iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** I **Month of publication:** January 2023

DOI: https://doi.org/10.22214/ijraset.2023.48878

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue I Jan 2023- Available at www.ijraset.com

A Survey on Issues and Challenges of Load Balancing Techniques in Cloud Computing Environment

Febina KS[1], Tanuja Hemchand[2], Shamimunnisabi[3], Roopa P[4], Aishwarya P[5]
VET First Grade College, Department of Computer Science, Bengaluru, India

Abstract: ***Cloud computing has transformed user services by introducing a new model that allows for pay-per-use access to information technology services at any time and from any location. Numerous businesses are moving their operations to the cloud as a result of the flexibility offered by cloud services, and additional data centers are being set up by service providers to provide users with services. However, it is essential to ensure efficient resource utilization and cost-effective task execution. Based on load balancing, task scheduling, resource management, quality of service, and workload management, several methods to improve performance and resource use have been reported in the literature. Data centers can use load balancing in the cloud to prevent virtual machines from being overloaded or under loaded, which is a problem in and of itself in cloud computing. As a result, developers and researchers must develop and implement a suitable load balancer for distributed and parallel cloud environments. In order for researchers to develop algorithms that are more effective, this survey provides a cutting-edge review of issues and obstacles associated with existing load-balancing methods.***
Keyword: cloud computing, Load Balancing, Load Balancer

I. INTRODUCTION

A. Cloud Computing

Cloud computing, sometimes known as the computing of the future, gives its users access to software and services through a virtualized network. No matter where the customer accesses the service, he is automatically forwarded to the available sources. Our machine occasionally hangs up or the printing of a page seems to take a very long time. All of this occurs because there may be a line of requests waiting in line to access resources that are shared among them. However, none of those requests can be fulfilled because the sources needed to fulfil each of them are controlled by another system.

In cloud platforms, resource allocation (or load balancing) takes place majority at two levels. Load balancing is a new method that assists networks and resources by providing a high throughput and minimal response time.

At the basic level: When an application is uploaded, the load balancer attempts to distribute the computational burden of multiple applications across physical computers by allocating the requested instances to physical computers.

At the next level: In order to balance the computational load across a number of instances of the same application, each request received by an application must be assigned to a specific instance[1]

The notion of load balancing, its requirements and objectives, kinds and comparisons between traditional computing environments and cloud computing environments, and various algorithms are covered in the sections that follow. The conclusion and references are then covered.

B. Load Balancing

By shifting the workload among the processors, load balancing improves the system's performance. A machine's workload is the total amount of time it takes to complete all tasks that have been given to it. By uniformly balancing the load of virtual machines, it ensures that no machine is idle or only partially loaded while others are heavily loaded. One important aspect of improving the cloud service provider's working performance is load balancing. Increased resource utilization is one of the advantages of distributing the workload, which further enhances overall performance and maximizes client satisfaction.

If the cloud provider is not configured with any good mechanism for load balancing, the increase in the number of users will result in poor performance in terms of resource usage, and the capacity of cloud servers will not be utilized appropriately. In cloud computing, if the number of users increases, the load will also increase.

The performance of a heavy-loaded node will be confiscated or seized as a result of this. We can maximize resource utilization if a good load balancing technique is used to equally divide the load (the term "equally" refers to low load on a heavily loaded node and more load on a node that is currently under less load). The dynamic division of the workload is one of the most important issues with cloud computing.

C. *Need of Load Balancing in Cloud Computing*

In clouds, load balancing is a way to evenly distribute the extra dynamic local workload among all nodes. It ensures that no single node is overwhelmed, thereby improving the system's overall performance, and is used to achieve a high user satisfaction and resource utilization ratio. Using load balancing correctly can help make the most of the resources at your disposal and reduce how much you use. Additionally, it facilitates failover implementation, enables scalability, prevents bottlenecks and overprovisioning, shortens response times, and other benefits.[2]

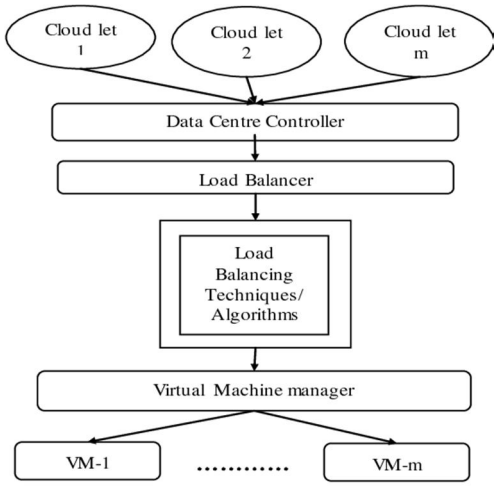


Figure 1:

II. CLASSIFICATION OF LOAD BALANCING ALGORITHM

Based on the current state of the system the y can be classified into the following-

A. *Static Algorithm*

Systems with extremely low load variations call for static algorithms. The entire traffic in a static algorithm is distributed equally among the servers. For better processor performance, this algorithm needs to know a lot about server resources, which is decided at the beginning of the implementation.

However, the decision to shift loads does not depend on the system's current state. The fact that load balancing tasks only function once they are created is a major drawback of the static load balancing algorithm; it cannot be applied to other load balancing devices.

B. *Dynamic Algorithm*

The Dynamic Algorithm chooses to load balance first by searching for the lightest server in the network. It necessitates real-time communication with the network, which may contribute to an increase in the system's traffic. The load is controlled here using the system's current state.

C. *Different Types of Load Balancing Algorithms in Cloud Computing*

1) *Round Robin Algorithm:* Time slicing is used in round robins. The algorithm's name suggests that it operates in a round-robin fashion, with each node receiving a time slice and waiting for their turn. Each node receives an interval and the divided time. Each node is given a time limit within which they must complete their task. Compared to the other two algorithms, this one has less complicity. The cloud analyst algorithm, which is used by default in an open source simulation, was used to carry out the simulation. This algorithm simply distributes the work in a round-robin fashion without taking into account the load on the various machines.[1]

2) *Minimum To Minimum Load Balancing Algorithm:* Under Minimum to Minimum Load Balancing Algorithms, the tasks with the shortest completion times are first identified. Among all the tasks, a minimum value is chosen from those. The machine schedules the task in accordance with that minimum amount of time. Similar to this, the machine has updated the other tasks, and the task is taken off that list. This procedure will continue until the final assignment is made. When there are a greater number of smaller tasks than large tasks, this algorithm appears to be the best.
3) *Maximum To Minimum Load Balancing Algorithm:* The Maximum to Minimum Load Balancing Algorithm is virtually identical to the one before it. However, this algorithm differs slightly. Here, the maximum value is chosen after determining the shortest possible implementation time. The task is then scheduled on the machine according to the maximum duration. All tasks' execution times are updated, and the assigned task is taken off the list.
4) *Weighted Round Robin Load Balancing Algorithm:* The weighted round-robin load balancing algorithm is designed to solve the most difficult problems among round-robin algorithms. In this algorithm, there is a predetermined weight and jobs are assigned based on the value of the weight. Processors with more features get better value. As a result, the server with the highest weight gets more tasks. When all weights are equal, the server will experience consistent traffic.
5) *Opportunistic Load Balancing:* According to Sang's proposal, Opportunistic Load Balancing is a static load balancing algorithm that aims to keep each cloud node busy . Opportunistic Load Balancing, on the other hand, does not calculate the node's execution time, so tasks will be processed more slowly and bottlenecks will occur because requests may be pending while nodes are free.

III. CHALLENGES & ISSUES OF LOAD BALANCING

It is important to figure out some important challenges and problems that affect how well load balancing algorithms work. The load balancer's performance can be improved by addressing the following major issues.

A. *Throughput*

Throughput is the total number of tasks that have been carried out to completion over a given time period. For the system to run better, it needs to have a lot of throughput.

B. *Associated Overhead*

It describes the amount of overhead involved in the load balancing algorithm's implementation. It is made up of tasks moving around, communication between processes, and between processors. There should be as little overhead as possible in order for the load balancing technique to work properly[3].

C. *Fault Tolerant*

It can be defined as the capacity to use the appropriate algorithm for load balancing without any arbitrary link or node failure. A good approach to fault tolerance should be included in every load balancing algorithm.

D. *Migration Time*

It's the amount of time it takes for a process to move from one system node to another so that it can run. This time should always be shorter for the system to work better.

E. *Response Time*

Response time is the amount of time it takes for a particular load balancing method to respond in a distributed system. To improve performance, this time should be reduced.

F. *Resource Utilization*

The data that determines how the resource is used currently comes from the parameter. The most effective resource should be used in the system for effective load balancing.

G. *Scalability*

It refers to the capacity of an algorithm for load balancing a system with any limited number of processors and machines. System performance can be enhanced by enhancing this parameter.

H. Performance

It refers to the system's overall effectiveness. The system's overall performance can be improved if all parameters are improved. Despite the widespread use of cloud computing, The field of cloud computing research is still in its infancy, and the scientific community has yet to resolve a number of problems, particularly load balancing issues.

I. Geographical Distributions Of The Nodes

Large-scale applications like Twitter, Facebook, and others make use of it. The DS of the processors in a cloud computing environment is very helpful for maintaining the system's efficiency and handling fault tolerance effectively. Any real-time cloud environment's overall performance is greatly impacted by the geographical distribution.

J. Emergence Of Small Data Centers For Cloud Computing

Compared to large datacenters, small ones can be more beneficial, cost less, and use less energy. Geo-diversity computing can be achieved through the provision of cloud computing services by small providers. In order to ensure an adequate response time and an optimal distribution of resources, load balancing will become a global issue.[4]

K. Stored Data Management

Even for businesses that outsource their data storage or for individuals, the management of data storage has become a major challenge for cloud computing. In the past decade, data stored across the network has increased exponentially. For optimal data storage and quick access, how can we distribute the data to the cloud?

L. Energy Management

The economies of scale are one of the benefits that encourage the use of the cloud. A key element that makes it possible for a global economy in which fewer providers support a collection of global resources rather than each one having its own resources is energy conservation. Therefore, how can we utilize a portion of the datacenter while maintaining acceptable performance?

M. Virtual Machines Migration

A virtual machine can be moved between physical machines to unload a heavily loaded physical machine. With virtualization, an entire machine can be seen as a file or set of files. The primary goal is to spread the load across a datacenter or cluster of datacenters. Therefore, in order to avoid bottlenecks in the Cloud computing system, how can we dynamically distribute the load when moving virtual machines?

N. Automated Service Provisioning

Elasticity, or the ability to automatically allocate or release resources, is a key feature of cloud computing. How then can we make use of the cloud's resources while still achieving the same level of performance as conventional systems and making the most of available resources?

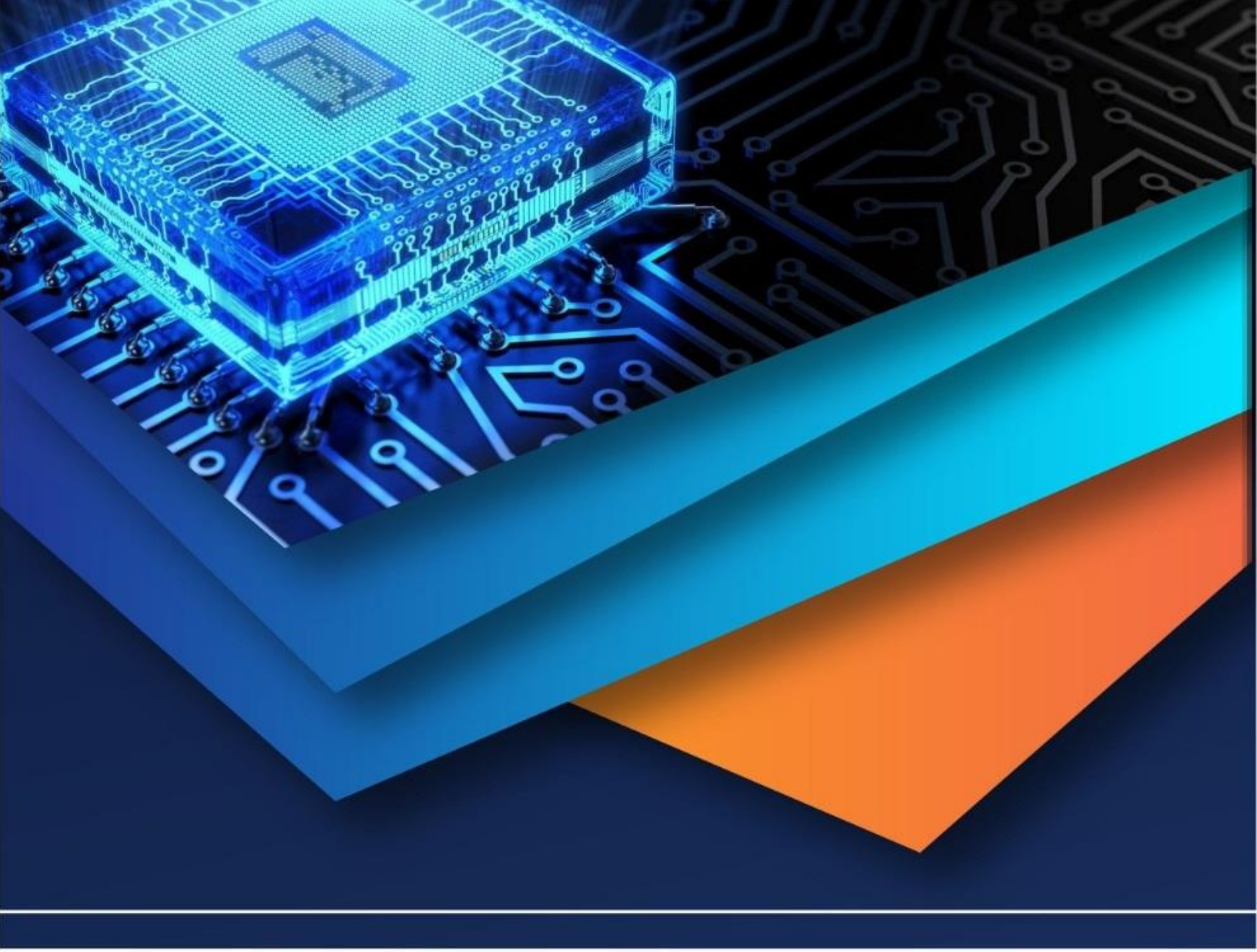
IV. CONCLUSION

In this paper, We looked at various algorithms and talked about the need for load balancing in cloud computing as well as metrics for cloud load balancing. We also talked about virtualization in the cloud. The main problem in cloud computing is load balancing. In order to achieve a high user satisfaction and resource utilization ratio, load balancing is required to evenly distribute the excess dynamic local workload across each node in the cloud. Additionally, it guarantees that each computing resource is distributed fairly and efficiently. Load balancing aims to improve customer satisfaction, maximize resource utilization, significantly improve cloud system performance, minimize response time, and reduce job rejections, all while lowering energy consumption and carbon emissions.

REFERENCES

[1] 1Foram F Kherani, 2Prof.Jignesh Vania, Load Balancing in cloud computing, 2014 IJEDR | Volume 2, Issue 1 | ISSN: 2321-9939
[2] A Survey of Load Balancing Techniques in Cloud Computing Namrata Swarnkar 1, Asst. Prof. Atesh Kumar Singh 2, Dr. R. Shankar, International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 8, August – 2013 ISSN: 2278-0181
[3] Pooja Kathalkar1, A. V. Deorankar2, Challenges & Issues in Load Balancing in Cloud Computing, International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887, Volume 6 Issue IV, April 2018- Available at www.ijraset.com

[4] Jyoti Thaman, Manpreet Singh, "Current Perspective in Task Scheduling Techniques in Cloud Computing: a review", DOI:10.5121/ijfcst.2016.6106
[5] Dr. Hemant S. Mahalle, Prof. Parag R. Kaveri ,Dr.Vinay Chavan," Load Balancing On Cloud Data Centres" , International Journal of Advanced Research in Computer Science and Software Engineering,Vol.3,Issue 1,January 2013
[6] https://www.semanticscholar.org/paper/A-Review-on-Different-Load-Balancing-Techniques-in-Ojha-Shrivastava/48e1eb8c696770dbae0b67c4736b1c1b8db3feac
[7] https://www.xcellhost.cloud/blog/load-balancing-different-types-load-balancing-algorithms-cloud-computing



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY