IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-5, Issue-1 E-ISSN: 2347-2693

Load Balancing and its Algorithms in Cloud Computing: A Survey

Sajjan R.S[1], Biradar Rekha Yashwantrao[2*]

[1,2]*Department of Computer Science and Engineering, Solapur University, Solapur, Maharashtra, India*

Available online at: www.ijcseonline.org

Received:28/12/2016 Revised: 06/01/2017 Accepted: 25/01/2017 Published: 31/01/2017

***Abstract*—**Cloud computing provides on-demand access to distributed resources on paid basis. Everybody wants to use these services to reduce the cost of infrastructure and maintenance, therefore the load on cloud is increasing day by day. Balancing the load is one of the most important issue that cloud computing is facing today. The load should be distributed fairly among all the nodes. Proper load balancing can reduce the energy consumption and carbon emission. This will helps to achieve Green Computing. There are many algorithms for load balancing. All these algorithms work in different ways and have some advantages and limitations. The most important for load balancing algorithms is to consider the characteristics like fairness, throughput, fault tolerance, overhead, performance, and response time and resource utilization. This paper mainly focuses on the concept of load balancing, literature survey on load balancing techniques and different measurement parameters.

***Keywords*—**Cloud Computing; Load Balancing; Static Algorithms; Dynamic Algorithms;Hierarchical Load Balancing

I. INTRODUCTION

With the ease of access to Internet, each individual, organization uses Cloud computing services. NIST defines Cloud computing is a computing model used everywhere and provides convenient, on-demand access to a shared pool of computing resources such as networks, servers, storage, applications, etc. These resources can be dynamically assigned and released with minimal management effort or service provider interaction [1]. It provides 3 services such as Software as Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). Different physical and virtual resources are provided to the users on demand. In Cloud computing, access to the resource is based on Virtualization [2]. Virtualization is an abstraction of real machines. Virtual Machine has ability to run applications like any real machine. Virtualization provides facilities similar to real machines. We can create larger number of less powered servers through virtualization, which in turn reduces the overall cost in space, power, and infrastructure. Cloud resources can be scaled rapidly using virtualization technique. Cloud resources are dynamically allocated to users on demand. As the number of user increases, the available resources decrease dynamically.

Allocation of cloud resources to users on demand gives rise to the problem of load balancing. If workload is not distributed properly, then some nodes in cloud will be heavily loaded and some nodes will be under loaded. In the same way if the resources provided by the cloud are not allocated efficiently, it leads to delay in providing service to the users [3]. Load imbalance may cause system bottleneck. To achieve resource utilization and no delay in providing service, resource allocation should be done in an efficient way [3].

Nodes of the system can be logically grouped into cluster and task of load balancing is distributed among clusters. Each individual cluster is going to allocate load to the nodes belonging to that cluster. This can be arranged in hierarchical form. For cloud environment various load balancing approaches have been implemented to provide efficient distribution of load among available machines. Such as Round Robin load balancing, Throttled load balancing, Min-Min load balancing, Min-Max load balancing, Honey Bee and Ant Colony behaviour based load balancing, etc. For effective load balancing, a single load balancing algorithm is not sufficient. Hence there is requirement of algorithm which combines features of two or more load balancing algorithms.

This paper is organized as follows. Section II describes what are load balancing and its measurement parameters? Section II and III describes types of Load Balancing algorithms. Section VI includes the comparison between various load balancing algorithms.

II. LOAD BALANCING

Load balancing is a process of distributing the workload dynamically and uniformly across all the available nodes in the cloud.

This improves the overall system performance by shifting the workloads among different nodes. Not properly utilized resources will sometimes overheated which causes Carbon emission. Carbon emission can be minimized by utilizing the resources properly [4]. Throughput, performance, scalability, response time, resource utilization and fault tolerance are some measurement parameters that can be used to evaluate the load balancing techniques. These parameters allow us to check whether the given technique or algorithm of load balancing is good enough to balance the load or not [5].
Through effective Load balancing, every virtual machine in the cloud system can process the same amount of work. Hence, load balancing will be needed to maximize the throughput by minimizing the response time. It also saves energy consumption which helps in clean and green environment. With the help of Load balancing, the energy consumption is reduced, hence reduced carbon emission. This helps in achieving Green computing. Efficient Load balancing will ensure [4]:

- Uniform distribution of load on nodes.
- Improves overall performance of the system
- Higher user satisfaction
- Faster Response
- System stability
- Reducing carbon emission

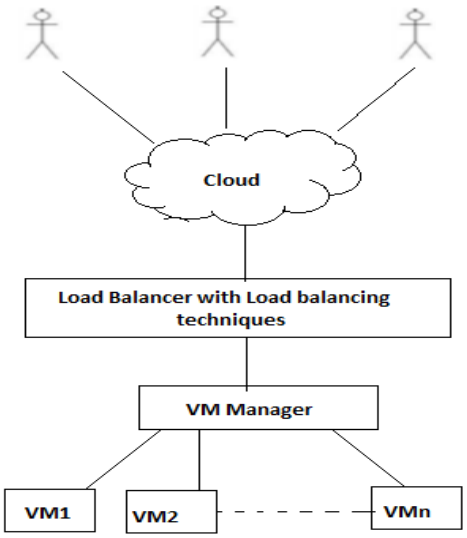

Fig -1: General Structure of Load balancing in Cloud Environment [4]

A) Load Balancing Measurement Parameter

There are some measurement parameters to evaluate the load balancing techniques which allow us to check whether the given technique is good enough to balance the load or not [5].

- *Throughput:* It is the amount of work to be done in the given amount of time.
- *Response time:* It is the amount of time used to start fulfilling the demand of the user after registering the request.
- *Fault tolerance:* It is the ability of the load balancing algorithm that allows system to work in some failure condition of the system.
- *Scalability:* It is the ability of the algorithm to scale itself according to required conditions.
- *Performance:* It is the overall check of the algorithms working by considering accuracy, cost and speed.
- *Resource utilization:* It is used to keep a check on the utilization of various resources.

B) Classification of Load Balancing Algorithms

There are many load balancing algorithms. Generally Load balancing algorithms are classified into two categories based on the present system state [6] [7]:

- *Static Algorithm:* Static Algorithms are good for homogeneous and stable environment.
- *Dynamic Algorithm:* Dynamic Algorithms are good for heterogeneous environment.

III. STATIC ALGORITHMS

Static algorithms are best in homogeneous and stable environments. However, static algorithms are not flexible and cannot consider the dynamic changes to the attributes. While assigning tasks to the nodes, static load balancing algorithms will not check the state and functionality of the node in previous tasks [5]. Some Static Algorithms are:

A) Round Robin Load Balancing Algorithm (RR)
B) Load Balancing Min-Min Algorithm (LB Min-Min)
C) Load Balancing Min-Max Algorithm (LB Min-Max)

A. Round Robin Load Balancing Algorithm

In this algorithm, fixed quantum time is given to the job. It allocates jobs to all nodes in a circular fashion. Processors are assigned in a circular order and hence there is no starvation [8]. This algorithm provides faster response in the case of equal workload distribution among processes. However, some nodes may be over loaded while others remain idle and under-utilized [6].

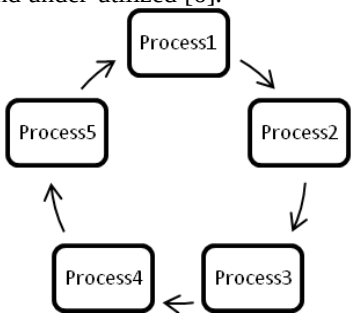

Fig- 2: Round Robin Load Balancer

B. MIN-MIN Load Balancing Algorithm

A list of task is maintained and minimum completion time is calculated for all the available nodes. A task with minimum completion time is assigned to the machine. Hence the name of the algorithm is min-min [5]. Update the list and running time of the machine. It provides good results when small task are more [8].

C. MIN-MAX Load Balancing Algorithm

A list of task is maintained and minimum completion time is calculated for all the available nodes. A task with maximum completion time is assigned to the machine. Hence the name of the algorithm is min-max [5]. Update the list and running time of the machine [8].

IV. DYNAMIC ALGORITHMS

Dynamic algorithms provide better results in heterogeneous and dynamic environments. These algorithms are more flexible. Dynamic algorithms can consider the dynamic changes to the attributes. However, these algorithms are more complex [5]. Main advantage of this is that selection of task is based on current state and this will help to improve the performance of the system.

Dynamic algorithms can be implemented in following two forms [9]:

1. Distributed System

Here all the nodes interact with each other and load balancing algorithm is executed by all the nodes in the system. The task of load balancing is distributed among all the nodes. Interaction among nodes can be cooperative or non-cooperative [9]. If any node fails in the system, it will not stop the functionality.

i) In cooperative distributed system, all node works together [9].
ii) In non-cooperative distributed system, each node works independently [9].

2. Non-distributed System

Non-distributed can be centralized or semi-distributed [9].

- *In centralized system*, central node is responsible for load balancing of the whole system. The other nodes interact with this central node. If central node fails, it will stop the functionality [9]. In case of failure, recovery will not be easy [12].
- *In semi-distributed system*, nodes are grouped to form a cluster. A central node of each cluster performs load balancing of whole system. If central node of cluster fails, it will stop the functionality of that cluster only [9]. Multiple central nodes manage the load balancing. Hence more accurate load balancing [12].

Some dynamic algorithms are:

A) Honeybee Foraging Behavior Load Balancing Algorithm
B) Throttled Load Balancing Algorithm
C) ESCE (Equally Spread Current Execution) Load Balancing Algorithm
D) Ant Colony Load Balancing Algorithm
E) Biased Random Sampling Load Balancing Algorithm
F) Modified Throttled Load Balancing Algorithm

A. Honeybee Foraging Behavior Load Balancing Algorithm

This algorithm was derived from the behavior of real honey bees in finding their food sources. After finding the food sources, the honey bees come back to the bee hive to inform the food source. They do this by performing group movement. This group movement is also known as "waggle Dance". They perform waggle dance to inform other bees of the exact location of the food source. This waggle dance shows the quality, quantity of the food and the distance of the food source from the bee hive [5] [8] [9].

B. Throttled Load Balancing Algorithm

Throttled load balancing algorithms is best suitable for virtual machines. Load balancer maintains the list of entire virtual machines in the system. When load balancer receives a request, it scans the indexing table. If virtual machine is available, then the job is assigned to that machine. Load balancer updates the indexing table after each allocation and de-allocation of resource [5] [8] [9].

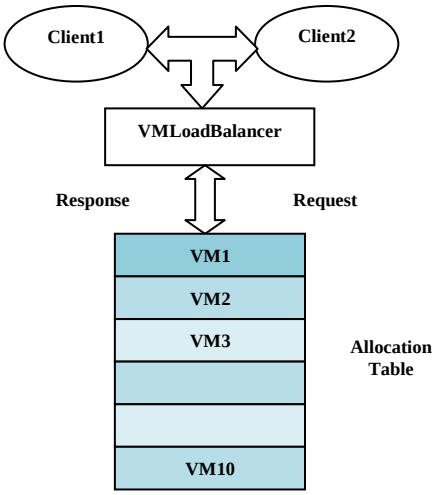

Fig-3: Throttled Load Balancing

C. ESCE (Equally Spread Current Execution) Load Balancing Algorithm

Load balancer maintains the list of entire virtual machines and jobs. When load balancer receives a request, it scans the list of VM's. If a VM is found which can handle the client's request, then the request is allocated to that particular VM. This algorithm distributes the equal load among all VM's [9] [11].

D. ANT COLONY Load Balancing Algorithm

Real ant selects a shortest path in search of its food [9] [10]. This algorithm is based on the behavior of real ants. When request is initiated ant starts its movement. Ant continuously checks whether the node is overloaded or under loaded. If ant finds any overloaded node, it turns back. And if ant finds any under loaded node, it proceeds. In this way behavior of ant is used to collect the information from different nodes [5].

E. Biased Random Sampling Load Balancing Algorithm

This algorithm balances the load through random sampling of system domain. Virtual graph of the system is constructed. In a directed graph, each node is represents a vertex and each in-degree represents free resources of that node. The load balancer allocates the job to the node which has at least one in-degree. The in-degree of the node is incremented and decremented when job is completed and when job is allocated respectively. This is done by the process of random sampling [9].

F. Modified Throttled Load Balancing Algorithm

This algorithm focuses on how jobs are allocated to the available VM's intelligently. This algorithm maintains an index table of VM's and also the state of VMs (BUSY/AVAILABLE). This algorithm initially selects a VM at first index depending upon the state of the VM. Available VM is assigned to the request. If the VM is not available -1 is returned. If the new request arrives, the VM at the previous VM index + 1 is chosen depending on the state of VM [12].

V. HIERARCHICAL LOAD BALANCING ALGORITHM

Hierarchical Load Balancing involves different levels in load balancing decisions. Every node is managed or balanced by its parent node [13]. Parent node is responsible for load balancing [13]. Hierarchical load balancing can be used in homogeneous as well as heterogeneous environment [13]. Cluster can also be used in hierarchical load balancing. Clustering is the process of organizing similar type of objects into groups. VM's having similar characteristics are logically grouped. VM's are at last level.

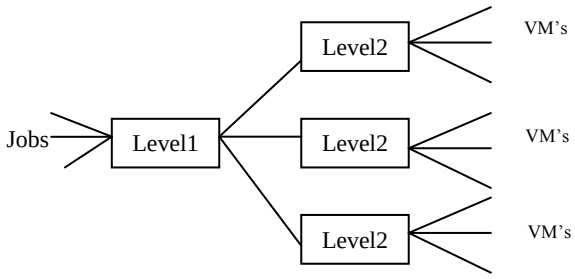

Fig-4: Hierarchical Load Balancing

Jobs are received by Level1 and dispatched to the level2 according to the resource specific requirement of jobs. Level2 selects VM and assigns Job to the VM for execution.

VI. COMPARISON BETWEEN VARIOUS LOAD BALANCING ALGORITHMS

Algorithm	Advantages	Disadvantages
Round Robin Load Balancing Algorithm	• It is Simple algorithm and emphasis is on fairness. • It works in circular fashion. • Fast response in the case of equal workload distribution • There is no starvation.	• Each node is fixed with a time slice. • It is not flexible and scalable. • Some node may possess heavy load and some nodes are idle. • Does not save the state of previous allocation of a VM. • Pre-emption is required.
MIN-MIN Load Balancing Algorithm	• It is simple and fast algorithm. • It works better for smaller task.	• Selects the task having minimum completion time. • There is starvation. Smaller tasks will get executed first, while the larger tasks keep on in the waiting stage. • Poor load balancing • Does not consider the existing load on a resource.
MIN-MAX Load Balancing Algorithm	• It is simple algorithm. • It runs short tasks concurrently.	• Selects the task having the maximum completion time • There is a starvation. Larger tasks will execute

		first, while the smaller tasks need to wait. • Poor load balancing.
Honeybee Foraging Behavior Load Balancing Algorithm	• Self-organizing, nature inspired algorithm. • Performance will be achieved by increasing the system size. • Suitable for heterogeneous environment.	• Increase in resources will not increase the overall throughput.
Throttled Load Balancing Algorithm	• List of VMs is maintained along with the status of each VM • Good performance • Better resource utilization	• Scans the entire list of VMs from the beginning • Does not consider the current load on VM.
ESCE Load Balancing Algorithm	• Maintains equal load at all VMs • Maximize the throughput	• Central point of failure • Not fault tolerant.
ANT COLONY Load Balancing Algorithm	• Under loaded node is found at beginning of the search • Decentralized .	• Network overhead • Delay in moving forward and backward.
Biased Random Sampling Load Balancing Algorithm	• Fully decentralized • Suitable in large network	• Performance is degraded with an increase in diversity
Modified Throttled Load Balancing Algorithm	• Index table is parsed from the index next to already assigned VM. • Faster response than throttled algorithm • Efficient usage of available resources	• Does not consider the current load on VM.
Hierarchical Load Balancing	• Faster response • Suitable for homogeneous and heterogeneous environment	• Less fault tolerant

CONCLUSION

Cloud computing allows wide range of users to access distributed, scalable, virtualized, hardware and software resources over the Internet. Load balancing is one of the most important issue of cloud computing. It is a mechanism which distributes workload evenly across all the nodes in the whole cloud. Through efficient load balancing, we can achieve a high user satisfaction and resource utilization. Hence, this will improve the overall performance and resource utility of the system. With proper load balancing, resource consumption can be kept to a minimum which will further reduce energy consumption and carbon emission rate. Through hierarchical structure of system, performance of the system will be increased.

ACKNOWLEDGMENT

Author would like to thank Ms Sajjan R.S. for her constant guidance and her parents and her siblings for their unconditional love and support.

REFERENCES

[1] Peter Mell Timothy Grance", The NIST Definition of Cloud Computing", National Institute of Standards and Technology Special Publication 800-145(September 2011),csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[2] Rajwinder Kaur and Pawan Luthra," Load Balancing in Cloud Computing", Association of Computer Electronics and Electrical Engineers, 2014, DOI: 02.ITC.2014.5.92

[3] Ashalatha R; J. Agarkhed, "Dynamic load balancing methods for resource optimization in cloud computing environment ", 2015 Annual IEEE India Conference (INDICON) , Pages: 1 - 6, DOI: 10.1109/INDICON.2015.7443148

[4] Garima Rastogi, Dr Rama Sushil, "Analytical Literature Survey on Existing Load Balancing Schemes in Cloud Computing", 2015 International Conference on Green Computing and Internet of Things (ICGCloT), pages:1506-1510

[5] R. Kanakala; V. K. Reddy; K. Karthik, " Performance analysis of load balancing techniques in cloud computing environment", 2015

IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Pages: 1 - 6, DOI: 10.1109/ICECCT.2015.7226052

[6] K. Garala; N. Goswami; P. D. Maheta, " A performance analysis of load Balancing algorithms in Cloud environment ", 2015 International conference on Computer Communication and Informatics (ICCCI), Pages: 1 - 6, DOI: 10.1109/ICCCI.2015.7218063

[7] A.N. Ivanisenko; T. A. Radivilova , "Survey of major load balancing algorithms in distributed system ", Information Technologies in Innovation Business Conference (ITIB), 2015 ,Pages: 89 - 92, DOI: 10.1109/ITIB.2015.7355061

[8] Sidra Aslam, Munam Ali Shah, "Load Balancing Algorithms in Cloud Computing: A Survey of Modern Techniques", 2015 National Software Engineering Conference (NSEC 2015)

[9] A. A. Jaiswal, Dr. Sanjeev Jain," An Approach towards the Dynamic Load Management Techniques in Cloud Computing Environment", 2014 International Conference on Power, Automation and Communication (INPAC)

[10] G.Punetha Sarmila,Dr.N.Gnanambigai, Dr.P.Dinadayalan," Survey on Fault Tolerant –Load Balancing Algorithms in Cloud Computing", IEEE Sponsored 2nd International Conference On Electronics And Communication System (ICECS 2015), Pages-1715-1720

[11] Surbhi Kapoor, Dr. Chetna Dabas," Cluster Based Load Balancing in Cloud Computing", 2015 Eighth International Conference on Contemporary Computing (IC3)

[12] Shridhar G.Domanal and G.Ram Mohana Reddy, " Load Balancing in Cloud Computing Using Modified Throttled Algorithm ", 2013 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)

[13] Aarti Vig, Rajendra Singh Kushwah and Shivpratap Singh Kushwah, "An Efficient distributed Approach for Load Balancing in Cloud Computing", 2015 International Conference on Computational Intelligence and Communication Networks

AUTHORS PROFILE

Ms. Sajjan R.S. received her M.Tech in Computer Science and Engineering. She has a working experience of 15 years and is currently the H.O.D. of the Computer Science and Engineering Department. She is currently pursuing Ph.D. Her research interest is in Cloud Computing.

Ms. Biradar Rekha Yashwantrao received Bachelor of Engineering in Information Technology from W.I.T., Solapur. She is currently working toward the M.E degree in Computer Science & Engineering from Solapur University, Solapur. Her research interests lies in area of programming & cloud computing.

