A Survey Of Various Load Balancing Algorithms In Cloud Computing

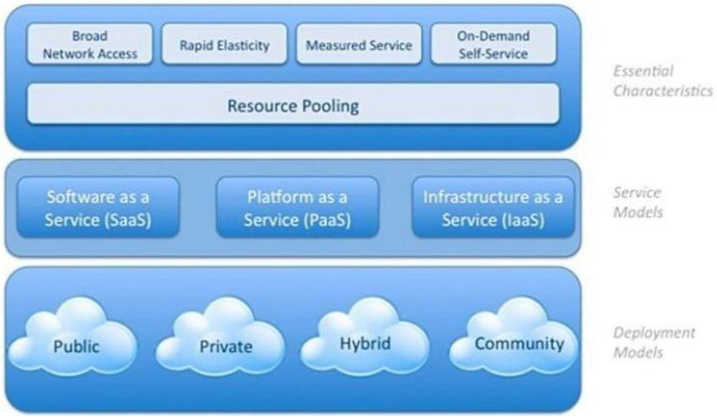
Dharmesh Kashyap, Jaydeep Viradiya

Abstract: Cloud computing is emerging as a new paradigm for manipulating, configuring, and accessing large scale distributed computing applications over the network. Load balancing is one of the main Challenges in cloud computing which is required to distribute the workload evenly across all the nodes. Load is a measure of the amount of work that a computation system performs which can be classified as CPU load, network load, memory capacity and storage capacity. It helps to achieve a high user satisfaction and resource utilization ratio by ensuring an efficient and fair allocation of every computing resource. Proper load balancing aids in implementing fail-over, enabling scalability, over- provisioning, minimizing resource consumption and avoiding bottlenecks etc. This paper describes a survey on load balancing algorithms in cloud computing environment along with their corresponding advantages, disadvantages and performance metrics are discussed in detail.

Index Terms: Cloud Computing, Virtualization, Load Balancer, Load Balancing, Load Balancing algorithm.

1 Introduction

The term "cloud" originates from the world of telecommunications when providers began using virtual private network (VPN) services for data communications .Cloud computing is an on demand service in which shared resources, information, software and other devices are provided according to the client's requirement at specific time.



Architecture of cloud computing [13]

There are four types of cloud deployment model in,

- Private Cloud(used by single organization)
- Public Cloud(anyone can access)
- Community Cloud(shared by several organizations)
- Hybrid Cloud(combination of two or more clouds)

- *Dharmesh Kashyap is currently pursuing masters degree program in computer science and engineering in Gujarat Technology University, India, PH-8672889988.*
  E-mail: dharmeshkashyap143@mail.com
- *Jaydeep Viradia is currently working as a Assistant professor in Parul Institute Engineering and Technology , India, PH-9909628300.*
  E-mail: jaydeepviradiya1@gmail.com

According to the National Institute of Standards and Technology (NIST), basic services provided by the Cloud Environment are as below [1]:

- Software as a Service (SaaS): Customers hire software hosted by vendor.
- Platform as a Service (PaaS): Customers hire infrastructure and programming tools hosted by vendor to create applications.
- Infrastructure as a Service (IaaS): Customers hire processing, networking, storage and other fundamental computing resources.

National Institute of Standards and Technology (NIST) defines characteristics of Cloud Computing are as below [1]:

- **Broad network access**: available through standard Internet- enabled devices
- **On demand self-service access** : consumers are provisioned services without other's help
- **Location independent resource pooling**: demands are balanced across a common infrastructure with no particular resource assigned to any individual user
- **Rapid elasticity**: Quality of service will be same as increase or decrease the number of consumers.
- **Pay per use**: consumers pay charges based on their usage of computing resources.

2 VIRTUALIZATION

Cloud computing uses virtualization as a base for provisioning services to the client. Multiple operating systems can be run on the single computer on the base of virtualization so utilization of resources is increased. For enhanced the productivity of server the hardware resources are combined. For the proper resource utilization the, computer architecture is uses software called Hypervisor. It is also called Virtual Machine Monitor (VMM) for running the multiple operating systems on the single host. Hypervisor provides the computing resources (memory, processor, bandwidth) to the virtual machine. There are two types of virtualization in cloud computing.

1) Full Virtualization

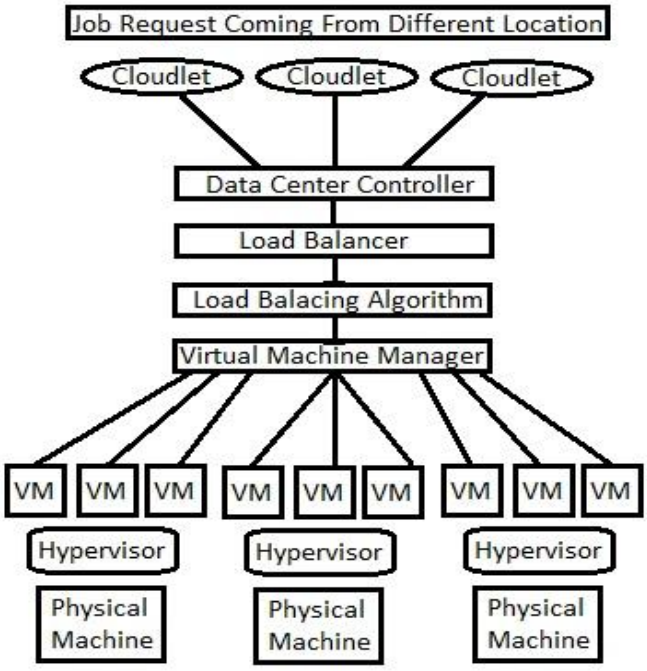
In Full Virtualization [2], the entire installation of one computer is done on the other computer. So the functionality of the actual machine can also be available in virtual machine.

2) Para Virtualization

In Para Virtualization [2], multiple operating systems can be run on a single machine. Here all the functionalities are not fully available, rather than the services are provided in a partial manner.

3 LOAD BALANCING

Load balancing in clouds is a technique that distributes the excess dynamic local workload evenly across all the nodes. It is used for achieving a better service provisioning and resource utilization ratio, hence improving the overall performance of the system Incoming tasks are coming from different location are received by the load balancer and then distributed to the data center ,for the proper load distribution



The aim of load balancing is as follows:

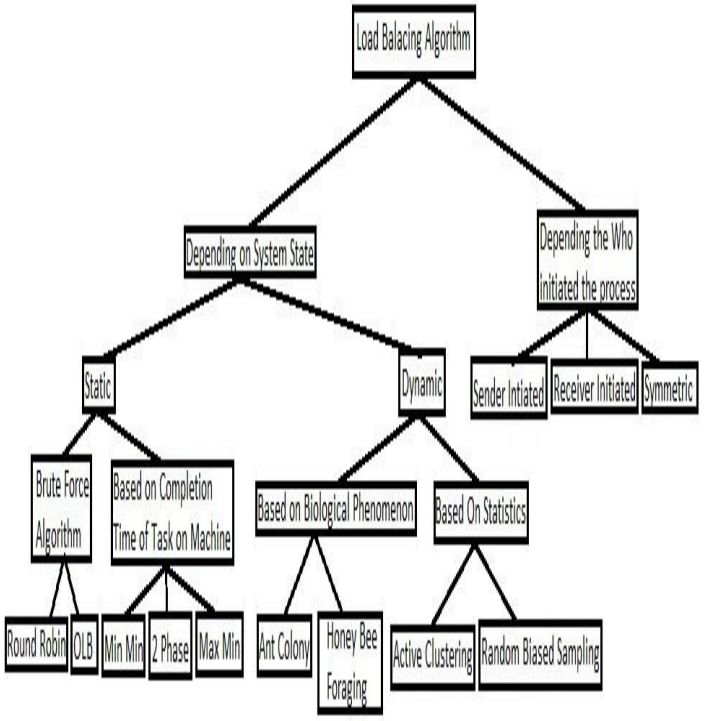
- To increase the availability of services
- To increase the user satisfaction
- To maximize resource utilization
- To reduce the execution time and waiting time of task coming from different location.
- To improve the performance
- Maintain system stability
- Build fault tolerance system
- Accommodate future modification

4 CHALLENGES OF LOAD BALANCING

- **Overhead Associated** -determines the amount of overhead involved while implementing a load-balancing system. It is composed of overhead due to movement of tasks, inter-process communication. Overhead should be reduced so that a load balancing algorithm performs well.
- **Throughput** – It is the number of task executed in the fixed interval of time. To improve the performance of the system, throughput should be high .
- **Performance** – It can be defined as the efficiency of the system. It must be improved
- **Resource Utilization** -is used test the utilization of resources. It should be maximum for an efficient load balancing system.
- **Scalability** - the quality of service should be same if the number of users increases. The more number of nodes can be added without affecting the service.
- **Response Time** – can be defined as the amount of time taken to react by a load balancing algorithm in a distributed system. For better performance, This parameter should be reduced.
- **Fault Tolerance** –In spite of the node failure, the ability of an system to perform uniform load balancing. The load balancing is the best fault-tolerant technique.
- **Point of Failure**: designed the system in such a way that the single point failure does not affect the provisioning of services. Like in centralized system, if one central node is fail, then the whole system would fail, so load balancing system must be designed in order to overcome this problem.

5 BASIC TYPES OF LOAD BALANCING ALGORITHMS

There is a extremely large need for load balancing in complex and large distributed systems,. Load balancer takes a decision to transfer the job to the remote server for load balancing. Load balancer can works in two ways: one is cooperative and non-cooperative. In cooperative way, to achieve the optimal response time, all the nodes work to gather. In non-cooperative way, response time is increase by the independently running the tasks. Some of the algorithms for load balancing are studied in this paper.



- Based on the current state of the system, load balancing algorithms can be classified into two types:

- Static algorithm: The current status of the node is not taken into consideration [3]. All the nodes and their properties are known in advance. Based on this prior knowledge, the algorithm works. Since it does not use

current system status information, it is easy to implement.

- Dynamic algorithm: This type of algorithm is based on the current status of the system [3]. The algorithm works according to the dynamic changes in the state of nodes. Status Table maintains the Current status of all the nodes in the cloud. Dynamic algorithms are complex to implement but it balances the load in effective manner.

❖ Based on the initiator of the algorithm, Load Balancing algorithms can be classified into three types [2]:

- Sender Initiated: Sender identifies that the nodes are in large number so that the sender initiates the execution of Load Balancing algorithm.
- Receiver Initiated: The requirement of Load balancing situation can be identified by the receiver/server in cloud and that server initiates the execution of Load Balancing algorithm.
- Symmetric: It is the combination of both the sender initiated and receiver initiated types.

6 LOAD BALANCING ALGORITHMS

Following load balancing algorithms are currently prevalent in clouds

6.1 Round-Robin Algorithm [4]

It is the static load balancing algorithm which uses the round robin scheme for allocating job. It selects the first node randomly and then, allocates jobs to all other nodes in a round robin fashion. Without any sort of priority the tasks are assigned to the processors in circular order. Because of the non uniform distribution of workload, this algorithm is not suitable for cloud computing .some nodes get heavily loaded and some nodes get lightly loaded because the running time of any process is not known in advance. This limitation is overcome in the weighted round-robin algorithm .In the weighted round-robin algorithm some specific weight is assigned to the node .on the basis of assignment of weight to the node it would receive appropriate number of requests .If there are equal assignment of weight, each node receive some traffic. This algorithm is not preferred because prior prediction of execution time is not possible

6.2 Opportunistic Load Balancing Algorithm [5]

This is static load balancing algorithm so it does not consider the current workload of the VM. It attempts to keep each node busy. This algorithm deals quickly with the unexecuted tasks in random order to the currently available node. Each task is assigned to the node randomly. It provides load balance schedule without good results. The task will process in slow in manner because it does not calculate the current execution time of the node.

6.3 Min-Min Load Balancing Algorithm[6]

The cloud manager identifies the execution and completion time of the unassigned tasks waiting in a queue. This is static load balancing algorithm so the parameters related to the job are known in advance. In this type of algorithm the cloud manager first deals with the jobs having minimum execution time by assigning them to the processors according to the capability of complete the job in specified completion time. The jobs having maximum execution time has to wait for the unspecific period of time. Until all the tasks are assigned in the processor, the assigned tasks are updated in the processors and the task is removed from the waiting queue. This algorithm performs betters when the numbers of jobs having small execution time is more then the jobs having large execution time. The main drawback of the algorithm is that it can lead to starvation.

6.4 Max-Min Load Balancing Algorithm[6]

Max Min algorithm works same as the Min-Min algorithm except the following: after finding out the minimum execution time, the cloud manager deals with tasks having maximum execution time. The assigned task is removed from the list of the tasks that are to be assigned to the processor and the execution time for all other tasks is updated on that processor. Because of its static approach the requirements are known in advance then the algorithm performed well. An enhanced version of max min algorithm was proposed in [7]. It is based on the cases, where meta-tasks contain homogeneous tasks of their completion and execution time, improvement in the efficiency of the algorithm is achieved by increasing the opportunity of concurrent execution of tasks on resources.

6.5 The two phase scheduling load balancing algorithm [8]

It is the combination of OLB (Opportunistic Load Balancing) and LBMM (Load Balance Min-Min) Scheduling algorithms to utilize better execution efficiency and maintain the load balancing of the system.OLB scheduling algorithm keeps every node in working state to achieve the goal of load balance and LBMM scheduling algorithm is utilized to minimize the execution of time of each task on the node thereby minimizing the overall completion time. This algorithm works to enhance the utilization of resources and enhances the work efficiency.

6.6 Ant Colony Optimization Based Load Balancing Algorithm [2]

Aim of the ant colony optimization to search an optimal path between the source of food and colony of ant on the basis of their behavior. This approach aims efficient distribution of work load among the node. When request is initialized the ant starts movement towards the source of food from the head node. Regional Load Balancing Node (RLBN) is chosen in Cloud Computing Service Provider (CCSP) as a head node. Ants keep records the every node they visits ant record their data for future decision making .Ant deposits the pheromones during their movement for other ants to select next node The intensity of pheromones can vary on the bases of certain factors like distance of food, quality of food etc. When the job gets successful the pheromones is updated. Each ant build their own individual result set and it is later on built into a complete solution. The ant continuously updates a single result set rather than updating their own result set. By the ant pheromones trials, The solution set is continuously updated.

6.7 Honeybee Foraging load balancing Algorithm [10]

It is a nature inspired decentralized load balancing technique which helps to achieve load balancing across heterogeneous virtual machine of cloud computing environment through local server action and maximize the throughput. The current workload of the VM is calculated then it decides the VM states whether it is over loaded ,under loaded or balanced .according to the current load of VM they are grouped. The priority of the

task is taken into consideration after removed from the overload VM which are waiting for the VM .Then the task is schedule to the lightly loaded VM. The earlier removed task are helpful for the finding the lightly loaded VM. These tasks are known as scout bee in the next step. Honey Bee Behavior inspired Load Balancing technique reduces the response time of VM and also reduces the waiting time of task.

6.8 Biased Random Sampling load balancing Algorithm[11]

Biased Random Sampling Load Balancing Algorithm is dynamic approach, the network is represented in the form of virtual graph. Each server is taken as a vertex of the node and the in degree represents the available free resources the nodes have. On the basis of the in degree the load balancer allocates the job to the node. The nodes have at least one in degree then load balancer allocates the job to that node. When the job is allocates to the node then the in degree is decrement by one, and it's get incremented again when job gets executed. Random sampling technique is used in the addition and deletion of the processes. The processes are centralized by the threshold value, which indicates the maximum traversal from one node to destination node. The length of traversal is known as walk length. The neighbor node of the current node is selected for the traversal. After receiving the request, load balancer selects a node randomly and compares the current walk length with the threshold value. If the current walk length is equal to or greater than the threshold value, the job is executed at that node. Otherwise, the walk length of the job is incremented and another neighbor node is selected randomly. The performance is decrease as the number of servers increase

6.9 Active Clustering load balancing Algorithm [12]

Active Clustering is works on the basis of grouping similar nodes and increase the performance of the algorithm the process of grouping is based on the concept of match maker node. Match maker node forms connection between its neighbors which is like as the initial node .Then the matchmaker node disconnects the connection between itself and the initial node. The above set of processes is repeating again and again. The performance of the system is increases on the basis of high availability of resources, because of that, the throughput is also increasing. This increase in throughput is because of the efficient utilization of resources

7 COMPARISON OF ALGORITHMS

	Round Robin	OLB	Min Min	2 phase	Min Max	Ant colony	Honey Bee	Biased Random Sampling	Active Clustering
Throughput	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No
Overhead	Yes	No	Yes	Yes	Yes	No	No	No	Yes
Fault tolerance	No	No	No	no	No	No	No	No	No
Response Time	Yes	No	Yes	Yes	Yes	No	No	No	No
Resource Utilization	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes
Scalability	No	No	No	No	No	Yes	Yes	No	No
Performance	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No

8 CONCLUSION

Cloud computing provides everything to the user as a service over network. The major issues of cloud computing is Load Balancing. Overloading of a system may lead to poor performance which can make the technology unsuccessful, for the efficient utilization of resources , the efficient load balancing algorithm is required. In this paper, we have surveyed various load balancing algorithms in the Cloud environment. We have discussed the already proposed algorithms by various researchers. The various load balancing algorithms are also being compared here on the basis of different types of parameter.

REFERENCES

[1] Maria Spinola, "An Essential Guide to Possibilities and Risks of Cloud Computing: a Pragmatic Effective and Hype Free Approach for Strategic Enterprise Decision Making". (white paper) 2009

[2] Ratan Mishra and Anant Jaiswal, "Ant Colony Optimization: A solution of Load Balancing in Cloud", International Journal of Web & Semantic Technology (IJWesT), April 2012

[3] Venubabu Kunamneni, "Dynamic Load Balancing for the cloud", International Journal of Computer Science and Electrical Engineering, 2012.

[4] Pooja Samal, Pranati Mishra, "Analysis of variants in Round Robin Algorithms for load balancing in Cloud Computing" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (3) , 2013, 416-419.

[5] Che-Lun Hung1, Hsiao-hsi Wang2 and Yu-Chen Hu2, "Efficient Load Balancing Algorithm for Cloud Computing Network". IEEE Vol. 9, pp: 70-78, 2012

[6] T. Kokilavani, Dr. D. I. George Amalarethinam "Load Balanced Min-Min Algorithm for Static Meta Task Scheduling in Grid computing" International Journal of Computer Applications Vol-20 No.2, 2011

[7] Upendra Bhoi, Purvi N. Ramanuj, "Enhanced Max-min Task Scheduling Algorithm in Cloud Computing" International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 4, April 2013.

[8] Karanpreet Kaur, Ashima Narang, Kuldeep Kaur, "Load Balancing Techniques of Cloud Computing", International Journal of Mathematics and Computer Research, April 2013.

[9] Nidhi Jain Kansal, Inderveer Chana, "Cloud Load Balancing Techniques: A Step Towards Green Computing", International Journal of Computer Science, January 2012

[10] Dhinesh B. L.D , P. V. Krishna, "Honey bee behavior inspired load balancing of tasks in cloud computing environments", in proc. Applied Soft Computing, volume 13, Issue 5, May 2013, Pages 2292-2303.

[11] Rahmeh OA, Johnson P, Taleb-Bendiab A.,"A Dynamic Biased Random Sampling Scheme for scalable and reliable Grid Networks", The INFOCOMP Journal of Computer Science, vol. 7, 1-10

[12] Ram Prasad Padhy ,P Goutam Prasad Rao, “Load Balancing in Cloud Computing Systems”, National Institute of Technology, Rourkela, India, 2011.

[13] Tushar Desai, Jignesh Prajapati ,” A Survey Of Various Load Balancing Techniques And Challenges In Cloud Computing” International Journal Of Scientific & Technology Research , Volume 2, Issue 11, November 2013