Study on Load Balancing Algorithms and Performance Metrics in Cloud Computing

Ananya Jain

Computer Science Student, Lotus Valley International School Gurugram, India

---***---

Abstract - *Cloud Computing is an emerging technology that has benefitted the individuals and organizations greatly. It offers several advantages like pooling of resources, cost efficiency, scalability and flexibility but can also exhibit some challenges, one of which can be performance related challenges due to poor management of the available resources and work load. Load balancing in cloud computing refers to the process of distributing workloads and computing resources in the cloud computing environment. On basis of these algorithms, client requests are directed to the servers. In this paper, an attempt has been made to study a few of the existing load balancing algorithms and techniques in cloud computing and critically analyze the various performance metrics that affect the load balancing decisions. On basis of this review, a proposal has been made to combine load balancing algorithms with task priority to further enhance the performance of cloud computing systems.*

Key Words*: *Cloud computing, Virtual Machines, Load Balancing Algorithms, Performance Metrics, Task Priority

1. INTRODUCTION

Cloud Computing is a computational model where all the resources like storage, processors and software are not present physically on the user's machine but are available virtually from servers dispersed over the network, typically the internet. These resources are available on an on-demand basis where the user utilizes what and how much he needs. The term cloud can be thought of as a metaphor or a synonym for the Internet [1].

Cloud computing architecture refers to the way the resources and services are organized and available to the user. It can be categorized in two ways - firstly on the basis of how the user gets access to the resources and secondly on the basis of the services provided.
Depending on how the user gets the access to the cloud, also known as the **deployment model** [2], we have three main types of cloud: **Public cloud** (here the computing services are provided by a third-party service provider and users who wish to avail these services need to subscribe via their accounts), **Private cloud** (here services and access to the cloud are reserved for a particular organization or business and only its employees or members have access to this particular cloud) and **Hybrid cloud** (which combines on-premise data centers with public clouds). Additionally, **Community cloud** is where several organizations or businesses, typically with common interests, have access to the same computing resources.

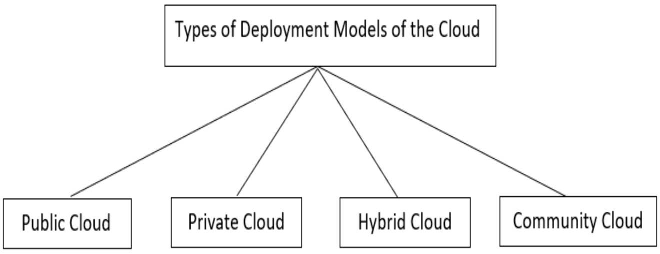

Fig -1: Types of Deployment Models of Cloud Computing

On the basis of what kind of services are being offered there are three **service models**: **IaaS - Infrastructure as a Service** which allows users to access physical resources like servers, storage and networks virtually, **PaaS - Platform as a Service** which allows users to access application development environments or platforms and **SaaS-Software as a Service** which allows users to directly access a service or an application for a particular purpose [3].

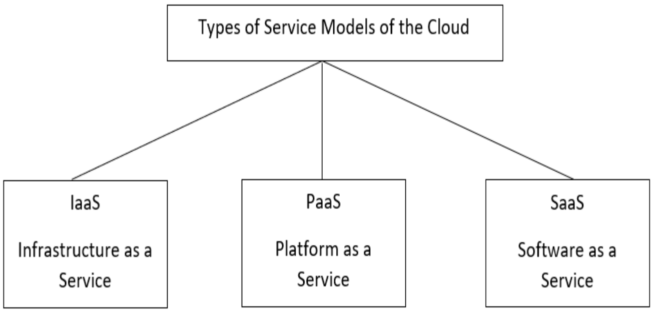
Types of Service Models of the Cloud
IaaS
Infrastructure as a Service
PaaS
Platform as a Service
SaaS
Software as a Service

Fig -2 Types of Service Models of Cloud Computing

2. LOAD BALANCING IN CLOUD COMPUTING

In cloud computing nomenclature, workload is the amount of work that would be assigned to a particular server or a virtual machine (VM). Load Balancing is the method to ensure that all devices or servers are assigned a balanced amount of workload [4]. In simplest words, it means that load balancing determines which task to assign to which server on the basis of several factors. Assigning too many requests to a particular server (i.e. overloading the system) or not properly utilizing the system capabilities (i.e. underloading the system) can causes the computing environment to have lower performance and inefficiency. Therefore, load balancing is required to overcome the problems of underloading and overloading the system [5]. It's a concept that finds application in real life as well. As an analogy, one can visualize the checkout counters at any supermarket where individual customers with their shopping carts can be equated with individual tasks and the billing counters can be equated with the servers or virtual machines (VMs). The ultimate goal of Load balancing will be to maximize the utilization of all billing counters while ensuring that the wait time for customers is minimized.

Similarly, in a cloud computing environment, by assigning servers to task requests in an on-demand basis, efficient resource utilization is ensured.

2.1. Load Balancing Techniques and Algorithms

There are various techniques that determine how the client requests or traffic is processed. The technique for load balancing may take into account the network status as well as nature of requests [6].

Depending on the current state of the system, the load balancing (LB) algorithms can be classified as static or dynamic algorithms [5]. **Static algorithms** assume that the load in the system typically remains the same and all the incoming requests are then distributed across the available servers based on a static policy without actively monitoring the current status of servers [5] [7]. **Dynamic algorithms** check the current network status and on the basis of certain criteria determine which is the server with the least workload or servicing the least number of requests. In this method, work load is distributed among the processors at runtime [7].

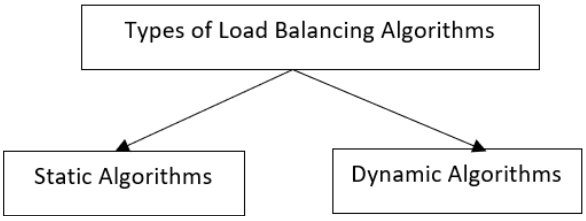

Fig -3 Classification of Load Balancing Algorithms

Depending on the level of requests to be serviced or the load available in the computing system, one of the following algorithms, methods or approaches of load balancing may be used [6][8] [9][10] [11] [12]:

1. Round Robin - It is a simple static algorithm that distributes incoming client requests to servers in rotation. Here the assignment of requests to servers is basically time based and it typically does not take under consideration the characteristics or load handling capacity of the servers. So here, every server receives approximately the same load. Provided that the workload is equally distributed, this is a simple and fast algorithm but its disadvantage is that it may lead to overloading or underloading of servers.

2. Weighted Round Robin - This is a variation of the round robin where every server is assigned a weight based on certain criteria which typically is the processing or computing power of the server. The client requests are assigned to the servers in round robin fashion, but the server with more weight or load-handling capacity will be assigned more requests versus the one with less weight. This helps overcome the drawbacks of the simple round robin technique of load balancing.

3. Opportunistic Algorithm – This is a static algorithm where the requests are randomly distributed to the nodes. No consideration is given to the workload of the system or the execution time of the nodes. Here the tasks are executed slowly and problem of overloading or underloading of nodes may be seen. Hence this algorithm does not provide a good result.

4. Min-Min Algorithm – In this algorithm, the completion times of all the queued tasks is calculated for each of the servers based on current loading of individual servers and the task with the minimum completion time (smallest task) is assigned to the server with minimum completion time (least loaded server). The tasks that require more time to complete need to wait for longer times. Overall, this is a simple and fast technique but may lead to the problem of starvation as the tasks requiring more time may fail to get scheduled.

5. Max-Min Load Balancing Algorithm – This algorithm is similar to min-min. Here, the execution time of all tasks is calculated but here the one with the maximum completion time (largest task) is assigned to the servers for execution first. This allows large tasks to be executed concurrently, and hence solves that problem of starvation.

6. Active Monitoring Algorithm – This algorithm checks the current status of the network as well as the current availability of the servers. The server with the least workload is assigned user request. Thus, before assigning a task to a server, its workload is calculated and post completion of the task the workload is updated.

2.2. Performance Metrics for Load Balancing

How does one determine the efficacy of the cloud computing environment? To analyze the performance of a Cloud Computing System, certain performance characteristics, criteria or metrics can be used. While there are several metrics like throughput, response time, associated overhead, fault tolerance, migration time, resource utilization, scalability, thrashing, reliability, accuracy, predictability and makespan, that can be measured to determine and enhance the load balancing performance of a Cloud Computing system [13] [14] [15] [16] [17] [18] [19], the following seem to be the most important:

- **Throughput**: Any system's throughput is the number of requests that it can service or execute in a given unit of time. As a simple example, if system A serves x requests in 1 second while system B serves y requests (where y>x), B demonstrates a higher throughput than A. Throughput value is an indicator of the overall system performance. High throughput indicates a good or robust system performance.
- **Response Time**: It typically refers to the time taken to complete a request. It comprises of transmission time, waiting time, and service time. Response time should be minimized for better performance of the system.
- **Resource Utilization**: It determines how efficiently the various resources in the system are being utilized. This indicates that there are no unnecessary wastages occurring anywhere. For efficient load balancing, resource utilization should be maximized.
- **Fault Tolerance and Reliability**: This parameter indicates how well does a system handle unexpected problems and conditions that might lead to disruptions like server hardware or link failure. Fault tolerance basically determines how a system that encounters such problems, still continues to reliably provide uninterrupted services. A high level of fault tolerance indicates that the load balancing system is robust and working reliably.
- **Scalability**: The expandability of a system is its scalability. It is the ability of a system to service more requests or handle an increased workload by supporting addition of more servers to the system. A system that can efficiently handle increased workloads is referred to as a scalable system. A higher value of this parameter indicates better system performance.
- **Accuracy**: It determines how perfectly or accurately the tasks are being executed. What it means is that does the actual output match the expected result? A greater accuracy is highly desirable to come up with robust systems.

The main aim of any good Load Balancing algorithm is to maximize the Throughput and minimize the Response Time while ensuring the proper Resource Utilization, Scalability, Accuracy and Reliability of the system. The overall performance of a cloud computing system can be enhanced if the performance metrics or parameters are improved or optimized.

3. PROPOSAL

From the study of above-mentioned load balancing algorithms and performance metrics, it is observed that all of these algorithms assume that all incoming tasks are of equal priority, while in real applications, there could be different level of prioritization required for different task types. For instance, banking related tasks could be high priority while simple browsing requests can be low priority. Hence, **Task priority** could be an important aspect of load balancing to meet the service level agreements and quality of service [20] [21]. It is proposed that combination of the concept of task priority with any of the above discussed algorithms will further enhance the effectiveness of the load balancing and scheduling of tasks. All incoming tasks can be assigned priorities like high, medium and low based on defined criteria. A new algorithm can be a combination of an existing algorithm, say Weighted Round Robin and the priority of the task to ensure that high priority tasks can be serviced first over low priority tasks. This will help with implementation of Quality of Service (QoS) and customer Service Level Agreements (SLAs) in cloud computing environment.

4. CONCLUSION & FUTURE WORK

The purpose of this paper was to critically review commonly used load balancing algorithms and the performance metrics in Cloud Computing environment. As Load Balancing is essential for optimal performance of the Cloud Computing system, an understanding of the performance metrics and the optimization of these metrics helps to ensure that the system performs to the best of its capability. It is proposed that in future we can visualize algorithms that combine the existing load balancing algorithms like Weighted Round Robin with task priority to come up with an efficient load balancing strategy. Further work needs to be done using simulation environments to test this proposal for efficiency and efficacy.

ACKNOWLEDGEMENT

I would like to express my gratitude towards Dr. Amit Kumar Dhingra, HOD Computer Science Department, Lotus Valley International School Gurugram, for his expert guidance, constructive suggestions and timely feedback during the course of the paper.

REFERENCES

[1] Trevor D Croker, Formation of the Cloud: History, Metaphor, and Materiality https://vtechworks.lib.vt.edu/handle/10919/96439

[2] Cloud Deployment Model https://www.geeksforgeeks.org/cloud-deployment-model/

[3] Types of Cloud Computing https://aws.amazon.com/types-of-cloud-computing/

[4] What is Load Balancing? Different types of algorithms in cloud computing https://www.xcellhost.cloud/blog/load-balancing-different-types-load-balancing-algorithms-cloud-computing

[5] Afzal, S., Kavitha, G. Load balancing in cloud computing – A hierarchical taxonomical classification. J Cloud Comp 8, 22 (2019). https://doi.org/10.1186/s13677-019-0146-7

[6] Load Balancing Algorithms and Techniques, https://kemptechnologies.com/load-balancer/load-balancing-algorithms-techniques/

[7] Mohammadreza Mesbahi, Amir Masoud Rahmani, Load Balancing in Cloud Computing: A State of the Art Survey, I.J. Modern Education and Computer Science, 2016, 3, 64-78 Published Online March 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2016.03.08

[8] https://www.cloudflare.com/en-in/learning/performance/types-of-load-balancing-algorithms/

[9] Vignesh Joshi Load Balancing Algorithms in Cloud Computing. International Journal of Research in Engineering and Innovation, 2019, 3, pp.530 - 532. ffhal-02884073f

[10] https://www.xcellhost.cloud/blog/load-balancing-different-types-load-balancing-algorithms-cloud-computing

[11] Sajjan, Rajani. (2017). Load Balancing and its Algorithms in Cloud Computing: A Survey.

[12] Bhawesh Kumawat, Rekha Kumawat, A Comparative Study of Load Balancing Algorithms in Cloud Computing Environment using Cloud Analyst, IJEC, Volume 7 Issue No 3

[13] Meenakshi, 2013, Comparative Study of Load Balancing Algorithms in Cloud Computing Environment, International Journal of Engineering Research and Technology (IJERT) Volume 02, Issue 10 (October 2013),

[14] Amit Kumar Dhingra, Dr Dinesh Rai, Study on the Performance Analysis Tools and Techniques of Cloud Computing, Journal of Critical Reviews, Volume 7, Issue 15, 2020

[15] Sushil Kumar, Deepak Singh Rana, Various Dynamic Load Balancing Algorithms in Cloud Environment: A Survey, International Journal of Computer Applications (0975 – 8887) Volume 129 – No.6, November2015

[16] Tushar Desai, Jignesh Prajapati, A Survey Of Various Load Balancing Techniques And Challenges In Cloud Computing, International Journal of Engineering Research & Technology Volume 2, Issue 11, November 2013

[17] Sambit Kumar Mitra, Bibhudatta Sahoo, Priti Paramita Parida, Load balancing in cloud computing: A big picture, Journal of King Saud University - Computer and Information Sciences, Volume 32, Issue 2, February 2020

[18] Wani, Vipin. (2018). A Comparative Study of Various Load Balancing Strategies for Performance Analysis in Distributed System https://www.researchgate.net/publication/326223922_A_Comparative_Study_of_Various_Load_Balancing_Strategies_for_Performance_Analysis_in_Distributed_System

[19] Sambit Kumar Mishra, Bibhudatta Sahoo, Priti Paramita Parida, Load balancing in cloud computing: A big picture, Journal of King Saud University - Computer and Information Sciences, Volume 32, Issue 2, 2020, Pages 149-158, ISSN 1319-1578,

[20] Sadiskumar Vivekanandhan, Priority weighted round robin algorithm for load balancing in the cloud, https://tamucc-ir.tdl.org/handle/1969.6/24399

[21] Kobra Etminani and M. Naghibzadeh, "A Min-Min Max-Min selective algorihtm for grid task scheduling," 2007 3rd IEEE/IFIP International Conference in Central Asia on Internet, 2007, pp. 1-7, doi: 10.1109/CANET.2007.4401694.