International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 11, Issue: 05 (Sept-Oct, 2025)

Dynamic Resource Management and Load Balancing in Cloud Computing using Efficient Hierarchical Clustering Techniques

Varsha Tiwari [1], A.C. Nayak [2], Gagan Sharma [3], Deepak pathak [4],
[2] Department of Computer Science & Engineering,
RKDF University, Bhopal, India

Abstract

The cloud computing environments demand efficient resource management and load balancing to optimize performance, scalability, and cost-effectiveness. This research proposes a novel approach to dynamic resource management and load balancing using efficient hierarchical clustering techniques. By leveraging the hierarchical clustering, the proposed framework dynamically groups cloud resources based on workload patterns, resource utilization, and system demands. This employs a multi-level clustering strategy to categorize virtual machines and tasks, which enables the adaptive resource allocation and load distribution. Simulations results demonstrate that the proposed technique significantly improves resource utilization, reduces response times, and enhances system scalability compared to traditional load balancing algorithms. The proposed approach also minimizes the energy consumption and operational costs, which makes it suitable for large-scale cloud infrastructures. This research contributes to advancing cloud computing efficiency, which offeres a scalable and robust solution for dynamic resource management in heterogeneous cloud environments.
Keywords: Load Balancing, Resource Allocation, Virtual Machine, Cluster, E-stab, Round Robin

1 Introduction

Cloud computing is a growing model of business IT infrastructure that delivers information and services over the Internet, which is easily accessible to users through a web browser [1]. It enables access to resources such as infrastructure, platforms, software, services, or storage in a flexible, scalable manner based on application needs. This cloud model basically reduces the requirements for organizations to invest in expensive computing hardware, software, and network bandwidth [2, 3].

Generally, cloud computing refers to the delivery of hosted services over the Internet. It mainly relies on the concept of shared computing resources, storage, networks, and applications provided by the third-party vendors as illustrated in the Figure 1. The cloud deployments generally acquire several forms, which includes the private, the public, and the hybrid clouds. In the private cloud, their infrastructure is basically dedicated to a single organization and those are controlled and managed either internally or by other external service providers. This cloud service is usually hosted on-premises or at the different remote locations [4].

A public cloud generally offers services to any user over the Internet with the infrastructure owned and operated by a company that provides cloud services. However, a private cloud has a dedicated network or data center that only delivers hosted services to a specific individual or group, which is generally within a single organization [5]. A hybrid cloud takes advantages of both models (private and public), a combined in-house resources with those from external providers. This hybrid setup allows the organizations to gain rich benefits from the scalability, efficiency, cost-effectiveness, and flexibility of the cloud computing system.

Cloud computing represents a shift from relying on individual computers or servers to using a “cloud” which is basically a collection of virtual servers that deliver computing resources on demand.The users do not require to manage or understand the underlying technological infrastructure behind it. The services and data are hosted in highly scalable data centers and which can be accessed globally from anywhere using internet-connected device [6].

Essentially, cloud computing delivers IT capabilities as services over the Internet, billed based on usage. This model has gained significant traction, with major providers like Microsoft, Google, IBM, Yahoo, and Amazon (a pioneer in the field) that offers cloud solutions. Even smaller businesses, such as SmugMug - a photo hosting platform, also use cloud services to manage data and power their operations. Cloud computing is being adopted across various domains, including web hosting, parallel graphics processing, financial modeling, web mining, and genomic analysis [7].

International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 11, Issue: 05 (Sept-Oct, 2025)

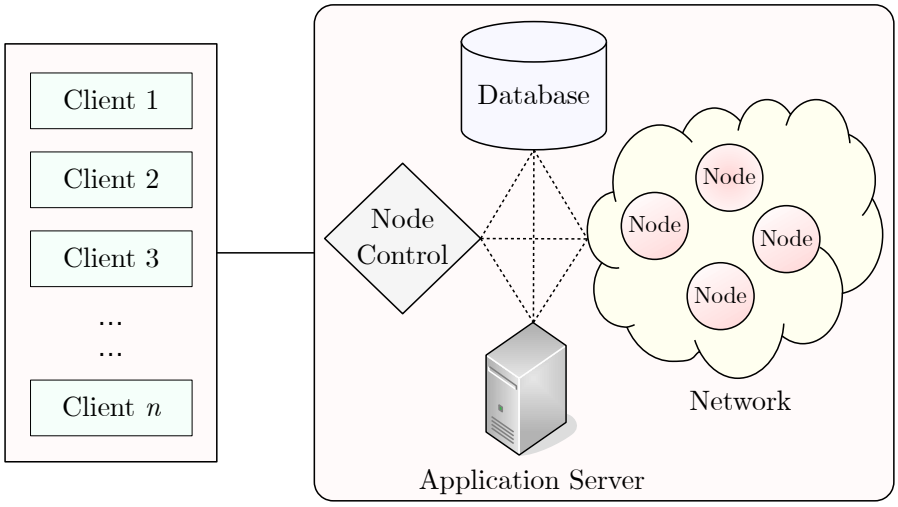

Figure 1: Cloud Network

1.1 Advantages of Cloud Computing

Computers have become an essential part of everyday life, which plays a critical role in nearly every field, from work to research, from customer to manufacturing, and from end user to server. As our reliance on computers grows, so it demands for high computing resources [8]. Large global companies such as Google, Amazon, IBM, and Microsoft can scale their resources easily as required, but the small businesses can face some key concerns such as accessibility and affordability [9]. For smaller organizations, hardware failures, such as machine breakdowns or hard drive crashes, and software bugs can create serious challenges. Cloud computing usually provides an effective solution to solve these issues [10, 11].

Cloud computing has seen a tremendous growth in both personal and business applications. It allows users to access and utilize resources online without any requirements for heavy local infrastructure [12]. Among its many benefits, ther are several key advantages that stand out as discussed below:

- **Scalability:** Scalability refers to a system capability to handle increased workload by expanding its resources, such as hardware, servers, storage, or network components. In the cloud computing services, the users can easily scale resources up or down based on the utilized requirements, without any essential demand to quickly buying the additional physical infrastructures.

- **Virtualization:** In cloud services, virtualization basically allows users to access and interact with resources as if they are dedicated, regardless of the actual physical location or hardware. This means services can be delivered efficiently using fewer physical resources, that offers a more cost-effective and flexible solution for users.

- **Mobility:** Cloud computing enables mobility, which allows users to access applications via the Internet anytime and from anywhere.

- **Low Infrastructure Cost:** The cloud supports a pay-per-use model, which can allow the organizations to only pay for the particular resources which they are actually using, rather than investing in and maintaining their own infrastructure.

Cloud computing also offers several advantages such as cloud service providers avoid infrastructure maintenance and upgrade costs, while users benefit from increased storage capacity compared to personal systems. This reduces the requirements to buy extra storage and enhances local system performance. Moreover, data and applications stored in the cloud are accessible anytime over the Internet [13, 14].

1.2 Limitations of Cloud Computing

Cloud computing can offer various benefits in the form of the elasticity, performance, availability and scalability on demand [15, 16]. However, there are some limitations or restrictions as listed below:

- **Low Latency:** Latency is a critical factor in telecommunications for voice, video, and data transmission. Since cloud services rely on Internet access, so the high latency can significantly impact communication between clients and providers.

- **Language Adaptation:** Compatibility with programming languages and platforms remains a challenge. Most cloud providers currently support only

specific languages or platforms, that often limits the interoperability with others. Therefore, establishment of the universal standards for language and platform adaptation is necessary step.

- **Resource Control:** Clients often have limited control over cloud resources, which can vary depending on the provider. Isolating specific resources can be difficult, and resource shortages may require migrating data or services to other systems. Effective resource management and load balancing, including dynamic migration, are essential challenges in cloud computing.

2 Related Work

This section presents general survey on the various methodologies regarding Cloud Computing. After Studying number of researches there is research gap of those work in the era of cloud computing, its approaches and load balancing in the cloud computing.

Hayyolalam and Özkasap [17] proposed a novel load-balancing method which is called the "CBWO", which combines "Chaos theory with the Black Widow Optimization" algorithm to address the challenges in workload distribution and resource allocation in the cloud computing. This method aims to enhance energy efficiency and resource utilization. Simulations using CloudSim show that CBWO outperforms existing approaches, with average improvements of 67.28% in makespan and 29.03% in energy consumption.

Liu *et al.* [18] introduced a "Load-Aware Switch Migration (LASM)" mechanism to improve controller load balancing in edge–cloud networks. Unlike existing methods, LASM considers both migration costs and the potential overload of target controllers. They used a knapsack-based model and a greedy algorithm to optimize switch selection and migration. Their experiments represented LASM significantly enhanced performance, reducing controller load, migration costs, and response times.

The growth of the Internet and the emergence of cloud computing are possible due to the rapid advancement in network bandwidth and hardware, which uses distributed, low-power resources to perform complex tasks efficiently. Cloud computing relies on dynamic, scalable, and virtualized resources provided as services over the Internet. Efficient task execution in such systems requires careful selection of service nodes based on task properties. Wang *et al.* [19] introduced a "two-phase scheduling algorithm" for a "three-level cloud computing network", combining "Opportunistic Load Balancing (OLB)" and "Load Balance Min-Min (LBMM)" to improve execution efficiency and maintain system load balance.

Hu *et al.* [20] proposed a genetic algorithm-based VM scheduling strategy that utilized both current and historical data to predict and minimize the impact of deployments. Their method improved the load balancing and reduced the migration costs. Their results showed it enhanced resource use and stability under varying system loads.

Khiyaita *et al.* [21] provided an overview of load balancing in cloud computing, which highlighted it as a key challenge to ensure acceptable response times and service quality. As cloud computing continues to grow rapidly, effective load balancing is essential for system availability and user trust. They also outlined major research challenges in this area.

Zhang *et al.* [22] highlighted the rise of the intelligent cloud, which uses machine learning, specifically deep reinforcement learning to optimize service configurations and resource allocation. It presented an architecture for intelligent cloud management and demonstrated its effectiveness through an example, which shows improved adaptability and efficiency in complex cloud environments.

Marques *et al.* [23] proposed an autonomous monitoring and management system for microservices and container clusters that predicts load changes and resource shortages. It uses customizable metrics to anticipate demand and proactively allocate or release resources, which ensures uninterrupted service. This proposed method was tested in the dynamic AWS environment. This solution improved scalability efficiency, reduced response time, and enhanced overall QoS/QoE.

Chen *et al.* [24] proposed a new algorithm, "RAA-PI-NSGAII," to efficiently allocate cloud resources for sudden and unclear demands. They used a multi-objective model to minimize server usage and resource mismatches. Their method improved speed, optimization, and resource balance as compared to the conventional approaches.

Fan *et al.* [25] proposed an optimized task offloading scheme for edge-cloud networks that minimizes processing delays and ensures queue stability. They integrated service placement, task scheduling, resource and transmission allocation using "Lyapunov optimization" and a multi-timescale algorithm. Their simulation results show that the method outperformed existing solutions in efficiency and performance.

Hu *et al.* [26] addressed efficient task offloading for energy-harvesting IoT devices in edge-cloud systems by minimizing service cost by balancing the energy use and service delay. They introduced a "hybrid deep reinforcement learning algorithm (DDPG-D3QN)" to manage both continuous (power control) and discrete (server selection) decisions. Their model respects constraints like delay bounds, resource limits, and error rates. Their simulation results show that the proposed method achieved better convergence, stability, and performance than existing DRL approaches, and that edge-cloud collaboration outperformed non-collaborative solutions.

Huang *et al.* [27] proposed a "computation offloading

and resource allocation (CORA)" algorithm for the Internet of Vehicles (IoV), which optimizes the "computation offloading and resource allocation" in a dynamic environment using "collaborative MEC and cloud computing." Their goal is to minimize system cost while meeting delay and transmission constraints. The problem is modeled as a "Markov decision process" and solved using "deep reinforcement learning" to handle complex, "high-dimensional" scenarios. Their simulation results represents that CORA adapts well to network changes and outperforms both DRL and non-DRL baselines in cost, processing delay, and training efficiency.

Liu *et al.* [28] proposed a method to reduce task handling latency in mobile edge computing by optimizing how tasks are split and resources are allocated across devices, edge, and cloud (DEC). They breaks the problem into two parts: task partitioning and resource allocation, solved using analytical and optimization techniques. Their real-world tests show that the approach effectively improves performance.

Wu *et al.* [29] explored the optimization of "cloud-edge-end computing" for handling "multi-source IoT data streams" in dynamic network environments (as illustrated in Figure 2). They modeled the problem as a "Markov decision process," which addressed two key sub-problems: offloading ratio assignment and resource allocation. To solve these efficiently, they combined "Proximal Policy Optimization (PPO)" for offloading decisions with "convex optimization" for resource distribution. Their results show that the approach effectively enhanced edge computing performance under varying conditions.

Zhou *et al.* [30] presented "Reverse Auction-based Computation Offloading and Resource Allocation Mechanism (RACORAM)," a reverse auction-based system where edge servers handle mobile device tasks to reduce cloud costs. They used efficient algorithms for resource allocation and task offloading, achieved near-optimal performance with low complexity and reduced CSC expenses as a result.

Goyal *et al.* [31] focused on reducing energy consumption and improving load balancing in cloud computing using optimization algorithms. They compared several methods such as PSO, CSO, BAT, CSA, and WOA etc. for efficient resource scheduling. Among them, the Whale Optimization Algorithm (WOA) showed the best performance in terms of response time, energy use, execution time, and throughput, especially in tests with seven and eight server setups.

Mobile devices struggle with running demanding applications due to limited resources. To help, tasks can be offloaded to nearby edge servers, but this alone insufficient for requirements of all applications. Dai *et al.* [32] proposed a novel approach called "end-edge-cloud orchestrated computing (EECOC)" which addresses this issue, but current research does not handle device mobility well. They proposed a deep reinforcement learning-based method that predicts device movement and optimizes task offloading and resource use, which shows strong performance improvements over existing solutions.

Thakur and Goraya [33] introduced "RAFL," a hybrid metaheuristic framework for resource allocation in cloud computing to achieve load balancing. They aimed to evenly distribute CPU and RAM usage across active physical machines, which prevents overload or underuse. The core algorithm, "PPSO-DA (a blend of phasor particle swarm optimization and the dragonfly algorithm)," generates optimal allocation plans. Their simulations using CloudSim show that PPSO-DA outperformed several existing algorithms in balancing load, with statistical tests confirmed its effectiveness.

Iqbal *et al.* [34] focused on improving energy efficiency (EE) in "cloud radio access networks (CRAN)" (as presented in Figure 3) by optimizing the on/off status and power use of "remote radio heads (RRHs)." Unlike traditional methods, it accounts for switching overhead between time intervals. The problem is modeled as a Markov decision process and addressed using deep reinforcement learning. A "Double Deep Q-Network (DDQN)" is proposed to avoid overestimating Q-values and achieve better energy efficiency than standard DQN and baseline methods. Their simulation results confirm the superior performance of the DDQN-based approach.

Delaram *et al.* [35] explored resource allocation in Cloud Manufacturing (CM), a key component of Industry 4.0 that offers manufacturing services on demand. They modeled provider and consumer behavior based on preferences and analyzes how platform type, matching algorithm, and resource availability affect utility. They proposed a decision framework recommending different matching algorithms depending on platform type and resource-demand balance. For public platforms, it suggested Consumer- or Provider-Proposer Deferred Acceptance algorithms; for private platforms, Consumer- or Provider-oriented Kuhn-Munkres algorithms, based on whether supply meets or falls short of demand.

Agomuo *et al.* [36] proposed an integrated approach to optimize resource allocation in cloud computing by combining multiple machine learning and optimization techniques. They used LSTM for accurate demand forecasting, PSO for efficient initial allocation, Q-learning for real-time dynamic adjustment, and Linear Regression to predict energy consumption. The ensemble method improved resource efficiency and adaptability while supporting energy-efficient cloud operations, with each component showed strong performance in their simulation results.

Alahdadi *et al.* [37] addressed the challenge of untruthful bidding in cloud computing double auctions, where users underbid and providers overbid, which harms market fairness and efficiency. Since achieving truthfulness, budget balance, and efficiency simultaneously is impossible, they proposed a reward mechanism to promote honesty. It tracked bidders' history and rewards consistent honest

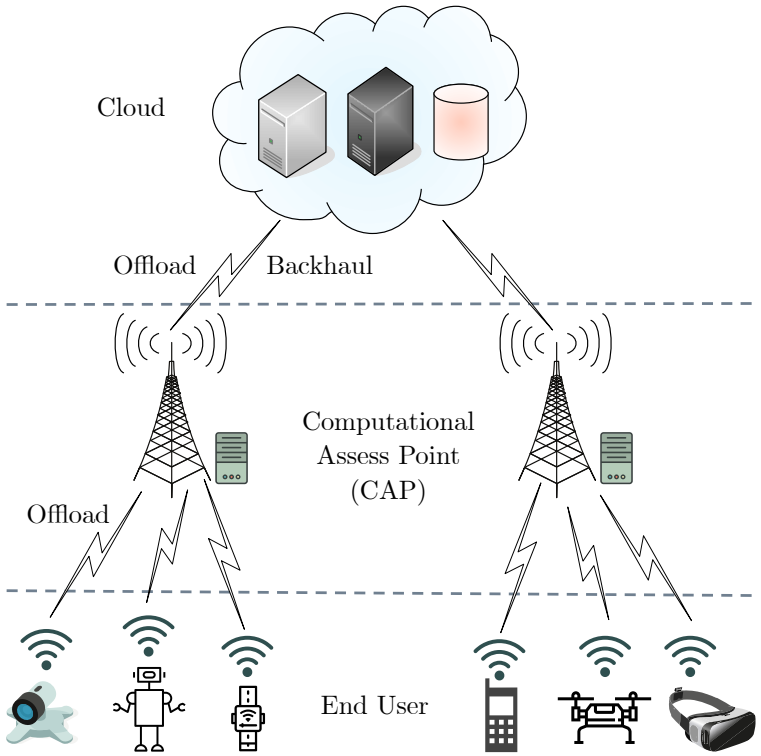

Figure 2: System Model of Collaborative 3-Rier Architecture – Cloud-Edge-End

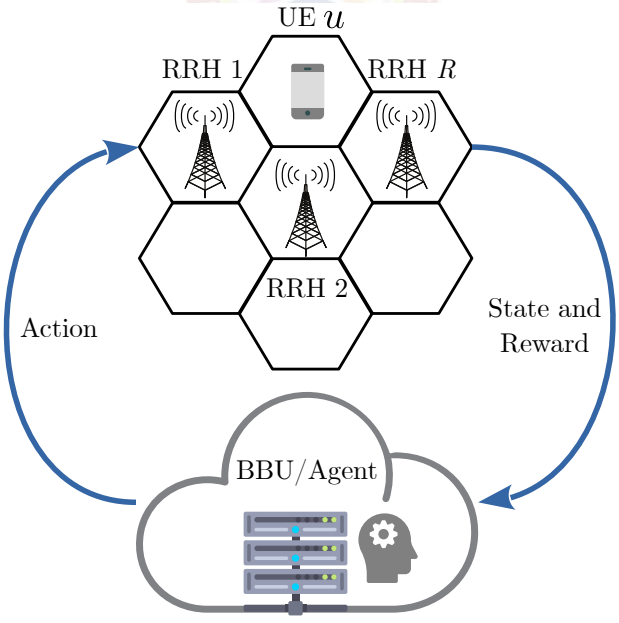

Figure 3: DRL-based C-RAN Scenario

behavior by ranking them and redistributing bid amounts among winners. Their results showed that the mechanism improved truthfulness and mitigates weak budget balance, encouraged fairer and more stable market participation.

3 System Model

We consider a cloud computing environment with:

- M physical hosts (servers), indexed by $m = 1, 2, \ldots, M$.

International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 11, Issue: 05 (Sept-Oct, 2025)

- N virtual machines (VMs), indexed by $n = 1, 2, \ldots, N$.
- J user-submitted tasks, indexed by $j = 1, 2, \ldots, J$.

Each host m has processing capacity C_m (e.g., in GHz), memory R_m (e.g., in GB). Each VM n is assigned resource vector (c_n, r_n) such that:

$$\sum_{n \in \mathcal{V}_m} c_n \leq C_m, \quad \sum_{n \in \mathcal{V}_m} r_n \leq R_m,$$

where $\mathcal{V}_m$ is the set of VMs hosted on server m.

Each task j has:

- Workload requirement w_j (e.g., CPU cycles).
- Memory requirement u_j.
- Arrival time t_j.

We define assignment variables:

$$x_{jn} = \begin{cases} 1, & \text{if task } j \text{ runs on VM } n, \\ 0, & \text{otherwise.} \end{cases}$$

Load on VM n at time t:

$$L_n(t) = \sum_{j \in \mathcal{J}(t)} w_j \, x_{jn}.$$

We aim to minimize the maximum completion time (makespan) and achieve balanced resource usage:

$$\min_{\{x_{jn}\}} \max_{n=1,\ldots,N} \left(\frac{L_n(T)}{c_n} \right)$$

subject to:

$$\sum_{n=1}^{N} x_{jn} = 1, \ \forall j, \tag{1}$$

$$\sum_{j} u_j \, x_{jn} \leq r_n, \ \forall n, \tag{2}$$

$$x_{jn} \in \{0, 1\}, \ \forall j, n. \tag{3}$$

4 Proposed Algorithm

The **Hierarchical Clustering Load Balancing (HCLB)** algorithm is proposed to efficiently allocate user tasks to virtual machines (VMs) in a cloud computing environment. It aims to achieve balanced resource utilization and minimize task completion time. The algorithm proceeds through the following stages:

1. **Feature Extraction:**
   Each task is represented as a feature vector $\mathbf{f}_j = [w_j, u_j]$, where w_j and u_j denote the CPU workload and memory requirement of task j, respectively. These vectors are used for clustering tasks based on similarity.

Algorithm 1 Hierarchical Clustering Load Balancing (HCLB)

Require: Set of active tasks $\mathcal{J}$, VM resource vectors $\{(c_n, r_n)\}$, known initial host capacity.
Ensure: Mapping x_{jn} of tasks to VMs.

Step 1: Feature Extraction
For each task j, build feature vector $\mathbf{f}_j = [w_j, u_j]$.

Step 2: Hierarchical Clustering
Use agglomerative clustering (e.g., Ward's method) to cluster $\{\mathbf{f}_j\}$ into K clusters.

Step 3: Cluster Ranking
For each cluster k, compute:

$$W_k = \sum_{j \in \mathcal{C}_k} w_j, \quad U_k = \sum_{j \in \mathcal{C}_k} u_j.$$

Step 4: VM Suitability Score
For each cluster k and VM n, define score:

$$S_{kn} = \alpha \cdot \frac{W_k}{c_n} + \beta \cdot \frac{U_k}{r_n},$$

where $\alpha, \beta \geq 0$ balance CPU vs memory fit.

Step 5: Assignment
For clusters in descending order of $\max_n S_{kn}$:

- Find VM $n^* = \arg\min_n S_{kn}$ such that $W_k \leq c_n - L_n^{\text{CPU}}$, $U_k \leq r_n - L_n^{\text{RAM}}$.
- Assign all tasks in cluster k to n^*.
- Update $L_{n^*}^{\text{CPU}}, L_{n^*}^{\text{RAM}}$.

Step 6: Refinement
If some tasks remain unassigned:

- Break large clusters, repeat assignment.
- Or assign greedily to the least-loaded compatible VM.

Step 7: Output
Return optimized mapping x_{jn}.

2. **Hierarchical Clustering:**
   Agglomerative hierarchical clustering (e.g., Ward's method) is applied to group similar tasks into K clusters. This step reduces complexity and helps assign tasks in bulk based on their resource profiles.

3. **Cluster Ranking:**
   For each cluster $\mathcal{C}_k$, the aggregate CPU and memory requirements are calculated as:

$$W_k = \sum_{j \in \mathcal{C}_k} w_j, \quad U_k = \sum_{j \in \mathcal{C}_k} u_j.$$

4. **VM Suitability Scoring:**

A suitability score S_{kn} is computed for assigning cluster k to VM n, defined as:

$$S_{kn} = \alpha \cdot \frac{W_k}{c_n} + \beta \cdot \frac{U_k}{r_n},$$

where c_n and r_n are the CPU and memory capacities of VM n, and α, β are weights to balance CPU and memory fit.

5. **Task Assignment:**
Clusters are processed in descending order of their scores. Each cluster is assigned to the VM with the lowest suitability score S_{kn}, provided that:

$$W_k \leq c_n - L_n^{\text{CPU}}, \quad U_k \leq r_n - L_n^{\text{RAM}},$$

where L_n^{CPU} and L_n^{RAM} denote the current CPU and memory load on VM n.

6. **Refinement Step:**
If certain tasks remain unassigned due to resource constraints:
   - Large clusters are split into smaller ones and reassigned.
   - Remaining tasks are greedily assigned to the least-loaded compatible VMs.

7. **Output Generation:**
The final output is an optimized mapping of tasks to VMs, achieving load balancing and efficient resource utilization.

Advantages:

- **Scalability:** Clustering reduces the complexity of handling large numbers of tasks.
- **Flexibility:** Supports diverse task types and resource profiles.
- **Efficiency:** Minimizes makespan and prevents VM overloading.
- **Adaptability:** Dynamically accommodates changing workloads by reapplying clustering.

5 Result Analysis

Fig. 4 illustrates the average response time of the proposed Hierarchical Clustering Load Balancing (HCLB) method compared with conventional scheduling approaches such as Round Robin (RR) and First Come First Serve (FCFS) as the number of tasks increases from 100 to 500.

As observed, the HCLB method consistently outperforms both baselines across all task loads. For instance, at 500 tasks, HCLB achieves an average response time of approximately 20 seconds, whereas RR and FCFS show much higher delays, around 33 and 34 seconds respectively.

The improvement in HCLB can be attributed to:

- **Task Clustering:** Grouping similar tasks allows more efficient resource allocation, avoiding resource fragmentation.
- **Suitability Scoring:** Assigning tasks based on a multi-resource fitness score ensures better load distribution across VMs.
- **Refinement Step:** Dynamic reassignment reduces bottlenecks for tasks that initially cannot be allocated efficiently.

Moreover, the gap between HCLB and the baseline methods widens as the system load increases, indicating that the proposed technique scales better under high task intensity scenarios.

This validates that HCLB not only reduces response time but also improves overall system throughput and VM utilization efficiency in cloud environments.

6 Conclusion

In this research, an efficient resource management and load balancing approach was proposed for cloud computing environments based on hierarchical clustering techniques. By leveraging task similarity through agglomerative clustering and assigning task clusters to virtual machines using a suitability score, the proposed method—Hierarchical Clustering Load Balancing (HCLB)—effectively reduces average response time and balances system load.

Experimental results demonstrated that HCLB outperforms traditional scheduling algorithms such as Round Robin and First Come First Serve, particularly under high-load conditions. The combination of task clustering, multi-resource scoring, and refinement strategies contributes to significant performance improvements in terms of response time and VM utilization.

While the proposed HCLB algorithm shows promising results, several directions can be explored in future work:

- **Dynamic Workloads:** Extend the model to support real-time dynamic task arrivals and departures in a streaming fashion.
- **Energy Efficiency:** Integrate energy-aware scheduling to reduce power consumption alongside performance optimization.
- **Multi-Objective Optimization:** Formulate the task assignment problem as a multi-objective optimization model considering latency, energy, and cost trade-offs.

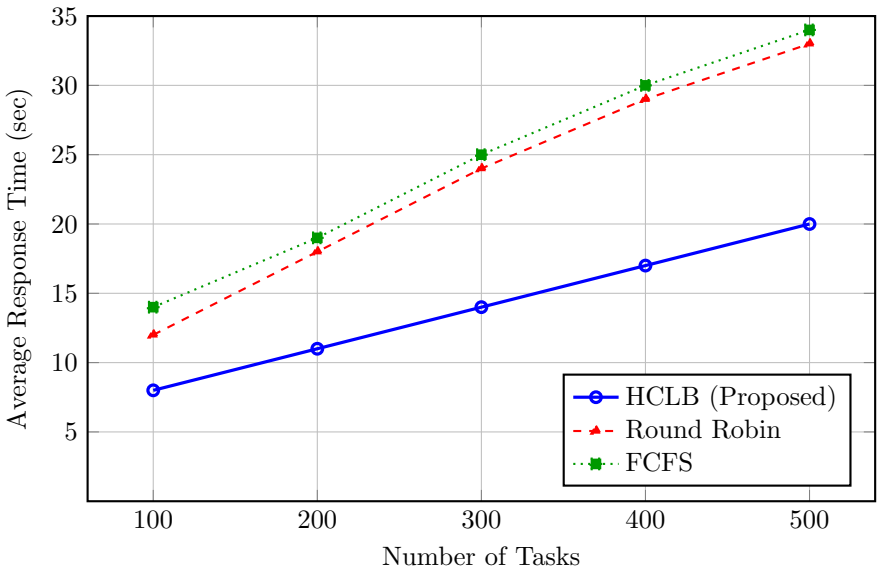

Figure 4: Comparison of Average Response Time vs. Number of Tasks for Different Scheduling Algorithms

- **Scalability in Large Data Centers:** Investigate distributed implementations of the clustering and assignment logic to maintain efficiency in large-scale cloud infrastructures.
- **Heterogeneous Resources:** Adapt the approach to heterogeneous environments with diverse hardware capabilities and network constraints.

References

[1] A. Mukherjee, D. De, and R. Buyya, "Cloud computing resource management," in *Resource Management in Distributed Systems*. Springer Nature Singapore, 2024, pp. 17–37. [doi: https://doi.org/10.1007/978-981-97-2644-8_2

[2] H. Klinke, "Cloud computing," in *Cultural Data Science: An Introduction to R*. Cham: Springer Nature Switzerland, 2025, pp. 145–153. [doi: https://doi.org/10.1007/978-3-031-88130-5_17

[3] T. Mastelic, A. Oleksiak, H. Claussen, I. Brandic, J.-M. Pierson, and A. V. Vasilakos, "Cloud computing: Survey on energy efficiency," *ACM Comput. Surv.*, vol. 47, no. 2, Dec. 2014. [doi: https://doi.org/10.1145/2656204

[4] A. Jyoti, M. Shrimali, and R. Mishra, "Cloud computing and load balancing in cloud computing – survey," in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2019, pp. 51–55. [doi: https://doi.org/10.1109/CONFLUENCE.2019.8776948

[5] C. Surianarayanan and P. R. Chelliah, "Cloud computing fundamentals," in *Essentials of Cloud Computing: A Holistic, Cloud-Native Perspective*. Cham: Springer International Publishing, 2023, pp. 39–71. [doi: https://doi.org/10.1007/978-3-031-32044-6_2

[6] A. Juan Ferrer, "Cloud computing," in *Beyond Edge Computing: Swarm Computing and Ad-Hoc Edge Clouds*. Cham: Springer International Publishing, 2023, pp. 21–42. [doi: https://doi.org/10.1007/978-3-031-23344-9_3

[7] K. Thakur, A.-S. K. Pathan, and S. Ismat, "Distributed cloud computing," in *Emerging ICT Technologies and Cybersecurity: From AI and ML to Other Futuristic Technologies*. Cham: Springer Nature Switzerland, 2023, pp. 185–197. [doi: https://doi.org/10.1007/978-3-031-27765-8_7

[8] M. Avram, "Advantages and challenges of adopting cloud computing from an enterprise perspective," *Procedia Technology*, vol. 12, pp. 529–534, 2014. [doi: https://doi.org/10.1016/j.protcy.2013.12.525

[9] M. N. Sadiku, S. M. Musa, and O. D. Momoh, "Cloud computing: Opportunities and challenges," *IEEE Potentials*, vol. 33, no. 1, pp. 34–36, 2014. [doi: https://doi.org/10.1109/MPOT.2013.2279684

[10] A. Gajbhiye and K. M. P. Shrivastva, "Cloud computing: Need, enabling technology, architecture, advantages and challenges," in *2014 5th International Conference - Confluence The Next Generation Information Technology Summit (Confluence)*, 2014, pp. 1–7. [doi: https://doi.org/10.1109/CONFLUENCE.2014.6949224

[11] G. Yan, "Application of cloud computing in banking: Advantages and challenges," in *Proceedings of the 2017 2nd International Conference on Politics, Economics and Law (ICPEL 2017)*. Atlantis Press, 2017, pp. 29–32. [doi: https://doi.org/10.2991/icpel-17.2017.8

[12] I. Nwobodo, "Cloud computing: Models, services, utility, advantages, security issues, and prototype," in *Wireless Communications, Networking and Applications*, Q.-A.

Zeng, Ed. Springer India, 2016, pp. 1207–1222. [doi: https://doi.org/10.1007/978-81-322-2580-5_110

[13] W. Voorsluys, J. Broberg, and R. Buyya, "Introduction to cloud computing," in *Cloud Computing*. John Wiley & Sons, Ltd, 2011, ch. 1, pp. 1–41. [doi: https://doi.org/10.1002/9780470940105.ch1

[14] S. Pal, D.-N. Le, and P. K. Pattnaik, "Introduction to cloud computing," in *Cloud Computing Solutions*. John Wiley & Sons, Ltd, 2022, ch. 2, pp. 21–38. [doi: https://doi.org/10.1002/9781119682318.ch2

[15] P. Wang, R. X. Gao, and Z. Fan, "Cloud computing for cloud manufacturing: Benefits and limitations," *Journal of Manufacturing Science and Engineering*, vol. 137, no. 4, p. 040901, 08 2015. [doi: https://doi.org/10.1115/1.4030209

[16] N. Pramod, A. K. Muppalla, and K. G. Srinivasa, "Limitations and challenges in cloud-based applications development," in *Software Engineering Frameworks for the Cloud Computing Paradigm*, Z. Mahmood and S. Saeed, Eds. Springer London, 2013, pp. 55–75. [doi: https://doi.org/10.1007/978-1-4471-5031-2_3

[17] V. Hayyolalam and Öznur Özkasap, "CBWO: A novel multi-objective load balancing technique for cloud computing," *Future Generation Computer Systems*, vol. 164, p. 107561, 2025. [doi: https://doi.org/10.1016/j.future.2024.107561

[18] Y. Liu, Q. Meng, K. Chen, and Z. Shen, "Load-aware switch migration for controller load balancing in edge – cloud architectures," *Future Generation Computer Systems*, vol. 162, p. 107489, 2025. [doi: https://doi.org/10.1016/j.future.2024.107489

[19] S.-C. Wang, K.-Q. Yan, W.-P. Liao, and S.-S. Wang, "Towards a load balancing in a three-level cloud computing network," in *2010 3rd International Conference on Computer Science and Information Technology*, vol. 1, 2010, pp. 108–113. [doi: https://doi.org/10.1109/ICCSIT.2010.5563889

[20] J. Hu, J. Gu, G. Sun, and T. Zhao, "A scheduling strategy on load balancing of virtual machine resources in cloud computing environment," in *2010* 3^{rd} *International Symposium on Parallel Architectures, Algorithms and Programming*, 2010, pp. 89–96. [doi: https://doi.org/10.1109/PAAP.2010.65

[21] A. Khiyaita, H. E. Bakkali, M. Zbakh, and D. E. Kettani, "Load balancing cloud computing: State of art," in *2012 National Days of Network Security and Systems*, 2012, pp. 106–109. [doi: https://doi.org/10.1109/JNS2.2012.6249253

[22] Y. Zhang, J. Yao, and H. Guan, "Intelligent cloud resource management with deep reinforcement learning," *IEEE Cloud Computing*, vol. 4, no. 6, pp. 60–69, 2017. [doi: https://doi.org/10.1109/MCC.2018.1081063

[23] G. Marques, C. Senna, S. Sargento, L. Carvalho, L. Pereira, and R. Matos, "Proactive resource management for cloud of services environments," *Future Generation Computer Systems*, vol. 150, pp. 90–102, 2024. [doi: https://doi.org/10.1016/j.future.2023.08.005

[24] J. Chen, T. Du, and G. Xiao, "A multi-objective optimization for resource allocation of emergent demands in cloud computing," *Journal of Cloud Computing*, vol. 10, no. 1, p. 20, Mar 2021. [doi: https://doi.org/10.1186/s13677-021-00237-7

[25] W. Fan, L. Zhao, X. Liu, Y. Su, S. Li, F. Wu, and Y. Liu, "Collaborative service placement, task scheduling, and resource allocation for task offloading with edge-cloud cooperation," *IEEE Transactions on Mobile Computing*, vol. 23, no. 1, pp. 238–256, 2024. [doi: https://doi.org/10.1109/TMC.2022.3219261

[26] H. Hu, D. Wu, F. Zhou, X. Zhu, R. Q. Hu, and H. Zhu, "Intelligent resource allocation for edge-cloud collaborative networks: A hybrid DDPG-D3QN approach," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 8, pp. 10 696–10 709, 2023. [doi: https://doi.org/10.1109/TVT.2023.3253905

[27] J. Huang, J. Wan, B. Lv, Q. Ye, and Y. Chen, "Joint computation offloading and resource allocation for edge-cloud collaboration in internet of vehicles via deep reinforcement learning," *IEEE Systems Journal*, vol. 17, no. 2, pp. 2500–2511, 2023. [doi: https://doi.org/10.1109/JSYST.2023.3249217

[28] F. Liu, J. Huang, and X. Wang, "Joint task offloading and resource allocation for device-edge-cloud collaboration with subtask dependencies," *IEEE Transactions on Cloud Computing*, vol. 11, no. 3, pp. 3027–3039, 2023. [doi: https://doi.org/10.1109/TCC.2023.3251561

[29] Y. Wu, C. Cai, X. Bi, J. Xia, C. Gao, Y. Tang, and S. Lai, "Intelligent resource allocation scheme for cloud-edge-end framework aided multi-source data stream," *EURASIP Journal on Advances in Signal Processing*, vol. 2023, no. 1, p. 56, May 2023. [doi: https://doi.org/10.1186/s13634-023-01018-x

[30] H. Zhou, T. Wu, X. Chen, S. He, D. Guo, and J. Wu, "Reverse auction-based computation offloading and resource allocation in mobile cloud-edge computing," *IEEE Transactions on Mobile Computing*, vol. 22, no. 10, pp. 6144–6159, 2023. [doi: https://doi.org/10.1109/TMC.2022.3189050

[31] S. Goyal, S. Bhushan, Y. Kumar, A. u. H. S. Rana, M. R. Bhutta, M. F. Ijaz, and Y. Son, "An optimized framework for energy-resource allocation in a cloud environment based on the whale optimization algorithm," *Sensors*, vol. 21, no. 5, 2021. [doi: https://doi.org/10.3390/s21051583

[32] B. Dai, J. Niu, T. Ren, and M. Atiquzzaman, "Toward mobility-aware computation offloading and resource allocation in end-edge-cloud orchestrated computing,"

IEEE Internet of Things Journal, vol. 9, no. 19, pp. 19 450–19 462, 2022. [doi: https://doi.org/10.1109/JIOT.2022.3168036

[33] A. Thakur and M. S. Goraya, "RAFL: A hybrid metaheuristic based resource allocation framework for load balancing in cloud computing environment," *Simulation Modelling Practice and Theory*, vol. 116, p. 102485, 2022. [doi: https://doi.org/10.1016/j.simpat.2021.102485

[34] A. Iqbal, M.-L. Tham, and Y. C. Chang, "Double deep q-network-based energy-efficient resource allocation in cloud radio access network," *IEEE Access*, vol. 9, pp. 20 440–20 449, 2021. [doi: https://doi.org/10.1109/ACCESS.2021.3054909

[35] J. Delaram, M. Houshamand, F. Ashtiani, and O. Fatahi Valilai, "A utility-based matching mechanism for stable and optimal resource allocation in cloud manufacturing platforms using deferred acceptance algorithm," *Journal of Manufacturing Systems*, vol. 60, pp. 569–584, 2021. [doi: https://doi.org/10.1016/j.jmsy.2021.07.012

[36] O. C. Agomuo, O. W. B. Jnr, and J. H. Muzamal, "Energy-aware AI-based optimal cloud infra allocation for provisioning of resources," in *2024 IEEE/ACIS 27th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 2024, pp. 269–274. [doi: https://doi.org/10.1109/SNPD61259.2024.10673918

[37] A. Alahdadi, A. A. Safaei, and M. J. Ebadi, "A truthful and budget-balanced double auction model for resource allocation in cloud computing," *Soft Computing*, vol. 27, no. 23, pp. 18 263–18 284, Dec 2023. [doi: https://doi.org/10.1007/s00500-023-08081-4