. 512–528, Jan. 2025, doi: 10.36548/jaicn.2024.4.009.

Kethu, S. S. D. R. Natarajan, S. Narla, D. T. Valivarthi, and S. Peddi, "Innovative AI Applications in Healthcare: Integrating SDOH, EHRs, Multi-Omics Data, and Resource Optimization Models for Geriatric Chronic Care," J. Inf. Technol. Digit. World, vol. 6, no. 4, pp. 401–417, Jan. 2025, doi: 10.36548/jitdw.2024.4.007.

"Cloud Resource Management Dataset." Accessed: Mar. 01, 2025. [Online]. Available: https://www.kaggle.com/datasets/bhagvendersingh/cloud-resource-management-dataset

****