EasyChair Preprint

№ 10405

A Review on Load Balancing Algorithms Used in Cloud Computing

Manisha Dewangan

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

June 15, 2023

A review on Load Balancing Algorithms used in Cloud Computing

Dr. Manisha Dewangan, Assistant professor, UTD,Sant Gahira Guru Vishwavidlaya ,Ambikapur Surguja Chhattisgarh(ranidevm@gmail.com)

ABSTRACT

In the realm of computer networking, cloud computing is an advanced networking concept in which computational resources (such as software, hardware, etc.) are made available as a service over the internet. The service providers oversee all the necessary systems that they wish to make available to users as needed. In order to make money, service providers must efficiently and quickly handle all consumer demands while keeping costs at a manageable level. However, managing resources to meet demand is difficult due to the complex load characteristics that depend on numerous variables and lead to significant variations in load (the total amount of data processed in the cloud) and service requirements (such as multimedia processing and database management methods). This essay discusses cloud functionality.

Keywords: ***Words*** Cloud Computing, Resource management, Algorithm,

1. INTRODUCTION

The cloud computing (CC) technology is responsible for the development of the idea of global network resource optimisation. It is common knowledge that massive, Internet-accessible computer and storage farms carry out central information processing, which is quicker and more efficient. When computing is done via remote data centres as opposed to local ones, it is referred to as network-centric computing or network-centric content. Two new computer models are widely utilised now thanks to this idea and advancements in Internet technology.Accepted: grid computing; cloud computing; utility computing, of which CC is the most recent.Cloud computing (CC) is a novel technique to using internet technology to set up computing infrastructure. It refers to on-demand applications and hardware that is offered as a service through the virtualization of hardware and systems software in datacenters. The hardware and operating system software present in datacenters are referred to as "Cloud" resources. Utility computing was created as a result of the consolidation of hardware and software into sizable data centres and the ability for consumers to pay for the computing, storage, and communication resources as they are used. On the basis of Internet technologies, it provides elastic services.Elastic computing refers to the ability to instantly modify resource consumption in response to fluctuating workloads. The cloud is poised to become the next dominant computing paradigm as a result of these appealing qualities, which are grabbing the attention of the industry. Security, virtualization, capacity planning, load balancing, energy optimisation, and assuring Quality of Service (QoS) are just a few of the issues that CC is battling, despite the fact that it is still in its early phases. While some of these challenges are carried over from parallel and distributed computing, cloud computing also encounters a lot of unique challenges.In order to effectively balance the load over the cloud, this paper reviews various load balancing techniques.

1.1 CLOUD COMPUTING

The term "cloud computing," sometimes known as "the cloud," is used to describe computer principles including several computing resources connected via a real-time communication network, most often the Internet"[2]. In science, the term "cloud computing" refers to distributed computing via a network and describes the capacity to simultaneously run a programme on numerous connected machines. The term is more frequently used to describe network-based services that appear to be delivered by actual server hardware but are actually offered by virtual hardware that is emulated by software operating on one or more actual machines. Such virtual servers can move around and scale up (or down) instantly without harming the end user because they do not actually exist. In this regard, they are comparable to clouds.

The usage of computer resources (hardware and software) that are provided as a service across a network (usually the Internet) is how cloud computing is more specifically defined. The name is derived from the use of a cloud-shaped symbol in system diagrams as a metaphor for the intricate architecture it comprises. Cloud computing entrusts the data, software, and processing of a user to remote services.

Public Cloud

In a public cloud, cloud services are made available to the general public across a network. Customers have no say in where the infrastructure is located. All users contribute to the cost, which is either free or takes the form of a licence scheme like pay per user.

Public clouds are fantastic for businesses that need to manage both the host application and the multiple user-facing applications.

Private Cloud

A private Cloud is a cloud infrastructure that is solely used by one organization.

- It gives organizations greater control over security and data safeguarded by a firewall and managed internally.
- It can be hosted internally or externally.
- Private clouds are great for organizations that have high-security demands, high management demands, and uptime requirements.

Hybrid Cloud

Hybrid Cloud uses both private and public clouds but can remain separate entities.

- Resources are managed and can be provided either internally or by external providers.
- A hybrid cloud is great for scalability, flexibility, and security.
- An example of this is an organization that can use the public cloud to interact with customers while keeping their data secured through a private cloud.

Community Cloud

It is an infrastructure that is mutually shared between organizations that belong to a particular community.

- The community members generally share similar privacy, performance, and security concerns.
- An example of this is a community cloud at banks, the government in a country, or trading firms.
- A community cloud can be managed and hosted internally or by a third-party provider.
- A community cloud is good for organizations that work on joint ventures that need centralized cloud computing ability for managing, building, and executing their projects.

1.2 CLOUD SERVICE MODELS

Cloud service models focus on providing some type of offering to their clients.

Cloud Software as a Service:

Cloud Software as a Service, is a type of cloud that offers an application to customers or organizations through a web browser.

- The data for the app runs on a server on the network, not through an app on the user’s computer.
- Software is usually sold via subscription
- Examples of SaaS are Salesforce, Google Docs, Office 365, Basecamp, etc.

Cloud Infrastructure as a Service

Cloud Infrastructure as a Service, provides the hardware and usually virtualized OS to their customers.

- Software is charged only for the computing power that is utilized, usually, CPU hours used a month.
- Examples of IaaS are Amazon EC2, Rackspace, Google Compute Engine, etc.

Cloud Platform as a Service

Cloud Platform as a Service, provides networked computers running in a hosted environment, and also adds support for the development environment.

- PaaS offerings generally support a specific program language or development environment.
- Deploying your app in this environment, you can take advantage of dynamic scalability, and automated database backups without the need to specifically code for it.
- PaaS is billed as an additional cost on top of the IaaS charges.
- Examples of PaaS are Google App Engine, Cloud Foundry, and Engine Yard Etc.

1.3. WHY LOAD BALANCING?

The key issue with any network or system is load balancing because it has an impact on the system's performance, functionality, and cost in the cloud [7]. Utilising all resources as little as possible is an approach for managing resources. It is a method for equally distributing the burden across the network's slug nodes. For instance, if we had four identical servers named A, B, C, and D, with relative loads of 80%, 60%, 40%, and 20% of their capacity, respectively, each would have 50% of the load under ideal load balancing. In distributed systems, LB middleware is frequently used to increase scalability and overall system throughput [8]. Why is LB such a big deal in the cloud when there are already so many scheduling methods available? The flexibility of it is the reason. Independent businesses frequently supply the resource provisioning. Therefore, these businesses can alter the number of resources they give in accordance with their needs or competitive strategy.Therefore, after receiving a specific request, it is the responsibility of the load balancer to determine which server component within the list of accessible server components delivers the most benefit.

1.4. BASIC LOAD BALANCING ALGORITHMS

1.4.1 INTRODUCTION & GOALS OF LOAD BALANCING

It involves reassigning the entire load to each individual node of the collective system in order to increase resource utilisation and task response times while simultaneously eradicating a situation in which some nodes are underloaded while others are overloaded. A dynamic load balancing algorithm relies on the system's current behaviour and does not take into account the system's previous state or behaviour. When creating such an algorithm, it is crucial to take into account factors like load estimation, load comparison, system performance, system stability, node interaction, the type of work that needs to be transferred, node selection, and many others.

The goals of load balancing are:

- To improve the performance substantially
- To have a backup plan in case the system fails even partially
- To maintain the system stability
- To accommodate future modification in the system

1.4.1 LOAD BALANCING METRICS FOR CLOUD

The existing load balancing techniques in clouds, consider various parameters “metrics” viz. response time, scalability, throughput, resource utilization, fault tolerance, migration time, associated overhead, energy consumption and carbon emission, etc [10]. Those are discussed below: a) Response Time: is the amount of time taken by computing system to give first response to the given task. It should be low. b) Throughput: It is the number of jobs completed in a given time period. It should be higher for better performance. It should be high. c) Resource Utilization: This means how many resources a system is using to complete given amount of load in optimized manner. All resources should be used to fulfill the request in context of minimizing response time and increasing the throughput of CC environment. d) Scalability: means the technique is able to manage load in a changing number of nodes environment. e) Fault Tolerance: is the ability of load balancing technique to balance the load uniformly of a failure node to other nodes. In CC environment this property is very important because CC came as a business model. f) Associated Overhead: determines the amount of overhead involved in calculating the movement of tasks, inter-processor and inter-process communication. It should be minimized so that a LB technique can work fast. g) Migration Time: is the time to migrate a cloudlet from one node to another. It should be minimized for better performance. h) Energy Consumption: is the amount of energy (i.e. electricity) consumed by resources for execution. It must be kept low from cost perspective as well as natures'. i) Carbon Emission: means the amount of carbon generated from the system. It is directly proportional to the energy consumption. The more energy consume the more carbon will be emitted.

1.4.2 Load balancing algorithms in the cloud computing

Round-Robin Algorithm The round and robin algorithm is amongst the easiest methods of load balancing since it has a very efficient and effectivescheduling 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.

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.

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.

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.

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.

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 Algorithem/ Performance parameters	Throughtput	Overhead	Fult/Tolerance	Response Time	Resource Utilization	Scalability	Performance
Round Robiin	yes	yes	yes	yes	yes	yes	Yes
Opportmisic	no	no	no	yes	no	no	No
Min-Min	yes	yes	no	yes	yes	no	Yes
Mx-Min	yes	yes	no	yes	yes	no	Yes
Achive Monitoring	yes	yes	no	yes	yes	yes	no

2. CONCLUSION

Applying various load balancing methods is important to ensure that all nodes or devices that want cloud computing services have the fastest connectivity possible. These algorithms will aid in enhancing the cloud system's throughput, fault tolerance, resource utilization, performance, and overhead. The Round andRobin (Weighted Round Robin) algorithm is the best appropriate algorithm for heterogeneous and homogeneous workloads, according to the aforementioned table. It is crucial to get the conclusion that load balancing algorithms are both a crucial component and a difficult task in cloud computing. The ideas of cloud computing, various cloud computing architectures, and load balancing methods have all been thoroughly examined in this study. The overall completion time of all the processes in the queue has been the focus of the aforementioned algorithmic study. The algorithms will therefore need to be improved in the future to produce more accurate findings from various angles.

3. 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

[13] Dan C Marinescu, Cloud Computing: Theory and Practice.: Newnes, 2013.

[14] Luis Rodero-Merino, Juan Caceres and Maik Lindner Luis M. Vaquero, "A Break in the Clouds: Towards a Cloud Definition," ACM SIGCOMM Computer Communication Review, vol. 39, no. 1, pp. 50-55, 2008.

[15] Pradeep, Kang G. Shin, Xiaoyun Zhu, Mustafa Uysal, Zhikui Wang, Sharad Singhal, Arif Merchant, and Kenneth Salem Padala, "Adaptive control of virtualized resources in utility computing environments," ACM SIGOPS Operating Systems Review,ACM, vol. 41, no. 3, pp. 289-302, 2007.

[16] Armando Fox, Rean Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, and I. Stoica. Michael Armbrust, "Above the clouds: A Berkeley view of cloud computing," vol. 58, no. 4, pp. 50-58, 2010.