ISSN Number: 2454-2024(Online)

International Journal of Technical Research & Science

DYNAMIC LOAD BALANCING FOR CLOUD: A REVIEW

Devendra Suthar[1], Jitendra Singh Chouhan[2]
E-Mail Id: dev_arya123@yahoo.com[1], chauhan.jitendra@live.com[2]
Department of Computer Engineering, Aravali Institute of Technical Studies, Udaipur, Rajasthan, India

Abstract- The newest internet-based technology that places a strong emphasis on business computing is called cloud computing. Load balancing aids in enhancing the centralized server's efficiency. Using an analytical tool called cloud analyst, numerous algorithms are analyzed in the current work. Load balancing algorithms are also compared. The cloud computing load balancing issue is a significant one, a crucial element for proper system performance, and it has the potential to slow the industry's rapid growth. Recently, there has been a fast increase in the number of customers from all over the world requesting various services.

Keywords: VM, cloud computing, task scheduling, priority scheduling, scheduling algorithms, virtual machines.

1. INTRODUCTION

In order to avoid any one server from becoming overloaded, load balancing in cloud computing manages huge workloads and distributes traffic among cloud servers. Performance is improved and downtime and latency are reduced as a result.

To reduce latency and increase server availability and dependability, advanced load balancing in cloud computing spreads traffic over several servers. Utilizing a variety of load balancing approaches, effective cloud load balancing implementations reduce server failure and enhance performance. Before rerouting traffic in the event of a failover, a load balancer, for instance, can assess the distance between servers or the load on those servers. Load balancers can be networked hardware-based devices or just software-defined operations. Hardware load balancers are typically not allowed to operate in vendor-managed cloud settings and are ineffective at controlling cloud traffic in any case. Because they may operate in any location and environment, software-based load balancers are more suited for cloud infrastructures and applications.

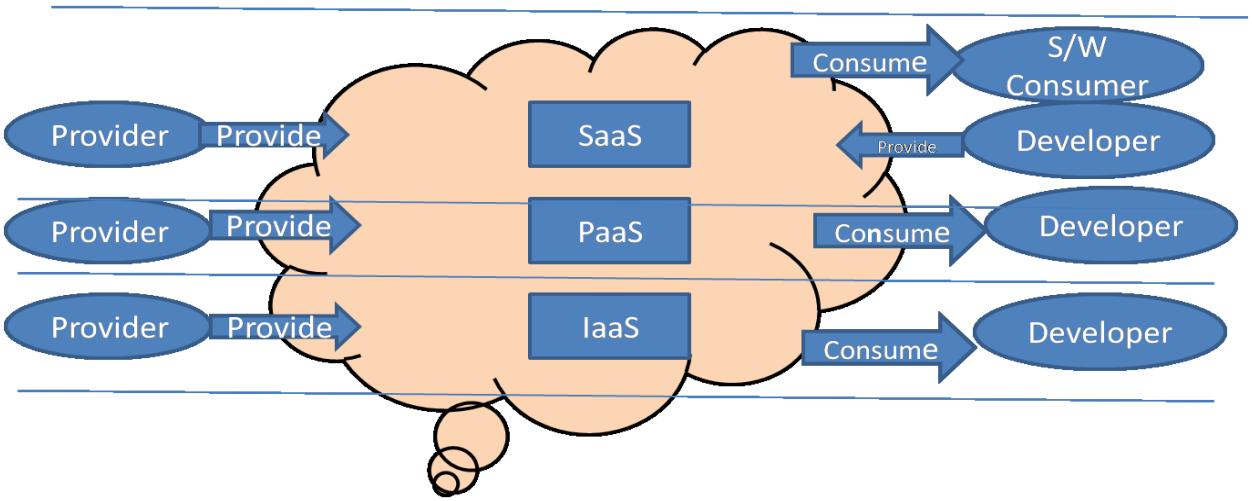

Fig. 1.1 Cloud Computing Architecture

The four different deployment models for cloud computing are as follows:

1.1 Private Cloud

Cloud-based applications and processing environments can only provide users with a private cloud's predetermined cloud computing paradigm.

1.2 Public Cloud

The user can use resources and application services that are made available by a third party that has computing infrastructure.

1.3 Community Cloud

The public cloud, which is a coalition, can be divided by various organizational infrastructures. It can be shared among different applications because of this.

1.4 Hybrid Clouds

Hybrid clouds are created by combining public and private clouds.

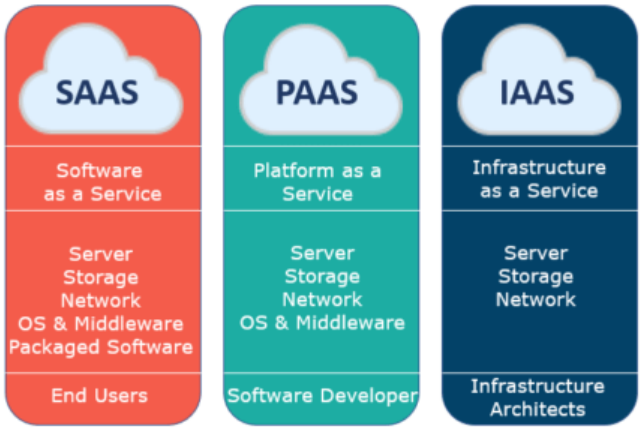

Fig. 1.2 Model of Cloud Computing

2. LITERATURE REVIEW

In the study by Liaqat, Misbah [1], in order to prevent service-level agreement (SLA) violations, users' increasing computational needs call for effective cloud resource management. Co-locating several virtual machines (VMs) on a single physical server allows for efficient resource management by sharing the available resources. However, the choice of "what" and "where" to place workloads has a big impact on how well hosted workloads function. Existing cloud schedulers only take into account a single resource (RAM) when co-locating workloads, which results in SLA violations because to improper VM placement. The current study has revised the nova scheduler to provide a multi-resource based VM placement solution to address this problem and enhance application performance in terms of CPU utilization and execution time. It has been demonstrated through experimentation that our suggested solution has reduced application execution time.

[3] Today, keeping the load under control in a cloud environment is the most difficult problem for a researcher. In the world of information technology, cloud computing is a fantastic technology. It uses computing resources to provide various services over the internet, and its fees are depending on how much of those resources are used. It offers users a wide range of cost-effective, dependable, and productive services. As a result, more people are using cloud computing today because businesses, governments, and educational institutions are embracing it. Therefore, we can use the load balancing strategy to meet user demand when there are a lot of requests for cloud resources.

To balance the workload across the cloud system, we have suggested a load balancing technique in this study by merging two algorithms. For priority-based activities, we employed a modified algorithm inspired by honey bee behavior, and for non-priority-based tasks, we used an improved weighted round-robin method. Our research is important because it will lead to better system performance, better resource utilization, and quicker completion times.

[25] According to Mishra, Kaushik, and Santosh Majhi, as rising businesses and research institutions look to benefit from the cloud's on-demand access, service models, and deployment patterns, it is becoming dangerously ubiquitous. It offers special features like self-accessible, dynamically scalable, and metered on-demand access to a shared pool of resources over the internet. It is frequently used due to its "pay-as-you-go" business strategy. These characteristics have made this paradigm a popular term in the high-performance distributed computing (HPDC) community. Even if this field is well-liked, improvements are still necessary to achieve optimal performance. According to the equilibrium load distribution, the problem of load balancing between virtual machines (VMs) is NP-hard.

This paper offers a thorough historical assessment and comparative analysis of the major load balancing (LB) literature currently in existence. Researchers can use the work that has been provided as a tool to create new, effective load balancing algorithms in the area of cloud computing.

According to Tong, Zhao [5], a computing technique known as "cloud computing" that uses the Internet and virtualization technology to share resources. Task scheduling is used to equitably distribute computing resources across numerous requests that are waiting to be completed. Deep reinforcement learning (DRL) offers a novel approach for more effectively resolving work scheduling issues in light of the quick evolution of computer hardware

and software. In order to lessen the load imbalance of virtual machines (VMs) and task rejection rate, we present a unique DRL-based dynamic load balancing task scheduling technique in this study. The suggested algorithm in this study performs best at balancing VM workloads and lowering job rejection rates, raising the general caliber of cloud computing services.

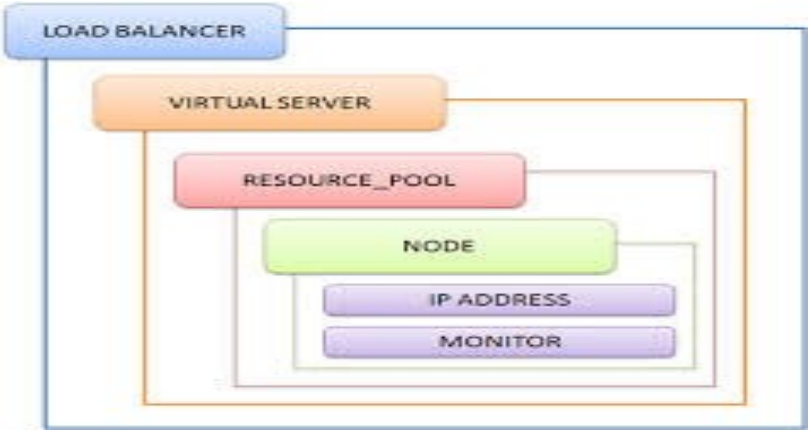

Fig. 2.1 Ant Colony Optimization

V., C. Sathiya Kumar, and Ramani Kannan stated that for virtualized file sharing in cloud infrastructure, cloud computing uses scheduling and load balancing. To accomplish efficient file sharing in a cloud computing environment, these two must be carried out in an optimised manner. In recent times, cloud data centres have created Scalable Traffic Management for traffic load balancing and quality of service provisioning. This work aims to present an integrated resource scheduling and load balancing method for effective provisioning of cloud services. The procedure creates a multidimensional resource scheduling model based on fuzzy logic to increase resource scheduling effectiveness in cloud architecture.

Ullah, Arif [8] in their research have stated that in the world of IT, cloud computing is a new technology. However, it still encounters issues like load balancing. It is a mechanism that evenly distributes work load among multiple nodes when some are overloaded and some are underloaded. The main benefits of load balancing are reduced resource usage and energy use. Swarm intelligence plays a significant part in the field of problems that require traditional and mathematical techniques and are difficult to solve. The algorithm was developed by Karaboga in 2005 and is inspired by an artificial bee colony's foraging behaviour. It features excellent robustness, rapid convergence, and high flexibility. A different researcher applied the ABC method to load balancing to improve it.

Comprehensive research on load balancing in cloud computing is presented in this review paper. In their research work [9], Khan, Farhan, et al have stated that the fundamental difficulty for the service provider in cloud computing is task scheduling. The assignment of jobs to virtual machines is one of the scheduling's most important goals in order to prevent overloading or overuse of specific machines. In order to accomplish this, load balancing is essential in the scheduling issue. Response time can be sped up and resources can be used more effectively by using the right load balancing technique. In order to improve load balancing and reliability in cloud computing, we offer in this work a dynamic approach for allocating a job to virtual machines. The suggested approach shortens the makespan, boosts load balancing, and boosts system dependability.

The proposed strategy has proven successful in increasing the accuracy of task scheduling based on the virtual machines' prior experience as well as the equitable allocation of workload among them.

In the study by Shafiq, Dalia Abdulkareem [15], despite extensive prior study in the subject of cloud computing, there are still certain issues with workload distribution in cloud-based applications, particularly in the Infrastructure as a Service (IaaS) cloud model. Because there are only so many resources and virtual machines available in cloud computing, efficient job allocation is a critical process. Task scheduling closely complies with the standards of the Service Level Agreement (SLA), a document made available to consumers by cloud developers, and it significantly contributes to load balancing. The LB algorithm takes into account crucial SLA criteria like Deadline. In light of the Quality of Service (QoS) job parameters, the priority of VMs, and resource allocation, the proposed approach aims to optimize resources and enhance load balancing.

As per Pourghaffari, Ali, Morteza Barari, and Saeed Sedighian Kashi[20], large-scale computing services and systems come together in a computational ensemble known as cloud computing technology. Resource management, task scheduling, and efficient resource sharing among users have recently been the subject of research. An improved strategy is needed to approach the optimal resource allocation in cloud computing due to the computational and resource management challenges. The goal of the current research is to improve computational cloud resource

allocation while maintaining load balance in cloud providers. It integrates evolutionary algorithms, fuzzy logic, and job scheduling techniques. The response time, task execution time, and energy consumption of the suggested technique are faster than those of other methods, according to simulation of the proposed model.

In the research by Nazir, Jaleel [22], the load balancing in cloud services can be managed using a variety of methods and techniques. In this work, a novel approach is introduced for load balancing at the database level in cloud computing. Companies of all sizes often use the database cloud services for business process and application development. It is possible to maintain an effective job scheduling procedure that also satisfies user requirements and boosts resource utilization by using load balancing for distributed applications. In order to prevent any one node from becoming overwhelmed, load balancing involves dispersing the load among several nodes. The load balancer distributes an equal amount of compute time to each node to prevent overloading. The outcomes of two distinct situations demonstrated how load balancer judgements on application traffic gateways might control cross-regional traffic and significantly increase restaurant income.

According to Kumar, Mohit, and Subhash Chander Sharma [14], a well-designed web-based tool or application on a pay-per-use basis provides users with on-demand services via the cloud computing service model. Due to the constant growth of users and applications in the cloud environment, load balancing has become a major issue for cloud service providers. In the past ten years, a number of load balancing algorithms have been presented for cloud computing, however none of them offer resource utilisation and flexibility along with load balancing. In order to equitably divide the load among all of the virtual machines (VM), we proposed a cloud resource broker. We have created a dynamic load balancing algorithm that not only effectively uses cloud resources and shortens job turnaround times, but also offers elasticity in a cloud environment.

CONCLUSION

In this research work we reviewed many papers on Dynamic Load Balancing for Cloud. To summarize, the effective distribution of website traffic to available servers is ensured by cloud load balancing. It guarantees that the apps are always accessible to the customer while preventing downtime or machine failure problems.

ACKNOWLEDGMENT

I would like to thank the Department of CSE at AITS for facilitating the development of the paper, making available resources and also for final deployment.

REFERENCES

[1] Liaqat, Misbah, et al. "Characterizing dynamic load balancing in cloud environments using virtual machine deployment models." IEEE Access 7 (2019): 145767-145776.

[2] Afzal, Shahbaz, and G. Kavitha. "Load balancing in cloud computing–A hierarchical taxonomical classification." Journal of Cloud Computing 8.1 (2019): 22.

[3] Patel, Karan D., and Tosal M. Bhalodia. "An efficient dynamic load balancing algorithm for virtual machine in cloud computing." 2019 International conference on intelligent computing and control systems (ICCS). IEEE, 2019.

[4] Jena, U. K., P. K. Das, and M. R. Kabat. "Hybridization of meta-heuristic algorithm for load balancing in cloud computing environment." Journal of King Saud University-Computer and Information Sciences 34.6 (2022): 2332-2342.

[5] Tong, Zhao, et al. "DDMTS: A novel dynamic load balancing scheduling scheme under SLA constraints in cloud computing." Journal of Parallel and Distributed Computing 149 (2021): 138-148.

[6] Annie Poornima Princess, G., and A. S. Radhamani. "A hybrid meta-heuristic for optimal load balancing in cloud computing." Journal of Grid Computing 19.2 (2021): 21.

[7] Priya, V., C. Sathiya Kumar, and Ramani Kannan. "Resource scheduling algorithm with load balancing for cloud service provisioning." Applied Soft Computing 76 (2019): 416-424.

[8] Ullah, Arif. "Artificial bee colony algorithm used for load balancing in cloud computing." IAES International Journal of Artificial Intelligence 8.2 (2019): 156.

[9] Ebadifard, Fatemeh, Seyed Morteza Babamir, and Sedighe Barani. "A dynamic task scheduling algorithm improved by load balancing in cloud computing." 2020 6th International Conference on Web Research (ICWR). IEEE, 2020.

[10] Golchi, Mahya Mohammadi, Shideh Saraeian, and Mehrnoosh Heydari. "A hybrid of firefly and improved particle swarm optimization algorithms for load balancing in cloud environments: Performance evaluation." Computer Networks 162 (2019): 106860.

[11] Mishra, Sambit Kumar, Bibhudatta Sahoo, and Priti Paramita Parida. "Load balancing in cloud computing: a big picture." Journal of King Saud University-Computer and Information Sciences 32.2 (2020): 149-158.

[12] Sefati, SeyedSalar, Maryamsadat Mousavinasab, and Roya Zareh Farkhady. "Load balancing in cloud computing environment using the Grey wolf optimization algorithm based on the reliability: performance evaluation." The Journal of Supercomputing 78.1 (2022): 18-42.
[13] Semmoud, Abderraziq, et al. "Load balancing in cloud computing environments based on adaptive starvation threshold." Concurrency and Computation: Practice and Experience 32.11 (2020): e5652.
[14] Kumar, Mohit, and Subhash Chander Sharma. "Dynamic load balancing algorithm to minimize the makespan time and utilize the resources effectively in cloud environment." International Journal of Computers and Applications 42.1 (2020): 108-117.
[15] Shafiq, Dalia Abdulkareem, et al. "A load balancing algorithm for the data centres to optimize cloud computing applications." IEEE Access 9 (2021): 41731-41744.
[16] Negi, Sarita, et al. "CMODLB: an efficient load balancing approach in cloud computing environment." The Journal of Supercomputing 77 (2021): 8787-8839.
[17] Ebadifard, Fatemeh, and Seyed Morteza Babamir. "Autonomic task scheduling algorithm for dynamic workloads through a load balancing technique for the cloud-computing environment." Cluster Computing 24 (2021): 1075-1101.
[18] Shafiq, Dalia Abdulkareem, N. Z. Jhanjhi, and Azween Abdullah. "Load balancing techniques in cloud computing environment: A review." Journal of King Saud University-Computer and Information Sciences 34.7 (2022): 3910-3933.
[19] Haris, Mohammad, and Rafiqul Zaman Khan. "A systematic review on load balancing issues in cloud computing." Sustainable Communication Networks and Application: ICSCN 2019 (2020): 297-303.
[20] Pourghaffari, Ali, Morteza Barari, and Saeed Sedighian Kashi. "An efficient method for allocating resources in a cloud computing environment with a load balancing approach." Concurrency and Computation: Practice and Experience 31.17 (2019): e5285.
[21] Gamal, Marwa, et al. "Osmotic bio-inspired load balancing algorithm in cloud computing." IEEE Access 7 (2019): 42735-42744.
[22] Nazir, Jaleel, et al. "Load balancing framework for cross-region tasks in cloud computing." Computers, Materials & Continua 70.1 (2022): 1479-1490.
[23] Abed, Marwa M., and Manal F. Younis. "Developing load balancing for IoT-cloud computing based on advanced firefly and weighted round robin algorithms." Baghdad Science Journal 16.1 (2019): 130-139.
[24] Junaid, Muhammad, et al. "A hybrid model for load balancing in cloud using file type formatting." IEEE Access 8 (2020): 118135-118155.
[25] Mishra, Kaushik, and Santosh Majhi. "A state-of-art on cloud load balancing algorithms." International Journal of computing and digital systems 9.2 (2020): 201-220.