e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:01/January-2022 Impact Factor- 6.752 www.irjmets.com

CHALLENGES OF CLOUD COMPUTE LOAD BALANCING ALGORITHMS

Mr. Gopala Krishna Sriram[*1]

[*1]Software Architect, Edge Soft Corp, Mckinney, TX USA.

ABSTRACT

Cloud computing reshaped the modern world by offering solutions to all problems faced by organizations. Cloud computing provides computational services for users in a pay-per-use fashion. Hence, they do not need to purchase these resources for their use. With the help of cloud computing services, users can use Software, Hardware, infrastructure, and many other computational resources without taking the pain of their maintenance. Because of their versatile services, cloud computing service providers face many challenges such as security, privacy, quality of service and load balancing. In this research, we focus on load-balancing issues and investigate significant challenges in the load-balancing domain of cloud computing. At the start, we introduce domain knowledge related to cloud computing technologies and then briefly discuss load balancing techniques. At last, we present some potential challenges in load balancing techniques.

Keywords: Cloud Computing, Load Balancing, Algorithms, Load Distribution, Heterogeneous Nodes, Single Point Of Failure, Etc.

I. INTRODUCTION

The excellent Cloud computing services reshape the world of Information Technology. Not only IT companies but also other business organizations and individual's leverages cloud computing technology. The more core concept of cloud computing is to provide users with computing services in a pay-per-use fashion[1]. Cloud allows users to use a shared pool of computing resources. These resources are hardware-based; they include Software, network, platform, and many other valuable resources. Several companies provide cloud services; the top of them are Amazon Web Service (AWS), Google Cloud, Microsoft Azure, RackSpace, IBM Cloud; the primary goal of the cloud is to provide distributed services to improve throughput and performance. Users task from all around the world are distributed among widely spread data centers to improve speed and performance for those jobs. Cloud computing distributed services must gain more popularity than other distributed technologies like grid computing peer-to-peer computing. One reason for this popularity is the vast cloud services[2]. Cloud computing can be categorized on two bases, one is location-based, and the other is offered services. The base of the location cloud can be categorized as public, private and hybrid cloud. The cloud has three basic categories based on services provided, but immense extended categories exist.

The essential services provide Software as a Service (SaaS), Platform as a Service (PaaS), Infrastructure as a Service (IaaS). There can be many other services such as Hardware as a Service (HaaS), Communication as a Service (CaaS), Database as a Service (DaaS), broadly we can tell that XaaS(anything as a service)

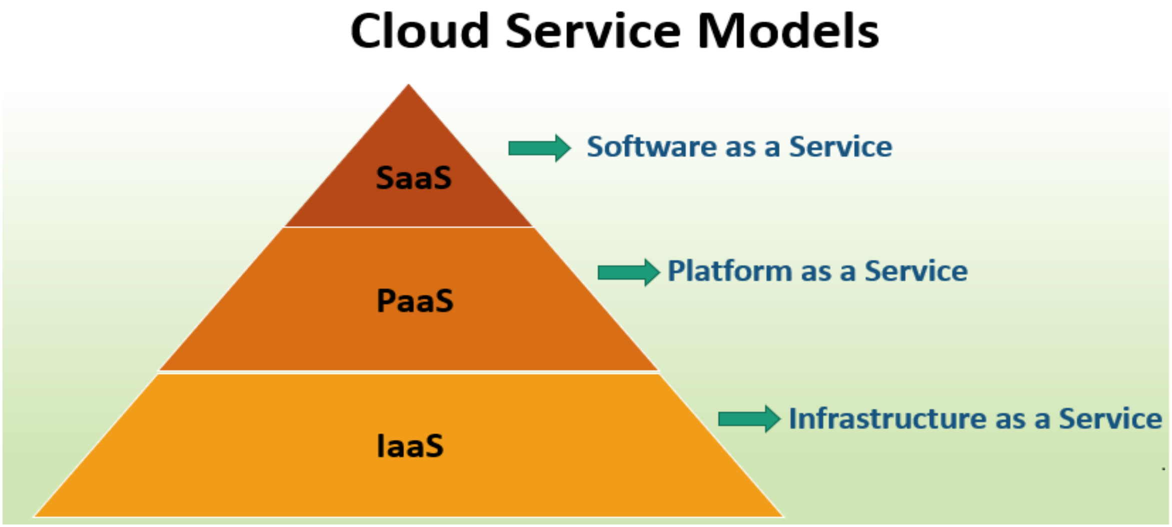

Fig 1: cloud computing service model

Since cloud computing offers immense services, they need Quality of Services (QoS) monitoring to evaluate their services for fulfilling user requirements. The cloud faces many challenges during this process, such as load balancing, performance analysis, monitoring, throughput, response time, security, and privacy. Load balancing is a significant challenge to avoid overloading or underloading virtual machines during service provision. Cloud companies need to identify these potential challenges and find effective load balancing techniques to improve their quality. This research has the following objectives

- State of the art related to different available load balancing techniques
- Taxonomies different load balancing techniques and investigate challenges of existing load balancing techniques
- Point out different research areas for future researchers.

Rest of this paper, we will provide domain knowledge of our target study area, which is load balancing in cloud computing. Afterwards, we will present currently available load balancing techniques and categorize them. At last, we will propose a research area in load balancing in cloud computing.

II. BACKGROUND

This section will provide preliminary knowledge about load balancing techniques, why we need load balancing, the used technologies for load balancing and other domain knowledge.

Load balancing techniques are used to balance the load on virtual cloud machines so that each machine can work equally according to its capacity. By load balancing, tasks can be equally distributed to each virtual machine and hence get the best performance from each. With the help of load balancing techniques, cloud services providers can manage workload on virtual machines. Virtual machines are the core of cloud computing infrastructure. Cloud service providers provide hardware access to users with the help of virtual machines. Virtual machines can act as one or more servers. One machine can act on more than one server.

In contrast, more than one machine can also manage one server machine[3]. Furthermore, through load balancing, cloud providers ensure that in case of failure of one machine to complete, operations will not disturb. In addition to this, load balancing also provides scalability for those applications whose size can grow or shrink with time. Since scalability is one of the significant characteristics of cloud computing, it cannot be achieved without proper load balancing techniques[2]. The other responsibilities of load balancing techniques are to provide energy efficiency, green cloud by reducing carbon emission, QoS requirement fulfilment and other services[4, 5].

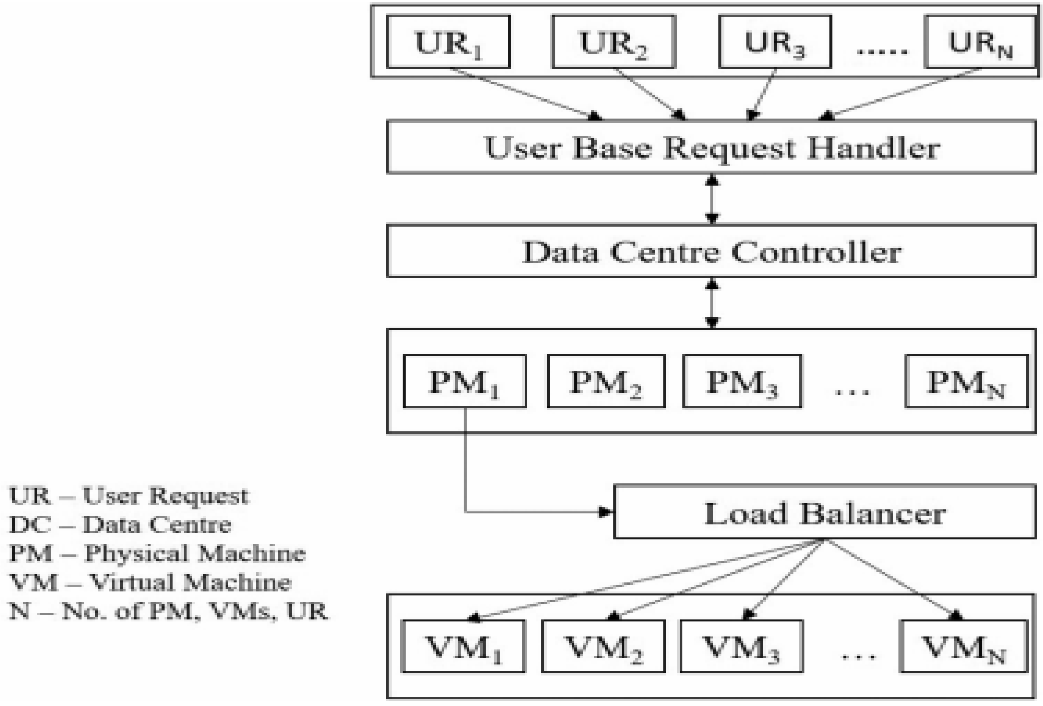

Fig 2: load-balancing model

III. CHALLENGES OF LOAD BALANCING

In this section, we will discuss the significant challenges cloud computing service providers face, particularly for load balancing and other security-related aspects since cloud computing has gained the attention of researchers due to its challenging nature. Cloud researchers work on virtual machine migrations, fail tolerance, virtual machine security, QoS satisfaction, and other related issues[6]. Load balancing is one of the significant challenges for cloud computing researchers. For the rest of this section, we will provide a few challenges in load balancing.

- **Geographically distributed nodes:**

Data centers are primarily located in geographically isolated locations for large cloud computing providers. These data centers act as a single server to solve user requirements. To make them act as a single machine is a potential challenge, particularly for the environment where this distance is considerable. There are many issues related to delay, such as communication speed, network infrastructure and many more like this. If this separation is not at a long distance, network delay is tolerated. Other challenges come even in geographically close data centers, such as choosing a suitable algorithm that works well for all virtual machines.

- **Single point of failure:**

Currently, available load balancing algorithms mostly use a single machine to control the load balancing of other virtual machines: Particularly, most virtual machines in the cloud environment work in a distributed fashion. Hence there is a need for distributed load balancing algorithms so that the failure of a single machine does not cause the failure of all computing resources.

- **Virtual machine migration:**

More than one virtual machine can be configured on a single physical machine or server. These virtual machines act as different systems. In the case underlying physical machines become overloaded, virtual machines need to be migrated from one machine to other physical machines. Load-balancing for such kinds of VM migrations is a challenge to be addressed.

- **Heterogeneous nodes:**

The early load-balancing researchers developed algorithms for homogeneous nodes. As user requirements are complex, there comes a concept of heterogeneous nodes in a cloud environment. New load-balancing techniques are much needed to meet the requirements of heterogeneous nodes.

- **Load balancing scalability:**

Scalability refers to the increasing or decreasing computational user resources according to their demand. Scalability and on-demand availability are essential features of cloud computing. To achieve scalability suitable load-balancing algorithm is mandatory.

- **Algorithm complexity:**

Load-balancing algorithms should make as efficient and straightforward as possible to overcome overhead on cloud computing services

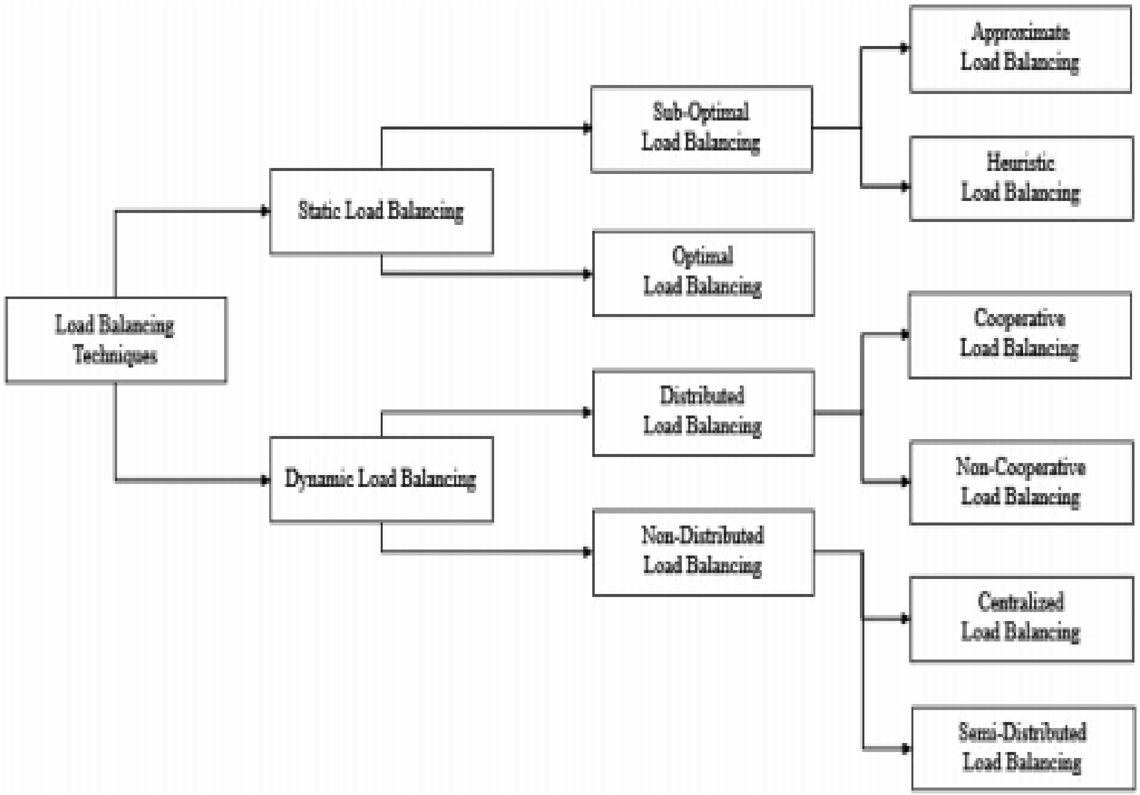
Load Balancing Techniques
Static Load Balancing
Dynamic Load Balancing
Sub-Optimal Load Balancing
Optimal Load Balancing
Distributed Load Balancing
Non-Distributed Load Balancing
Approximate Load Balancing
Heuristic Load Balancing
Cooperative Load Balancing
Non-Cooperative Load Balancing
Centralized Load Balancing
Semi-Distributed Load Balancing

Fig 3: classification of load-balancing algorithm

Below we mentioned some load balancing techniques with their pros and cons.

Table 1: some general load-balancing techniques.

System State	Technique	Concept	Pros	Cons
Dynamic Dynamic Dynamic Dynamic Dynamic	Workload balancing and resource management framework Predictive load-balancing approach using machine learning Modified active VM load-balancing technique Novel	Live VM Migration Load balancing VMs using end of service time Identifying overloaded and underloaded machine using machine learning Use of reservation table for uniform allocation of requests	Low task execution Reduced response time, reduced processing time High resource use, low migration overhead, reduced number of migrations Minimum response time for tasks, load balancing among VMs improves elasticity Improved resource use, makespan, and QoS	Homogeneous VM, Independent task Actual instant processing power calculation is difficult, more power consumption Not tested on a real cloud Allocates tasks uniformly on single data centre Higher response time and few load-balancing Parameters.

	load-balancing approach for minimizing load on servers	Dynamic annexed method over static load balancing		

IV. CONCLUSION

Cloud computing is an ever-spading paradigm of the modern eras computing world. Cloud computing researchers put their intensive care in various aspects; load-balancing is one of them. Load-balancing refers to evenly distributing the computational load on virtual machines of the cloud. In this research, we discussed some essential aspects of load-balancing techniques. We focused mainly on the challenges of load-balancing techniques for future researchers. This research can play a positive role for new researchers in cloud-computing load-balancing to narrow their research problem.

V. REFERENCES

[1] P. Kumar and R. Kumar, "Issues and challenges of load balancing techniques in cloud computing: A survey," ACM Computing Surveys (CSUR), vol. 51, no. 6, pp. 1-35, 2019.

[2] D. A. Shafiq, N. Jhanjhi, and A. Abdullah, "Load balancing techniques in cloud computing environment: A review," Journal of King Saud University-Computer and Information Sciences, 2021.

[3] S. Dhahbi, M. Berrima, and F. A. Al-Yarimi, "Load balancing in cloud computing using worst-fit bin-stretching," Cluster Computing, pp. 1-15, 2021.

[4] A. Bisong and M. Rahman, "An overview of the security concerns in enterprise cloud computing," arXiv preprint arXiv:1101.5613, 2011.

[5] M. Armbrust et al., "A view of cloud computing," Communications of the ACM, vol. 53, no. 4, pp. 50-58, 2010.

[6] J. S. Ward and A. Barker, "A Cloud Computing Survey: Developments and Future Trends in Infrastructure as a Service Computing," arXiv preprint arXiv:1306.1394, 2013.