INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 06, June 2023, pp : 123-126

LOAD BALANCING ALGORITHMS IN CLOUD COMPUTING ENVIRONMENT: A SURVEY

Mandeep Kaur Gulati[1]

[1]Assistant Professor, PG Deptt. of Computer Science, Khalsa College for Women, Amritsar, Punjab, India

DOI : https://www.doi.org/10.56726/IRJMETS31629

ABSTRACT

Cloud computing is emerging as a new standard model for enabling ubiquitous network access, computing resources, deploying, organizing, and accessing vast distributed computing applications over the network. In Cloud Computing, Load balancing is one of the important techniques used to make sure that there is an equal and dynamic distribution of workload and efficient resource utilization. Thus, it is important to solve issues regarding load balancing and to enhance the performance of cloud-based applications. This paper emphasizes the review of various load balancing algorithms with their advantages and disadvantages that will help improve the load balancing performance of cloud systems.

Keywords: Cloud Computing, Load balancing, static load balancing, dynamic load balancing, load balancing algorithm

1. INTRODUCTION

"Cloud is a parallel and distributed computing system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements (SLA) established through negotiation between the service provider and consumers." [1]. Figure 1 shows the cloud computing architecture. The US National Institute of Standards and Technology (NIST) [2] characterizes cloud computing as ". . . a pay-per-use model for enabling available, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." This definition describes Cloud Computing using [2], [3]:

- **Three service models:** Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS).
- **Four deployment models:** Private Clouds, Community Clouds, Public Clouds, and Hybrid Clouds.
- **Five characteristics:** on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service.

Any cloud computing system consists of three major components which are:

Client: Client is end users, which interact with the clouds to manage information related to the cloud. Clients can be Mobile client, Thin client and Thick client.

- **Datacenter:** Datacenter is the collection of servers hosting different applications and it may exist at a large distance from the clients.
- **Distributed Servers:** Distributed servers are the part of a cloud which actively checks services of their hosts and available throughout the internet hosting different applications.

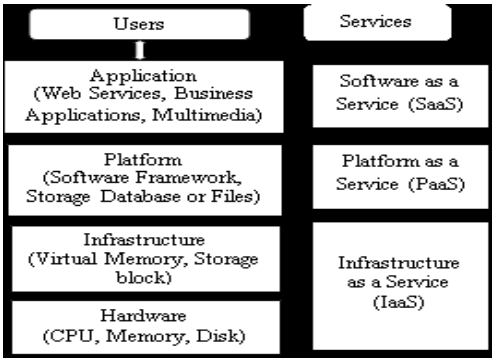

Fig 1: Cloud Computing Architecture

In cloud environment, Load balancing is a technique that distributes the excess dynamic local workload evenly across all the nodes. Load balancing is used for achieving a better service provisioning, resource utilization and improving the overall performance of the system. For the proper load distribution a load balancer is used which received tasks

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
www.ijprems.com
editor@ijprems.com
Vol. 03, Issue 06, June 2023, pp : 123-126
e-ISSN : 2583-1062
Impact Factor : 5.725

from different location and then distributed to the data center. A load balancer is a device that acts as a reverse proxy and distributes network or application load across a number of servers [4][5]. Figure 2 presents a framework under which various load balancing algorithms work in a cloud computing environment.

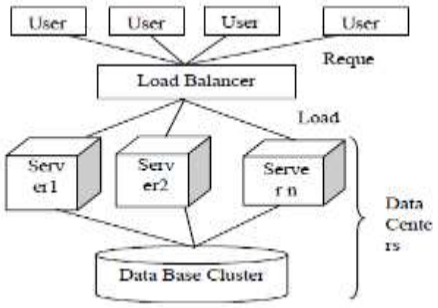

Fig 2: Framework for working of Dynamic Load Balancing

Load Balancing plays a vital role in maintaining activities in a cloud computing environment. It minimizes the response time in order to avoid system overload; also, it maximizes throughput as well as obtaining optimal resource utilization. The main aim of introducing algorithms in load balancing is to avoid overloading and idleness of nodes in a cloud system. Load balancing algorithms are necessary because they provide continuous services to users without service breaking. Static load balancing is found in a static environment where algorithms' performance does not consider the current state of the system. Therefore user requirements do not change during the run-time. In dynamic load, balancing the performance of the algorithms highly depends on the state of the system. In a dynamic environment, algorithms efficiently perform load balancing since resources are flexible in nature. Different algorithms are designed for different purposes e.g. some algorithms are intended to attain maximum throughput; others intend to have the least response time, or have a maximum usage of resources or target to have a trade-off in all the system measurements. This paper presents survey of some of the load balancing algorithms with their advantages and disadvantages.

2. LOAD BALANCING ALGORITHMS IN CLOUD COMPUTING

Load balancing is a technique of distributing the total load to the individual nodes of the collective system to the facilitate networks and resources to improve the response time of the job with maximum throughput in the system. In literature, there are various load balancing algorithms some of which have been surveyed in this section along with their advantages and disadvantages.

Round-Robin Load Balancing Algorithm - This is a static load balancing algorithm and its implementation is the simplest of all algorithms. In these algorithms, the allocation of jobs to processors is done circularly.Initially, it selects any random node and allocates a job to it, then it moves to other nodes to allocate in a round-robin approach, without showing any priority [6]. Here, each node is assigned with some time quantum in which it has to execute the job, if the job is not finished it has to wait for the next slot to resume its execution. The advantage of this algorithm is the fastest response time of the processes. Also it doesn't lead to starvation. The process need not wait for a long time to execute its job. However, due to the uneven distribution of workload, some of the nodes get overloaded and underloaded as the execution time of the process is not determined earlier.

Opportunistic Load Balancing Algorithm- This is a static load balancing algorithm that does not consider the current workload of each system. Therefore it keeps each node busy by randomly distributing all uncompleted tasks to the available nodes. This makes the algorithm to provide poor results on load balancing [7]. It fails to calculate the node's implementation time, which then lowers the performance of the processing task. Also, when there are nodes in the idle state, then there will be bottlenecks in the cloud system.

Min-Min Load Balancing Algorithm-The algorithm is concerned with those tasks which take minimum time to complete. It is simple and fast and provides improved performance [8]. The process starts by calculating the minimum completion time of all the loads. The minimum value is then selected, and as per that minimum time, the task is scheduled in the machine. After updating the current execution time on the machine, the task is then removed from the available task set. This process continues until all the tasks in the set are allocated to the equivalent machine. However, several shortcomings of Min-Min Algorithm include: inability of running tasks simultaneously, the algorithm gives high priority to smaller tasks which leads to starvation for larger tasks and in turn results in imbalanced virtual machine load.

Max-Min Load Balancing Algorithm- The Max-Min algorithm is identical to the above Min-Min algorithm, once the minimum completion time of all the available tasks is computed, then among these, the task which has maximum completion time among all the tasks as assigned to the corresponding node that has minimum completion time [9].

Then all the remaining tasks on that node are updated and that allocated task is deleted from the record. Similarly, all the remaining tasks are allocated with a resource. In this algorithm, smaller jobs (less execution time) are executed simultaneously along with the larger jobs (large execution time), so the makespan (total time taken for executing all the tasks) is reduced and resources are utilized efficiently unlike in the Min-Min algorithm. However, this algorithm is applicable only to small-scale distributed systems.

Active Clustering Load Balancing Algorithm- Active Clustering algorithm [10] works on the principle of grouping the similar nodes and work together on the available groups. A set of processes is iteratively executed by each node on the network. Initially any node can become an initiator and selects another node from its neighbours to be the matchmaker node satisfying the criteria of being a different type than the former one. The matchmaker node then forms a connection between neighbours of it which are similar to the initiator. The matchmaker node then removes the connection between itself and the initiator The advantage of the algorithm is that resources are utilized efficiently as the virtual machines are grouped as a cluster with similar properties. However, the system's performance is decreased when the variety of nodes increases.

Ant Colony Optimization Load Balancing Algorithm- The main goal of this load balancing algorithm [11] is to explore an optimal path between the food source and colony of ants according to the behavior of the ant. Its objective is to efficiently distribute theworkload among all the nodes. Firstly, when the request is made, the ant begins moving in the direction of the food source from the head node. While moving ahead, ants keep a record of every node they have visited for making future decisions. During their movement ants deposit the pheromones so that it helps further ants to choose the next node. The strength of pheromones depends on the components such as food quality, the distance of food etc. Denser pheromone is attracted by many ants. The pheromones are updated when the jobs are executed. The advantage of this algorithm is that it overcomes heterogeneity and is adjustable for dynamic environments. Also it enhances the performance of the system. However, the drawback of the protocol is that the network overhead is increased.

Throttled Load Balancing (TLB) Algorithm: Throttled load balancing algorithm is a dynamic load balancing algorithm [12] in which the client first requests the load balancer to find a suitable virtual machine to perform the required operation. In Cloud computing, there may be multiple instances of virtual machine. These virtual machines can be grouped based on the type of requests they can handle. Whenever a client sends a request, the load balancer will first look for that group, which can handle this request and allocate the process to the lightly loaded instance of that group. The advantage of this algorithm is that the resources are utilized efficiently and good performance is obtained. However, the drawback of this protocol is that the current workload of VM is not considered.

Equally Spread Current Execution (ESCE): According to [12], ESCE is a dynamic load balancing algorithm, which handles the process with the priority. It determines the priority by checking the size of the process. This algorithm distributes the load randomly by first checking the size of the process and then transferring the load to a virtual machine, which is lightly loaded. The load balancer spreads the load on different nodes, and hence, it is known as spread spectrum technique. However, common problem of ESCE is that it causes overhead when updating the index table due to the communication that occurs between the Data Center controller and the load balancer

Honey Bee Algorithm (HB): The idea of this is that a group of bees known as the foraging bees, spread to look for food sources and send location information to the other bees. This is known to solve decision making and classification problems with more robust and flexible ways) [13]. The fitness of the bees is evaluated every time and the searching for food process is repeated. Similarly, in a cloud environment, there are various changes of demand on the servers, and services are allocated dynamically and VMs should be utilized to the maximum limit reducing the waiting time. Honey bee behavior can be mapped to a cloud environment. The problem with this algorithm is that tasks with low priority tasks will be waiting in the queue.

Genetic Algorithm (GA): According to [14], Genetic Algorithm has been used as a soft computing approach, which uses the mechanism of natural selection strategy. A simple Genetic Algorithm is composed of three operations: genetic operation, selection, and replacement operation. The advantage of this technique is that it can handle a vast search space applicable to complex objective function and can avoid being trapped in locally optimal solution. A generation is a collection of artificial creatures (strings). In every new generation, a set of strings is created using information from the previous ones. Occasionally, a new part is effort for good measure. According to [15] Genetic Algorithms are randomized, but they are not simple random walks. They adept exploit historical information to speculate on new search points with expected improvement. The effectiveness of the GA depends in appropriate mix of exploration and exploitation.

Particle Swarm Optimization (PSO) Algorithm: Particle Swarm Optimization (PSO) as a meta-heuristics method is a self-adaptive global search based optimization technique introduced by Kennedy and Eberhart [16]. The PSO algorithm is alike to other population-based algorithms like Genetic algorithms (GA) but, there is no direct

recombination of individuals of the population. The PSO algorithm focuses on minimizing the total cost of computation of an application workflow. The objective is to minimize the total cost of execution of application workflows on Cloud computing environments. Results show that PSO based task-resource mapping can achieve at least three times cost savings as compared to Best Resource Selection (BRS) based mapping for application workflow. In addition, PSO balances the load on compute resources by distributing tasks to available resources.

3. CONCLUSION

In the cloud computing environment, load balancing is one of the main issues, which is required to distribute workload to all the nodes in the cloud to improve the performance and maximize resource utilization. This paper explains overview of cloud computing and types of load balancing. Different load balancing algorithms in cloud computing have been surveyed. The strengths and weaknesses of these protocols have also been provided so as to explore the future areas of research. Therefore in the future, it will be necessary to fine-tune the algorithms to achieve better consistent results from different perspectives.

4. REFERENCES

[1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic Cloud computing and emerging IT platforms: Vision, Hype and reality for delivering computing as the 5th utility," Future Generation Computer Systems, 25: 599-616, 2009.

[2] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," National Institute of Standards and technology, Information Technology Laboratory, Technical Report Version 15, 2009.

[3] Rimal, Bhaskar Prasad, Eunmi Choi, and Ian Lumb. "A taxonomy and survey of cloud computing systems." INC, IMS and IDC, 2009. NCM'09. Fifth International Joint Conference on. IEEE, 2009.

[4] L. M. Vaquero, L. Rodero-Merino, J. Caceres and M. Lindner, "A break in the clouds: towards a cloud definition," SIGCOMM ACM Computer Communication Review, Vol. 39, December 2008, pp. 50–55.

[5] Rahman, Mazedur, Samira Iqbal, and Jerry Gao. "Load Balancer as a Service in Cloud Computing." In Service Oriented System Engineering (SOSE), 2014 IEEE 8th International Symposium on, 2014, pp. 204-211.

[6] A. Aditya, U. Chatterjee, S. Gupta, "A comparative study of different static and dynamic load balancing algorithm in cloud computing with special emphasis on time factor," Int. J. Curr. Eng. Technol. **5**(3), 2015, pp.1898–1907.

[7] A. Jyoti, M. Shrimali and R. Mishra, "Cloud Computing and Load Balancing in Cloud Computing -Survey," 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 51-55

[8] G. Liu, J. Li, J. Xu, "An Improved Min-Min Algorithm in Cloud Computing. In: Du Z. (eds) Proceedings of the 2012 International Conference of Modern Computer Science and Applications. Advances in Intelligent Systems and Computing, vol 191. Springer, Berlin, Heidelberg, 2013.

[9] N. Sharma, S. Tyagi, S. Atri, "A comparative analysis of min-min and max-min algorithms based on the makespan parameter. Int. J. Adv. Res. Comput. Sci. 8(3), 2017, pp. 1038–1041.

[10] S. K.Dhurandher, M. S. Obaidat, I. Woungang, P. Agarwal, A. Gupta, and P. Gupta, "A cluster-based load balancing algorithm in cloud computing" in Communications (ICC), 2014 IEEE International Conference on, IEEE, 2014, pp. 2921-2925.

[11] D. Kashyap, J. Viradiya, "A survey of various load balancing algorithms in cloud computing," Int. J. Sci. Technol. Res 3(11), 2014, pp. 115–119.

[12] S. G. Domanal, and G. Ram Mohana Reddy Load Balancing in Cloud Computingusing Modified Throttled Algorithm Cloud Computing in Emerging Markets CCEM 2013 IEEE International Conference on IEEE, 2013.

[13] P.B. Kiritbhai, N. Y. Shah, "Optimizing Load Balancing Technique for Efficient Load Balancing. Int. J. Innov. Res. Technol. 4 (6), 2017, 39–44.

[14] Ye Zhen, Xiaofang Zhou and Athman Bouguettaya. Genetic algorithm based QoS-aware service compositions in cloud computing." Database systems for advanced applications. Springer Berlin Heidelberg, 2011

[15] Dam, Scintami, et al. "Genetic algorithm and gravitational emulation based hybrid load balancing strategy in cloud computing." Computer, Communication, Control and Information Technology (C3IT), 2015 Third International Conference on. IEEE, 2015.

[16] Pandey, Suraj, Lin Wu, Siddeswara Mayura Guru, and Rajkumar Buyya. "A particle swarm optimization-based heuristic for scheduling workflow applications in cloud computing environments." In Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on, pp. 400-407. IEEE, 2010.