Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 03, March -2018

Load Balancing Algorithms in Cloud Computing: Review

Yogesh Banyal[1], Jawahar Thakur[2]

[1]Dept. of Computer Science, Himachal Pradesh University
[2]Dept. of Computer Science, Himachal Pradesh University

ABSTRACT-*Cloud computing is a generic term used for the delivery of hosted services over the Internet. It provides an infrastructure for resource sharing, software hosting and service delivering in a pay as you go model which makes it very easy and economical. It is a challenge in cloud computing to distribute work load among all incoming requests and balance those requests.To tackle this Load balancing technique is used which uses multiple nodes and distributes dynamic workload among the nodes so that no single node is under loaded or over loaded. Load balancing allows the resources to be used aptly which enhance the performance of the system and minimize carbon emission. This paper studies various load balancing algorithms and gives a comparison among these algorithms on the basis of different qualitative metrics like overhead,performance, reliability, scalability, throughput etc.*

Keywords - *Cloud Computing; Load balancing;Virtual machines; Load Balancing Algorithms; Green Computing*

I. INTRODUCTION

Cloud computing is the on-demand delivery of applications, compute power, database storage, and other IT resources through a cloud services platform via the internet with pay-as-you-go pricing [1].It is an emerging field for research and study. It is a pool of multiple configurable computing resources available on demand to user. It has evolved from past technologies like web services, hardware virtualization, grid and utility computing, system management.Cloud computing has many issues such as load balancing, cloud security, management of energy, privacy which hinder its growth. Load balancing issue is handled using algorithms which helps to distribute load between all the nodes. It also ensures that every computing resource is distributed efficiently and fairly. It helps in preventing bottlenecks of the system which may occur due to load imbalance. Load balancing provides better response time and high resource utilization.This paper studies various load balancing algorithms in cloud computing i.e. round robin, throttled, equally spread current execution, active clustering, ant colony ,biased random sampling, honeybee foraging ,max-min, min-min and power aware load balancing. A comparison of these algorithms is done on the basis of different qualitative metrics like overhead, performance, reliability, scalability, throughput etc.

The organization of the paper is as follows: In section II, load balancing is discussed. In section III, review of some related work is done. In section IV, some load balancing algorithms are studied. In section V some metrics for comparing load balancing algorithms are given. In section VI Load Balancing algorithms are compared on the basis of the metrics discussed in previous section. Section VII, concludes the paper.

II. Load Balancing in Cloud Computing

Load balancing is defined as an approach to increase and improve the performance of two or more nodes or links connected nodes by the redistribution or the reassignment of load. Load balancing is used to distribute a larger processing load to smaller processing nodes for enhancing the overall performance of system [3]. The main task of Load Balancing Algorithms is how to select next server node and to transfer incoming request to that distinct node. It is considered as a mechanism of lifting the entire load of a whole system to the nodes which are idle or having lesser load. It increases the effectiveness of resources and decreases the response time. Different load balancing policies are used by datacenters to balance load of requests between virtual machines.

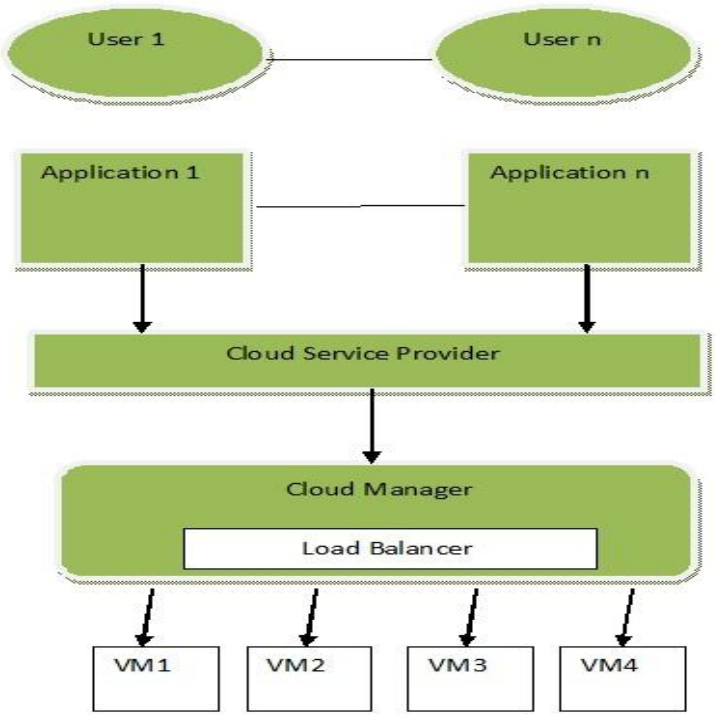

Figure 1: Load Balancing

Figure 1 explains the load balancing process. Users interact with cloud service providers through applications. Load balancer is in direct control of cloud manager has the task of balancing the workload over the entire system. Load balancer allocates and de-allocates virtual machines to the nodes. Virtual machines perform certain tasks.

A. Need for Load Balancing in Cloud Computing:

Load Balancing in clouds provide a mechanism to evenly distribute load among all nodes. It is used for achieving high user satisfaction and minimizes resource consumption which improves overall performance of the system. Load balancing avoids over heating of nodes due to excessive workload thereby reducing energy consumption which automatically reduces carbon emission.

B. Types of Load Balancing Algorithms:

Load Balancing Algorithms are of two types:

1. **Static Load Balancing Algorithms** – Algorithms which do not depend upon the current state of the system and need previous knowledge of the system[4]. They are non-pre-emptive and aim to minimize execution time, communication overhead and delays. It has a drawback that the task cannot be shifted at the time of execution to any other node to balance the load.

2. **Dynamic Load Balancing Algorithms** – This Algorithm does not need prior knowledge of the system. It improves the overall performance of the system by shifting the load dynamically during execution time because its decision for balancing the load is based on the current state of the system.

III.LITERATURE SURVEY

Jadeja, Y. and K. Modi[2] reviews the architecture of cloud computing, its benefits and issues such as security, privacy etc. and some of its major applications.

Jing Yao &Ju-hou He[5] discusses about architecture plan of cloud computing where cloud computing framework are divided in to two parts that is front-end & back-end. Both are connected through the internet. Front end is visible to users and back end is for cloud framework. Front end consist of client's computer accessed by the cloud, where as back end gives the 'cloud computing services' like storage, computers etc. It also discusses about the services and layers provided by cloud

computing design which are Software as a Service, Platform as a Service, and Infrastructure as a Service and some issues related to privacy, security, reliability etc.

Khiyaita et al. [6] provides the definition and taxonomy of load balancing. They described the different implementations of load balancing in most used distributed systems. They also mentioned the major challenges of load balancing in cloud computing.

Martin Randles et al. [7]investigate three possible distributed solutions proposed for load balancing; approaches inspired by Honeybee Foraging Behaviour, Biased Random Sampling and Active Clustering.This paper also presented a comparative study of three distributed load-balancing algorithms for Cloud computing scenarios.

Shreya Purohit [9] provides an in depth study of the factors favoring cloud computing, reviewing various cloud deployment and service models. It considers security, privacy, and internet dependency and availability as challenges of cloud computing. The author inspects certain benefits of cloud computing over traditional IT service environment including adaptability, higher resource usage, reduced capital, and scalability which are considered as reasons for switching to cloud computing environment. It considers vertical scalability as technical challenge in cloud computing.

J. M. Galloway et al. [10] proposed a load balancing algorithm that could be applied to the cluster controller of a local cloud that is power aware. This load balancer maintains the utilization of all computing nodes and distributes virtual machines in a way that is power efficient. The goal of this algorithm is to maintain availability to computing nodes while reducing the total amount of power consumed by the cloud. On the basis of utilization percentages Power Aware Load Balancing algorithm (PALB) decides the number of compute nodes that should be operating.

Bhathiya, Wickremasinghe[11] has discussed the detailed functioning of GUI based tool called as Cloud Analyst which was developed to simulate large-scale Cloud applications for studying the behaviour of such applications under various deployment configurations. Cloud Analyst helps developers in understanding how to spread applications among Cloud infrastructures and value added services such as performance optimization of applications and providers incoming with the use of Service Brokers.

Mohapatra, S. et al. [12] discussed a performance comparison for different load balancing algorithms of virtual machine and policies in cloud computing. In this study four well known load balancing algorithms have been considered. Performance of Execution Load, First Come First Serve, Round Robin and Throttled Load Balancing Algorithms have been analyzed based on the average response time, average datacenter request servicing time and total cost. The simulation results according to the CloudAnalyst simulator show that round robin has the best integration performance.

Nitika, M. et al. [13] addressed execution of three load balancing algorithms examined the inadequacies and researched why it is unrealistic to have Centralized Scheduling policy during the cloud condition. Author inspected three possible solutions which are Honeybee Foraging Behaviour algorithm, Random Sampling algorithm and Active Clustering algorithm proposed for load balancing.

IsamAzawiMohialdeen[14] discusses about various scheduling policies. Author does a comparative study of scheduling algorithms in cloud computing and explains there requirement in cloud environment.

Singh, A. et al. [15] develops an alternative method for round robin scheduling which improves the CPU efficiency in real time and time sharing operating system. The algorithm proposed by author improves all the snags of simple round robin architecture. He has also done a comparative analysis of simple round robin scheduling algorithm with proposed algorithm. Theproposed algorithm increases the system throughput and solves the problem faced in simple round robin architecture by decreasing the performance parameters to desirable extent.

Behal, V. and A. Kumar [16] provide a comparative study of Round Robin and Throttled virtual machine load balancing algorithms has been proposed. Both the algorithms are used with optimized response time service broker policy and simulation is performed to calculate overall response time, datacentress hourly average processing times, response time according to region,datacentress request servicing time, user base hourly response times and total cost which has significant effect on performance. According to the simulation results, the combination of the proposed strategy of throttled and optimized response time service broker policy has the better performance than round robin load balancing algorithm in heterogeneous cloud computing environment.
K Nishant et al. [19] have proposed an algorithm which is a modified approach of ant colony optimization that has been applied from the view of cloud network systems with the purpose of load balancing of nodes. It is different from the original

approach in which each ant builds own result set and later builds a complete result set. However in their approach the ants update a single result set rather than their individual result set.This approach detects overloaded and under loaded nodes and thereby perform operations based on the identified nodes.The task of each ant is specialized rather than being general and the task depends on the type of first node which was encountered whether it was overloaded or under loaded.

IV. LOAD BALANCING ALGORITHMS

A. Round Robin Algorithm (RR):

Round Robin is one of the traditional widely used algorithms. In round robin policy, the time slices are allotted to each task in uniform proportion and in circular fashion .Each task is allotted to available virtual machine in circular order .This policy is not considered as priority intended scheduling policy. In it, situation occurs where some nodes are massively loaded and some are slightly loaded. This leads to situation where system load gets imbalance [17].

B. Throttled Algorithm:

Throttled algorithm initiates by assigning favourable virtual machine when customer sends request to load balancer .The role of load balancer is to look after an index table of all virtual machine together with their states depicting busy and available mode. At start, all virtual machines are set to available mode. The datacentre controller consults balancer for next virtual machine allocation, when it receives a new request. The balancer start checking table thoroughly until a relevant match of virtual machine is found. If favourable virtual machine is found then the balancer returns id of that particular virtual machine to datacentres controller. At that instant, datacentres controller sends request to virtual machine identified by that particular id. After that, datacentres controller sends notification to the balancer of new allocation so that it can update the table. If there's a case, when virtual machine is not found, then the balancer returns -1 value and datacentres queues the request. As soon as virtual machine finishes with the processing of the assigned request, later the datacentres controller receives a response cloudlet and it sends the notification to balancer to virtual machine de-allocation[8][11].

C. Equally Spread Current Execution Algorithm (ESCE):

ESCE algorithm balances the tasks among available Virtual machines in a way to even out the number of active tasks at any given time on each Virtual Machine. ESCE algorithm handles the system workload with priorities [13]. ESCE distributes the datacentres workload randomly by checking the size and transfer the load to that virtual machine which is lightly loaded. This algorithm finds the VM with least number of allocations and in a way that the number of active tasks on each VM is kept evenly distributed among the VMs.

D. Active Clustering

This algorithm works on the principle of grouping similar ones and working on them group wise. The performance of the system is enhanced with high resources thereby increasing the parameter outcome using the algorithm. This algorithm is degraded with an increase in system diversity [7]. A process is initiated by the node and another node is selected known as matchmaker node from its neighbours, satisfying the condition that it should be a different type from the earlier one. The following sets of processes are executed one by one up to process end.

1. The match maker algorithm performs mechanism to form a connection between matchmaker node and neighbour of it which is of the same type as the initial node.
2. The matchmaker node then detaches the connection b/w itself and the initial node.

E. Min-Min Algorithm

It begins with a set of all unassigned tasks. First of all, minimum completion time for all tasks is found. Then among these minimum times the minimum value is selected which is the minimum time among all the tasks on any resources. Then according to that minimum time, the task is scheduled on the corresponding machine. Then the execution time for all other tasks is updated on that machine by adding the execution time of the assigned task to the execution times of other tasks on that machine and assigned task is removed from the list of the tasks that are to be assigned to the machines. Then again the same procedure is followed until all the tasks are assigned on the resources. But this approach has a major drawback that it can lead to starvation [3].

F. Max-Min Algorithm

Max-Min is almost same as the min-min algorithm.The core difference among Min-Min and Max-Min algorithm is following: in this algorithm first finding out minimum execution times, then first the most extreme value is chosen, then the performance time for all tasks is updated on that machine, this is done by adding the performance time of the assigned task to the performance times of other tasks on that machine. Then all assigned task is erased from the list that executed the system [3].

G.Biased Random Sampling

It is a dynamic approach which is based on random sampling of the system domain to balance the load in all available nodes. A virtual graph is used which is connected each node which shows the load on the server. Whena node executes a job, it deletes the incoming edge which indicates the resource is occupied. Resources are freed after the completion of the job.Random sampling algorithm is used for addition and deletion of processes.It can be started from any node and at every step a neighbour is chosen randomly. The previous node is selected for allocation of load. Alternatively a node can be selected on the basis of computing efficiency or the node which is under loaded. A node upon receiving a job, will execute it only if its current walk length is equal to or greater than the threshold value. Otherwise, the walk length of the job under examination is incremented and another neighbour node is selected arbitrarily. When, a job is executed by a node then in the graph, an incoming edge of that node is deleted. After the job is completed, an edge is created from the node initiating the load allocation process to the node which executed the job. Finally we get a directed graph. This is a fully decentralized load balancing scheme, hence making it apt for cloud computing systems.[7]

H. Power Aware Load Balancing (PALB)

This algorithm calculates utilization percentage of each computing node which is estimated for the working module, then decides the number of operating computing nodes while other nodes are completely shut down or are not in working condition. This algorithm has three sections in working module: balance section, upscale section and downscale section. Balance section is responsible for determining initialization process where virtual machine is going to start. The second section power-up the additional computing nodes and the third downscale section shut-downs the idle compute node in the process participant. [10]

I. Ant Colony

Ant colony optimization is Meta heuristic approach for load balancing system. The heuristic algorithm has guaranteed for optimal solution with any number of jobs and machines that are used in it. The approach ACO is based on nature of real ants which form the network in order to process the job. It is proposed by Dorigo at [1991].The ants are moving for searching food from source to nest in a path. The ants communicate with each other using a liquid evaporating content named as pheromone. During the path, other ants follow the same path with the help of pheromone. If the intensity of pheromone is high, ants follow that path otherwise no optimal solution. In the proposed strategy ants are moved in the graph where all the nodes are connected and randomly moved until an optimal solution has found.

J. Honey Bee Foraging

This load balancing algorithm [18] is similar to the behavior of honey bees finds and reaps their food. There is a category of bees called forager bees. They search for food and after getting it they come back for announcement. They announce it by doing a dance called waggle dance. This dance illustrates the availability of metadata food. After collecting the info scout bees follow the searcher bees towards the location of food for the purpose of storing food. Again returning to beehive they do a waggle dance which gives the information of available food to be occupied and then more food can be consumed by the honey bee. With the decreasing and increasing web server's demand in load balancing, the services are assigned dynamically to map the changing user demands. The virtual servers are clustered,each having its own virtual service queue. Like the quality that bee shows by their waggle dance each server also calculate a profit or reward from the request queues. This reward can be measured by the amount of time that the CPU spends on the processing of a request. In case of honey bees the dance floor is analogous to an advert board here. This mechanism in virtual server and load balancing is also useful while occupy the server for a process.

V. METRICS FOR LOAD BALANCING ALGORITHMS

1. Fault tolerance: It is the ability of the algorithm to perform correctly even in conditions of failure at any arbitrary node in the system.
2. Migration time: The time taken in transfer of a task from one machine to other machine in the system. This time should be least for improving the performance of the system.
3. Overhead: The amount of overhead that is produced by the execution of the load balancing algorithm. Minimum overhead is expected for the algorithm to be successful.
4. Performance: It represents the effectiveness of the system after performing load balancing. If all the metrics described are satisfied optimally then it will positively impact the performance of the system.
5. Resource Utilization: It is the extent to which the resources of the system are utilized. A good load balancing algorithm provides maximum resource utilization.
6. Response time: It is the time taken by a distributed system to respond for executing a specific load balancing algorithm.
7. Scalability: It determines the ability of the system to accomplish load balancing algorithm with a restricted number of processors or machines. Algorithm with higher scalability is preferable.

8. Throughput: It is the total amount of work done by all the nodes in a given time period. High throughput is necessary for better system performance.

VI. COMPARISON OF LOAD BALANCING ALGORITHMS

The table compares load balancing algorithms on the basis of metrics defined in previous section.

Parameters> Algorithms\/	Fault Tolerance	Migration Time	Overhead	Performance	Resource Utilization	Response Time	Scalability	Throughput
Round Robin[17]	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Throttled[8]	Yes	Yes	No	Yes	Yes	Yes	Yes	No
ESCE[13]	No	Yes	No	Yes	Yes	No	Yes	No
Active Clustering[17]	No	Yes	Yes	No	Yes	No	No	No
Min-Min[3]	No	No	Yes	Yes	Yes	Yes	No	Yes
Max-Min[3]	No	No	Yes	Yes	Yes	Yes	No	Yes
Biased Random Sampling[7]	No	No	Yes	Yes	No	Yes	No	No
PALB[10]	Yes	Yes	Yes	No	Yes	Yes	No	Yes
Ant Colony [19]	No	Yes	No	Yes	Yes	No	No	No
Honey Bee Foraging[18]	No	No	No	No	Yes	No	No	No

VII.CONCLUSION

Cloud Computing is very useful for managing large data but due to load, performance may be affected. Thus it is important to manage the load. Load balancing is important for resource utilization because it improves system performance and therefore is important for research. In this paper, work done by various authors on Load Balancing is reviewed. Load Balancing and various load balancing algorithms in cloud computing environment are discussed and and comparative analysis is performed on the basis of different metrics parameters like fault tolerance, migration time, overhead, performance, resource utilization, response time, scalability, throughput. Future work can be done on exploring new efficient load balancing algorithm which will maintain better trade off among parameters and also helps to reduce carbon emission and achieve green computing.

REFERENCES

[1] "What is Cloud Computing?" Amazon Web Services. Retrieved 2018-01-19
[2] Y.Jadeja andK. Modi, “Cloud Computing - Concepts, Architecture andChallenges”,International Conference on Computing, Electronics and Electrical Technologies (ICCEET), 2012.
[3] RajwinderKaur andPawanLuthra, “Load balancing in cloud computing”, Proc. of Int. Conf. on Recent Trends in Information, Telecommunication and Computing, ITC, 2012.
[4] Yogesh Kumar, AnishTalwar andAnuRathi, “Aspects of Security in Cloud Computing”,International Journal of Engineering and Computer Science ISSN:2319-7242, Volume 2 Issue 4, pp. 1361-1363, April, 2013.
[5] Jing Yao andJu-hou He, “Load Balancing Strategy of Cloud Computing based on Artificial Bee Algorithm”, ISBN: 978-1- 4673-0893-9, Pages: 185-189, IEEE, April 2012.
[6] A. Khiyaita, M.Zbakh, H. El Bakkali and Dafir El Kettani, “Load Balancing Cloud Computing : State Of Art”, IEEE,2012.
[7] Martin Randles, David Lamb and A. Taleb-Bendiab,”A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing”, IEEE 24th International Conference on Advanced Information Networking and Applications Workshops, 2010.
[8] Ram Prasad Padhy and PGoutam Prasad Rao ,”Load balancing in cloud computing system”, Department of Computer Science and Engineering National Institute of Technology, Rourkela-769 008, Orissa, India May,2011.
[9] ShreyaPurohit, “An Exhaustive Study on Cloud Computing,” Volume II, Issue I, ISSN NO. 2454-5678, 2016.

[10] J. M. Galloway, K. L. Smith, and S. S. Vrbsky, "Power aware load balancing for cloud computing", Proceedings of the World Congress on Engineering and Computer Science, vol. 1, pp.19–21, 2011.
[11] Bhathiya, Wickremasinghe,"Cloud Analyst: A Cloud Sim-based Visual Modeller for Analysing Cloud Computing Environments and Applications", IEEE, 2010.
[12] S.Mohapatra, K. SmrutiRekha, and S. Mohanty, "A Comparison of Four Popular Heuristics for Load Balancing of Virtual Machines in Cloud Computing", International Journal of Computer Applications, Volume 68-No 6, pp. 33-38, April2013.
[13] Nitika, M., M. Shaveta, and M.G. Raj, "Comparative analysis of load balancing algorithms in cloud computing". International Journal of Advanced Research in Computer Engineering & Technology, Volume 1 Issue 1,pp.120-124,2012.
[14] IsamAzawiMohialdeen, "Comparative Study of Scheduling Algorithms in Cloud Computing Environment", Journal of Computer Science, Volume 9 Issue 2, 2013.
[15] A. Singh, P. Goyal, and S. Batra, "An Optimized Round Robin Scheduling Algorithm for CPU Scheduling",(IJCSE) International Journal on Computer Science and Engineering, Volume 2 Issue 7,pp. 2383-2385,2010.
[16] V.Behal and A. Kumar "Cloud computing: Performance analysis of load balancing algorithms in cloudheterogeneous environment", Confluence The Next Generation Information Technology Summit (Confluence), 5th International Conference, 2014.
[17] Ajit Singh, Priyanka Goyal and SahilBatra, "An optimized round robin scheduling algorithm for CPU scheduling", International journal of computer and electrical engineering (IJCEE), vol. 2, No. 7, pp. 2383-2385, 2010.
[18] M. Randles, D. Lamb, and A. Taleb-Bendiab, "Experiments with Honeybee Foraging Inspired Load Balancing" Proceedings IEEE Second International Conference on Developments in eSystems Engineering (DESE), pp.240 – 247, 2009.
[19] KumarNishant, Pratik Sharma, Vishal Krishna, ChhaviGupta, KuwarPratap Singh, Nitin, and Ravi Rastogi,"Load Balancing of Nodes in Cloud Using Ant Colony Optimization",14th International Conference on Modelling and Simulation,UKSim, pp. 3–8, IEEE,2012.

