International Journal of Research in Science and Technology

Volume 11, Issue 02: April-June 2024

ISSN 2394 - 9554

AN ADVENT LOAD BALANCING TRANSMISSION MEASURES IN CLOUD COMPUTING

Sivakumar Dhandapani[1*]
Department of Computer Science and Engineering, AMET Deemed to be University, Kanathur, Chennai, Tamilnadu, India-603112

***E-mail: sivakumar.d@ametuniv.ac.in**

Abstract
Energy conservation is a crucial issue in virtualized cloud computing systems since it may provide several significant benefits, including lower operating costs, higher structure efficiency, and environmental protection. The users of the cloud have continuous access to services from anywhere in the globe. As the emergence of next-generation information technologies like bitcoin, the internet of things, and big data, diverse and enormous volumes of data are produced. The resources such as virtual machines, networks, retrieval, etc. are made available online through cloud computing. In order to accomplish and maintain the application performance, it is essential to locate the virtual machines that host the programme in the appropriate locations and to move those virtual machines when unexpected network delay or obstruction develops. The paper demonstrates ways research from prior approaches to application performance enhancement could possibly be used to optimize data transmission between the virtual machine and data. In order to provide a broad overview of the most recent methods in this field of study, numerous suggested algorithms tackling the problem of load balancing in cloud computing are analyzed and compared in this work.

Keywords: cloud computing, virtual machines, load balancing, and virtualization

1. INTRODUCTION:

Cloud computing, which is founded on the underlying virtualization technologies, has significantly changed the approach to providing information technology services as well as their underlying hardware infrastructures. In order to host the virtual machines (VMs) that are handling the end-user's requests, thousands of computers are combined into a pool of computation resources. Mobile cloud computing has recently been offered as a viable method for enhancing the abilities of low-resource mobile devices [14]. A significant advancement in computing is cloud computing [11], which offers pooled computer power on demand. Services and the data they contain can be hosted by several networked virtual machines (VMs) under this situation. Energy conservation is a crucial issue in virtualized cloud computing systems since it may provide several significant benefits, including lower operating costs, higher structure efficiency, and environmental protection. According to the fast expansion of distributed cloud computing network services, the amount of data in various sectors, including scientific computing, signal processing, bioinformatics, and applications for the Internet of Things (IoT) have expanded. These applications encompass the billions of operations carried out by thousands of high- performance devices installed computers in cloud data centers. As virtual machines (VMs), the cloud provides a range of services, and virtualization is one of the finest advantages for users who may use these many services. A crucial component of cloud computing is load balancing which prevents the scenario where some nodes are overburdened while others are idle or insufficiently utilized. The key factor that makes so many businesses choose cloud computing over other areas of study like HPC and grid computing is the economic trend. The QoS (Quality of Service) parameters, such as reaction time, cost, throughput, performance, and resource use, can be enhanced by load balancing [4] [9]. This paper also aims to

ISSN 2394 - 9554

provide a comprehensive overview of cloud computing features that will aid in the creation and uptake of this quickly developing field of study.

In the world of cloud computing, load unbalancing is a serious problem that cloud providers must deal with because it affects both the QoS (Quality of service) promised in the Efficiency and effectiveness covered by the Service Level Agreement (SLA) between the client and the supplier. of the resources used for computing. Load balancing is perceived as an issue, however, it is fundamentally a solution to the problem of load unbalancing, which has two unpleasant aspects - overloading and under loading. The surplus tasks must be moved from over- flowing to under-loaded machines during the load-balancing process, and in order to do that, the cloud provider must incur additional costs. These costs are referred to as migration costs and are essentially a penalty in terms of a decrease in revenue or Gaining cash while playing the migration operation. Virtual machines employ specific task planning and resource allocation methods to complete user tasks that have been scheduled with the appropriate resources allotted. As a result of the underutilization of resources, the assigned tasks will be carried out with the maximum level of availability in an under loading condition [6]. Cloud computing is connected to the internet network protocol that has shown tremendous growth in the advancements of communications technology by delivering services to consumers with varying needs through the use of online computing resources. IaaS is the most important service paradigm in which consumers receive actual physical computer resources in the way of virtual machines (VMs) because of virtualization methods [1] [3]. Modern technology for computers called cloud computing makes it possible to provide clients with services at any time. Resources in systems for using the cloud are dispersed globally for quicker customer servicing.

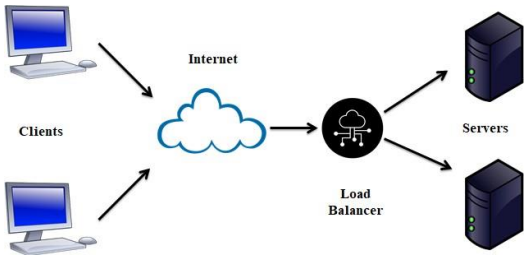

Fig. 1. Load Balancing in Cloud

Web Services are beginning to be able to reach their first anticipated possibilities owing to cloud computing. In recent years, a lot of study has been devoted to various existence- inspired networking and computational models in an effort to find distributed solutions to these systems' growing complexity and extent. As a result of the size and complexity of these systems, it is impossible to regulate individual servers centrally, necessitating efficient distributed solutions. The goal embodied in the Internet of Things (IoT) paradigm of service and hardware provision requires networked oversight, which is delivered through personal experience.

II. Related Work

In the area of Cloud Computing, there have been many different kinds of studies. Some of the general problems are load balancing, organising resources, service broker rules, allocating resources, and so on. In this study, we talk about some LB and CSB problems in cloud computing. Here is a list of the polls that have already been done on LB and service brokering in Cloud Computing.

Singh and Chana (2016) chose 110 research papers from a large number of research papers, workshops, etc. to study resource sharing problems in the cloud [17-19]. The main goal of the resource schedule algorithm is to choose the most efficient and right algorithm for a given job from the methods that are already in place. The main goal of their study was to find a broad analytical analysis of resource management in the cloud as a whole and resource planning in the cloud. By doing research and analysing it in a thorough way, you can find out what the requirements are the requirements are regarding several scheduling methods and choose the best one for planning a certain task. In their review of the literature, they found that there are 13 different capacity arranging techniques and 8 different resource-sharing schemes [14] [17]. The results have been looked at in many different ways, including the grouping of resources, the history of resource planning, proportions of different scheduling approaches and their related QoS parameters, an extensive categorization of the scheduling algorithms of resources and their subtypes, a comparison of the time management algorithms of assets aspects of organizing resources, distribution policies for supplies and organizing them. Prior to enrolling in a resource planning graduate program, you need to discover progress in the same cloud search. The material already exists in the form of a structured development of resource planning. For better resource planning in CC, we need a self-contained resource scheduling system that takes into consideration all-important QoS factors like security, availability, and runtime, SLA failure rate, etc. The efficient utilization of resources can be improved by allocating them based on how similar or different the jobs are. When work and tools are matched up well, success can be greatly improved. In this study work, there are also suggestions for further research.

Milani, along with Navimipour (2016), published a comprehensive overview of the literature on extant LB techniques through a comprehensive evaluation of more than 15 essential studies. The study identifies various LB techniques such as natural language processing, machine learning, and data mining, highlighting their strengths and weaknesses and providing valuable insights for researchers and practitioners in the field among the 726 fundamental articles in their study query. Detailed classifications of various parameters were also included based on an analysis of existing methodologies [15,20]. In addition, the benefits and drawbacks of various LB algorithms were discussed, as well as the primary challenges affiliated with these algorithms. The discipline was divided into two sub- domains based on the available literature: dynamic LB studies and hybrid LB studies. The computer relocation approach is meant to alter a virtual machine (VM) in the public internet, whereas the combination, split, and connect technique is intended to change physical hardware in the cloud (CC). Cloud consumers and cloud developers can utilize hybrid algorithms. From the perspective of the cloud user, makes pan and response time are extremely important parameters for Cloud computing load balancing Scalability has been raised from 9% in dynamic approaches to 33% in suggested hybrid techniques in order to improve reaction time. This is a key component in the development of cloud-based systems, which are now expanding quickly. The main disadvantage of the hybrid approaches chosen is migration time. Researchers can better comprehend the state of the art in load balancing thanks to the study's broad data valuable insights into hybrid load balancing techniques and their impact on reaction time improvement, aiding cloud computing researchers and practitioners in making informed decisions. It also emphasizes the importance of understanding the trade-off between scalability and migration time for optimizing system performance.

III. CLOUD COMPUTING

A. FUNDAMENTALS OF CLOUD COMPUTING

A well-known label for the service itself is" s o f t w a r e as a service" (SaaS). Software as a service will become more and more popular, and how IT equipment is developed and purchased will change

ISSN 2394 - 9554

as a result of cloud computing, the long-awaited realization of computing as a utility [2]. Cloud computing has changed how businesses access software by enabling on-demand subscriptions without expensive infrastructure investments, providing cost savings and more flexibility for businesses of all sizes. Modern internet service providers no longer need to spend a lot of cash on costly technology or staff it with a large crew. In order to avoid squandering funds while losing out on potential customers and revenue- generating prospects, businesses have no worries regarding providing a deal that is highly sought after insufficiently or excessively. Grid calculating, virtualization of computers, and HPC are only a few of the computer research areas on which cloud computing is based. A TCP/IP-based ecosystem with significant development and the incorporation of technology for computers, including quick microprocessors that have large amounts of memory, high-speed networks, and dependable system design. Cloud computing is unlikely to have been a reality without standardized interconnection protocols and developed data center assemblage technologies. As a new paradigm in computing, cloud technology seeks to provide end users with dynamic computational settings that are dependable, customizable, and QoS-assured. In the cloud, load balancing is the process of distributing workload evenly across virtual ma- chines to optimize resource utilization. Various load-balancing algorithms are highlighted in this survey based on various metrics [5] [7]. The fact that load balancers facilitate the equitable to find load-balancing problems and strive toward their resolutions by the need to allocate resources fairly to activities in order to achieve the best resource utilization, user happiness, and cost-effectiveness. Due to the ongoing demand for cloud resources and the increasing workload in this sector, we have determined the necessity of distributing the load among them. As a result, the present investigation has been recognized and meticulously presented in a way that highlights problems and obstacles for further study on the basis of the existing research. Load balancing is crucial for optimal resource utilization, user satisfaction, and cost-effectiveness. Due to increasing cloud resource demand and workload, distributing the load among activities is necessary due to the persistent demand and rising workload in the cloud computing industry. On the basis of the available research, therefore, the existing work has been identified and systematically summarized in a manner that depicts issues and challenges for future research [13] [16].

B.THE CRUCIAL ELEMENTS OF CLOUD COMPUTING

The concept of cloud computing is for users to buy computer platforms or IT facilities via these clouds, then run their programs there. The SOA idea in grid computing is comparable to the customer-focused notion, although the latter is more useful. To connect to Computer clouds offer customers the services that follow in an open way, using software, data, and hardware resources, as well as a single computer platform as a service:

- **IAAS: INFRASTRUCTURE AS A SERVICE**

Infrastructure as a service (IaaS) is one of the distributed computing industry's most significant and increasing segments. Online providers give individuals and machines the use of assets including virtualized computers, raw (block) storage, antivirus programs, balancers for load, and network components under this service delivery model. IaaS offers hosting, equipment configuration, and essential services needed to operate in a cloud

- **SAAS: SOFTWARE AS A SERVICE**

Customers can access software or an application online thanks to hosting as a service. In this mode, the client's local machines don't need to be installed or used to run the programme. SaaS relieves the customer's responsibility for software upkeep and lowers the cost of software acquisition through streaming services. pricing.

ISSN 2394 - 9554

- **PAAS: PLATFORM AS A SERVICE**

Platform-as-a-A service often offers a number of application programming interfaces for cloud-based applications and abstracts about the infrastructures. It serves as the intermediary link between both software and hardware. Due to the significance of platforms, many large corporations seek to seize the opportunity to dominate the realm of cloud computing, just as Microsoft does in the era of personal computers.

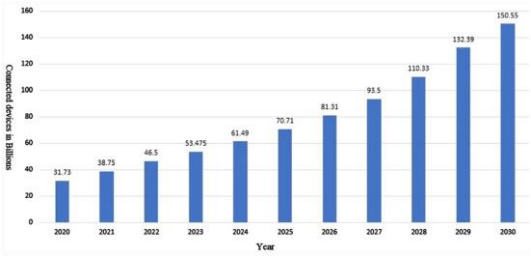

Fig. 2. Increase in the number of connected devices from 2020 to 2030

IV. LOAD BALANCING

Load balancing is a technique for spreading a larger processing load among fewer processing nodes in order to improve overall system performance. In cloud computing, load balancing is essential to distribute the changing local workload evenly across all nodes. Load balancing promotes efficient resource use, avoiding system failures and downtime by ensuring that no one node is overwhelmed with excessive processing demands, enhancing efficiency and decreasing response times. Reduced resource usage is made possible by proper load balancing and high resource utilization. It aids with scaling, establishing failover, and averting bottlenecks. A method known as load balancing benefited networks and resources by offering the highest throughput with the quickest reaction times. Data may be transferred and received instantly with load balancing since it divides traffic among all servers. As it is essential to estimate the right load, compare all the loads, take into account the stability of all the different systems, the effectiveness of the desired system, the connections across every single node, and the type of work to be sent while establishing a load balancing algorithm. The selection of the nodes, which also includes several others, is of utmost importance. Calculating the machine load takes into account both the CPU load and the needed quantity of memory [10]. Loading unbalancing, an OA multi-variant, multi-constraint issue, lowers the efficiency and performance of the resources available to computers. Weight-balancing methods address the two undesirable effects of load unbalancing, overloaded and under loading [8].

If assets are used as efficiently as feasible, a system that uses cloud computing will be productive. This may be done by utilizing and maintaining sufficient cloud resource administration. A technique called the use of clouds enables users access a variety of services and share data [4] [12]. Users only pay for the resources they really consume. In a distributed setting, cloud computing maintains data and dispersed resources, and the rate of data storage growth is rapid [18]. Therefore, the primary duty in the cloud domain is load balancing. In order to prevent any node from becoming overburdened, load shifting helps to balance the dynamic workload among several nodes.

V. NEED OF LOAD BALANCING

A load balancing algorithm must accurately predict the load, evaluate all the loads, evaluate the stability of all the various systems, the desired system's performance, interactions among all the nodes, and the

ISSN 2394 - 9554

sort of work to be transmitted. With the help of load balancing, it is possible to distribute the workload equally across the available resources.

Fig. 3. Service as Load Balancing in Cloud Computing

By providing and de provisioning application instances, it seeks to maintain service in the event that a service component fails while also making efficient use of resources. Load balancing also seeks to provide precedence to activities that need to be executed right away in comparison to other jobs and to offer adaptability and adaptability for programs whose size can expand in the years to come and need more resources. Other goals of load balancing include lowering energy use and carbon emissions, eliminating bottlenecks, allocating resources, and meeting QoS standards to enhance load balancing. Techniques for load balancing and task mapping that are appropriate and take various parameters into account are required.

VI. CHALLENGES FOR LOAD BALANCING

Some qualitative measures in cloud computing can be enhanced for better load balancing.

- Throughput - Throughput is the overall amount of tasks that are executed effectively throughout a specific time period. High throughput is necessary for the system to operate at its best.
- Scalability - Scalability refers to an algorithm's capacity to balance the load on any system with a finite number of machines and processors. This setting can be tweaked to improve system performance.
- Performance - It is the system's total effectiveness. The performance of the entire system can be enhanced if every single one of the variables are improved.
- Migration Time - Migration Time is the length of time it takes for a process to be moved for performance from one system node to another. The amount of time should constantly be lower for the system to operate more efficiently.
- SLA Violation - SLA Violation indicates the number of SLA violation causes that have been reduced in terms of deadline restriction, priority, etc. When resources (VMs) are unavailable because they are overloaded, breaching service level agreements occur. A minimum SLA is required for improved client satisfaction..

ISSN 2394 - 9554

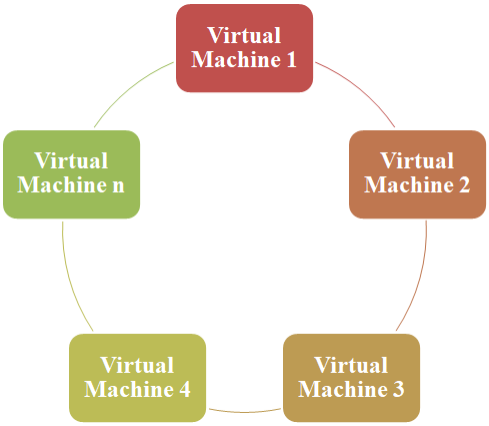

Fig. 4. Virtual Machines in Cloud System

VII. CONCLUSION

In this paper, we have reviewed a number of load-balancing methods for cloud computing. The basic goal of load balancing is to maximize resource utilization by dynamically reassigning the overall load to each individual node while also satisfying customer requirements. Additionally, it is determined that more study has to be done on several unresolved problems. To better optimize the resources scientists ought to consider the study's suggestions for enhancing the distribution of the load on algorithms in their next work on cloud computing.

REFERENCES

[1] 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.

[2] Ghomi, Einollah Jafarnejad, Amir Masoud Rahmani, and Nooruldeen Nasih Qader. "Load-balancing algorithms in cloud computing: A survey." Journal of Network and Computer Applications 88 (2017): 50-71.

[3] Kumar, Pawan, and Rakesh Kumar. "Issues and challenges of load balancing techniques in cloud computing: A survey." ACM Computing Surveys (CSUR) 51.6 (2019): 1-35.

[4] 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.

[5] Tunguturi, Mahesh. "Comparative Analysis of Balancing Techniques in Cloud Computing." International Journal of Management Education for Sustainable Development 2.2 (2019): 41-50.

[6] Shah, Nadeem, and Mohammed Farik. "Static load balancing algorithms in cloud computing: Challenges & solutions." International Journal of Scientific & Technology Research 4.10 (2015): 365-367.

[7] G. Huang, S. Song, J. Gupta and C. Wu, "Semi-Supervised and Unsupervised Extreme Learning Machines", IEEE Transactions on Cybernetics, vol. 44, no. 12, pp. 2405-2417, 2014.

[8] R. Kothari, N. Choudhary and K. Jain, "CP-ABE Scheme with Decryption Keys of Constant Size Using ECC with Expressive Threshold Access Structure", Studies in Autonomic, Data-driven and Industrial Computing, pp. 15-36, 2021.

[9] Tong, Liang, Yong Li, and Wei Gao. ”A hierarchical edge cloud architecture for mobile computing.” IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications. IEEE, 2016.

[10] Ageed, Z., et al. ”Cloud computing resources impacts on heavy-load parallel processing approaches.” IOSR Journal of Computer Engineering (IOSR-JCE) 22.3 (2020): 30-41.

[11] G. Kaur and H. Singh, ”Data Mining Techniques for Text Mining”, Indian Journal of Science and Technology, vol. 9, no. 44, 2016.

[12] R. Rathore, R. Sharma, R. Bhanawat, P. Soni, P.R. Soni, and P. Sachdev, ”Inter-linked platform for Campus Placement in Higher Educational Institutions of India.” International Journal of Advanced Research in Computer Science 13 (2022).

[13] F. Khan, R. Kothari, and M. Patel, ”Advancements in Blockchain Technology With the Use of Quantum Blockchain and Non-Fungible Tokens.” In Advancements in Quantum Blockchain With Real-Time Applications, pp. 199-225. IGI Global, 2022.

[14] Shafiq, Dalia Abdulkareem, et al. ”A load balancing algorithm for the data centres to optimize cloud computing applications.” IEEE Access 9 (2021): 41731-41744.

[15] 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.

[16] Kazeem Moses, Abiodun, et al. ”Applicability of MMRR load balancing algorithm in cloud computing.” International Journal of Computer Mathematics: Computer Systems Theory 6.1 (2021): 7-20.

[17] Singh, Aarti, Dimple Juneja, and Manisha Malhotra. ”Autonomous agent based load balancing algorithm in cloud computing.” Procedia Computer Science 45 (2015): 832-841.

[18] Kothari, R. (2023). Integration of Blockchain and Edge Computing in Healthcare: Accountability and Collaboration. Transdisciplinary Journal of Engineering & Science, 14. https://doi.org/10.22545/2023/00230.

[19] N.Noor Alleema and D.Sivakumar “volunteer nodes of ant colony optimization routing for minimizing delay in peer to peer MANETs”, Peer to Peer Networking and Applications, 2019.

[20] D.Sivakumar and J.P Srividhya, “PQ Indices signal analysis using empirical wavelet transform and rational dilation wavelet transform ,” Sensor Letters, American Scientific Publisher, ISSN:1546-1971, vol. 17, issue 1-14, pp. 1-13, 2019.