International Journal of Science, Strategic Management and Technology
Volume **02** Issue **03** March-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

AI-Based Resource Allocation Techniques in Cloud Computing: A Comparative Study

Dharshini K[1], Dr.D.Geethamani[2]

[1]Undergraduate Student,[2]Assistant Professor, Department of Computer Technology, Dr. NGP Arts and Science College Coimbatore, Tamil Nadu, India

Email: dharshinikannan2206@gmail.com, geethamani.d@drngpasc.ac.in

https://doi.org/10.55041/ijsmt.v2i3.168

Cite this Article: K, D. (2026). AI-Based Resource Allocation Techniques in Cloud Computing: A Comparative Study. International Journal of Science, Strategic Management and Technology, *02*(03). https://doi.org/10.55041/ijsmt.v2i3.168

Abstract

Cloud computing is recognized as a vital technology for on-demand and flexible computing resources via the internet [1]. The allocation of resources is a key challenge in cloud computing because inefficient allocation of resources results in wastage of resources and poor performance of systems. The traditional approach for efficient allocation of resources is based on heuristic and rule-based algorithms. However, these approaches are not efficient in adapting dynamically to changes in workload. Recently, AI techniques are also being explored for efficient allocation of resources in cloud computing systems. This paper presents a comparative study of traditional, optimization-based, and AI-based approaches for efficient allocation of resources in cloud computing systems. The paper also presents the advantages and disadvantages of different approaches and discusses the challenges and future directions of efficient allocation of resources in cloud computing systems.

Keywords- Cloud Computing, Resource Allocation, Artificial Intelligence, Machine Learning, Scheduling Algorithms, ComparativeStudy.

1. Introduction

Cloud computing allows users to access computing services like processing, memory, storage, and networking through the internet [1]. Due to its scalability and cost-effectiveness, cloud computing has gained popularity in different fields like industries, education, and research areas [11]. Cloud service providers need to manage the allocated cloud resources efficiently so that multiple users can access at any time. Resource allocation is the process of efficiently allocating the available cloud resources to the incoming virtual machines. If the allocation is not done efficiently, it can lead to server overload and increased response time. Hence, efficient management of cloud resources has become an important area of research [2]. Initially, cloud computing has adopted some basic scheduling algorithms for efficient allocation of cloud resources.

These techniques are not capable of adapting dynamically according to the changing environment. Recently, the application of Artificial Intelligence and Machine Learning techniques has gained popularity for efficient cloud resource allocation. This paper aims at presenting the different techniques of cloud resource allocation and understanding the application of Artificial Intelligence techniques in improving the performance of cloud computing.

2. Background of Cloud computing

Cloud computing is the delivery of computing services over the internet on a pay-as-you-use basis [1]. The primary aim is to offer flexible resource usage without the need for physical infrastructure.

2.1 Cloud Service Models:

Cloud computing offers a number of service models that allow users to use various computing services via the internet. The first service model is Infrastructure as a Service (IaaS), which allows users to use basic computing infrastructure such as virtual machines, storage devices, and network devices. The second service model is Platform as a Service (PaaS), which allows users to use a platform to develop, test, and deploy various applications. The third service model is Software as a Service (SaaS), where users can use various software applications via the internet using web browsers.

2.2 Importance of Resource Allocation

Resource allocation is an essential part of the cloud computing environment, as it ensures the effective use of the available resources. The allocation of the resources efficiently ensures the performance of the system, maximizes the utilization of the resources, and minimizes the time taken for the execution of the tasks. In addition, the

allocation of the resources ensures the maintenance of the Service Level Agreements (SLA) between the cloud service provider and the cloud users. The allocation of the resources also ensures the minimization of the operation costs and the energy consumption in the cloud computing environment. The cloud computing environment is highly dynamic; therefore, intelligent resource allocation strategies are essential for the performance of the system.

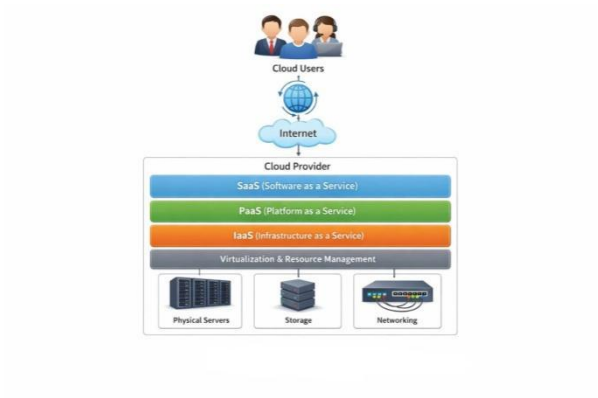

Fig. 1. AI-Based Resource Allocation Architecture

3. AI-Based Resource Allocation in Cloud Computing

"AI-based resource allocation" is the term given to the allocation of cloud resources like CPU, memory, and storage using Artificial Intelligence technologies. AI-based allocation is different from other allocation strategies like the traditional rule-based allocation method.

Generally, the allocation of cloud resources is done using the traditional allocation method. However, the allocation of cloud resources is a complex task in a cloud environment where the workload is highly dynamic and unpredictable. AI-based allocation strategies use the past data and the current information collected through the monitoring of the system resources to make the allocation decisions.

The overall workflow of the AI-based resource allocation is shown in Fig. 2. The overall workflow starts with the user task request and then proceeds with the monitoring and analysis of the system resources. Then, the AI decision module decides the allocation strategy, and the allocation is done accordingly.

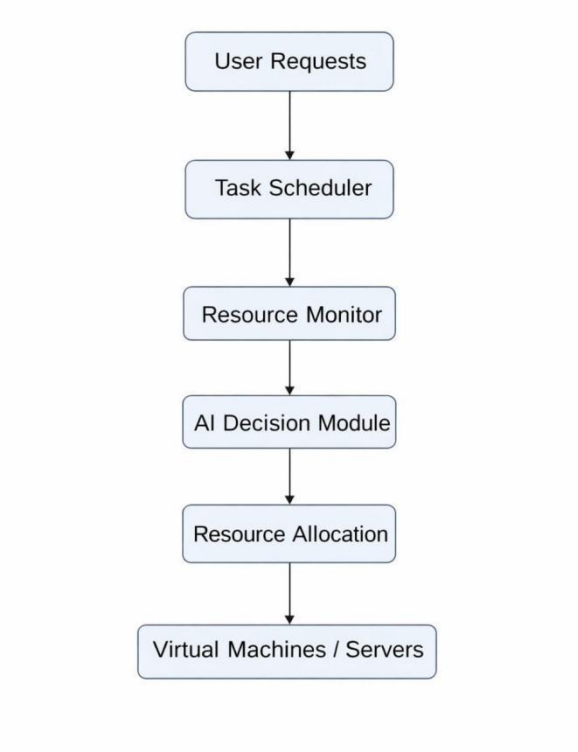

Fig. 2. Workflow of AI-Based Resource Allocation in Cloud Computing

3.1 Traditional Resource Allocation Techniques

In traditional techniques, rules are fixed.

3.2 Round Robin Algorithm

The Round Robin algorithm is another

popular technique used in the distribution of workload. In this technique, the workload is distributed in a cyclic manner. This technique provides fair distribution of workload. However, this technique is quite simple and does not consider the workload of the servers. It simply distributes the workload in a cyclic manner. It is quite possible that the servers might be loaded unevenly.

3.3 First Fit and Best Fit Algorithms

The First Fit algorithm uses the approach of allocating the task to the first available server that has the capacity to perform the task. On the other hand, the Best Fit algorithm uses the approach of allocating the task to the server that has the closest match of the resource requirements of the task. Although these approaches provide the benefits of resource utilization compared to other scheduling algorithms such as the Round Robin algorithm, these approaches are still faced with the problem of lack of adaptability and resource fragmentation.

3.4 Priority-Based Scheduling

In the priority-based scheduling method, the allocation of resources takes place based on the level of priority for each task. The higher the priority level, the sooner the task will get executed compared to other tasks. This method is particularly useful for environments where certain tasks must get executed quickly, considering their level of importance. However, the major disadvantage of the priority scheduling method is that there are chances for starvation, i.e., the lower-priority tasks might get delayed for a long time.

4. Optimization-based Approach

Optimization-based methods strive to find near-optimal solutions for resource allocation problems in cloud computing systems. Various optimization algorithms like Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), etc., can be employed to increase the overall performance of the systems through better resource allocation. However, the optimization-based methods usually suffer from high computational complexities that do not allow them to be effectively applied to the highly dynamic cloud computing environments for real-time decision-making.

5. AI-Based Resource Allocation Techniques

Techniques of resource allocation in cloud computing, based on AI, use intelligent algorithms to study the behavior of the system in order to allocate resources efficiently in the cloud. The techniques monitor the performance of the system in real time, making adjustments to the resources as needed. The AI model learns from historical data, as well as the current state of the system, in order to predict future requirements of the workload. This way, it prevents overload of the system, making efficient use of computing resources. AI-based techniques improve the stability of the system, its overall performance, as well as Quality of Service (QoS) in cloud computing. Moreover, AI techniques can

facilitate quick and adaptive resource management in dynamic environments like clouds.

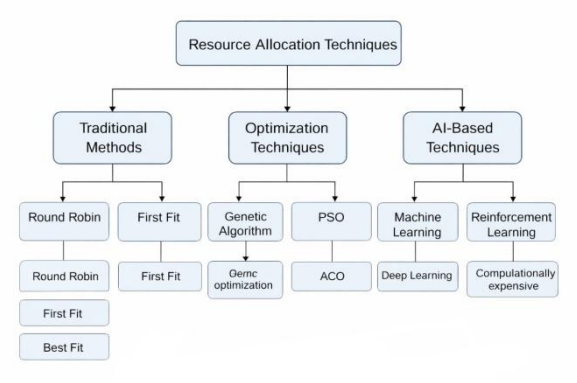

Fig. 3. Classification of Resource Allocation Techniques

5.1 Machine Learning Approaches

Machine learning techniques examine the past workload data and help in the identification of patterns in the usage of the resource and the prediction of future resource requirements. Such predictions help in the efficient allocation of the resource in cloud systems. Some of the machine learning techniques that can be used in the allocation of the resource in cloud computing systems include regression analysis, decision tree, random forest, and clustering.

5.2 Deep Learning Approaches

Deep learning models make use of the neural networks in analyzing the complex patterns in the large sets of data. For the cloud resource allocation problem, models like ANN and LSTM are used for the prediction of the workloads and the allocation of the resources [13]. These models provide high accuracy in the prediction and are capable of handling large sets of data. However, the models require high computational resources and are associated with a complex training process.

5.3 Reinforcement Learning Approaches

Reinforcement learning allows the system to learn the optimum strategy for resource allocation through continuous interaction with the environment [3]. In this method, the system learns from the reward and penalty obtained from different resource allocation strategies. With this learning, the system improves continuously. Deep Q Networks and Policy Gradient are the most commonly used methods in reinforcement learning for resource management. However, this method involves a high training period and high computational cost [5].

6. Comparative Analysis

Technique	Advantages	Limitations
Round Robin	Simple	Poor adaptability
First Fit	Fast allocation	Limited intelligence
Optimization	Better performance	High complexity
Machine Learning	Predictive allocation	Requires training data
Deep Learning	High accuracy	Resource-intensive
Reinforcement Learning	Dynamic decisions	Complex training

Table 1. Comparative Analysis of Resource Allocation Techniques

Both traditional and optimization techniques are easy to implement and provide effective resource allocation. However, these techniques do not offer the flexibility that is necessary for dealing with a highly dynamic cloud computing environment. On the other hand, AI techniques are more appropriate for modern cloud computing systems due to their predictive abilities. Although the computational cost is high for these techniques, they provide effective decision-making and resource allocation.

7. Challenges In AI-Based Resource Allocation

However, despite the advantages of using AI-based techniques, certain challenges have been identified in the effective implementation of these techniques in cloud computing environments. Firstly, the management of dynamic workloads is a major challenge in the effective implementation of AI-based techniques in cloud computing environments [14]. The second major challenge is the high energy consumption of the large-scale data centers. Scalability is another major challenge

in the effective implementation of AI-based techniques in cloud computing environments, particularly for large-scale cloud infrastructures. In addition, security and privacy challenges need to be considered while using AI-based techniques in cloud computing environments.

8. Research Gaps And Future Directions

Currently, the research in cloud resource allocation is mainly focused on the development of advanced AI models, which require high computational power for their execution. However, the future of cloud resource allocation is in the development of lightweight AI models, which can be implemented in small and medium-scale cloud environments. Another area of interest in cloud resource allocation is the integration of hybrid models with AI models for better results. Researchers are also working on energy-efficient techniques for cloud resource allocation, which can reduce the energy consumption of the data center. The integration of real-time adaptive allocation mechanisms and edge computing with cloud computing is another area of interest in cloud resource allocation [15].

9. Case Study: Application of AI in Cloud Resource Management

Some cloud companies have begun to adopt AI-based monitoring tools to enhance the efficiency of their cloud infrastructures. Machine learning techniques are used to analyze historical usage patterns and make predictions regarding peak usage times [8].On the basis of such predictions, cloud platforms can dynamically allocate their resources to ensure that the system is not overburdened.

For example, predictive models can make predictions regarding increased usage during particular time slots and can scale up the virtual machines accordingly.

10. Conclusion

In this paper, a comparative study on resource allocation techniques was conducted on cloud computing environments, particularly on AI-based techniques. Conventional techniques like Round Robin and First Fit provide simple resource allocation methods. However, they do not support adaptability. Optimization techniques can enhance resource allocation. However, the complexity of the process is higher. Machine learning, deep learning, and reinforcement learning techniques provide intelligent and adaptive resource allocation methods. These techniques analyze the pattern of workloads and can estimate future requirements. There are challenges in implementing AI-based resource allocation techniques. However, resource allocation techniques play a vital role in enhancing the efficiency of cloud computing.

11. References

[1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation Computer Systems*, vol. 25, no. 6, pp. 599–616, 2009.

[2] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 13, pp. 1397–1420, 2012.

[3] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in *Proc. ACM HotNets*, 2016, pp. 50–56.

[4] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *Proc. International Conference on Learning Representations (ICLR)*, 2016.

[5] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conference on Artificial Intelligence*, 2016.

[6] Z. Wang, T. Schaul, M. Hessel, H. Hassabis, and D. Silver, "Dueling network architectures for deep reinforcement learning," in *Proc. International Conference on Machine Learning (ICML)*, 2016.

[7] S. Tuli, S. S. Gill, M. Xu, P. Garraghan, R. Bahsoon, S. Dustdar, and O. Rana, "HUNTER: AI-based holistic

resource management for sustainable cloud computing," *Journal of Systems and Software*, vol. 184, 2022.

[8] N. Liu et al., "A hierarchical framework of cloud resource allocation and power management using deep reinforcement learning," in *Proc. IEEE ICDCS*, 2017.

[9] W. Shi et al., "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[10] E. Masanet et al., "Recalibrating global data center energy-use estimates," *Science*, vol. 367, no. 6481, pp. 984–986, 2020.

[11] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[12] Q. Zhang, M. Chen, L. T. Yang, and Z. Chen, "A survey on cloud resource allocation strategies," *Journal of Network and Computer Applications*, vol. 33, no. 3, pp. 334–350, 2010.

[13] Y. Xu, M. Bailey, F. Jahanian, K. Joshi, and M. Hiltunen, "An exploration of LSTM-based resource allocation in cloud computing environments," *IEEE Transactions on Cloud Computing*, 2019.

[14] T. Lorido-Botran, J. Miguel-Alonso, and J. A. Lozano,"A review of auto-scaling techniques for elastic applications in cloud environments," *Journal of Grid Computing*, vol. 12, no. 4, pp. 559–592, 2014.

[15] J. Dean and L. A. Barroso, "The tail at scale," *Communications of the ACM*, vol. 56, no. 2, pp. 74–80, 2013.