JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Comprehensive Survey on Dynamic Resource Allocation for Virtual Machine in Cloud

[1]Viplav Soni, [2] Dr. Rakesh Kumar Khare

[1]Research Scholar, [2] Associate professor
[1,2] Computer Science Engineering,
[1,2] Shri Shankaracharya Technical Campus, Bhilai chhattisgarh, India

Abstract: Cloud computing has had a significant impact on how people deploy and manage IT resources. Fundamentally, cloud computing entails the online supply of computing resources including storage, processing power, or apps. This model offers users the flexibility to use and utilize resources as needed, paying only for what they consume, without the need for extensive on-premises infrastructure. The aim of this review article is to present a comprehensive survey of the dynamic resource allocation techniques employed in virtual machines within cloud computing environments. The paper aims to shed light on the various approaches, optimization algorithms, and performance metrics used in this domain. The scope of this review encompasses a wide spectrum of techniques, including heuristic algorithms, machine learning-based approaches, and feedback control mechanisms. It also covers key performance metrics such as resource utilization, response time, energy efficiency, and cost-effectiveness, which are crucial in evaluating the effectiveness of dynamic allocation strategies.

***IndexTerms* – Cloud Computing, Virtualization, Dynamic Resource Allocation(RA), Virtual Machine, Optimization.**

I. INTRODUCTION

The term "cloud computing" refers to a method of delivering computing services over the internet ("the cloud"), involving servers, storage, databases, networking, software, analytics, or information. Businesses may utilize these services on demand, pay only for the services they actually use, rather than purchasing, owning, and servicing real servers as well as additional IT equipment. Virtualization is a crucial component of how cloud computing functions. Multiple applications and operating systems can operate on the same physical hardware thanks to a method known as virtualization. For each operating system or application, a virtual machine (VM) is created to accomplish this. Each VM is isolated from the others, so they cannot interfere with each other.

Virtual machine resource allocation is the method of allocating computing resources, including CPU, memory, and storage, to VMs. This can be done manually or automatically. There are 2 types of VMRA; static & dynamic. Static resource allocation is the process of assigning a fixed amount of resources to each VM at the time it is created. This is the simplest approach to resource allocation, but it can be inefficient. For example, if a VM needs more resources than it is allocated, it will not be able to perform at its full potential. Whereas dynamic resource allocation is the process of assigning resources to VMs as needed. This is a more complex approach to resource allocation, but it can be more efficient. Dynamic resource allocation can help to ensure that VMs have the resources they need when they need them, while also avoiding over-provisioning of resources [1]. Table 1 summarizing the key differences among static & dynamic resource allocation.

Table 1: Differences between static and dynamic resource allocation

Features	Static Resource Allocation	Dynamic Resource Allocation
Complexity	Simple	Complex
Efficiency	Less efficient	More efficient than static resource allocation
Stability	Not scalable	Scalable
Flexibility	Not flexible	Flexible

Dynamic resource allocation is the process of assigning computing resources to virtual machines (VMs) in a cloud computing environment in a manner that gathers the requires of the VMs while optimizing the use of the underlying physical resources. This is in contrast to static resource allocation, where VMs are assigned a fixed amount of resources at the time they are created. There are many factors that can affect the need for resources by VMs, such as the type of application running on the VM, the number of users accessing the VM, and the amount of data being processed. Dynamic resource allocation can help to ensure that VMs have the resources they need when they need them, while also avoiding over-provisioning of resources, which can lead to wasted resources and increased costs [2].

The choice of a dynamic resource allocation approach will depend on the specific needs of the cloud environment [3]. There are a number of different approaches to dynamic resource allocation, every with its own benefits & drawbacks. A few of the most popular approaches contains:

- **Demand-based allocation**: This approach allocates resources to VMs based on their current demand. This is the simplest approach to dynamic resource allocation, but it can lead to over-provisioning of resources during periods of low demand.
- **Predictive allocation**: This approach uses historical data to predict the future demand for resources by VMs. This can help to avoid over-provisioning of resources, but it can also lead to under-provisioning of resources during periods of high demand.
- **Proportional allocation**: This approach allocates resources to VMs in proportion to their importance. This can help to ensure that critical VMs always have the sources they need, but it could also conduct to low-priority VMs being starved of resources.

There are several challenges and complexities associated with static resource allocation [4]. First, it can be difficult to predict the amount of resources that each VM will need. This could conduct to either under-provisioning or over-provisioning of resources. Second, static resource allocation does not enable for the dynamic scaling of sources to meet changing demand. Third, static resource allocation can be inflexible and difficult to change. Dynamic resource allocation can address the challenges and complexities of static resource allocation by dynamically assigning resources to VMs as needed. This can help to enhance the efficiency, scalability, & flexibility of cloud computing environments [5, 6]. Through efficiently allocating sources, cloud providers can:

- **Reduce costs**: By avoiding over-provisioning of resources, cloud providers can save money on the hardware and power needed to run their cloud infrastructure.
- **Improve performance**: By ensuring that VMs have the resources they need when they need them, cloud providers can improve the performance of their cloud applications.
- **Increase scalability**: By dynamically allocating resources to VMs, cloud providers can easily range their cloud infrastructure up or down to meet changing demand.
- **Improve reliability**: By avoiding resource contention, cloud providers can improve the reliability of their cloud infrastructure.

The aim of this review paper is to:

- To present a comprehensive overview of the different approaches to dynamic RA for VM in cloud computing.
- To highlight the significant of efficient RA in optimizing cloud infrastructure.
- To discuss the challenges and opportunities in DRA for cloud computing.

The following is how the paper is set up: Section II presents the literature review, and Section III discusses DRA strategies. Section V wraps up the paper by presenting optimization strategies for dynamic resource allocation.

II. LITERATURE REVIEW

Mengistu et al.[7] To satisfy the goals and limitations particular to Volunteer Cloud Computing, 3 heuristic-based methods were developed to describe the VM placement issue as a limited 0-1 multi-dimensional knapsack issue. A real Volunteer Cloud Computing test-bed provides direct proof of these techniques' successful competitive outcomes.

Savitha et al.[8] provided the P-PAVA approach, a perceptive priority aware VM allocation policy that takes into account an application's priority as well as its calculate, memory, and bandwidth requirements. Using a ML-based prediction model, the programme distributes the applications according to the priority it receives. Additionally, parallelization is utilized before allocating distinct workloads in way to lower the overhead of the allocation method. The method uses the First fit method as a baseline for the requests allocation with a less priority criterion in order to accomplish this. PPAVA outperforms the most recent approach for VM allocation for priority aware apps in terms of a number of metrics, including average response time, execution time, & power usage.

Vikrant Sharma et al. [9] By utilizing the enhanced emperor penguin optimization technique, authors offer a revolutionary method for assigning VM to satisfy the demands of specific consumers. Focusing on tried-and-true optimization approaches for effective VM allocation further highlights the benefit of the proposed strategy. The Java programming language was used to create this application, which was then installed on the Netbeans IDE 12.4.

V Pallavi et al. [10] For effective VM allocation with load balancing, the TDRB method is presented. It compares the standard deviation of all the accessible VMs with the threshold value. Only when the VM's standard deviation is small than or equal to the threshold value are the jobs assigned to it. The results of the experiments show that the suggested TDRB conducts more effectively than the current techniques to distribute workloads to the VMs in the cloud environment.

Feng Shi et al. [11] introduces a novel multiobjective resource allocation strategy for the stability of multivirtual machine distribution. By merging the current state & future anticipated data of every application demand, the cost of transferring VM and the reliability of the new VM placement state are carefully taken into considerations. The issue was resolved using the multiobjective optimization genetic algorithm (MOGANS). The VM distribution method generated by MOGANS has a longer stability time as compared to the GA-NN for energy conservation and multivirtual machine redistribution overhead. This research suggests a multiobjective optimization DRA approach (MOGA-C) based on MOEA/D for VM distribution in an effort to address this shortfall. By using empirical simulation, it is demonstrated that moGA-D could converge more quickly & produce comparable multiobjective optimization outcomes at the same calculation size.

Vadivel et al. [12] proposed enhnaced HPSO-MGA to provide a dynamic resource allocation mechanism that disperses work or requests across virtual machines. The feature outcome was likewise estimated by the rule generating process. The user or consumer starts the process by gathering data from numerous online resources. Task Manager will receive these data. Authors can extract parameters from the task manager such as the task's cost, speed, data size, weight, etc. Similar to this, may collect information from the cloud storage about system resources like CPU Utilization, Memory Usage, Processing Speed, & Process Cycle, Bandwidth, Set of Requests, load on VMs, and Disc space. To reduce the processing time needed for dynamic resource allocation, Hybrid PSO and Modified Genetic Algorithm (HPSO-MGA) is used here to choose the necessary features. Only essential features will be used out of all the features. As a result, resource allocation takes less time and is less accurate. As a consequence, the execution time of the suggested HPSO-MGA is decreased while resource allocation performance is improved. When compared to the current approaches, the performance of our proposed RA methodology is superior.

Khan et al. [13] the optimal distribution of the computational workload to the resources in heterogeneous CPU-GPU systems as the basis for a power-conscious scheduling technique. In order to lower the peak power, the scheduler controls the resources of numerous computing nodes. The technique can be used in conjunction with software services that allow for flexible power states to further cut the cost of computing during times of heavy demand. Though the approach in our study is limited to GPU workloads, it can be applied to other heterogeneous systems. A genuine CPU-GPU heterogeneous system has been used to implement the algorithm. During studies, it has been shown that the suggested method reduces peak power by 10% compared to the system without any power-aware rules and by up to 24% relative to the most severe case with a rise in implementation time of about 2%.Costs associated with the system and services are decreased as a result.

Chhabra et al. [14] created the concept of choosing the best host allocation plan that could minimize energy usage and increase resource utilization. For parallel & distributed applications, we provide an Optimal VM Placement for Load Balancing (OPLB) employing Maximum Likelihood estimate. In order to account for increases in throughput and failure rate, the problem is phrased in a speculative paradigm based on CPU, Memory, and Energy predictions. The performance evaluation shows that the suggested method significantly outperforms the sequential, random, and LB-BC virtual machine placement heuristics in terms of traffic scalability by up to 49.54%, 32.63%, and 19.23%, etc.

Belgacem et al. [15]suggested a dynamic resource allocation approach that can more quickly and effectively respond to client resource demands. In order to reduce both the makespan & the expense of employing VM, it also suggests the multi-objective search technique known as the Spacing Multi-Objective Antlion approach (S-MOAL). Its effects on energy usage & fault tolerance were also investigated. The experiment showed that, especially in terms of makespan, suggested method outperformed the PBACO, DCLCA, DSOS, and MOGA methods.

Praveenchandar et al. [16] suggested that an efficient dynamic resource allocation technique use enhanced job scheduling & an ideal power minimization method. It is possible to attain efficiency in RA in regards of job completion & reaction time by utilizing a prediction method or a dynamic resource table update method. This structure lowers the power usage in data centers, which produces an efficient result with regard to of power reduction. The suggested approach presents precise values for changing the resource table. A more efficient job scheduling approach & decreased power consumption strategy result in efficient resource allocation. When compared to other methods currently in use, a simulation result produces outcomes that are 8% better.

Tao et al. [17] developed a schedule technique for allocating global container resources depend on the FIS, which evaluates nodes' statuses. By executing sample scenarios, we demonstrate how to create a containerized test environment & confirm the efficiency of the resource allocation rules. The outcomes of the experiments demonstrate that the architecture and schema that are currently being used produce the best resource combinations and greatly enhance cluster efficiency.

Girish Kalele et al. [18] provided a Secured Dynamic Priority Weighted Scheduling technique that utilizes Virtualized Subnet Transmission Support Factor. Additionally, Secure Time variant Traffic evaluation, packet transfer on networks with virtualized security, and cloud scheduling procedure are all improved by Secure Time. In comparison to the alternative system, the suggested enhances scheduling efficiency as well as the security and trust aspect.

Dhiraj Singh et al. [19] implemented the Targetive dynamic secured resource allocation (TDSRA) method, suggest a virtualized secure cloud computing system. Load-balancing technique is used in distributed computing to improve resource utilization security in computer systems. Particularly, rather than problems like energy economy, throughput, or resource planning, the emphasis is on enhancing virtualization app planning relative to other platforms or boosting cloud user security effectiveness.

Yongqiang Gao et al. [20] suggested a dynamic resource allocation system for MMOGs in heterogeneous cloud computing systems that utilizes both dynamic VM consolidation between PMs as well as STM based VM resizing at physical machine degree. Suggested plan takes into account various resource kinds, AFK gaming characteristics, diverse PMs & VMs, rigorous QoE demands, overheads associated with VM migration. Additionally, a brand-new hybrid technique for dynamic VM consolidation in heterogeneous cloud data centres is provided. It depends on differential evolution as well as an improved first-fit heuristic. The results of the study demonstrate that, as contrasted to the conventional over-provisioning policy, suggested resource allocation strategy may preserve MMOG players' quality of experience while achieving energy savings of up to 44.8%.

Gao et al. [21] suggested PORA, an effective plan that uses predictive data to address the issue. The suggested mathematical investigation and simulation findings demonstrate that PORA delivers minimal latencies & nearly ideal power consumption. Additionally, PORA is resilient to forecasting inaccuracy and effectively decreases latencies with only a small amount of predictive data.

Yu et al.. [22] developed the JRP issue of the RAUs & CUs as the solution to the energy-saving issue in C-RANs. The JRP issue resembles a unique bin-packing issue in that the active RAU selection and VM consolidation are connected. The amount of things and the sizes of the items in this challenge are connected & both movable. This issue cannot be resolved directly using any method currently in use. Therefore, in order to dynamically choose active RAUs or consolidate VMs to CUs, authors presented a powerful low-complexity method together with a context-aware approach. By doing this, authors can drastically lower the energy usage of C-RANs while minimizing the overhead associated with VM migration. The simulation findings show that suggested approach is efficient and workable for big networks.

III. DYNAMIC RESOURCE ALLOCATION TECHNIQUES

These techniques play a crucial role in optimizing resource utilization within cloud computing environments. These techniques ensure that resources are allocated and reallocated efficiently based on varying workload demands. They enhance the responsiveness, scalability, and overall efficiency of cloud infrastructure. Figure 1 shows the classification of dynamic resource allocation techniques.

Classification Based on Optimization Methods

Heuristic Algorithms

Heuristic approaches, like First-Fit, Best-Fit, & Worst-Fit, are simple and quick resource allocation methods. They aim to find reasonably good solutions without exhaustive computation. While they're easy to implement, they might not always lead to optimal resource utilization.

Genetic Approaches

Evolutionary biology serves as an inspiration for GA. They create and modify solutions iteratively to find optimal resource allocations. Genetic algorithms offer the advantage of exploring a wide range of possible solutions, but they might require more computational resources.

Machine Learning Approaches

Machine learning techniques, like reinforcement learning and neural networks, are employed to predict workload patterns and optimize resource allocation accordingly. These techniques adapt over time and can handle complex scenarios, but they require significant training data and computational power.

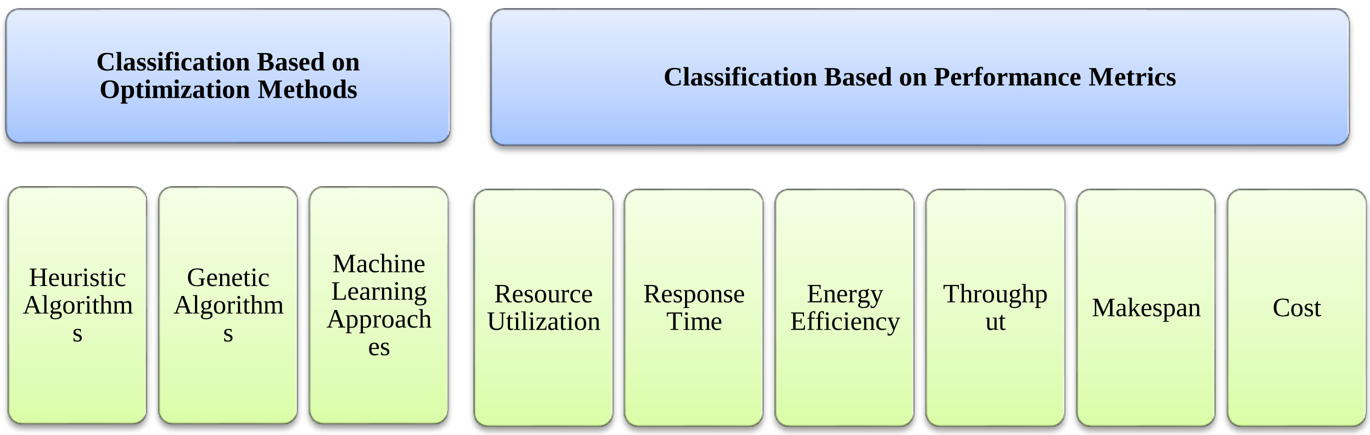

Figure 1: Classification of dynamic resource allocation techniques

Classification Based on Performance Metrics

Dynamic resource allocation techniques can also be classified based on the performance metrics they optimize. Some of the most common performance metrics include:

Makespan: Makespan is the whole set of time it takes to complete all the tasks in a workload.

Resource Utilization: Techniques focused on resource utilization ensure that resources are efficiently distributed among VMs to minimize waste. This metric is crucial for maximizing the use of available resources.

Cost: This is the amount of money that is spent on resources.

Response Time: Some techniques prioritize minimizing the response time for user requests. They allocate resources to VMs in a way that reduces waiting times and enhances user experience.

Energy Efficiency: Energy-efficient techniques optimize resource allocation to reduce power consumption, contributing to cost savings and environmental sustainability.

Throughput: Throughput is the set of tasks that could be completed per unit of time.

Classification Based on Architectural Approaches

Dynamic resource allocation techniques can also be classified based on their architectural approaches. Some of the most common architectural approaches include:

Centralized: In centralized approaches, a central controller monitors the resource usage across VMs and makes allocation decisions. This can be a more efficient approach to resource allocation, but it can also be more complex to implement and manage.

Decentralized: Decentralized techniques distribute decision-making among VMs themselves. This approach does not use a central controller. Instead, VMs communicate with each other to negotiate resource allocation. This can be a more scalable approach to resource allocation, but it can also be less efficient.

Hybrid: This approach combines centralized and decentralized approaches. They aim to harness the benefits of both approaches while mitigating their limitations. This can be a more efficient and scalable approach to resource allocation.

Table 2 summarizes the advantages and limitations of the different DRA techniques mentioned in this section.

Table 2: Advantages and limitations of the different DRA techniques

Techniques	Advantages	Disadvantages
Heuristic algorithms	Quick and easy implementation	Might not always find optimal solutions, especially in complex scenarios
Genetic algorithms	Can explore diverse solutions, suitable for complex problems	Requires substantial computational resources and parameter tuning
Machine learning approaches	Adapt to changing workload patterns, handle intricate scenarios	Demand large training datasets, resource-intensive training phase
Resource utilization-oriented techniques	Maximizes resource usage, reduces waste	May overlook other performance aspects
Response time-oriented techniques	Improves user experience, reduces waiting times	Might lead to underutilization of resources
Energy-efficient techniques	Lowers energy costs, environmentally friendly	Balancing energy efficiency and performance can be challenging
Centralized approaches	Suitable for small to medium-scale environments	May become a bottleneck in large-scale setups

Decentralized approaches	Scalable, reduced central control dependency	Potential for suboptimal solutions, coordination challenges
Hybrid approaches	Combines benefits of centralized and decentralized methods	Complexity in design and implementation

IV. OPTIMIZATION ALGORITHMS IN DYNAMIC RESOURCE ALLOCATION

DRA is a critical challenge in various fields, including manufacturing, transportation, and computer networking. Efficiently distributing resources based on changing demands is crucial for achieving optimal performance. In this section, three common optimization algorithms used for dynamic resource allocation: Genetic Algorithms (GAs), Particle Swarm Optimization (PSO), and Reinforcement Learning (RL).

Genetic Algorithms

Genetic algorithms (GA) are inspired by the process of natural selection [23]. They work by generating a population of solutions, called chromosomes, and then iteratively mutating and recombining them to produce better solutions. GAs are well-suited for issues with a huge set of variables and constraints. To adapt to changing resource demands, GAs can be configured to focus on finding new solutions that are significantly better than the current best solution, or they can be configured to explore the search space more widely. For example, if the resource demands are increasing, the GA can be configured to focus on finding solutions that use fewer resources.

PSO

PSO is a SI approach [24] that models the behavior of flocks of birds or schools of fish. In PSO, every particle indicates a possible approach, & the particles move through the search space by following the best-performing particles. PSO is relatively simple to implement & could be effective for a huge range of issues. In PSO [25] , each potential solution is represented as a particle in a multidimensional search space.

- **Initialization**: Particles are assigned random locations & velocities in the search space.
- **Update**: Every particle adjusts its velocity depend on its best-known position & the best-known position of its neighbors. This movement guides particles toward promising areas.
- **Adaptation**: As particles explore the search space, they adapt to changing resource demands by dynamically shifting their movement patterns. Particles that were once converging towards a particular solution might change direction if new resource demands suggest a more optimal region.

Reinforcement learning (RL)

RL [26, 27] is a ML algorithm that learns to make decisions by trial and error. In RL, an agent interacts with an environment and receives rewards for taking actions that lead to desired results. The agent learns to take actions that maximize the expected reward. RL is a strong algorithm that could be utilizes to solve a huge scale of problems, including dynamic resource allocation [28]. RL algorithms adapt to changing demands by continuously learning and updating their policies based on the observed outcomes and rewards. When resource demands shift, the agent adjusts its allocation strategy by favoring actions that lead to higher rewards under the new conditions [29].

V. CONCLUSION

This paper has provided a valuable overview of the state-of-the-art in DRA for virtual machines in cloud computing. The paper has identified the key challenges and opportunities in this area such as the need to balance user demands, ensure QoS, and optimize resource utilization, and it has highlighted the importance of efficient resource allocation in optimizing cloud infrastructure. Effective RA could help to enhance the utilization of sources, reduce costs, and enhance the performance of cloud applications. The paper is a valuable resource for researchers & practitioners who are interested in dynamic resource allocation for cloud computing. The paper has concluded by saying that DRA is a key part of cloud computing, and that it will continue to be important in the future.

REFERENCES

[1] A. Belgacem, "Dynamic resource allocation in cloud computing: analysis and taxonomies," *Computing*, vol. 104, no. 3, pp. 681–710, 2022. https://doi.org/10.1007/s00607-021-01045-2

[2] J. Vergara, J. F. Botero, and L. A. Fletscher, "A comprehensive survey on resource allocation strategies in FOG/Cloud environments," *Sensors*, vol. 23, no. 9, 4413, 2023. https://doi.org/10.3390/s23094413

[3] J. Bachiega, B. Costa, L. R. De Carvalho, M. Rosa, and A. Araujo, "Computational Resource allocation in Fog Computing: A Comprehensive survey," *ACM Computing Surveys*, vol. 55, no. 14s, pp. 1–31, 2023. https://doi.org/10.1145/3586181

[4] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing," *Future Generation Computer Systems*, vol. 28, no. 5, pp. 755–768, 2012. https://doi.org/10.1016/j.future.2011.04.017

[5] L. Zohrati, M. N. Abadeh, and E. Kazemi, "Flexible approach to schedule tasks in cloud-computing environments," *IET Software*, vol. 12, no. 6, pp. 474–479, 2017. https://doi.org/10.1049/iet-sen.2017.0008

[6] A. Kaur, B. Kaur, and D. Singh, "Optimization Techniques for resource provisioning and load balancing in cloud Environment: A review," *International Journal of Information Engineering and Electronic Business*, 2017. https://doi.org/10.5815/ijieeb.2017.01.04

[7] T. M. Mengistu, D. Che, and S. Lu, "Multi-Objective Resource Mapping and Allocation for Volunteer Cloud Computing," *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*, pp. 344–348, April 2019. https://doi.org/10.1109/cloud.2019.00063

[8] S. Savitha, and S. Salvi, "Perceptive VM Allocation in Cloud Data Centers for Effective Resource Management," *2021 6th International Conference for Convergence in Technology (I2CT)*. April 2021. https://doi.org/10.1109/i2ct51068.2021.9417960

[9] V. Sharma, and S. Bhardwaj, "Enhanced Resource Control in Cloud Environment using Optimized VM Allocation in Data Centers," *IEEE Xplore*, Dec. 2022. https://ieeexplore.ieee.org/document/10046918

[10] V Pallavi, and P. Vinothiyalakshmi, "Threshold based Dynamic Resource Balancing (TDRB) algorithm in Cloud Computing," *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, Aug. 2022

[11] F. Shi, and J. Lin, "Virtual Machine resource allocation optimization in cloud computing based on multiobjective genetic algorithm," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–10, March 2022. https://doi.org/10.1155/2022/7873131

[12] R. Vadivel, and T. S. Muthu, "An effective HPSO-MGA Optimization Algorithm for Demand based Resource Allocation in Cloud Environment," *6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, April 2020.

[13] H. Khan, M. U. Hashmi, I. ud, Janjua, K., Sikandar, A., Qazi, M. W., and Z. Hameed, "An Efficient Scheduling based cloud computing technique using virtual Machine Resource Allocation for efficient resource utilization of Servers," *2020 International Conference on Engineering and Emerging Technologies (ICEET)*, March 2020.

[14] S. Chhabra, and A. K. Singh, "Optimal VM Placement Model for Load Balancing in Cloud Data Centers," *2019 7th International Conference on Smart Computing & Communications (ICSCC)*. Sept. 2019.

[15] A. Belgacem, K. Beghdad-Bey, K., H. Nacer, and S. Bouznad, "Efficient dynamic resource allocation method for cloud computing environment," *Cluster Computing*, Feb. 2020.

[16] J. Praveenchandar, and A. Tamilarasi, "Dynamic resource allocation with optimized task scheduling and improved power management in cloud computing," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 3, pp. 4147–4159, March 2020.

[17] Y. Tao, X. Wang, X. Xu, and Y. Chen, "Dynamic Resource Allocation Algorithm for Container-Based Service Computing," *IEEE 13th International Symposium on Autonomous Decentralized System (ISADS)*. June 2017.

[18] K. Girish, "Time Variant Resource Allocation With Secured Dynamic Priority Weighted Scheduling Process Based On Virtualized Subnet Transmission Support Factor in Cloud Computing," *International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, June 2023.

[19] S. Dhiraj, and C. Mansi, "Virtualized Secure Cloud Computing Using Targetive Dynamic Secured Resource Allocation Algorithm," *International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*. June 2021.

[20] G. Yongolang, W. Lin, X. Zhulong, Q. Zhengwei, and Z. Jiantao, "Energy- And Quality Of Experience-Aware Dynamic Resource Allocation For Massively Multiplayer Online Games In Heterogeneous Cloud Computing Systems," *IEEE Transactions On Services Computing*, vol. 16, no. 3. pp. 1793 – 1806, July 2023.

[21] X. Gao, X. Huang, S. Bian, Z. Shao, and Y. Yang, "PORA: Predictive Offloading and Resource Allocation in Dynamic Fog Computing Systems," *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*. vol.7, no. 1, pp. 72 – 87, Oct. 2019.

[22] N. Yu, Z. Song, H. Du, H. Huang, and X. Jia, "Dynamic Resource Provisioning for Energy Efficient Cloud Radio Access Networks. IEEE Transactions on Cloud Computing, vol. 7, no. 4, pp. 964 – 974, June 2018.

[23] D. E. Goldberg, “Genetic algorithms in search, optimization, and machine learning,” *Choice Reviews Online*, vol. 27, no. 02, 27–0936, 1989. https://doi.org/10.5860/choice.27-0936

[24] A. Ewees, A., Nabil, N., R. A. Ibrahim, and R. M. ElEraky, “A Survey of Swarm Intelligence for Task Scheduling in Cloud Computing,” *In Swarm Intelligence for Cloud Computing*, pp. 1–37, 2020. https://doi.org/10.1201/9780429020582-1

[25] Y. Zhang, S. Wang, and G. Ji, “A comprehensive survey on particle swarm optimization algorithm and its applications,” *Mathematical Problems in Engineering*, vol. 2015, pp. 1–38, 2015. https://doi.org/10.1155/2015/931256

[26] Sutton, R. S., & Barto, A. G., “Reinforcement Learning: An Introduction,” *IEEE Transactions on Neural Networks*, vol. 16, no.1, pp. 285–286,2005. https://doi.org/10.1109/tnn.2004.842673

[27] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. , “Human-level control through deep reinforcement learning”, Nature, vol. 518, no. 7540, pp. 529–533,2015. https://doi.org/10.1038/nature14236

[28] Sepideh Goodarzy, Maziyar Nazari, Richard Han, Eric Keller, Eric Rozner, “Resource Management in Cloud Computing using Machine Learning: A survey,” *IEEE Conference Publication*, *IEEE Xplore*, pp. 811-816, 2020. https://ieeexplore.ieee.org/abstract/document/9356180

[29] Zhi, Y., Tian, J., Deng, X., Qiao, J., and D. Lu, “Deep reinforcement learning-based resource allocation for D2D communications in heterogeneous cellular networks,” *Digital Communications and Networks*, vol. 8, no. 5, pp. 834–842.,2022. https://doi.org/10.1016/j.dcan.2021.09.013