Load Balancing Algorithms in Cloud Computing

Vignesh Joshi

Load Balancing Algorithms in Cloud Computing

Vignesh Joshi

Department of Computer Science, University of Delhi, India

Abstract

Cloud computing plays an important role in enhancing the effective sharing of resources in virtual machines. Scheduling and load balancing are the two common concepts that cloud computing relies on ensuring that a prescribed task is assigned to the most appropriate virtual machine. Also, cloud computing should have the ability to handle multiple independent tasks that are arriving and execute them in the same or multiple nodes. In a heterogeneous environment, static and dynamic scheduling plays an important role by enhancing the allocation of tasks to the appropriate resources to satisfy the internet users' requests and making cloud computing technology more efficient. This work aims to evaluate and discuss important algorithms that will help improve the load balancing performance of cloud systems.

Keywords: Cloud Computing, Algorithms, Computer Science, Software Engineering

1. Introduction

Load balancing is a key aspect of cloud computing, and it has grown tremendously since the inception of cloud computing [1]. Cloud computing started to become apparent in the year 2007, and it has been providing large infrastructure, scalability, storage, virtualization, resource pooling, and a wide range of services via the internet. All cloud computing users enjoy different types of technologies provided by cloud service providers, at any instance of time. As technology is advancing, the demand for cloud computing services by internet users is increasing day by day [7]. This has led to various technical challenges, including virtual machine migration, high availability, fault tolerance, scalability, and server consideration, which must be addressed to meet the growing demand. However, the main centralized issue that has been affecting cloud computing is load balancing. Therefore, load balancing in a cloud system is expected to provide a proper balance of workload among various available nodes of a distributed system [2]. The reason why load balancing was introduced was to improve the nodes' speed and performance in the cloud system and to keep saving individual devices from hitting their threshold by dropping down their performance.

2. Cloud Computing

It can be applied in either parallel or distributed systems, and it has brought a great positive impact in a computing environment. It is most effective where a shared pool of computing resources is configured together since the model is convenient. It allows on-demand network access with minimal or no interaction of a service provider. Therefore cloud computing is an internet-based model that allows users to access services over a network with a reliable data center. A client is the end-user who can access the cloud to manage his or her information. A data center represents a collection of servers that holds different types of applications. Distributed servers store information and respond to clients' requests.

Different types of cloud computing are categorized according to two approaches that are, capability and accessibility. Under accessibility, there are three approaches, including private cloud, public cloud, and hybrid cloud [3]. Basing on the cloud computing capabilities, then cloud system provides three different services, as discussed below.

- Software as a Service (SaaS)- this approach allows internet users to use applications built on another system by someone else without maintaining the overhead or purchasing the application. For example, google documents and web-based e-mail.
- Platform as a Service (PaaS)- this approach allows users to develop an application from a cloud service provider that provides a development environment with web-based tools and libraries.
- Infrastructure as a Services (IaaS)- these are online services that provide virtualized computing resources over the internet. Its main aim is to provide highly scalable resources, including hardware, storage, servers, software, and infrastructure in a virtualized manner.

3. Load Balancing

It 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 [4]. The main of introducing algorithms in load balancing is to avoid overloading and idleness of nodes in a cloud system. Therefore, algorithms will ensure that all the nodes are assigned to the same amount of workload in a cloud system. An increase in cloud computing platforms such as Windows Azure Platform, Amazon S3, etc., and usage in Artificial Intelligence [5] will enhance the development of many web searches with distinctive features, from cloud service providers.

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.

Load balancing algorithms in the cloud computing

3.1 Round-Robin Algorithm

The round and robin algorithm is amongst the easiest methods of load balancing since it has a very efficient and effective scheduling policy that is time triggered. It uses the round-robin method for assigning jobs to the devices in a cloud environment. The algorithm randomly selects the nodes when performing load balancing. Data centers are the main components that these algorithms rely on. Internet users will send a request to the cloud system, and then the data center controller will receive the request and pass it to the round-robin algorithm. The algorithm is mostly based on time-sharing, where it divides time into slice and quantum.

The process starts by storing all the processors in a circular queue where the scheduler allocates the server according to the defined time slot, among all processes in the list set. The algorithm schedules the processes so that when a new process comes in, it will be added at the further end of the queue. The algorithm will randomly select the first process from the queue using the scheduler, and when the time slot of the process is over, the algorithm will forward the process to the end of the queue. Also, when the process ends before the defined time slot, the algorithm will voluntarily release the process [11]. Therefore all the processes have different loading times, and it is possible to have some nodes being overloaded while others being underutilized. This makes the performance of the load balancing to decrease and to solve this issue; then, a Weight round-robin load balancing algorithm was introduced to provide a better allocation technique.

Weight round-robin balancing algorithm ensures that it has distributed the prescribed weight and jobs as per the values of the weight. Therefore the algorithm assigns the processors that have a greater ability with a bigger value of the weight. The servers with the highest weight value will hold more tasks, and when the entire weight comes in level, servers will get steady traffic.

3.2 Opportunistic Algorithm.

This 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 [12]. 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.

3.3 Min-Min 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.

3.4 Max-Min Algorithm

The max-min algorithm calculates maximum value after searching out the minimum implementation time for all available tasks [9]. The algorithm then selects a task with high completion time and assigns the task to the equivalent machine. Then the algorithm updates the execution time of all the tasks and later after execution task is removed from the list. The difference of this algorithm from the min-min algorithm is that it has only one long task in a set that runs in parallel with many shorter tasks.

3.5 Active Monitoring Algorithm

This is a dynamic load balancing algorithm that finds out the least loaded, or the idle virtual machine assigns a load to them [10]. Controllers in load balancing maintain all the servers and requests in the server’s index table. Therefore when the system receives a new request, the data center expects the index table to identify the servers that are least loaded or are idle. That is, the algorithm uses first come first serve technique when assigning load to the servers. The task is identified using server-id, and when a load is allocated to the server, its state increases in the index table. Similarly, when a task is completed, the data center and the controllers receive the information, reducing the server state in the index table. When

an internet user sends a request, the load balancer will scan the index table again and allocate the processes accordingly.

3.6 Equally Spread Current Execution Algorithm

This is a dynamic load balancing algorithm that distributes an equal amount of load to all the servers in data centers. The algorithm will select all the processes in the list, assign priority to them, and then calculates the size and capacity of the processes. The algorithm will then find the server that will use less time to handle the load. In order to identify the best server, the capacity of the virtual machine is measured as well as estimating the load. Therefore, the algorithm assigns the load to the matching virtual machine regarding the size and capacity.

Various measured parameters have used to compare the performance of the above algorithms, as shown in the table below [6].

Load Balancing Algorithms/ Performance parameters	Throughput	Overhead	Fault-Tolerance	Response Time	Resources Utilization	Scalability	Performance
Round Robin	yes	yes	yes	yes	yes	yes	yes
Opportunistic	no	no	no	yes	no	no	no
Min-Min	yes	yes	no	yes	yes	no	yes
Max-Min	yes	yes	no	yes	yes	no	yes
Active Monitoring	yes	yes	no	yes	yes	yes	no

Figure 1: LB Algorithms Comparison chart

4. Conclusion

To provide the fastest connectivity for all the nodes or devices that require cloud computing services, then it is necessary to apply different kinds of load balancing algorithms. These algorithms will help in improving the cloud system's performance, scalability, resource utilization, fault-tolerance, throughput, and overhead. The above table proved that the most suitable algorithm for heterogeneous and homogeneous tasks is the Round and Robin (Weighted Round Robin) algorithm. Therefore it is important to conclude that load balancing algorithms play a key aspect as well a challenging task in cloud computing. This paper has extensively evaluated the concepts of cloud computing, different types of cloud computing, and also the load balancing algorithms. The above research on the algorithms has majored on the overall completion time of all the processes in the queue. Therefore in the future, it will be necessary to fine-tune the algorithms to achieve better consistent results from different perspectives.

References

[1] Jyoti, A., & Shrimali, M. (2019). Dynamic provisioning of resources based on load balancing and service broker policy in cloud computing. Cluster Computing, 23(1), 377-395. doi: 10.1007/s10586-019-02928-y

[2] Singh, P., Baaga, P., & Gupta, S. (2016). Assorted Load Balancing Algorithms in Cloud Computing: A Survey. International Journal Of Computer Applications, 143(7), 34-40. doi: 10.5120/ijca2016910258

[3] Shukla, S., & Arora, D. (2015). A Hybrid Optimization Approach for Load Balancing in Cloud Computing. International Journal Of Private Cloud Computing Environment And Management, 2(2), 11-22. doi: 10.21742/ijpccem.2015.2.2.02

[4] Afzal, S., & Kavitha, G. (2019). Load balancing in cloud computing – A hierarchical taxonomical classification. Journal Of Cloud Computing, 8(1). doi: 10.1186/s13677-019-0146-7

[5] Tadapaneni, N. R. (2017). Artificial Intelligence In Software Engineering. Available at SSRN: 3591807 or doi: 10.2139/ssrn.3591807

[6] Kumar, R., & Prashar, T. (2015). Performance Analysis of Load Balancing Algorithms in Cloud Computing. International Journal Of Computer Applications, 120(7), 19-27. doi: 10.5120/21240-4016

[7] Tadapaneni, N. R. (2018). Cloud Computing: Opportunities and Challenges. Available at SSRN Electronic Journal. 10.2139/ssrn.3563342.

[8] Liu G., Li J., Xu J. (2013) 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

[9] Patel, G., Mehta, R., & Bhoi, U. (2015). Enhanced Load Balanced Min-min Algorithm for Static Meta Task Scheduling in Cloud Computing. Procedia Computer Science, 57, 545-553. doi: 10.1016/j.procs.2015.07.385

[10] Rai, S., Sagar, N., & Sahu, R. (2017). An Efficient Distributed Dynamic Load Balancing Method based on Hybrid Approach in Cloud Computing. International Journal Of Computer Applications, 169(9), 16-21. doi: 10.5120/ijca2017914876

[11] Tadapaneni, N. R. (2017). Different Types of Cloud Service Models. Available at SSRN 3614630.

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