A LOAD BALANCING ALGORITHM FOR THE DATA CENTRES TO OPTIMIZE CLOUD COMPUTING APPLICATIONS

PAVITHRA S[1] , SANTHOSH A[2], PRAKASH K[3],SARAN SUJAI T[4]

Final Year of B-tech – Information Technology [1,2,3]

Assistant Professor, Department of Information Technology[4]

K.S.R College of Engineering (Autonomous)

Tiruchengode-637 215, Tamilnadu, India

ABSTRACT

Virtual Machines (VMs) in Cloud systems are scheduled to hosts based on their instant resource consumption (e.g., hosts with the greatest accessible RAM), rather than their overall and long-term utilisation. Furthermore, the scheduling and placement operations are often computationally intensive and have an impact on the performance of deployed VMs. In this paper, we provide a Cloud VM scheduling method that considers existing VM resource consumption over time by assessing previous VM utilisation levels in order to schedule VMs while maximising performance using the PSO technique. Because Cloud management activities such as VM placement have an impact on previously deployed systems, the goal is to minimise such performance deterioration. Furthermore, because overcrowded VMs tend to grab resources from neighbouring VMs, the task enhances the VMs' true CPU consumption.

The results reveal that our method refines traditional Instant-based physical machine selection as it learns and adapts to system behaviour over time. The idea of VM scheduling based on resource monitoring data taken from previous resource utilisations (VMs). The PSO classifier reduces the physical machine count by four.

Keywords: Cloud Data Centre, Virtual Machine, Energy Consumption, Resource Management

1.INTRODUCTION

Cloud computing is the computational paradigm of the future. It is quickly solidifying its position as the future of distributed on-demand computing. Cloud Computing is growing as a critical backbone for various online enterprises by utilising the notion of virtualization. On the other hand, Internet-enabled business (e-Business) is quickly becoming one of the most successful company models in the modern era. To meet the needs of internet-enabled businesses, computing is being turned into a paradigm of commoditized services offered in a way again to conventional utilities such as water. Customers may access services according on their needs, regardless of where they are housed or how they are provided. Many computing paradigms have pledged to provide utility computing. One such dependable computing paradigm is cloud computing. A front end and a back end comprise the architecture of cloud computing. These two ends are linked through the Internet or an intranet. Client devices such as thin clients, fat clients, and mobile devices are included in the front end. Clients require an interface and apps in order to access the cloud computing infrastructure. The many servers and data storage systems comprise the back end. There is also a "Central Server" server. The cloud system is managed by a centralised server. It also monitors general traffic and responds to client requests in real time.

1.1 CLOUD DATA CENTER

A cloud data centre is a state-of-the-art setup created to efficiently and scalable store and handle massive volumes of digital data, apps, and services. Cloud data centers, as opposed to conventional on-premises data centers, use virtualization and cloud computing technologies to offer remote resource access via the internet. This eliminates the need for substantial physical hardware upkeep and lets companies and individuals to access, develop, and expand their IT infrastructure and services as needed. The digital age is mostly powered by cloud data canters, which provide dependable, adaptable, and affordable solutions for a variety of computing requirements, including processing and storing large amounts of data as well as managing intricate applications and services.

1.2 VIRTUAL MACHINE

A virtual machine, sometimes known as a VM for short, is an essential part of contemporary computing that has completely changed how we organize and use hardware resources. A virtual machine is essentially a software-based simulation of a real computer system, replete with an operating system and software of its own. It permits the operation of numerous virtualized instances on a single physical server, facilitating flexible computing environment management, effective resource usage, and task segregation. Because they allow multiple operating systems and applications to run on a single piece of hardware, virtual machines are widely used in data centers, cloud computing, and development environments. This facilitates hardware consolidation and makes it easier to

test, develop, and deploy a wide range of software solutions. This technology, which improves computing environments' scalability, security, and general efficiency, has emerged as a key component of contemporary IT infrastructure.

1.3 ENERGY CONSUMPTION

Energy consumption, which includes the quantity of energy used to power our homes, businesses, transportation, and numerous technological equipment, is an essential component of both our everyday lives and the global economy. It is a vital component of economic growth and a crucial determinant of the sustainability and environmental impact of our actions. The way we generate, distribute, and use energy is a critical factor in tackling environmental issues like resource conservation and climate change. Achieving a more sustainable future requires understanding and controlling energy use, which entails making decisions about energy sources, efficiency, and conservation tactics that can lessen their negative effects on the environment and build a more robust energy infrastructure.

1.4 RESOURCE MANAGEMENT

A key idea that supports the effective distribution and use of valuable resources in a variety of settings, such as enterprises, governmental organizations, and environmental conservation initiatives, is resource management. It includes organizing, allocating, and keeping an eye on resources including money, people, time, materials, and technology in order to meet goals and increase output. Organizations that want to improve overall performance, cut waste, and optimize operations must practice effective resource management. It is a broad discipline that includes strategic decision-making, coordination, and ongoing assessment of resource use to guarantee that resources are used prudently and in accordance with the objectives of the business. Furthermore, resource management is becoming a vital factor in today's global economy as it is acknowledged as a vital instrument for attaining sustainability and responsible stewardship of our planet's natural resources.

2.LITERATURE REVIEW

This work was proposed by Yong Yu [1] et.al. Remote data integrity checking (RDIC) allows a data storage server, such as a cloud server, to demonstrate to a verifier that it is honestly storing a data owner's data. A number of RDIC protocols have been presented in the literature to date, however the most of the constructs suffer from the issue of sophisticated key management, that is, they rely on the expensive public key infrastructure (PKI), which may impede RDIC implementation in practise. In this research, we present a novel identity-based (ID-based) RDIC protocol that uses key-homomorphic cryptographic primitives to decrease system complexity and costs associated with creating and administering the public key authentication framework in PKI-based RDIC schemes. We formalise ID-based RDIC and its security model, which includes protection against a rogue cloud server and zero knowledge privacy from a third-party verifier. Throughout the RDIC procedure, the proposed ID-based RDIC protocol does not provide any information about the stored data to the verifier. In the generic group model, the novel design has been shown secure against the malicious server and provides zero knowledge privacy against a verifier. Comprehensive security research and implementation results show that the suggested protocol is both provably safe and applicable in real-world situations.

This paper was proposed by UsmanWazir [2] et.al. Cloud computing makes scattered resources available to people all over the world. Cloud computing has a scalable design that delivers on-demand services to enterprises across several disciplines. Yet, there are several obstacles with the cloud services. In the cloud services, several solutions have been presented to address various types of issues. This study examines the many models presented for SLA in cloud computing in order to tackle the issues that exist in SLA. Problems with performance, customer satisfaction, security, profit, and SLA violations. SLA architecture in cloud computing is discussed. Next we review existing SLA models presented in various cloud service formats such as SaaS, PaaS, and IaaS. With the use of tables, we explore the benefits and limits of current models in the following section. We summarise and offer a conclusion in the last part. In this research, we examined several SLA models utilised in cloud computing environments. Some models can provide high-level security safeguards for customer data, while others can impose penalties for SLA violations. Some of these promote consumer trust, while others improve their performance as compared to other models. To construct a SLA between a customer and a cloud service provider, we must first define the function of the cloud service provider.

This paper was proposed by PritiNarwal [3] et.al. Cloud computing is a rapidly evolving and dynamic platform that employs virtualization technologies. Virtualization isolates the hardware system resources in software in the Cloud computing environment so that each application may operate in an isolated environment called the virtual machine, and the hypervisor allocates virtual machines to various users who are hosted on the same server. Although it has several advantages like as resource sharing, cost-efficiency, high-performance computing, and lower hardware costs, it also has a variety of security risks. Threats can exist directly on Virtual Machines (VMs) or indirectly on Hyper-visors via virtual machines hosted on them. This paper provides an overview of all potential security risks as well as responses using Game Theoretic techniques. Because of the autonomous and strategic rational decision making nature of cloud users, where each player competes for the best feasible answer in a safe manner, Game Theory may be employed as a protective mechanism. Because various users have varying resource requirements in a cloud environment, it should focus on other concerns such as efficiency and optimization in addition to security and privacy. As a result, a comparative examination of many models that employ game theory is conducted in order to gain a fundamental knowledge of security concerns that occur for users and previously used approaches for fair resource allocation.

This work was proposed by Nitin Kumar Sharma [4] et al. Attribute Based Access Control (ABAC) models are intended

to solve the problems of traditional access control models (DAC, MAC, and RBAC) while integrating their benefits. ABAC provides access control based on generic properties of entities. Many organisational security rules make access decisions dependent on qualities. OWL may be used to officially describe and process security policies represented by ABAC models. With OWL, we created models, domains, data, and security regulations, and then used a reasoned to determine what is permissible. We offer a method for representing the ABAC model using Web Ontology Language in this work (OWL). The EYE reasoned infers the logical link and deduces the access permit for each requested activity to enforce policies. We demonstrated how the Attribute Based Access Control model may be expressed using Web Ontology Language in this study (OWL)

This work was proposed by ZiadIsmail [5] et al. Cloud computing advancements have created substantial security difficulties in ensuring the confidentiality, integrity, and availability of outsourced data. Typically, a Service Level Agreement (SLA) is signed between the cloud provider and the customer. It is critical to check the cloud provider's compliance with data backup standards in the SLA for redundancy considerations. There are several security methods in place to ensure the integrity and availability of outsourced data. This work can be completed by the client or assigned to an independent body known as the verifier. Nevertheless, evaluating data availability incurs additional fees, which may dissuade customers from doing data verification on a regular basis. Game theory may be used to capture the interaction between the verifier and the cloud provider in order to determine the best data verification approach. This problem is formulated as a two-player non-cooperative game in this paper. We explore the scenario where each piece of data is copied a number of times depending on a set of factors such as its size and sensitivity. We investigate the cloud provider's and verifier's tactics at the Nash Equilibrium and deduce the predicted behaviour of both participants.

3. RELATED WORK

The power and battery life of Smart Mobile Devices (SMDs) hinder the development of computational-intensive and delay-sensitive applications as technology advances. Mobile edge computing (MEC) has the potential to meet application needs and reduce workload on SMDs through compute offloading. Offloading computations in a multi-server and multi-task context is challenging due to device mobility and server status fluctuation. To address these difficulties, we introduce a parallel task offloading mo del and a small area-based edge offloading strategy in MEC. We create an optimization problem to minimize task completion time and use a deep reinforcement learning-based offloading technique with a Markov decision strategy. We provide a deep deterministic policy gradient (DDPG) technique for determining offloading strategy. Experimental results show that the DDPG-based offloading strategy outperforms previous strategies by at least 19% in terms of ultra-low latency, server efficiency, and SMD mobility.

4.METHODOLOGY

The goal is to offer the idea of VM scheduling based on resource monitoring data taken from previous resource utilisations and to assess previous VM usage levels using two classification techniques such as PSO in order to schedule VMs while maximising performance. The suggested VM scheduling technique improves the VM selection phase by collecting real-time monitoring data and analysing physical and virtual resources. Our goal is to improve VM scheduling by including factors relating to real VM use levels, so that VMs may be deployed while reducing the penalization of overall performance levels. The optimization approaches incorporate analytics on previously deployed VMs in order to (a) maximise utilisation levels and (b) minimise performance losses. Users have underused VMs and do not have the same resource usage pattern throughout the day. Lastly, cloud management activities such as VM placement have an impact on previously deployed systems (for example, performance loss in a database cluster) since heavily loaded VMs tend to steal CPU time from neighbouring VMs. These are basic examples that show the need for more precise VM scheduling that might increase performance. VM Id, CPU, RAM, and BW are the input datasets. The cloud, which generates CPU, Memory, and bandwidth. The VM machine of the allocation has been allocated to the host of that particular host id. The VM is migrated so that the assigned host may be scheduled.

A. Vm Scheduling

The proposed technique improves the VM selection phase by collecting real-time dataset monitoring data and analysing physical and virtual resources. Our goal is to improve VM scheduling. To add criteria relating to real VM use levels, so that VMs may be put while reducing the penalization of overall performance levels. The optimization approaches use analytics to previously deployed VMs in order to (a) maximise utilisation levels and (b) minimise performance losses. A monitoring engine that collects data about VM resource utilisation monitoring online. The engine may gather system data at regular intervals and store it in an online cloud service where it can be processed. Data is gathered at regular intervals (e.g., every 1 second) and saved in a temporary local file.

B. Classification Algorithm

The classification algorithm is an important component used to categorize and arrange data in the context of VM scheduling. This module will most likely use machine learning approaches to classify virtual machines based on their resource use characteristics. The algorithm's ability to discriminate between different types of virtual machines (VMs) is critical for making informed judgments about their placement and resource allocation within cloud infrastructure.

C. Particle Swarm Optimization

Particle swarm optimization (PSO) is a basic bio-inspired technique for searching for an optimal solution in the solution space. It differs from other optimization techniques in that it requires only the objective function and is not affected by the gradient or any differential form of the goal.

It also features a small number of hyper parameters. Using an example, you will discover the logic for PSO and its algorithm

D. Optimization Scheme

The goal of these optimization approaches is to define the weight of the Vm based on the vms' resource utilizations. This will give information about the state of previously deployed vms, such as whether or not a workload is executing. We present two optimization strategies to do this. The PSO is used to classify the Vm state in terms of its current resource utilizations. The virtual machine resource utilization information is initially gathered and monitored, and the acquired data is subsequently categorized using machine learning algorithms such as PSO.

5. ALGORITHM DETAILS

Particle swarm optimization

The basic principles in "classical" PSO are very simple. A set of moving particles (the swarm) will be set into the search area. Each particle has a position vector of Xi and a velocity vector Vi. The position vector Xi and the velocity vector Vi of the i th particle in the n-dimensional search area could be represented as Xi= (xi1, xi2, xin) and Vi= (vi1, vi2, vin) respectively. The memory with which each particle finds the best position is called Pbest and best location is known as Gbest. Assume Pbest = (xi1Pbest, xi2Pbest, ..., xin(Pbest) and Gbest = (x1Gbest, x2Gbest, ..., xn(Gbest)) be the best positions of the individual i and all the individuals. At each level, the velocity of the i th particle will be updated according to the following equation in the PSO algorithm.

$Vik+1=\omega Vik+c1r1\times Pbestik-Xik + c2r2\times Gbestk-Xik$ ……... (1)

In this velocity updating process, the acceleration coefficients c1, c2 and the inertia weight ω are predefined and r1, r2 are uniformly generated random numbers in the range of [0, 1]. In general, the inertia weight ω is set according to the following equation:

$\omega= \omega max-\omega max-\omega minitermax\times iter$…… (2)

The approach used by Eq (2) is called the “inertia weight approach. With the help of above equation diversification characteristic is gradually decreased and a specific velocity, which gradually moves through the current searching point close to Pbest and Gbest, can be calculated. Each individual moves from the present position to the next one by the modified velocity in Eq (1) using the following equation:

$Xik+1=Xik+Vik+1$…… (3)

Initialize particle;
End;
Do for each particle;
Calculate fitness value;
If the fitness value is better than the best fitness value (pbest) ;
Set current value as the new pbest;
End;
Choose the particle with the best fitness value of all the particles as the gbest;
For each particle;
Calculate particle velocity according equation; Update particle position according equation; End;
While maximum iterations or minimum error criteria are not attained;

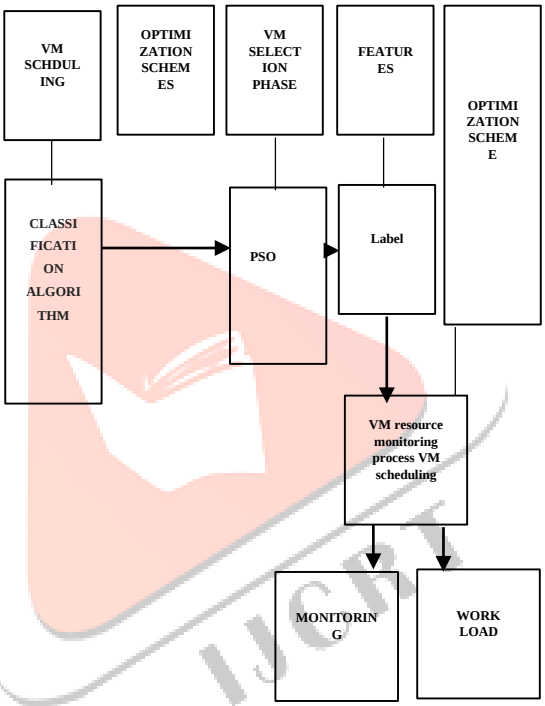

Figure. 1 .block diagram

6. RESULTS AND DISCUSSION

The emphasis is on CloudSim, an open source programme for creating private and public clouds. CloudSim default setup involves installing virtual machines by picking the host with the greatest available ram until the number of virtual machines exceeds the limit. Virtual Machines (VMs) are assigned to hosts based on their immediate resource consumption (e.g., hosts with the greatest available RAM), without regard for their overall and long-term utilisation. Furthermore, the scheduling and placement operations are often computationally intensive and have an impact on the performance of deployed VMs. As a result, the standard VM placement method does not take into account previous VM resource use levels. This is addressed by implementing a VM scheduling method. The notion of VM scheduling based on resource monitoring data derived from previous resource utilisations (including VMs and VMs), and the resource data

are categorised using the optimization techniques PSO, thus scheduling is performed. The programme analyses previous resource consumption levels and classifies them based on overall resource usage. Finally, a list of possible hosts is generated, and the resources are sorted accordingly. In detail, VMs are re-ranked using this method based on the optimization technique chosen and their VM consumption.

algorithm	accuracy
DDPG	75
Ant Lion	81

Table 1. Comparison table

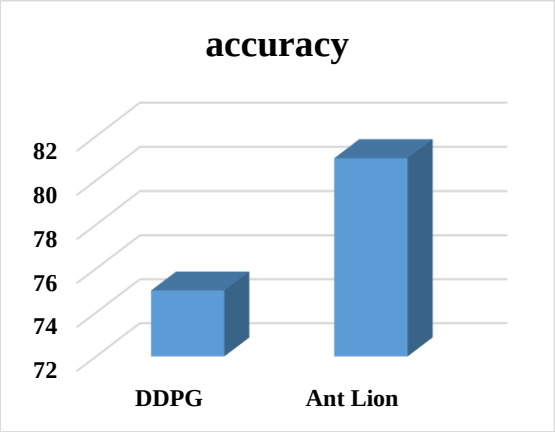

Figure 2. Comparison graph

The table compares algorithmic accuracy results for two distinct techniques: Deep Deterministic Policy Gradients (DDPG) and Ant Lion. DDPG has a commendable accuracy level of 75%, which reflects its usefulness in specific applications. Ant Lion, on the other hand, achieves an even higher accuracy of 81%, indicating superior performance in the settings or datasets under consideration. These accuracy measures are useful standards for evaluating and selecting algorithms based on their predictive skills, demonstrating the different levels of success obtained by DDPG and Ant Lion in their respective fields.

7.CONCLUSION

Various virtual machine placement techniques were employed for scheduling by selecting physical computers based on system statistics (i.e. CPU, memory, and bandwidth utilisation) in a cloud system. The current VM placement does not consider real-time VM resource use levels. In this section, we present a novel VM placement technique based on previous VM usage experiences. The VM consumption is then observed, and the data is trained using machine learning models (PSO) to anticipate VM resource utilisation and arrange VMs accordingly. It was proposed an algorithm that permits VM placement based on PM and VM consumption levels, as well as a computational learning approach based on the notion of assessing previous VM resource utilisation based on historical data to optimise the PM selection phase. A virtual machine placement technique based on real-time virtual resource monitoring was introduced, with machine learning models used to train and learn from prior virtual machine resource utilisation. As a result, a monitoring engine providing resource utilisation statistics is assumed. Using the PSO classifier instead of the Support Vector Machine (SVM) classifier reduces the physical machine count by four. The work was completed using ten virtual machines.

8. REFERENCES

[1] Y. Yong, M. H. Au, and G. Ateniese, Identity-based remote data integrity verification for cloud storage with complete data privacy preservation, IEEE Transactions on Information Forensics and Security, vol. 12, no. 4, pp.767-778, 2020.

[2] U. Wazir, F. G. Khan, and S. Shah A review of service level agreements in cloud computing, International Journal of Computer Science and Information Security, vol. 14, no. 6, p. 324, 2019.

[3] P. Narwal, D. Kumar, and M. Sharma, A review of game-theoretic techniques for safe virtual machine resource allocation in the cloud, Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, 2019.

[4] N. K. Sharma and A. Joshi, Modeling attribute-based access control rules in owl, IEEE Tenth International Conference on Semantic Computing (ICSC), 2020, pp. 333-336.

[5] Z. Ismail, C. Kiennert, J. Leneutre, and L. Chen, A game theoretical examination of a cloud provider's compliance with data backup requirements, IEEE Transactions on Information Forensics and Security, vol.11, no.8, pp. 1685-1699, 2019.

[6]. H. Kaur and S. Ajay, “Renewable Energy-based Multi-Indexed Job Classification and Container Management Scheme for Sustainability of Cloud Data Centers,” Next Generation Computing Technologies(NGCT), pp. 516–521, 2016.

[7] K. Toutanova and C. Cherry, "A placement architecture for a container as a service (CaaS) in a cloud environment," in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1- Volume 1, Association for Computational Linguistics, 2009, pp. 486–494.

[8] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Energy Consumption Optimization of Container-Oriented Cloud Computing Center," at the 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2015, pages 4580–4584.

[9]T. Mikolov and G. Zweig, "An energy, performance efficient resource consolidation scheme for heterogeneous cloud datacenters," in 2012 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2012, pages 234-239.

[10]. Rizky, W. M., Ristu, S., & Afrizal, D. "HeporCloud: An energy and performance efficient resource orchestrator for hybrid heterogeneous cloud computing environments". Scientific Journal of Informatics, Volume 3(2), Pages 41-50, November 2016.

[11] In J. Appl. Sci. Technol. Trends, vol. 1, no. 3, pp. 98–105, 2020, A. Salih, "Cloud Computing Virtualization of Resources Allocation for Distributed Systems," doi: 10.38094/jastt1331.

[12] Cloud computing: A paradigm shift in computing, M. Agarwal and D. G. M. S. Srivastava, Int. J. Mod. Educ. Compute Sci., vol. 9, no. 12, pp. 38–48, 2020, doi: 10.5815/ijmecs.2017.12.05.

[13] Int. J. Cloud Comput. Serv. Archit., vol. 5, no. December, pp. 1–9, 2015; doi: 10.5121/ijccsa.2015.5602; N. Zanoon, "Towards Cloud Computing: Security and Performance."

[14] Benefits and Difficulties of Cloud Computing Adoption in Business, C. T. S. Xue and F. T. W. Xin, Int. J. Cloud Computing Serv. Archit., vol. 6, no. 6, pp. 01–15, 2022, doi: 10.5121/ijccsa.2016.6601.

[15] "Proposing A Load Balancing Algorithm For The Optimisation Of Cloud Computing Applications," D. A. Shafiq, N. Jhanjhi, and A. Abdullah, MACS, 2020, pp. 1–6, doi: 10.1109/MACS48846.2019.9024785.
Reference:

[16] S. K. Mishra, B. Sahoo, and P. P. Parida, "Cloud computing load balancing: A broad overview," Comput. Inf. Sci. J. King Saud Univ., 2021; doi: 10.1016/j.jksuci.2018.01.003.

[17] Cloud Computing Architecture: A Critical Analysis, I. Odun-Ayo, M. Ananya, F. Agono, and R. Goddy-Worlu, in ICCSA, 2023, pp. 1–7, doi: 10.1109/ICCSA.2018.8439638.

[18] "Cloud Computing and Load Balancing in Cloud Computing -Survey," A. Jyoti, M. Shrimali, and R. Mishra, Confluence, 2022, pp. 51–55, doi: 10.1109/MTAS.2004.1371634.

[19] "Performance comparison of heuristic algorithms for task scheduling in IaaS cloud computing environment," PLoS One, vol. 12, no. 5, 2019, doi: 10.1371/journal.pone.0176321, S. H. H. Madni, M. S. Abd Latiff, M. Abdullahi, S. M. Abdulhamid, and M. J. Usman.

[20] Adhikari, M., and Amgoth, T., "Heuristic-based load-balancing algorithm for IaaS cloud," in Future Generation Computing Systems, vol. 81, pp. 156–165, 2020, doi: 10.1016/j.future.2017.10.035.

[21] A study on virtualization and hypervisor in cloud computing by B. and G. Singh, vol. 6, no. 1, pp. 17–22, 2021.A complete review for scheduling strategies in cloud computing,

[22] M. Kumar, S. C. Sharma, A. Goel, and S. P. Singh, J. Netw. Comput. Appl., vol. 143, no. June, pp. 1–33, 2019 doi: 10.1016/j.jnca.2019.06.006.
Optimal Performance Versus Fairness Tradeoff for Resource Allocation in Wireless Systems, IEEE Trans. Wirel. Commun., vol. 16, no. 4, pp. 2587–2600, 2021, doi: 10.1109/TWC.2017.2667644,

[23] F. Zabini, A. Bazzi, B. M. Masini, and R. Verdone.Int. J. Comput. Appl., vol. 42, no. 1, pp. 108–117, 2020 ,doi:10.1080/1206212X.2017.1404823;

[24] M. Kumar and S. C. Sharma, "Dynamic load balancing algorithm to minimise the makespan time and utilise the resources effectively in cloud environment. "In Procedia Computer Science, vol. 57, pp. 545–553, 2021,

[25] G. Patel, R. Mehta, and U. Bhoi, "Enhanced Load Balanced Minmin Algorithm for Static Meta Task Scheduling in Cloud Computing," doi: 10.1016/j.procs.2015.07.385.