Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

Resource Allocation Techniques in Cloud Computing: A Comprehensive Review

Shriya Pingulkar
Information Technology
K. J. Somaiya College of Engineering

Aryaman Tiwary
Information Technology
K. J. Somaiya College of Engineering

Shruti Tyagi
Computer Engineering
K. J. Somaiya College of Engineering

***Abstract*– Cloud computing has become an essential technology for various industries due to its cost-effectiveness, scalability, and flexibility. The efficient management of cloud resources is crucial to ensure high performance and minimal cost. Resource allocation is the process of assigning and managing assets in a manner that supports an organization's strategic planning goals. In cloud computing, resource allocation is necessary for maximizing resource utilization while upholding service-level agreements (SLAs) and reducing energy consumption. This paper reviews twelve research articles that propose different approaches for resource allocation in cloud computing. These approaches include energy-aware resource allocation, dynamic resource allocation, optimal resource management, resource allocation using game theory and many more. The review compares and contrasts these approaches based on their objectives, methodologies, advantages, and limitations.**

***Keywords*– Cloud computing; Resource allocation; Resource scheduling; Resource utilization.**

I. INTRODUCTION

Cloud computing has revolutionized the way businesses and individuals use computing resources. It provides on-demand access to a shared pool of computing resources, including networks, storage, applications, and services, without the need for upfront capital expenditure. Cloud computing is cost-effective, scalable, and flexible, making it a popular choice for various industries, including healthcare, finance, and e-commerce. However, cloud computing requires efficient resource allocation to achieve optimal utilization of resources while maintaining service-level agreements and minimizing energy consumption.

Resource allocation in cloud computing involves distributing computing resources, such as CPU, memory, and storage, to different tasks and applications in a manner that optimizes performance and minimizes energy consumption. Resource allocation in cloud computing is a complex problem due to the dynamic nature of workloads, the heterogeneous nature of resources, and the varying priorities of tasks and applications. To address this problem, researchers have proposed various approaches for resource allocation in cloud computing including but not limited to dynamic resource allocation, heuristic methods etc.

II. RESOURCE ALLOCATION:

Resource allocation in cloud computing refers to the process of assigning appropriate resources to fulfill customer-required tasks efficiently. This involves designating virtual machines that meet the consumers' specified properties. Users submit their tasks, which may require different time frames for execution. Effective resource allocation in the cloud also involves managing workloads and assigning them to virtual machines appropriately. The key factors in determining when a computational operation should start or end include the allocated resources, time spent, actions of predecessors, and relationships with previous tasks. Additionally, resource allocation encompasses activities such as disclosing available resources, selecting the right resources, provisioning them, planning their application, and overall resource management. As illustrated in Fig. 1, resource allocation in cloud computing involves deciding when, what, where, and how much resources should be allocated to each customer's tasks.

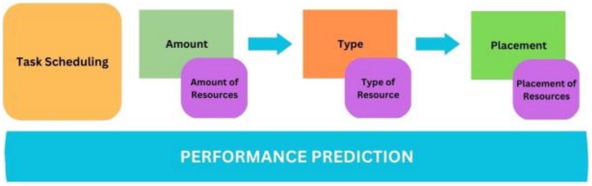

Fig. 1. Cloud resource allocation basic elements

In the resource allocation process, as depicted in Fig. 2, the following general steps are followed:

1. The consumer submits a request to the resource allocator.
2. The request is added to the queue list.
3. The resource allocator informs the allocation unit about the received request.
4. The allocation unit contacts the Infrastructure as a Service (IaaS) to obtain the requested resources.
5. If the required resources are available, the IaaS confirms the availability.
6. The allocation unit creates a Virtual Machine (VM) from the VM pool based on the request.
7. Once the VM is created, the resource allocator is notified.
8. The queues of pending requests are cleared.
9. Finally, the resources are allocated to the user[16].

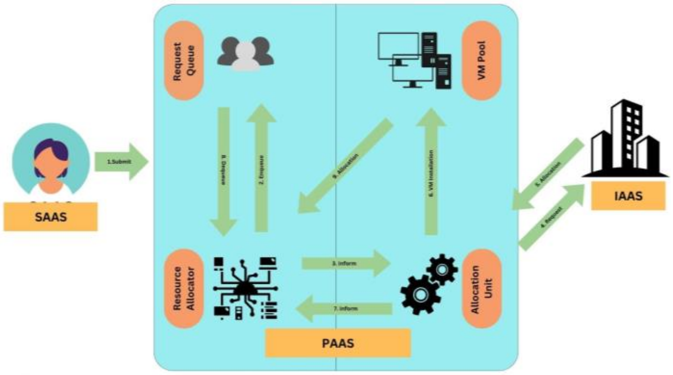

Fig. 2. The basic flow of resource allocation in cloud computing

III. RESOURCE ALLOCATION TECHNIQUES

Cloud computing has gained significant attention as a promising technology to provide cost-effective and scalable services for a wide range of applications. However, resource allocation in cloud computing is a critical task that requires efficient and effective utilization of resources to meet the growing demand for cloud services. In this regard, several approaches have been proposed to improve the efficiency of task scheduling and resource allocation in cloud computing environments with respect to various parameters like cost, resource utilization, energy, workload, SLA and QoS.

1. Energy-aware resource allocation

In cloud computing, resource distribution and energy use are closely connected. It is expected that effective resource management would result in both financial gains and environmental harmony. Techniques for allocating resources that take energy conservation into account have shown to be particularly effective in handling issues caused by data centers' excessive use of electricity.

Matre et al. (2016) conducted a survey on energy-aware resource allocation for cloud computing. It emphasizes that resource allocation in cloud computing is an NP-hard problem, and therefore, several exact and approximate solutions have been proposed. The paper highlights the enabling technologies of cloud computing, including virtualization and load balancing. Additionally, the article emphasizes that cloud computing is scalable, dynamically configurable, and driven by economies of scale [1].

2. Dynamic resource allocation

The amount of work that consumers are submitting to the cloud infrastructure fluctuates continuously. The cloud service provider must employ dynamic approaches to assign resources in order to meet the unique requirements of each task.

Mousavi and Várkonyi-Kóczy (2017) proposed a dynamic resource allocation technique that optimizes energy consumption while maintaining the performance requirements of the workload. It explores the application of two relatively new optimization algorithms and further proposes a hybrid algorithm for load balancing which can contribute well in maximizing the throughput of the cloud provider's network. The proposed algorithm is a hybrid of teaching-learning-based optimization algorithm (TLBO) and grey wolves optimization algorithm (GW). The hybrid algorithm performs more efficiently, balances the priorities and effectively considers load balancing based on time, cost, and avoidance of local optimum traps, which consequently leads to minimal amount of waiting time[2].

In contrast, An et al. (2014) presented an automated negotiation strategy that takes into account de-commitment when allocating resources. A framework based on negotiation that allows for dynamic resource allocation in cloud computing environments was proposed by the authors. In order to negotiate and allocate resources among multiple cloud service providers, the proposed framework makes use of automated negotiation strategies. Decommitment is taken into account in the framework, which enables cloud service providers to withdraw their offer in the event that better offers become available. According to the findings of the experiments, the proposed strategy performs better than more conventional ones in terms of cost reduction and energy efficiency[8].

3. Optimization-based resource allocation

The main aim of optimization is to improve the throughput by increasing the use of physical and virtual resources. This would enable cloud service providers to attract the greatest number of users while minimizing operational costs by spreading out the workload among fewer computers.

Tsai et al. (2019) proposed an improved differential evolution approach for optimizing job scheduling and resource allocation in the context of cloud computing. This method improves the performance of cloud computing systems by allocating resources in a way that minimizes job execution time and maximizes resource consumption. The proposed algorithm can be used to improve the performance and cost effectiveness of cloud computing systems[10].

Choi et al. (2018), on the other hand proposed an optimization mechanism for resource allocation in cloud computing for IoT. The focus is on reducing penalty costs for SLA violations by considering execution time constraints as an SLA constraint. The analysis of the mechanism's performance shows that it reduces penalty costs and increases the provider's profit compared to conventional methods[9].

Rajput and Pant (2014) analyzed the critical components of resource management in cloud computing, including cloud resources, components of cloud resources management and many more. The paper also presents the basic attributes of cloud computing, and explains the four basic deployment models along with the three service models of cloud computing. It states that optimal resource management in the cloud environment is a new domain of research that requires further exploration[3].

In another approach, Shi and Lin (2022) proposed a multiobjective optimization approach, MOGA-D (Multiobjective Optimization Genetic Algorithm based on Decomposition), for optimizing virtual machine resource allocation in cloud computing. The paper addresses the challenge of achieving efficient resource utilization, reducing user costs, and saving computing time in the Infrastructure as a Service (IaaS) mode of cloud computing. MOGA-D combines current and predicted application load data, considering the cost of virtual machine relocation and the stability of new virtual machine placement. The MOGA-D algorithm decomposes the multiobjective problem into single-objective optimization problems, resulting in faster convergence and similar multiobjective optimization results compared to MOGANS, a previous algorithm. Experimental simulations demonstrate the superiority of MOGA-D in terms of resource utilization, revenue generation, and overall performance in cloud computing scenarios.
The paper also introduces the MOGANS algorithm, which provides a virtual machine distribution method with longer stability time compared to the genetic algorithm (GA-NN) for energy-saving and multivirtual machine redistribution overhead. However, MOGANS suffers from limited computational performance, especially with large data sets. MOGA-D addresses this limitation and improves computational performance while achieving similar optimization results at the same calculation scale[14].

4. Resource allocation scheduling algorithms

The demand on servers has grown along with the daily growth in cloud computing usage. The primary objective of researchers has been to maintain low power usage while making the most of available resources. Algorithms enable efficient resource scheduling and allocation.

Cloud service providers take uncertain resources into consideration while arranging and executing tasks. This study divided resource scheduling methods into three categories: 1) Priority-based scheduling; 2) Round Robin; 3) Heuristic approach

A. Priority-based scheduling

Cloud service providers determine the priority among the different consumer demands while allocating the resources.

Pawar et al. (2015) proposed a priority-based dynamic resource allocation algorithm that considers multiple Service Level Agreement (SLA) parameters, such as memory, network bandwidth, and required CPU time, and executes preemptable tasks to improve resource utilization. The use of multiple SLA parameters and preemption allows for better utilization of resources, even in situations where resource contention is high. The algorithm outperforms existing methods, providing an efficient approach to dynamic resource allocation[5].

B. Round Robin

In recent times, Round Robin has been the most common and widely used scheduling algorithm which makes it suitable to allocate resources to the task efficiently.

Pradhan et al. (2017) proposed an algorithm that aims to optimize resource allocation by reducing waiting time and meeting customer and application requirements. It used the round robin algorithm to allocate resources in a circular manner to all the customers in a fair and balanced way. The proposed modification in the algorithm is to divide the resources into multiple segments and allocate them to customers according to their needs. This allocation helps to improve the overall efficiency of the cloud computing system[7].

C. *Heuristic approach*

A heuristic approach is a problem-solving method that uses practical experience and rules of thumb to find solutions quickly. In cloud computing, heuristic approaches are often used to allocate resources and schedule tasks efficiently.

Gawali et al. (2015) proposed a heuristic method that combined several methods, such as the modified analytic hierarchy process (MAHP), bandwidth-aware divisible scheduling (BATS) + BAR optimization, longest expected processing time preemption (LEPT), and divide-and-conquer methods. In this, each task is processed using the MAHP process before its actual allocation. The system preempts resource-intensive tasks using LEPT preemption and the divide-and-conquer approach improves the proposed system. The proposed approach outperforms existing frameworks and can be used to achieve better performance in cloud computing[6].

D. *Hungarian Optimisation Algorithm*

The Hungarian Algorithm is a combinatorial optimization algorithm used to solve the assignment problem. In the context of cloud computing resource allocation, the Hungarian Algorithm can be used to allocate resources to different tasks and applications in an optimal way.

Murali et al. (2023) proposes a novel resource allocation model for cloud computing using Hungarian optimization in AWS. The proposed methodology comprises a resource allocation model that comprises resource discovery, task scheduling, and resource allocation using the Hungarian optimization technique. The paper aims to provide better quality of service and profits to cloud service providers by overcoming the challenges of over-demand and under-availability of resources[11].

5. Reinforcement Learning based resource allocation

Multi-agent reinforcement learning (MARL) is a machine learning technique that involves multiple agents learning to interact with each other and the environment in order to optimize a common objective. In the context of resource allocation in cloud computing, MARL can be used to allocate resources, such as CPU, memory, and network bandwidth, to different tasks and applications running in the cloud.

Belgacem et al. (2022) proposes a new model called the Intelligent Multi-Agent Reinforcement which is a combination of a multi-agent system and the Q-learning process to dynamically allocate and release resources while responding to changing consumer demands. It uses a reinforcement learning policy to make virtual machines move to the best state according to the current state environment. IMARM proved to be a comprehensive solution to the essential concerns of cloud service providers in resource allocation, which include energy conservation, fault tolerance, and workload balancing[12].

On the other hand, Cen and Li (2022) propose a resource allocation strategy that utilizes deep reinforcement learning to optimize system delay in a cloud-edge collaborative computing environment. They construct a collaborative mobile edge computing (MEC) system model, combining the core cloud center with MEC, to enhance network interaction. By considering both the communication and computation models, they formulate the goal of minimizing system delay as a Markov decision process. To achieve this, they employ a deep Q network (DQN) improved with hindsight experience replay (HER). The simulation results demonstrate the effectiveness of their approach, achieving a maximum user delay of 1150 ms when 80 user terminals are involved, outperforming other comparison strategies in complex environments. Their proposed method optimizes computational resource allocation, significantly reducing network delay and enhancing overall system performance.

6. Game Theory based resource allocation
The game theory-based approach for resource allocation in cloud computing involves using principles of coalition formation and uncertainty to create more efficient allocation strategies. Game-theoretic approaches in general help simplify complex problems to a great extent.

Pillai et al. (2015) proposed a resource allocation mechanism for cloud computing based on the principles of coalition formation and the uncertainty principle of game theory. It aims to allocate resources in a way that minimizes wastage and configures services ahead of actual requests, thus leading to better resource utilization and higher request satisfaction, and highlights the need for coalitions of pre-configured VMs to address scalability concerns and shorten response times. The proposed method leads to better resource utilization and higher request satisfaction[4].

Ficco et al. (2018) present a meta-heuristic approach for optimizing elastic cloud resource allocation using a combination of Coral-Reefs Optimization (CRO) and Game Theory. Their approach addresses the complex multi-objective problem of maximizing demand satisfaction and minimizing costs and resource consumption in cloud scenarios. The CRO algorithm simulates the evolution processes of reefs to model resource reallocation dynamics, while Game Theory optimizes the VM reallocation schema considering both cloud provider's objectives and customer requirements expressed through Service Level Agreements (SLAs). The proposed approach offers promising results in achieving convergence towards global optima[13].

Xu and Yu (2014), in their research propose a game theoretic resource allocation algorithm for cloud computing, aimed at achieving fair and efficient utilization of computational resources. The algorithm considers both fairness among users and resource utilization to optimize the allocation of virtual machines to physical servers. Their experiments and simulations demonstrate that the FUGA algorithm outperforms the Hadoop scheduler in terms of fairness, while also reducing resource wastage and achieving better resource utilization compared to other allocation mechanisms. By incorporating game theory principles, the FUGA algorithm enables cloud providers to allocate resources fairly and efficiently, minimizing resource fragmentation and ensuring that no user receives significantly better resources than others. This approach enhances the overall performance and resource utilization in cloud computing environments[15].

IV. CONCLUSION

This research presents a structured literature survey based on resource allocation techniques in cloud computing. This study helps understanding different resource allocation techniques on the basis of their schemes, the problems addressed, and the results of their approaches that are used by the different researchers in a contextualized manner. Apart from presenting a summary of the selected articles under proper heads, it also presents promising future directions in the field of resource allocation in cloud computing. It also concludes that efficient resource allocation techniques should meet criteria like cost, energy, response time, execution time, workload, resource utilization, user satisfaction, and SLA. The techniques discussed in this research paper must be beneficial to the cloud users in terms of quality of service and also to the cloud service providers in terms of profit.

Future research directions involve improved usage of artificial intelligence in scheduling and optimization of resource allocation strategies. It also recommended that extensive research is needed on energy-aware resource allocation schemes, especially with regard to green optimization. Lastly, it is envisaged that the services of cloud computing will become an integral part of almost all types and scales of information systems.

V. REFERENCES

[1]Matre, P., Silakari, D. S., & Chourasia, U. (2016). A Survey on Energy Aware Resource Allocation for Cloud Computing. International Journal of Computer Applications, 142(1), 38-44.

[2]Mousavi, S., & Várkonyi-Kóczy, A. R. (2017). Dynamic Resource Allocation in Cloud Computing. Acta Polytechnica Hungarica, 14(2), 75-93.

[3]Rajput, R. S., & Pant, A. (2014). Optimal Resource Management in the Cloud Environment-A Review. International Journal of Advanced Research in Computer Science and Software Engineering, 4(7), 435-440.

[4]Pillai, P. S. (2015). Resource Allocation in Cloud Computing Using the Uncertainty Principle of Game Theory. International Journal of Advanced Research in Computer Engineering and Technology, 4(7), 2260-2267.

[5]Pawar, C. S. (2015). Priority Based Dynamic Resource Allocation in Cloud Computing. International Journal of Science and Research, 4(7), 1649-1655.

[6]Gawali, M. B. (2015). Task scheduling and resource allocation in cloud computing using a heuristic approach. International Journal of Computer Science and Mobile Computing, 4(8), 193-201.

[7]Pradhan, P. (2017). Modified Round Robin Algorithm for Resource Allocation in Cloud Computing. International Journal of Computer Applications, 173(3), 29-33.

[8]An, B., Lesser, V., & Irwin, D. (2014). Automated Negotiation with Decommitment for Dynamic Resource Allocation in Cloud Computing. In Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems (pp. 1425-1426).

[9]Choi, Y. (2018). Optimization Approach for Resource Allocation on Cloud Computing for IoT. In 2018 4th International Conference on Information Management (ICIM) (pp. 129-133). IEEE.

[10]Tsai, J. T. (2019). Optimized task scheduling and resource allocation on cloud computing environment using improved differential evolution algorithm. Cluster Computing, 22(1), 347-360.

[11]Murali, J. A., & Brindha, T. (2023). Efficient Resource Allocation in Cloud Computing Using Hungarian Optimization in Aws.

[12]Belgacem, A., Mahmoudi, S., & Kihl, M. (2022). Intelligent multi-agent reinforcement learning model for resources allocation in cloud computing. Journal of King Saud University-Computer and Information Sciences

[13]Ficco, M., Esposito, C., Palmieri, F., & Castiglione, A. (2018). A coral-reefs and game theory-based approach for optimizing elastic cloud resource allocation. *Future Generation Computer Systems*, *78*, 343-352.

[14]Feng Shi, Jingna Lin, "Virtual Machine Resource Allocation Optimization in Cloud Computing Based on Multiobjective Genetic Algorithm", Computational Intelligence and

Neuroscience, vol. 2022, https://doi.org/10.1155/2022/7873131

[15]Xin Xu, Huiqun Yu, "A Game Theory Approach to Fair and Efficient Resource Allocation in Cloud Computing", *Mathematical Problems in Engineering*, vol. 2014, Article ID 915878, 14 pages, 2014. https://doi.org/10.1155/2014/915878

[16] "Resource Allocation Techniques for Improving QoS in Cloud Computing ", International Journal of Science & Engineering Development Research (www.ijrti.org), ISSN:2455-2631, Vol.7, Issue 6, page no.821 - 827, June-2022, https://www.ijrti.org/papers/IJRTI2206136.pdf

[17]Junjie Cen, Yongbo Li, "Resource Allocation Strategy Using Deep Reinforcement Learning in Cloud-Edge Collaborative Computing Environment", Mobile Information Systems, vol. 2022, Article ID 9597429, https://doi.org/10.1155/2022/9597429