LOAD BALANCING IN CLOUD COMPUTING: A REVIEW STUDY

[1]Manpreet Kaur, [2]Usvir Kaur
[1]Student, [2]Assistant Professor
[1]Department of Computer Science and Engineering
[1]Sri Guru Granth Sahib World University, Fatehgarh Sahib, India

Abstract : Information technology has a developing region known as cloud computing that comes into consideration of researchers because of its wide zone of applications. Load balancing challenges confine cloud computing being received effectively in every research field. It is a system which is required to convey the dynamic workload over numerous nodes to guarantee that no single node is over-burden. Appropriate load balancing techniques results in optimal utilization of resources and enhance the performance of the system. The objective of load balancing is to limit the resource consumption which will additionally reduce energy consumption and carbon discharge rate that is the critical need of distributed computing. There is a need for deployment of new metrics in cloud computing and it requires efficient load balancing algorithm to boost up its performance and for optimal utilization of resources. The survey mostly centered on the idea of load balancing techniques in cloud computing, the existing load balancing methods, and demonstrates future directions for the researchers.

IndexTerms - **Load balancing, Techniques, Cloud computing, Overload, Challenges.**

I. INTRODUCTION

Cloud computing is very well known as a internet-based service which provides on-demand access to shared pool of resources to their clients. Clients just need to center around their essential objectives instead of stressing over their processing framework needs. The computing infrastructure needs of the buyers are satisfied by the cloud service providers as per "pay-as-you-use" rule [1]. One of the significant issues of cloud computing is load balancing. It is a mechanism which distributes the dynamic neighborhood workload uniformly over every one of the hubs in the entire cloud. This will keep away from the circumstance where a few hubs are vigorously loaded while others are sit idle or doing little work. It accomplishes a high client fulfillment and resource usage proportion. Henceforth, this will enhance the general execution and resource utility of the framework. It likewise guarantees that each registering resource is dispersed productively and reasonably. It additionally avoids bottlenecks of the framework which may happen because of load irregularity [3].

When at least one segments of any administration fail, load balancing helps in continuation of the administration by actualizing reasonable over, i.e. in provisioning and de-provisioning of occasions of utilizations without come up short. It additionally guarantees that each computing resource is appropriated productively and reasonably. Utilization of resources and protection of energy are not generally a prime focal point of exchange in distributed computing. Nonetheless, resource utilization can be kept to a base with legitimate load adjusting which helps in decreasing expenses as well as making endeavors greener. Versatility which is one of the imperative highlights of distributed computing is additionally empowered by load balancing. Consequently, enhancing resource utility and the execution of a circulated framework in such a way will reduce the energy utilization and carbon impressions to accomplish Green computing [1].

A. *Load Balancing*

Assigning jobs or shifting workload among the nodes is a basic issue in distributed computing. Sometimes, it tends framework to most noticeably awful circumstances if not dealt with precisely. Load balancing is the way toward enhancing the execution of the framework by moving of workload among the multiple processors. Workload of a machine means the total processing time it requires to execute all the tasks assigned to the machine. Resource allotment and adjusting of workload in the datacenter are vital parts of productive resource usage. Poor resource allocation and adjusting of workload in the datacenter prompts imbalanced resource usage and therefore a portion of the resources may get over-burden though some others remain under-loaded. Resource over-burdening corrupts the administration execution though resource under-loading brings about the wastage of resources [3].

In this way, over-burdening/under-loading of resources in the datacenter should be controlled with a specific end goal to accomplish the desired QoS and productive resource use. Load balancing is done as such that each virtual machine in the cloud framework does likewise measure of work all through subsequently expanding the throughput and limiting the response time. Load adjusting is one of the essential components to uplift the working execution of the cloud service provider. Adjusting the load of virtual machines consistently implies that anybody of the accessible machine isn't sit idle or partially loaded while others are vigorously loaded. One of the critical issue of distributed computing is to separate the workload progressively. The advantages of conveying the workload incorporates expanded resource utilization ratio which additionally prompts upgrading the general execution in this way accomplishing greatest customer fulfilment [6].

B. *What is the need of load balancing in cloud computing?*

Load balancing is used as a mechanism in cloud environment to evenly distribute workload across all the nodes. Load balancing recommend higher client fulfillment and boost up resource utilization ratio, making sure that no single node is sit idle or no single node is over-burdened, results in improved performance of the system. It can help in optimally utilize the available resources, and minimize the resource consumption. It additionally helps in executing fail-over, empowering adaptability, maintaining a strategic distance from bottlenecks and over-provisioning, diminishing response time and so on [2].

The factors in charge of it are:

Limited Energy Consumption: By avoiding over heating of nodes or virtual machines with excessive workload, amount of energy consumption can be reduced by load balancing techniques [3][7]

Reduced Carbon Emission: Acting as two sides of same coin energy utilization and carbon emission are directly proportional to each other. Load balancing helps in decreasing energy consumption which will consequently diminish carbon emission and therefore accomplish Green Computing [3].

C. *Goals of Load Balancing*

The goals of load balancing are as followed by:

- To enhance the execution generously.
- To have a reinforcement design in the event that the framework fails even halfway.
- To keep up the framework security.
- To suit future adjustment in the framework.

II. CATEGORIZE LOAD BALANCING ALGORITHMS

Load balancing algorithms can be categorized into two ways as follows:

A. *Depending upon the initiator of the algorithm:* In this case, Load balancing algorithms can be categorized into three types because the initiator can be sender, receiver or the combination of both the sender and receiver initiates the load balancing algorithm.

B. *Depending upon the current state of the load balancing algorithm***:** In case of current state, load balancing algorithm can be categorize into two types:

a. *Static algorithms*

In this approach load balancing is accomplished by giving priori data about the framework. The performance of the node is resolved at the initiation of execution. Nodes compute their designated work and present the result to remote node. At that point contingent upon the execution work load is dispersed in begin without thinking about the present load. The objective of static load balancing strategy is to reduce the execution time of a simultaneous program while limiting the correspondence delays. The fundamental downside of this approach is that it doesn't take current condition of the framework while settling on allotment choices. This has the real effect on the general execution of the framework because of load fluctuation in distributed framework [4][7].

b. *Dynamic algorithm*

It varies from static algorithms in that the workload is distributed among the nodes at run-time. The remote node allocates new procedures to the slaves based on the new data gathered. Dynamic algorithms assigned legitimate weights on servers and via looking in entire system a lightest server preferred to adjust the traffic. Be that as it may, choosing a fitting server required ongoing Communication with the systems, which will prompt additional traffic included framework. Dynamic algorithms predicated on inquiry that can be made much of the time on servers, however sometimes traffic load will prevent these inquiries to be replied, and correspondingly more included overhead can be recognized on system [7].

III. EXISTING LOAD BALANCING TECHNIQUES

Depending on their source of inspiration and operational model, load balancing algorithms sort into two general classifications, specifically nature inspired and statistic-based algorithms. As the name specifies, nature inspired techniques are derived from the phenomenon occurring in the nature [8]. Through description of these categories and techniques under both the categories are given in this section.

A. Nature-inspired techniques

In nature, different exercises and procedures, for example, swarm foraging, genetic hybrid, and advantageous interaction are going ahead in a very advanced way. The exercises, for example, swarm foraging, rushing of flying creatures, and tutoring by angles utilize group and agreeable way to deal with achieve the desired targets. While, different organic procedures e.g., genetic crossover and change, normally utilize transformative ways to deal with create the ideal arrangements. In the present time, different researches are motivated from smart exercises and procedures happening in nature. Nature-propelled approaches have demonstrated better execution for complex and expansive sample space issues. Various nature-motivated meta-heuristic enhancement algorithms have been produced and sent in the cloud condition to adjust the workload crosswise over data center computing resources. Nature-inspired load balancing techniques reviewed in this paper are characterized into two primary classes: Swarm behavior-based and Evolution-based [12][13].

a. Swarm behavior-based algorithm

Swarm behavior based algorithms are planned by numerically displaying the exercises of various species that live in settlements and work in gatherings to pursuit and assemble sustenance e.g., ants, bees, and feathered creatures. Advancement calculations like Ant Colony Optimization, Artificial Bee Colony streamlining, and Particle Swarm Optimization are created by numerically demonstrating the group and agreeable conduct of ants, bumble bees, and feathered creatures, separately. Swarm behavior-based calculations have indicated generally better effectiveness for complex and huge example space issues [18]

A meta-heuristic technique, ant colony optimization is widely used to locate the ideal solutions. Ants like to take after the ways with higher pheromone considerations. This conduct draws in more number of ants to take after this short way with higher pheromone thickness. Bit by bit, pheromone thickness of this way turns out to be generally substantially higher when contrasted with different ways. Henceforth, all ants will begin following this way to gather food. However, every ant is fit for building a total arrangement, potentially an alternate one, yet a great arrangement will develop just from the global collaboration among the member from the colony. Similarly, particle Swarm Optimization (PSO) reproduces the social conduct of life forms, for example, rushing in winged animals and tutoring in fishes. On the other hand, artificial bee colony optimization is inspired from the intelligent and co-operative foraging behavior of honey bee swarm [12][18].

b. Evolution-based algorithms

Evolution-based algorithms are ordinarily utilized for unpredictable and substantial example space issues whose example space isn't obviously characterized.

In Evolution-based algorithms the underlying arrangement of applicant arrangements is advanced to produce an ideal solution. Evolution-based algorithms for instance genetic algorithms are fundamentally composed by numerically demonstrating the key systems (e.g., genetic crossover and mutation) in charge of biological advancements. It works in an iterative design to produce ideal answer for a given issue in understanding to the goal work [4].

Genetic Algorithm works better for huge and complex pursuit space issues. In load balancing, GA decides the most reasonable arrangement of machines for the organization of VMs/Tasks to accomplish the desired results. For instance, GA can be utilized to locate the most suitable processors to execute the given arrangement of assignments for enhancing energy preservation, resource use, response time, throughput, and so on [2][4].

B. Statistic-based algorithms

In statistics-based load balancing algorithms, processing condition is always observed to assemble runtime information with respect to different key execution markers e.g., resource use, energy utilization, response time, and throughput. This information is then investigated and deciphered into data to help choices relating to load balancing. Powerful and productive information investigation approaches are utilized as a part of the refined load balancing algorithms to better anticipate and gauge diverse parameters of premium. This precision in estimation helps diverse algorithms to perform successfully and productively [12].

a. Resource-aware algorithm

These kinds of algorithms play out their tasks by observing and analyzing resource parameters such as % of resource utilization, energy consumption rate for the prediction of the availability and resource requirements. Bin-packing, agent-based, and dynamic clustering are well known techniques for load balancing. In bin-packing, tasks of same volumes are packed into a bin in a way that limits the quantity of bins used. Every computer device has different capacities. These capacities are required to be used in an efficient manner to effectively utilize the system. Tasks are assigned to resources until the point that their accessible resource limits wind up deficient to execute any given undertaking or until the point when framework execution fall underneath a built up edge, keeping in mind the end goal to limit energy utilization and maximize resource use [13].

b. Performance-based algorithms

To enhance the system performance in different circumstances, performance-based algorithms are utilized. These algorithms analyze different key execution markers and investigate them to come up with a decision to the defined strategies. Adaptive and QoS-based algorithms are summed up types of performance-based algorithms. Adaptive algorithms are particularly intended for the dynamic computing environments, to be adaptive to various changes in the environment conditions. It works productively in processing situations whose framework/workload behavior changes always. Adaptive algorithms find out about the present framework state from the current perceptions and roll out essential improvements in their mechanisms appropriately, to enhance execution of the framework. Adaptive algorithms are particularly desirable to balance load in distributed computing situations, particularly when execution should be upgraded. On the other side, QoS-based algorithms intend to aim QoS by effectively using the datacenter resources. Methodologies like effective usage of the idle slots of the provisioned resources, ideal planning and resource provisioning accomplishes desired QoS [4].

Table 1 Merits and Demerits of load balancing algorithms

Load BalancingTechniques	Pros	Cons
Static load balancing	Does not consider current state Less complex [2]	Do not have the ability to handle more load variations [2]
Round Robin	Fixed time slice Better performance for short CPU bursts [1]	Larger tasks takes more time for completion [1]
Max-Min	It works better as the requirements are known in prior [4]	Takes long time for task completion [4]
Min-Min	Smallest completion time Gives best results for small tasks [4]	Starvation [4]
Opportunistic Load Balancing	Improved performance Resource utilization [13]	Takes more time for task completion [13]
Dynamic load balancing	It requires current state of the system Fault tolerance [12]	More Complex Requires constant check of the nodes [12]
Ant Colony Optimization	Computationally fast Minimizes makespan [10]	Search takes long time Complex [10]
Honey Bee Foraging	Reduced response time Increases throughput [18]	Low priority load takes more time [18]
Biased Random Sampling	Improved performance Improved resource utilization [19]	Response time is more [19]
Resource Allocation Scheduling	Increases performance Less execution time [2][4]	Less fault tolerance [2][4]

IV. LOAD BALANCING CHALLENGES

Because of fast increment in the interest for cloud services, proficient usage of energy and processing resources has turned into a matter of great concern for researchers. In spite of the fact that cloud computing has been broadly embraced. Research in cloud computing is still in its beginning periods, and some scientific difficulties stay unsolved by mainstream researchers, especially load balancing challenges [2]. Some of the challenges are as follows:

a. *The Automated service provisioning*

Elasticity is a noteworthy segment in cloud computing because of which allocating and discharging of the resources occurs as default. The challenge with utilizing ideal resource is the means by which cloud flexibility can be utilized and how function with customary frameworks performance should be possible all the while?[1][7]

b. Energy management

Energy Management is additionally a noteworthy feature that grants clients utilize the resources from worldwide center. It is very essential to efficiently manage resource energy by effective load balancing techniques [1][2].

c. Migration of virtual machines

The thought is to envision a machine as a set of records or a document. It is conceivable to diminish the load on over-burdened machine by moving the virtual machine among them in powerful way. The primary motive is to convey the all kind of load in a data center. The challenge is to eliminate disadvantages of cloud computing framework when the load is progressively distributed by virtual machine [15]19].

d. Data management

Another key necessity in cloud computing is the storage of data. The challenge with data management is the means by what method would data is able to be circulated in the cloud framework with most proper storage and quick access?[9]

e. Point-of-failure

A few algorithms (centralized algorithms) give viable systems to handling load balancing in a specific pattern. However, the issue is that there is just a single controller for the whole framework. In such condition, if the controller fails, at that point the whole framework fails [2].

V. CONCLUSION

Because of fast increment in the interest for cloud services, proficient usage of energy and processing resources has turned into a matter of great concern for researchers. Cloud computing, in which distinctive resources are access by numerous clients over the web in on request premise. These resources are quickly increasing and furthermore expanding employments of heterogeneous framework in dynamic environment. In any case, there are several research challenges in distributed computing. Load balancing is real challenge (issue) in distributed computing. The key points of load balancing is to ful fill client's need by distributing work load among different nodes in framework, and boost resource usage and enhances framework execution. So productive load balancing is essential for system performance, resource usage, stability, amplifies the throughput and limits the response time that are the fundamental goals of the cloud environment. The study shows the outline of cloud computing, distributed computing engineering, load balancing and existing load balancing algorithms, and a few difficulties or challenges identified with balancing load in cloud computing are demonstrated for the future research directions.

REFERENCES

[1] Thakur, A. and Goraya, M.S., 2017. A taxonomic survey on load balancing in cloud. Journal of Network and Computer Applications.

[2] Sreenivas, V., Prathap, M. and Kemal, M., 2014, February. Load balancing techniques: Major challenge in Cloud Computing-a systematic review. In 2014 International Conference on Electronics and Communication Systems (ICECS),1-6.

[3] Joshi, S. and Kumari, U., 2016, December. Load balancing in cloud computing: Challenges & issues. In 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), 120-125.

[4] Dillon, T., Wu, C. and Chang, E., 2010, April. Cloud computing: issues and challenges. In 2010 24th IEEE International Conference on Advanced Information Networking and Applications (AINA), 27-33.

[5] Shariati, S.M. and Ahmadzadegan, M.H., 2015, November. Challenges and security issues in cloud computing from two perspectives: Data security and privacy protection. In 2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), 1078-1082.

[6] Thakkar, U. and Rajput, I., 2016, December. A novel approach for dynamic selection of load balancing algorithms in cloud computing. In International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), 1-4.

[7] Khara, S. and Thakkar, U., 2017, July. A novel approach for enhancing selection of Load Balancing algorithms dynamically in cloud computing. In 2017 International Conference on Computer, Communications and Electronics (Comptelix), 44-48.

[8] Ahmad, M.O. and Khan, R.Z., 2018. Load Balancing Tools and Techniques in Cloud Computing: A Systematic Review. In Advances in Computer and Computational Sciences, 181-195.

[9] Gutierrez-Garcia, J.O. and Ramirez-Nafarrate, A., 2015. Collaborative agents for distributed load management in cloud data centers using live migration of virtual machines. IEEE transactions on services computing, 8(6) :916-929.

[10] Zhang, Z. and Zhang, X., 2010, May. A load balancing mechanism based on ant colony and complex network theory in open cloud computing federation. In 2nd International Conference on Industrial Mechatronics and Automation (ICIMA), 2:240-243.

[**11]** Mishra, S.K., Khan, M.A., Sahoo, B., Puthal, D., Obaidat, M.S. and Hsiao, K.F., 2017, July. Time efficient dynamic threshold-based load balancing technique for Cloud Computing. In 2017 International Conference on Computer, Information and Telecommunication Systems (CITS), 161-165.

[12] Dave, A., Patel, B. and Bhatt, G., 2016, October. Load balancing in cloud computing using optimization techniques: A study. In International Conference on Communication and Electronics Systems (ICCES), 1-6.

[13] Al Nuaimi, K., Mohamed, N., Al Nuaimi, M. and Al-Jaroodi, J., 2012, December. A survey of load balancing in cloud computing: Challenges and algorithms. In2012 Second Symposium on Network Cloud Computing and Applications (NCCA), 137-142.

[14] Nadaph, A. and Maral, V., 2015, February. Methodical analysis of various balancer conditions on public cloud division. In 2015 International Conference on Computing Communication Control and Automation (ICCUBEA), 40-46.

[15] Pavithra, B. and Ranjana, R., 2016, March. A comparative study on performance of energy efficient load balancing techniques in cloud. In IEEE International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 1192-1196.

[16] Kapoor, S. and Dabas, C., 2015, August. Cluster based load balancing in cloud computing. In IEEE 2015 Eighth International Conference on Contemporary Computing (IC3), 76-81.

[17] Razali, R.A.M., Ab Rahman, R., Zaini, N. and Samad, M., 2014, June. Virtual machine migration implementation in load balancing for Cloud computing. In IEEE 2014 5th International Conference on Intelligent and Advanced Systems (ICIAS), 1-4.

[18] Zhang, Q., Cheng, L. and Boutaba, R., 2010. Cloud computing: state-of-the-art and research challenges. Journal of internet services and applications, 1(1): 7-18.

[19] Srivastava, S. and Singh, S., 2018. Performance Optimization in Cloud Computing Through Cloud Partitioning-Based Load Balancing. In Advances in Computer and Computational Sciences, 301-311.