DOI: http://dx.doi.org/10.26483/ijarcs.v9i2.5820

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

ANALYTICAL REVIEW OF LOAD BALANCING TECHNIQUES IN CLOUD COMPUTING

Abhikriti Narwal and Sunita Dhingra
U.I.E.T Maharshi Dayanand University Rohtak
abhikritiin@gmail.com

Abstract: Cloud computing is another pattern rising concept in IT industry with immense inevilability of framework furthermore, resources. Cloud computing is build up by conglomerating two terms in the area of innovation. Initial is Cloud and second is computing. Cloud comprises of heterogeneous assets. It is a work of immense framework with no significance with name "Cloud". Load Balancing is a vital part of distributed computing scenario. Productive load adjustment plan guarantees effective asset usage by provisioning of assets to cloud client's on-request premise in pay-as-you-use. Load Balancing may indeed, even help organizing clients by applying proper planning criteria. This paper presents a review on load balancing methods in cloud computing with its various techniques. The objective of this paper is to study the various existing load balancing techniques on various parameters utilized to compare the current techniques with each other.

Keywords: cloud computing, load balancing, virtual machine, pay-as-you-use.

I. INTRODUCTION

Cloud computing becomes the main commercial computing paradigm. Technologies of cloud computing are developing at a fast pace and have acquired much popularity in the past few years [1]. Cloud computing is a distributed computing that provides ability to its users to execute an application on various interconnected systems simultaneously. On demand technological resources are accessed through utility service provided by cloud computing. Cloud computing shifts the focus of its users from infrastructure provisioning to their businesses, so indirectly enhanced the production efficiency. It provides dependable and reliable on-demand services and infrastructure that minimizes expenses and time [2].

Cloud computing consists of two terms cloud and computing. Cloud is a collection of various distinct resources. It is a network of very vast communications and is not relevant to its name "Cloud". Infrastructure is applications provided to clients in terms of services on internet and software, hardware in datacenters, who are responsible to provide services. Computing is performed on the basis of particular criteria mentioned in service level agreement (SLA) to effectively use these resources and to ensure that the resources are available to users. Cloud computing provides infrastructure to users based on pay-as-you-use. Cloud computation is performed to attain highest resource utilization with high availability at much less cost [3].

Various issues are considered before shifting services to cloud such as security, availability, reliability, load balancing etc [4]. Cloud computing provides services to users by using virtual machines. Each task submitted by the user in the cloud environment is assigned to a virtual machine. It is important to run applications of users independently from each other. To ensure this, every physical host should load on one or more than one virtual machines. Hence, cloud task scheduling is performed among the virtual machines [5]. Task scheduling is the vital aspect of cloud computing so that cloud performance should get increased [1]. The rest of paper is organized as follows. Section 2 discusses about the literature review of load balancing algorithms. Section 3 presents all the load balancing techniques in the tabular method. Finally Section 4 concludes the paper.

II. LITERATURE REVIEW

Agarwal et al. [2] examines the current load adjusting calculations in a cloud based condition. Distributed computing, a developing processing worldview expects to share information, estimation and administrations straightforwardly finished a versatile system of hubs. In distributed computing, stack adjustment is main problem. Load is a measure of work that a calculation framework performs which can be delegated CPU stack, memory limit and system stack. Load adjustment is the way toward allocating the heap between various hubs of a circulated method to update asset use and occupation response time while maintaining a strategic distance from a circumstance where a part of the hubs are vigorously stacked while others are sit out of gear. Load adjustment guarantees that every hub in the framework does roughly measure up the estimate of work (according to their ability) at any moment of time.

Katyal et al. [3] presents different load adjustment plans in various cloud condition in view of prerequisites determined in Service Level Agreement (SLA). Distributed computing is pattern rising concept in IT condition with immense requirements of foundation and resources. Load Balancing is an vital portion of distributed computing condition. Proficient load adjustment plan guarantees effective asset usage by provisioning of assets to cloud client's on-request premise in pay-as-you-use. Load Balancing help organizing clients by relating suitable planning criterion.

Randles et al. [6] explored three conceivable circulated arrangements proposed for stack adjustment; approaches propelled by Honeybee Foraging Behavior, Biased Random

Sampling and Active Clustering. The expected take-up of Cloud processing, on basis of established innovation in Web Administrations, frameworks, utility enrolling, passed on figuring and virtualisation, will procure various purposes of intrigue cost, versatility and openness for advantage customers. These focal points are depended upon to furthermore drive the enthusiasm for Cloud organizations, growing both the Cloud's customer base and the span of Cloud foundations. This has recommendations for some specific issues in Service Oriented Architectures and Internet of Services (IoS)- type applications; including adjustment to non-basic disappointment, high availability and adaptability. Essential to these issues is the establishment of suitable load modifying frameworks. Unmistakably scale and eccentrics of these systems makes fused assignment of occupations to specific servers infeasible; requiring a capable flowed course of action.

Pandey et al. [7] display a Particle Swarm Optimisation (PSO) based heuristic arrangement applications to cloud resources that considers both count cost and data transmission cost and investigate distinctive roads with respect to a work procedure application by fluctuating its figuring and correspondence costs. Examination is done on the cost venture stores while using PSO and existing 'Best Resource Selection' (BRS) figuring.

Ghanbari et al. [8] proposed another need based employment planning calculation (PJSC) in distributed computing. These days distributed computing has turned into a well known stage for logical applications. Distributed computing plans to share a substantial scale assets and types of gear of calculation, stockpiling, data and information for logical looks into. Occupation booking calculations is a standout amongst the most difficult hypothetical issues in the distributed computing territory. Some concentrated looks into have been done in the region of occupation planning of distributed computing. The proposed calculation depends on different criteria based on leadership model.

Parsa et al. [9] planned novel task scheduling technique called RASA, considering the appropriation and adaptability attributes of network assets. The calculation is worked through an extensive report and investigation of two surely understood errand booking calculations, Min-min and Max-min. RASA utilizes the upsides of the two calculations and spreads their impediments. To accomplish this, RASA right off the bat evaluates the fulfillment time of the undertakings on every one of the accessible framework assets and afterward applies the Max-min and Min-min calculations, on the other hand.

Sahu et al. [10] given a diagram of cloud innovation where their parts are additionally concentrating on stack adjustment of distributed computing with a portion of the current load adjusting strategies, which are dependable to deal with the heap when some node of the cloud system is over-weight and others are under stacked. While registering the resources, the heap might be of different kinds like memory stack, CPU load and system stack and so forth. Load adjusting is the way toward looking over-burden hub and exchanging the additional heap of the over-burden hub to different hubs which are under stacked, for enhancing asset use and diminishing server reaction time of the occupations.

Babu et al. [12] planned a bee colony based algorithm to effectively balance the load on the basis of foraging behavior of honey bees for balancing load among VMs. In this technique, tasks from over loaded virtual machines are removed and treated as honey bees and under loaded virtual machines are treated as food sources. This method considered task properties in virtual machines waiting queue and tried to attain reduced response time and minimum tasks migration. Significant enhancement is shown in the QoS.

Kaur et al. [13] proposed novel task scheduling technique based on honey bee algorithm to enhance throughput and optimize execution time of virtual machines of assigned tasks to properly utilize resources in minimum cost. These technique balances load among virtual machines in manner such that total waiting time of tasks is reduced.

Gupta et al. [14] planned a technique called as HBB-LB that targets to attain a balanced load among VMs. Results demonstrate that this technique is more efficient than the old techniques. Waiting time and execution time is reduced. This paper briefly described about current load balancing techniques.

III. LOAD BALANCING IN CLOUD COMPUTING

Load balancing is the main issue of cloud computing, so that cloud computing should match the increasing requirements of its users. Load balancing is a process to equally divide the load among all the participating nodes. It is important to balance the load among the virtual machines such that no machine should be over-loaded or under-loaded so that resources are utilized effectively in less response time. A proper load balancing technique assist to prevent the bottleneck of the framework caused due to load imbalance[11].Due to lack of proper load balancing technique some virtual machines may remain idle and some may be overloaded [4]. Hence, load balancing is an vital feature of cloud computing. Load balancing could be classified as static and dynamic [2].

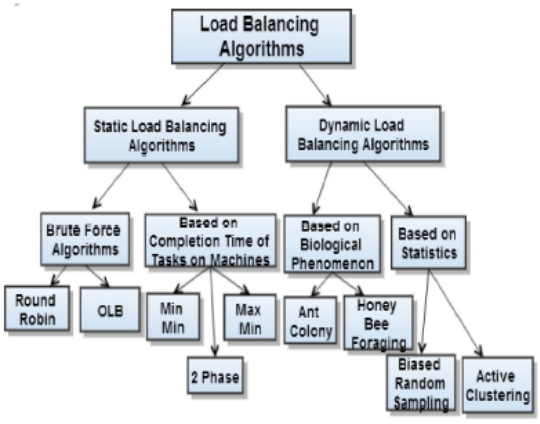

Fig. 1. Load Balancing Algorithms [2]

A. Static Load Balancing

A static load balancing algorithm does not consider the past stage or conduct of a hub while disseminating the load. It can be of various types:

a) Brute Force Load Balancing Algorithms

- *Round Robin Load Balancing Algorithm*
  This algorithm uses the round robin conspire for dispensing employments. It selects the primary hub arbitrarily and then designates occupations to all single other node in a round robin mold. Processors are

appointed to each process in a roundabout request without requirement and thus there is no starvation. By the round robin technique, the movement on the servers will be hated effectively and thusly it will tilt the circumstance towards more imperfectness. Subsequently, weighted round robin was produced to enhance the basic problems of round robin. In weighted round robin calculation every server is allocated a weight and as indicated by the estimations of the weights, the employments are appropriated. Processors with more prominent limits are relegated a bigger esteem. Henceforth the most elevated weighted servers will get more undertakings. In a circumstance, when every one of the weights wind up parallel, servers will do the adjusted movement. In distributed computing framework, exact expectation of execution time isn't conceivable; subsequently static calculation isn't favored. Henceforth, dynamic varieties of the round robin calculation have been proposed.

- *Opportunistic Load Balancing Algorithm (OLB)*
  This method events to keep all nodes busy. In this manner it does not think about the current workload of each PC. OLB dispatches unexecuted errands to the now accessible hubs in the arbitrary request, paying little heed to the hub's present workload. Each activity is relegated to the hub in a subjective request. It gives stack adjustment plan however it brings about exceptionally poor make-traverse. Since OLB does not compute the implementation time of the hub, the undertaking needs to get handled.

b) Static Load Balancing Algorithms on the basis of Completion Time of Tasks on Machines

- *Min Min Load Balancing Algorithm*
  The Min Min method begin with a course of action of each and every unassigned task. As an issue of first significance, minimum satisfaction time for all endeavors is found. By then among these accessible conditions the base regard is picked which is the base time among the each errand shows on any advantage. According to that base time, the task is set up for the looking at machine. The execution time for each and every other endeavor is revived on that machine and the errand is ousted from the once-over. This method is repeated until every errand is allocated to the available assets. In situations where the quantity of little errands is more than the quantity of huge undertakings, this calculation accomplishes better execution. Notwithstanding, this approach has a disadvantage that is, it can prompt starvation. This calculation does not consider low and high machine and assignment heterogeneity.
- *Max Min Load Balancing Algorithm*
  Max Min is similar to the min min algorithm but the going with: resulting to finding minimum execution times, the best regard is picked which is the most extraordinary time among the errands on the benefits. By then according to that most outrageous time, the endeavor is moved toward the relating machine. The execution time for each other endeavor is invigorated on that machine and alloted errand is removed from the once-over of the endeavors that are to be doled out to the machines. Since the necessities are known ahead of time, the estimation is depended upon to perform well.
- *Two Phase Load Balancing Algorithm*
  The two phase scheduling algorithm combines Opportunistic Load Balancing and Min-Min scheduling algorithms to use better executing productivity and keep up the heap adjustment of the framework. Deft planning calculation keeps each hub in working state to accomplish the objective of load adjustment. Min booking calculation is used to limit the execution time of each errand on the hub.In this way it will limits the general finishing time of the errands. This joined approach thus helps in an effective usage of assets and improves the work productivity.

B. Dynamic Load Balancing

A dynamic load balancing algorithm checks the past condition of a hub while conveying the load.

a) Based on Biological Phenomenon

- *Ant Colony Optimization (ACO) Based Load Balancing Algorithm*
  ACO is a probabilistic strategy for dealing with computational issues which can be limited by discovering the beneficial courses through outlines. The purpose of underground bug region change is to filter for a perfect route in a chart on the direct of ants searching for a path between their settlement and wellspring of sustenance. The approach goes for productive circulation of the heap among the hubs and with the end goal that ants never experience a deadlock for developments to hubs to build ideal arrangement set.

Table 1. Load Balancing Algorothms Comparison

Algorithm	Static	Dynamic	Possible Starvation	Optimal Resource Utilization	Handle Fallover	Handle Distinct Jobs
Round Robin	Yes	Variations supportive	No	No	No	No
Opportunistic	Yes	Variations Supportive	No	No	No	No
Min-Min	Yes	No	Yes	Yes	No	No
Max-Min	Yes	No	Yes	Yes	No	No
2 Phase Load Balancing	Yes	No	Yes	Yes	No	No
Ant Colony Optimization	No	Yes	No	Yes	Yes	Yes
Honey Bee Foraging	No	Yes	No	Yes	No	Yes
Biased Random Sampling	No	Yes	Yes	No	Yes	Yes
Active Clustering	No	Yes	Yes	Yes	Yes	Yes

- *Honey Bee Foraging Load Balancing Algorithm*

  This is a people based chase count. A region of bumble bees can extend itself over a detachment and from numerous points of view in the meantime to mishandle endless resources. Sprouts with bounteous nectar or clean grains are moving more as often as possible than the ones with fewer grains. There are scout bumble bees which are searching for sustenance return. They play out a waggle move demonstrating the course in which the nourishment was discovered. The scout honey bees alongside supporters go to those specific patches to accumulate sustenance rapidly and effectively. Thus, the activity planning can be performed – a total booking of undertakings is resolved in the issue. Each game plan can be thought as a path from the hive to the support resource.

b) Based on Statistical Models

- *Biased Random Sampling Load Balancing Algorithm*

  In Biased Random Sampling method everything is represented with the help of virtual diagram. Every server is symbolized as a hub in the diagram, with every level of the hub connected to the free assets of the server. The augmentation and decrement of hub's in-degree is done by means of Biased Random Sampling (BRS). Arbitrary examining can be characterized as the procedure wherein the servers are haphazardly chosen. The inspecting starts at some settled hub, and at each progression, it moves to an adjoining hub, picked haphazardly.
- *Active Clustering Load Balancing Algorithm*

  Active Clustering works with the collection of comparative hubs together and

  taking a shot at these hubs. A node initializes the procedure and selects another node known as go between node and its neighbors. The hub fulfilling the criteria of an unexpected sort in comparison to the previous one. The alleged relational arranger hub at that point shapes an association between one of its neighbors which is of an indistinguishable sort from the underlying hub. The go between hub at that point disengages the association amongst itself and the hidden center point. The above course of action of methods is taken after iteratively. The execution of the structure is updated with high openness of benefits, among these lines and increasing the throughput. This expansion in throughput is because of the effective use of assets.

IV. CONCLUSION

In this paper, a comparative study is done on various techniques of load balancing in cloud computing. Various load balancing techniques are analyzed in this paper. Load Balancing is a basic errand in Cloud Computing condition to accomplish greatest usage of assets. On one hand static load adjusting plan give least demanding reproduction and checking of condition yet neglect to display heterogeneous nature of cloud. Then again, dynamic load adjusting calculation are hard to reproduce in any case, are most appropriate in heterogeneous condition of cloud computing.

REFERENCES

[1] Hongli Zhang, Panpan Li, Zhigang Zhou, Xiangzhan Yu, "A PSO-Based Hierarchical Resource Scheduling Strategy on Cloud Computing", Springer, 2013, pp. 325-332.

[2] Aayush Agarwal, Manisha G, Raje Neha Milind, Shylaja S S, "A Survey of Cloud Based Load Balancing Techniques", 2014, pp. 9-13.

[3] Mayanka Katyal, Atul Mishra, "A Comparative Study of Load Balancing Algorithms in Cloud Computing Environment", International Journal of Distributed and Cloud Computing, Vol. 1, Issue 2, December 2013, pp. 5-14.

[4] Sheeja S Manakattu, Madhu Kumar S D, "An Improved Biased Random Sampling Algorithm for Load Balancing in Cloud Based Systems", International Conference on Advances in Computing, Communications and Informatics, 2012, pp. 459-462.

[5] Zhanghui Liu, Xiaoli Wang, "A PSO-Based Algorithm for Load Balancing in Virtual Machines of Cloud Computing Environment", Springer, 2012, pp. 142-147.

[6] Martin Randles, David Lamb, A. Taleb-Bendiab, "A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing", IEEE, International Conference on Advanced Information Networking and Applications Workshops, 2010, pp. 551-556.

[7] Suraj Pandey, LinlinWu, Siddeswara Mayura Guru, Rajkumar Buyya, "A Particle Swarm Optimization-based Heuristic for Scheduling Workflow Applications in Cloud Computing Environments",IEEE, 24th International Conference on Advanced information networking and applications,2010,pp. 400-407.

[8] Shamsollah Ghanbari, Mohamed Othman, "A Priority based Job Scheduling Algorithm in Cloud Computing", International Conference on Advances Science and Contemporary Engineering, 2012, pp. 778 – 785.

[9] Saeed Parsa, Reza Entezari-Maleki, "RASA: A New Task Scheduling Algorithm in Grid Environment", World Applied Sciences Journal, 2009, pp. 152-160.

[10] Yatendra Sahu, R.K. Pateriya, " Cloud Computing Overview with Load Balancing Techniques", International Journal of Computer Applications, Vol. 65, No.24, March 2013, pp. 40-44.

[11] Nidhi Jain Kansal,Inderveer Chaana, "Existing Load Balancing Techniques in Cloud Computing: A Systematic Review",Journal of Information Systems and Communication,Vol. 3,Issue 1,2012,pp. 87-91.

[12] K R Remesh Babu, Amaya Anna Joy, Philip Samuel, "Load Balancing Of Tasks In Cloud Computing Environment Based On Bee Colony Algorithm", IEEE, International Conference on Advances in Computing and Communications, 2015, pp. 89-93.

[13] Anureet kaur, Bikrampal Kaur, "Load Balancing in tasks using Honey bee Behavior Algorithm in Cloud Computing", IEEE, 2016.

[14] Harshit Gupta, Kalicharan Sahu, "Honey Bee Behavior Based Load Balancing of Tasks in Cloud Computing", International Journal of Science and Research, 2014, pp. 842-846.