Load Balancing Challenges in Cloud Computing: A Survey

Rafiqul Zaman Khan and Mohammad Oqail Ahmad

Abstract Cloud computing has broadly been put into practice by business sector, however, there are several actual issues including load balancing, virtual machine migration, automated service provisioning, algorithm complexity, etc., that have not been completely resolved. Each of these are the main challenges of load balancing, that is likely to distribute the unwanted dynamic local workload smoothly to all the nodes in the entire cloud to gain a remarkable consumer fulfillment and resource utilizing ratio. It also makes sure that every computing resource is distributed proficiently and reasonably. This paper describes a thought of cloud computing along research challenges in load balancing.

Keywords Cloud computing · Load balancing · Challenges of load balancing · Goals of load balancing

1 Introduction

Cloud computing has grown to be incredibly favorite during the last couple of years. Due to the part of its service, it provides a flexible to retrieve data as well as an easy way to keep files, incredibly for making huge data sets and files accessible for the dispersing number of consumers to the entire world [1, 2]. Managing these types of large data sets call for several approaches to enhance and simplify operations and provide perfect levels of efficiency for the consumers [1].

Load balancing is a technique that distributes the workload all through various nodes in the presented workspace such that it makes sure no more nodes in the system is overloaded or idle for each moment of time (refer to Fig. 1). An efficient load balancing algorithm would clarify that each and every single node in the

R.Z. Khan · M.O. Ahmad (✉)
Department of Computer Science, Aligarh Muslim University, Aligarh, India
e-mail: oqail.jmu@gmail.com

R.Z. Khan
e-mail: rzk32@yahoo.co.in

D.K. Lobiyal et al. (eds.), *Proceedings of the International Conference on Signal, Networks, Computing, and Systems*, Lecture Notes in Electrical Engineering 396, DOI 10.1007/978-81-322-3589-7_3

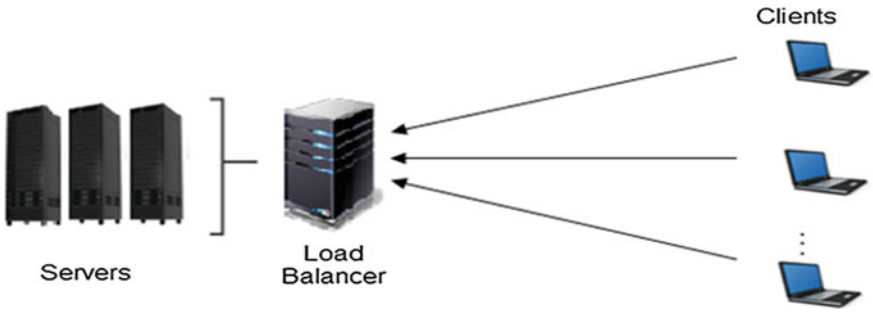

Fig. 1 Diagram for load balancing

system may have more or less identical amount of work. The accountability of the load balancing algorithm is that is really to manage the assignments which are put ahead to the cloud area for the unused services. Therefore, the entire accessible reactions time is enhanced and in addition it gives proficient resource use. Balancing the workload continue to be one of the significant worries in cloud computing since we are unable to figure out the quantity of demands that are released inside of every second in a cloud environment. The uncertainty is credited to the constantly varying tendency of the cloud. The fundamental consideration of load balancing in the cloud platform is in distributing and appointing the load dynamically throughout the nodes with a specific end goal to satisfy the consumer necessities and to give optimal resource use only by arranging the entire obtainable load to diverse nodes [3].

2 Cloud Load Balancing

Load balancing is the fact of dispersing the load all through several resources in every system. In this manner, load should to be distributed across the resources in a cloud-based construction modeling, so that each resource does around the identical quantity of task at every aspect of time. Elementary require is to deliver you some approaches to stabilize demands to give the choice of the application quicker [4]. Basically, load balancing method that makes each processor similarly busy as well as to complete the works around at the same time [5].

A diagrammatic representation of cloud load balancing shown in Fig. 2 and may be summarized as follows [6]:

- The consumer connects with the Internet and demands a service (e.g., a site).
- DNS puts the consumer with an exact open location which is linked to the total uptime technologies open for any activate the network.
- The consumer is linked to the nearby, native total uptime technologies node.

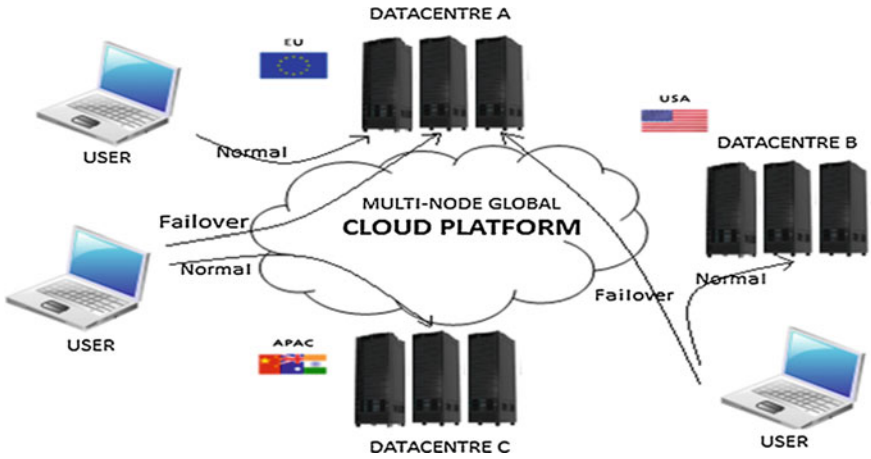

Fig. 2 Diagram for cloud load balancing

- Customer specified policy that allows the total uptime technology node to connect and decides which of the consumer's datacenter to send the user to.
- Users containing the desired application content they are directed to the customer's datacenter.
- Material is supplied to the consumer alternatively using direct server return, or through the nearby total uptime technologies cloud node. If a membership to progression is dynamic, material is optimized everywhere throughout the Internet back to the client.

2.1 Goals of Load Balancing

Goals of load balancing as discussed by the authors of [5, 7] include:

- Stability of the system remains on track.
- To have the capability to alter it as per any modifications or extend in system setup.
- Increase adoptability of the system to adjust to the modifications.
- Promote a fault tolerant system in case of performance, stamina under the partial failure of the system.
- To achieve tremendous improvement in performance.

2.2 Demand of Load Balancing

Load balancing is a technique such that it delegates to accomplish the task equally throughout every attainable node that is shown in the system. High-user fulfillment definitely is the aspect around it. Due to large number of client's and also their needs usually are growing time to time, the clouds will need to supply the products to the visitors with their personal at most fulfillment [3].

A desirable load balancing process guides to utilize of those available resources most positively, thus verifying no more node is over loaded or under loaded. Load balancing facilitates scalability, prevents difficulties as well as also minimizes the time period consumed to actually give the responds. A large number of load balancing algorithm has been developed during order to plan the load among most of the machines. But still so far at this time there is no this type of best load balancing algorithm has been introduced that will set aside the load smoothly all over the system [3].

2.3 Types of Load Balancing

Load balancing is classified as the current state of the system, like static and dynamic load balancing.

2.3.1 Static

Static load balancing usually call for previous knowledge related to the software and resources of the system. The choice to move the workload exactly should not rely on the present status of the system. Static load balancing relates to load balancing that distributes the workload derived exactly relating to a constant set of policies associated with qualities of the workload. Static load balancing are not defensive; thus every single system has at least one assignment allotted for itself [7]. Static algorithms usually do not take into consideration dynamic modifications at run-time [5].

Some static load balancing algorithms are Min–Min, Min–Max, and Round Robin algorithm.

2.3.2 Dynamic

In the course of the static load balancing, a lot more detail with regard to the system and tasks must be recognized prior to the execution. Each of these resources cannot be obtained in prior. A detailed survey of the system status as well as the tasks

Table 1 Static versus dynamic

Parameters	Static	Dynamic
Nature	Compile time	Run time
Performance	Less	More
Flexibility	More	Less
Implementation	Less	Difficult
Resource Utilization	Less	More
Communication Overhead	Less	More
Stability	More	Less
Adaptability	Less	More
Response Time	Easy	More
Reliability	Less	More
Processor thrashing	No	Considerable
Predictability	Easy	Difficult
Complexity	Less	More
Cost	Less	More

pretty boring technique in advance. Due to this fact, dynamic load balancing originated into the real world [5].

A dynamic load balancing does not take into consider the prior state of the system and no past understanding is required, i.e., it relies around the present status of the machine. It permits usual method to be relocated from heavily loaded machine to a lightly loaded machine dynamically to obtain rapid execution [5, 7]. In this conditions, the case communication over minds arises and turns into a lot more if a variety of processors enhance.

Some dynamic load balancing are active clustering, honey-bee foraging, biased random sampling, joint-idle queue, etc.

Various comparisons between static load balancing and dynamic load balancing are listed in Table 1 [8, 9].

2.4 Need of Load Balancing

Most people can equalize the work of a system by dynamically relocating the workload nearby to the system to faraway nodes or systems that are less widely used. Doing this enhances the client fulfillment, maximizing resource utilization, reducing reaction time, decreasing the set of amount of task refusals, as well as boosting the efficiency stability of the system. Additionally, green computing in the clouds can be achieved using load balancing. Some aspects with regards to it are [6]:

A. *Reduced Energy Absorption* Load balancing can easily cut down the capacity of power utilization by staying away from overheating of machines due to extreme workload.

B. *Minimizing Carbon Discharge* Energy absorption and carbon discharge are the two sections of the same stage. Both are specifically relative to one another. Minimizing energy utilization using load balancing which will automatically cut down carbon discharge and for that reason develop Green Computing.

3 Challenges for Load Balancing in Cloud Computing

In spite of the certainty, cloud computing has been broadly put into practice. Research in cloud computing is yet in its beginning phases, and many technical challenges continue to be uncertain in the scientific society and they are [1, 7].

3.1 Automated Service Provisioning

The core element related to cloud computing is flexibility; services may be assigned or delivered automatically. How then we would be able to utilize or discharge the resources of the cloud, just retaining the identical efficiency as traditional systems and utilizing ideal resources.

3.2 Spatial Distribution of the Cloud Nodes

A number of algorithms are formulated to be beneficial actually for an intranet in which communication waiting time are minor. Nonetheless, it is a test to layout a load balancing algorithm which can easily function with regards to divided nodes. This is because alternative aspects will have to be accepted into consideration like the network links speed amont most of the nodes, the length between the consumer and the task filtering nodes, and the ranges among the nodes concerned with delivering the service.

3.3 Virtual Machines Migration

Using virtualization, the whole machine are able to be seen like a file or group of files, to empty a physical machine, over loaded, it is achievable to swap a virtual machine inside physical machines. The most important goal is to disperse the load in a data center or group of data centers.

3.4 Algorithm Complexity

Load balancing algorithms are targeted to be much easier in conditions of application and implementation. The better execution barrier could possibly manage to much more challenging task that may reason few risky performance problems. Additionally, if the algorithms will need more detail and better connection for inspecting and to manage, waiting times may provokes additional issues and the affectivity will fall.

3.5 Energy Management

The advantages of energy management that proposes the usage of the cloud is the economy of scale. Power saving is the most important aspect that permits a worldwide economy where the pool of worldwide resources is going to be supported by minimal providers rather that every one possesses its private services.

4 Conclusion

In this paper, we have surveyed cloud load balancing, discussed it types and comparisons of static and dynamic load balancing, demand as well as its goals. We also take into consider some technical challenges continue to be unsolved in the scientific society, particularly challenges of load balancing. We desire this work may provide a much better guide to the critical issues that need to be addressed, while the design of load balancing algorithm in cloud computing.

References

1. Klaithem Al Nuaimi, Nader Mohamed, Mariam Al Nuaimi and Jameela Al-Jaroodi, "A Survey of Load Balancing in Cloud Computing: Challenges and Algorithms ", 2012 IEEE Second Symposium on Network Cloud Computing and Applications.
2. Mohammad Oqail Ahmad, Rafiqul Zaman Khan, "The Cloud Computing: A Systematic Review", International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), Vol. 3, Issue 5, May 2015.
3. Foram F Kherani, Prof. Jignesh Vania, "Load Balancing in cloud computing", International Journal of Engineering Development and Research, Volume 2, Issue 1 | ISSN: 2321-9939.
4. Soumya Ray and Ajanta De Sarkar, "Execution Analysis Of Load Balancing Algorithms In Cloud Computing Environment", International Journal on Cloud Computing: Services and Architecture (IJCCSA), Vol. 2, No. 5, October 2012.

5. Ali M. Alakeel, "A Guide to Dynamic Load Balancing in Distributed Computer Systems", IJCSNS International Journal of Computer Science and Network Security, VOL. 10 No. 6, June 2010.
6. http://totaluptime.com/solutions/cloud-load-balancing/cloud-load-balancing-101/.
7. Shoja, H.; Nahid, H.; Azizi, R., "A comparative survey on load balancing algorithms in cloud computing," Computing, Communication and Networking Technologies (ICCCNT), 2014 International Conference on, vol., no., pp. 1, 5, 11–13 July 2014.
8. Firoj Ali and Rafiqul Zaman Khan, "The Study On Load Balancing Strategies In Distributed Computing System", International Journal of Computer Science & Engineering Survey (IJCSES) Vol. 3, No. 2, April 2012.
9. GeethuGopinath P P, Shriram K Vasudevan, "An in-depth analysis and study of Load balancing techniques in the cloud computing environment.", Procedia Computer Science 50 (2015) 427–432.