A Study on Load Balancing Techniques in Cloud Computing Environment

Hirenkumar Parmar[1] Akshar Patel[2]
[1,2]Department of Computer Engineering
[1,2]Devang Patel Institute of Advance Technology and Research (DEPSTAR)
[1,2]Faculty of Technology and Engineering (FTE)
[1,2]Charotar University of Science and Technology (CHARUSAT), Changa, Anand, Gujarat, India

Abstract— Cloud computing is turning into the most developed and well-known innovation giving the world improved strategies for storage, the viability of the common resources like storage, computation power, dynamic allotment of resources based on the customer's interest. Using the cloud service shares information and give different preferences for clients. With Tremendous increment in the clients and their interest in various administrations on the cloud computing stage, productive or effective use of assets in the cloud climate turned into a basic concern. This developing innovation has a bunch of difficulties to share the resources, increment the accessibility and keep up the load between the resources. Cloud load balancing is one of the principal challenges which will disperse the dynamic workloads at hand and processing resources in the cloud environment between the resources uniformly. Load balancing is holding an essential function in keeping up the cadence of Cloud computing. The evaluation metrics of load balancing algorithms in the cloud are response time and waiting time to the client's request. In this paper, we review variegated existing load balancing algorithms.

Keywords: Cloud Computing, Load Balancing, Load Balancing Parameters, Static Load Balancing Algorithms, Dynamic Load Balancing Algorithms, Types of Load Balancing Algorithms

I. INTRODUCTION

Cloud computing developed as top choices recently. Because of the service's part, this gives flexibility in recovering information and simple route for saving records for making enormous documents and sets of information available for various customers to in the general world. Overseeing such sorts of tremendous arrangements of information call for some methodologies for upgrading and improving activities just as to give amazing efficiency levels to customers. Load Balancing is a procedure which distributes workload among a few nodes inside introduced workspace so this guarantees no nodes inside framework are inert or over-burden for each second. Effective algorithms of load balancing may explain every single node inside a framework may have less or more indistinguishable amount of work. In Cloud Computing the fundamental concerns include effectively relegating assignments to the Cloud nodes with the end goal that the exertion and solicitation handling is done as productively as could reasonably be expected, while having the option to endure the different influencing limitations, for example, heterogeneity and high correspondence delays.

Liability of load balancing algorithms is observing the tasks that are in front of cloud territory of unused services. Subsequently, in general, availability time for responses could be upgraded. Moreover, this gives efficient utilization of assets. Adjusting workloads proceed as one of the stresses inside cloud computing as the amount of the requests couldn't be sorted out which are delivered in a cloud environment. Load Balancing algorithms are divided as static and dynamic. Static algorithms are generally appropriate for homogeneous and stable conditions and can create excellent outcomes in these conditions. Notwithstanding, they are normally not adaptable and can't coordinate the dynamic changes to the traits during the execution time. Dynamic algorithms are more adaptable and take over various kinds of properties in the framework both before and during run-time. These algorithms can adjust to changes and give better outcomes in heterogeneous and dynamic conditions. Be that as it may, as the appropriation credits become more perplexing and dynamic. Therefore a portion of these algorithms could get wasteful and cause more overhead than would normally be appropriate bringing about a general debasement of the execution of the services.

II. LOAD BALANCING

Load Balancing is a technique that assigns the workloads at hand among different nodes in the given cluster with the end goal that it guarantees no node in the cluster is over stacked or sits inert for any moment. A productive load balancing algorithm will ensure that each node in the cluster accomplishes pretty much same volume of work. The obligation of load balancing algorithm is that to plan the tasks which are gone ahead to the cloud space to the abandoned assets so the general availability reaction time is improved just as it gives proficient resource usage. Adjusting the loads got one of the vital worries in cloud computing since we can't foresee the number of demands and requests that are given at each second in cloud climate. The unusualness is expected to the ever-evolving conduct of the cloud. The primary focal point of load balancing in the cloud area is in distributing the load dynamically among the nodes to fulfil the client prerequisites and to give most extreme asset use by grouping the generally accessible burden to particular nodes.

III. SIGNIFICANCE OF LOAD BALANCING

This question is pointed toward perceiving the significance of Load Balancing in cloud computing. Load Balancing is a basic piece of any cloud environment. It holds an essential place in keeping the straightforward entry for clients, colleagues, and end-clients of your cloud-based applications and services. Load Balancing is extraordinarily favourable for the cloud environment, where gigantic remaining tasks at hand could quickly overpower a solitary server, rising availability of service and reaction times are significant to some business activities or are allowed by SLAs. Without Load Balancing, recently turning virtual servers will be not

able or at all to acknowledge the approaching traffic in an organized manner. Hardly any virtual servers may likewise be left to deal with zero traffic while others may have been overpowered. Load Balancing is likewise ready to distinguish absurd servers and divert traffic to those that are still in working condition.

IV. CHALLENGES OF LOAD BALANCING

Cloud Computing faces numerous difficulties; with Load Balancing as one of the most basic issues requiring explicit consideration. This incorporates issues, for example, (VM) relocation, virtual machine security; client QoS comfort and asset/resources use get equivalent regard for looking for a superior answer for improving cloud asset use. The following is a rundown of some of the LB issues:

1) Distributed Geographical Nodes: Cloud server farms are regularly circulated for processing at unique areas. Powerfully dispersed hubs in these farms are utilized as a centralized network for productive handling of client demands. A few Load Balancing approaches are accessible in the literature with a restricted reach and where conditions, for example, network delay, correspondence delay, the reach inside the dispersed processing hubs, space inside client and assets are not mulled over. Hubs in extremely far off regions are challenging because specific algorithms sometimes fall short for this environment.
2) Single Point of Failure: Explicit Load Balancing algorithms are proposed in literature where decision-making isn't dispersed over different hubs, and Load Balancing choices are made by the centralized hub. On the off chance that the key gadgets glitch this will affect the general processing framework.
3) VM Migration: Virtualization takes into consideration the structure of different virtual machines on one physical unit. Those virtual machines have various settings and are self-governing in design. On the off chance that an actual machine is over-burden, it is proper to move all VMs to a distant area utilizing a Load Balancing strategy to migrate the VM.
4) Nodes Heterogeneity: The creators have proposed homogeneous hubs in the cloud load balancing in the introductory request. Cloud computing consumers need a powerful switch, which needs execution on heterogeneous hubs for a proficient organization and lessens reaction time.
5) Handling Data: Cloud computing tended to the issue of old traditional storage gadgets which requested tremendous asset and gear costs for equipment. The cloud permits buyers to heterogeneously hold the information, with no control issues. Capacity is expanding step by step and requires duplication of stored data for viable accessibility and information coherence.
6) Load Balancing Scalability: Accessibility and on-request versatility cloud administrations permit individuals to get to assets for fast downscaling or scale-up whenever required. A solid burden equilibrium ought to think about quickly changing necessities in computational conditions, memory, gadget geography, and so forth.
7) Complexity of Algorithm: Cloud Computing algorithms should be fast and easy to accomplish. The goal of a robust algorithm is to reduce cloud framework effectiveness and quality.
8) Automated Service Provisioning: The key perspective related to distributed computing is adaptability; assets can be consequently designated or circulated. How at that point do we use or release the cloud's services, just keeping up similar profitability as traditional frameworks and utilizing the best asset.
9) Energy Management: The advantages of energy management, which advocates cloud use, are the economies of scale. Saving power is the main thing that takes into consideration a worldwide economy where limited organizations will help the pool of global capital, as opposed to each offering its private types of assistance.

V. GOALS OF LOAD BALANCING

The goal of load balancing is to enhance the utilization of resources accessible, amplify throughput, limit reaction time, and keep away from over-burden of any single resource. It also needs to take care that systems remain stable, have the ability in changing this according to expansion or adjustment inside arrangement of the system, and promote arrangement of fault tolerance as for endurance, execution under system's halfway disappointment, attain enormous improvement inside the execution of tasks, complete usage of shared resources and increment system’s adaptability for acclimating to adjustments.

VI. PARAMETERS OF LOAD BALANCING

The parameters concerning cloud Load Balancing in a considerably more practical sense won't just improve output processing by Load Balancing cycle yet also make the hypothetical reason for contemplating productive algorithms to help Load Balancing effectiveness on Cloud Computing. Load Balancing alludes to the productive strategies utilized for cloud workload assignment between VMs. Inside a cloud network, the adaptability of the VMs relies upon the level of burden dispersed across existing assets. A nice scheduler considers a solid technique for load control. The performance measurements perceived in the Load Balancing techniques are separated into two significant quantitative and qualitative parameter characterizations. Truth be told, the boundaries can likewise be either dependent or independent. The following parameters are consolidated behind a typical classification called service quality metrics (QoS):

A. Load balancing Performance Parameter with Qualitative Attributes & Dependent Nature.

1) Overhead: The overhead connected with any Load Balancing algorithm underpins the additional expense of coordinating the algorithm.

B. Load balancing Performance Parameters with Qualitative Attributes & Independent Nature.

1) Scalability: Inside the unpredictable progression of traffic, a device can execute customer tasks. Separately, the Load Balancing algorithm should have the option to

build assets in top periods, and downscale in off-top occasions. This demonstrates the endurance rate for a working project, regardless of whether the sum or volume of the work or remaining task at hand is raised. The quantities of a hub in a process do not affect the algorithm's adaptation to fault tolerance power.

2) Fault Tolerance: The ability of a system to work reliably during any time of system failure which eventually brings about improved strength and accessibility. A fault-tolerant Load Balancing algorithm would ensure least network misfortune because of network over-burden or other. This shows the capacity of the algorithm to deal with the fault tolerance and its quality of recuperating from failures.

C. Load balancing Performance Parameters with Quantitative Attributes & Dependent Nature.

The results parameter which can be measured and which depending on certain variables in each form or another is described as follows

1) Throughput: The parameter ascertains the number of tasks executed in a unit of time while doing Load Balancing. This characterizes the level at which processing position is performed utilizing a Load Balancing algorithm. The target of the Load Balancing algorithm is to increase more noteworthy execution. Tasks that have achieved their satisfaction inside a predefined period and greatest no. of the dead (or served) work every unit of time.
2) Migration Time: The time of migration is the time expected to move tasks through imbalanced devices. This could likewise be the time needed to move the over-burden VMs through one Physical Machine (PM) to the following physical machine, as in the virtual machine move Load Balancing.
3) Resource Utilization Factor: It mirrors a part of the services open for the complete assets available. This decides how much a VM utilizes the instruments. On the off chance that a VM gets overpowered, the tasks devour a significant part of the energy, however, this is an undesirable marvel because the errands are impossible rapidly. More prominent asset use implies more prominent asset utilization which recommends exhausted assets which thus implies not many free services. In this manner, a powerful Load Balancing algorithm makes the best the greater part of the assets.
4) Power Saving: The measurement characterizes the degree of intensity and quality that the VM burns-through after the cycle of Load Balancing is completed. A viable algorithm for Load Balancing reduces power and energy utilization in a VM.

D. Load balancing Performance Parameters with Quantitative Attributes & Independent Nature.

1) Response Time: This is the general time frame a device needs to respond to a client demand & is numerically equivalent to the complete of time in service and stack stand by time while staying away from the sending time keeping up the unwavering quality property. It is tallied by deducting an errand's completion time from the beginning time of the task's delivery.
2) Performance: It is the mean duration that a PC needs to react to a customer request and is quantitatively equivalent to the entirety holding up time in activity and stacked in this way getting away from transmission time along these lines holding the effective Load Balancing algorithm for the efficiency parameter increases system consistency. Great accuracy guarantees the superior nature of the service through SLA protection.

VII. TYPES OF LOAD BALANCING ALGORITHMS

Broadly, Load Balancing algorithms can be classified into two types namely (i) Static Load Balancing algorithms and (ii) Dynamic Load Balancing Algorithms. First, we shall discuss static load balancing followed by dynamic load balancing.

A. Static Load Balancing Algorithms

These algorithms dole out the assignments to the nodes dependent on the capacity of the node to deal with new demands. The cycle depends exclusively on earlier information on the node's properties and abilities. These would incorporate the node's handling force, memory and capacity limit, and latest known correspondence execution. Although they may incorporate information on the correspondence earlier execution, static algorithms, by and large, don't consider dynamic changes of these characteristics at run-time. Also, these algorithms can't adjust to load changes during run-time.

B. Dynamic Load Balancing Algorithms

It doesn't consider the system's earlier state and no past comprehension is required. It relies upon machine's current status. Dynamic load balancing algorithms consider the various credits of the node's abilities and traffic handling capacity. The vast majority of these algorithms depend on a blend of information dependent on earlier accumulated data about the nodes in the Cloud and run-time properties gathered as the chosen nodes process the traffic's segments Regular technique is allowed by it for moving from the machines that are vigorously stacked dynamically for getting fast execution. These algorithms assign the tasks and may dynamically reassign them to the nodes dependent on the traits accumulated and determined. Such algorithms require steady checking of the nodes and assignment progress and are typically difficult to execute. Nonetheless, they are more exact and could bring about more effective load balancing.

VIII. LOAD BALANCING ALGORITHMS

A. Min-Min Load Balancing Algorithm

This Algorithm takes up with a task set which is at first not appointed to any of the hubs. At first, the base finish time is determined for all the accessible hubs. When this estimation gets finished, the task having the finish time least is picked and appointed to the particular hub. The execution time of all different tasks which are presently accessible in that machine is refreshed and the task gets disposed of from the accessible task set. The routine is done consistently until all the undertakings have been allotted to the identical machines. The algorithm works better when the circumstance resembles where the little assignments are more noteworthy in several

than the enormous tasks. The algorithm has a burden that it prompts starvation.

Min-Min is a straightforward and quick algorithm fit for giving improved execution. Min-Min plans the ideal errands from the start which brings about best schedules and improve the general makespan. Relegating little assignment initially is its disadvantage. Accordingly, more modest errands will get executed first, while the bigger errands keep on in the holding up or the waiting stage, which will at long last outcomes in helpless machine use. Min-Min displays least finishing time for errands which are unassigned and later apportioning the jobs with least finish time (thus min-min) to a hub that is fit for taking care of it.

B. Round Robin Algorithm

It is an algorithm for static load balancing which uses the design of round-robin to assign occupations. This planning is very proficient and powerful time booking strategy. The algorithm arbitrarily chooses hubs for load balancing. Here, the fundamental job is played by server farms in taking care of the cycle of the load adjusting inside distributed computing. At the point when data centre's regulators get demand from the client, at that point, this passes the request to the algorithm of the round-robin. Inside the algorithm, there is a division of time in little units which is known as time cut. Consequently, the algorithm is uniquely intended for sharing of time.

Initially, every processor which could be run, are put away inside the round queue. In the characterized opening of time, the server is apportioned by the scheduler to each process inside the queue. When there are new processes, this would be included at the queue's end. The first process is chosen by a scheduler from queue haphazardly. As there is the end of time allotment of process, the process is passed on from server and afterwards joined a queue's tail. On the off chance that this process is finished before time slot, the process is deliberately released by it. The server is relegated by a scheduler to prepare the process inside a queue. In such a manner, there is the handling of client's demand roundly by utilizing the algorithm. Nonetheless, because of the server's irregular choice a few times, scarcely any servers could be over-burden that outcomes in a decrement of load balancing's exhibition. For conquering this issue, a better strategy of designation is presented and is called weight round-robin load-balancing algorithm.

C. Max-Min Load Balancing Algorithm

The max-min algorithm is a lot of equivalent to a min-min algorithm. From the start for all the accessible assignments are submitted to the system and least finish time for every one of them is determined, at that point among these jobs the one which is having the finish time, most extreme is picked and that is dispensed to the corresponding machine. This algorithm outflanks than Min-Min algorithm where when short jobs are in high numbers when contrasted with that of long ones. For example, if in an assignment set just a solitary long job is introduced, at that point, Max-Min calculation runs short jobs simultaneously alongside a long job. The makespan centre around how much small jobs will get executed simultaneously with the enormous ones. Max-Min is practically indistinguishable from Min-Min, aside from it chooses the job having the greatest finish time and apportions to the corresponding machine. The algorithm experiences starvation where the jobs having the greatest finishing time will get executed first while abandoning the jobs having the minimum finish time.

D. Active Monitoring Load Balancing Algorithm

It is an algorithm of dynamic load adjusting where the burden is assigned to the virtual machine through discovering least loaded virtual machine or inert virtual machine in the list. At first, there is a look for an empty virtual machine if no empty virtual machine is there. Further, the most un-burdened virtual machine is picked. Here record table for each request and servers which are appointed to servers right now is kept up with the assistance of load balancer. When there is a new request, the servers' record table is filtered by the data centre which is least stacked or inactive. The algorithm utilizes the idea of the early bird gets the worm to allot load to a server having least index number for more than two servers.

E. Throttled Load Balancing Algorithm

The algorithm is about the virtual machine. Throttled Load Balancer (TLB) keeps up each process just as screens deal with servers. Henceforth, in the algorithm, the best virtual machine is found by load balancer for customer demand which could handle the load in a viable manner and without any problem. Diverse virtual machines have various properties and limit for taking care of various burdens. Henceforth, according to load, the right virtual machine ought to be chosen for a load. There is the support of the index table for each server and when the data centre receives a request by the customer, data centre's regulator forward requests to a throttled load balancer. To discover inert worker that is accessible, the index table is examined by TLB and send back server id to the data centre and the job is allotted to the servers. Index table after allotment is refreshed. At whatever point regulator of data centre gets data of job finishing there is decrement again in the index table. In the algorithm, if no worker is there in the inactive state, demand stays in the queue.

F. Active Clustering Algorithm

The algorithm characterizes the virtual machine's grouping to adjust load inside cloud computing. For the algorithm, the grouping is gathering of the articles together that have comparable sort of properties. Consequently, virtual machines having identical properties are together assembled in the group for giving sort of load.

G. Equally Spread Current Execution Algorithm

It is an algorithm of dynamic load balancing where exertion is made by load balancer for appropriating the equivalent amount of burden between each server which is accessible in a data centre. The processes are allocated priority at beginning of the algorithm, it at that point checks limit and size for moving burden to the server that could deal with the load in a more modest period. At such point, there is a measure of a limit of virtual machine and assessment of burden. The load is apportioned according to limit and measure of the matching virtual machine.

IX. CONCLUSION

Cloud computing is a rising pattern inside IT's time having gigantic prerequisites for frameworks, stockpiling and assets. Load Balancing is cloud computing's basic viewpoint for adjusting load in the system. Various clients are permitted in getting to appropriated, equipment, programming, virtualized and versatile assets over the web by cloud computing. Load Balancing is a significant issue for cloud computing. This is an approach that disseminates the remaining burden over each hub inside the vast cloud. This would improve asset utilization and generally the execution speed of the system. This paper examines about load balancing and its objectives, requests, types also, needs. This paper likewise examines algorithms of load balancing inside distributed/cloud computing. These algorithms of load balancing guarantee resource's usage through dispersing load between a few hubs inside the system by utilization of errand planning efficiently.

REFERENCES

[1] N. Ajith Singh, M. Hemalatha, "An approach on semi-distributed load balancing algorithm for cloud computing systems" International Journal of Computer Applications Vol-56 No.12 2012
[2] Sujit Tilak, and Patil.D, "A Survey of Various Scheduling Algorithms in Cloud Environment", International Journal of Engineering Inventions,September 2012,pp.36-39.
[3] A. Sidhu, S. Kinger, "Analysis of load balancing techniques in cloud computing", INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY 4 (2) (2013) pages737–741.
[4] Arun Pratap Singh, Prof. Pritesh Jain, Upendra Singh, "Survey of Load Balancing Algorithms in Cloud Computing", International Journal of Scientific Development and Research (IJSDR), Volume 2, Issue 9, September 2017, pp. 209-215.
[5] Randles, M., D. Lamb and A. Taleb-Bendiab, "A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing," in Proc. IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA), Perth, Australia, April 2010.
[6] Foster, I., Y. Zhao, I. Raicu and S. Lu, "Cloud Computing and Grid Computing 360-degree compared," in proc. Grid Computing Environments Workshop, pp: 99-106, 2008.
[7] Radojevic, B. and M. Zagar, "Analysis of issues with load balancing algorithms in hosted (cloud) environments." In proc.34th International Convention on MIPRO, IEEE, 2011.
[8] A. Garg, K. Patidar, G. K. Saxena, and M. Jain, ``A literature review of various load balancing techniques in cloud computing environment," Int. J. Enhanced Res. Manag. Comput. Appl., vol. 5, no. 2, p. 11_14, 2006.
[9] M. Mesbahi and A. M. Rahmani, ``Load balancing in cloud computing: A state of the art survey," Int. J. Modern Edu. Comput. Sci., vol. 8, no. 3, p. 64, 2016.
[10] V. R. Kanakala, V. K. Reddy, and K. Karthik, ``Performance analysis of load balancing techniques in cloud computing environment," in Proc. IEEE Int. Conf. Electr., Comput. Commun. Technol. (ICECCT), Mar. 2015, pp. 1_6.
[11] Survey on Fault Tolerant_Load Balancing Algorithms in Cloud Computing_IEEE Conference Publication. Accessed: May 10, 2020.
[12] M. Alam and Z. A. Khan, ``Issues and challenges of load balancing algorithm in cloud computing environment," Indian J. Sci. Technol, vol. 10, no. 25, pp. 1_12, 2017.
[13] Tejinder Sharma, Vijay Kumar Banga," Efficient and Enhanced Algorithm in Cloud Computing", International Journal of Soft Computing and Engineering (IJSCE)ISSN: 2231-2307, Volume-3, Issue-1, March 2013.
[1] Shanti Swaroop moharana, Rajadeepan d. Ramesh & Digamber Powar," Analysis of load balancers in cloud computing" International Journal of Computer Science and Engineering (IJCSE)ISSN 2278-9960 Vol. 2, Issue 2, May 2013, 101-108.

