ISSN 2278-7763

A Survey of Load Balancing Algorithms using VM

B.KalaiSelvi[1] and Dr.L.Mary Immaculate Sheela[2]

[1] Research Scholar, Mother Teresa Women's University, Kodaikanal.

jkalaimca@gmail.com

[2] Professor, R.M.D Engineering College, Chennai.

drsheela09@gmail.com

ABSTRACT

Cloud Computing is an emerging computing paradigm. It aims to share data, calculations, and service transparently over a scalable network of nodes. Since Cloud computing stores the data and disseminated resources in the open environment. So, the amount of data storage increases quickly. In the cloud storage, load balancing is a key issue. It would consume a lot of cost to maintain load information, since the system is too huge to timely disperse load. Load balancing is one of the main challenges in cloud computing which is required to distribute the dynamic workload across multiple nodes to ensure that no single node is overwhelmed. It helps in optimal utilization of resources and hence in enhancing the performance of the system. A few existing scheduling algorithms can maintain load balancing and provide better strategies through efficient job scheduling and resource allocation techniques as well. In order to gain maximum profits with optimized load balancing algorithms, it is necessary to utilize resources efficiently. The main aim of this paper is to discuss some of the existing load balancing algorithms in cloud computing environment.

Keywords : Cloud Computing, Load Balancing, Virtualization, Hypervisor

1 INTRODUCTION

Today, network bandwidth and hardware technology advance continuously to keep pace with the vigorous development of the Internet. Cloud computing is a new concept in distributed systems. Cloud computing is currently used many commodity nodes that can cooperate to perform a specific service together. It is currently used mainly in business applications in which computers cooperate to perform a specific service together. As network bandwidth and quality outstrip computer performance, various communication and computing technologies previously regarded as being of different domains can now be integrated. Thus, applications associated with network integration have gradually attracted considerable attention. In a cloud-computing environment, users have access to faster operational capability on the Internet, and the computer systems must

ISSN 2278-7763

have high stability to keep pace with this level of activity. In a cloud computing environment, users can access the operational capability faster with internet application and the computer systems have the high stability to handle the service requests from many users in the environment. However, the internet infrastructure is continuous grow that many application services can be provided in the Internet. In a distributed computing system, components allocated to different places or in separate units are connected so that they may collectively be used to greater advantage. In addition, cloud computing has greatly encouraged distributed system design and application to support user-oriented service applications. Furthermore, many applications of cloud computing can increase user convenience, such as YouTube. The Internet platform of cloud computing provides many applications for users, just like video, music et al. Therefore, how to utilize the advantage of cloud computing and make each task to obtain the required resources in the shortest time is an important topic.

2 Cloud Computing

In the cloud computing environment to allocate resources to achieve the purpose of the dynamic adjusting of resources. Cloud provides resources and a variety of services based on the needs of cloud users. Cloud computing provides a variety of computing resources,from servers and storage to enterprise applications like email, yahoo all delivered over the Internet. Fig 1 depicts this function. The Cloud delivers several resources that is flexible, scalable, secure, and available while saving corporations money and time. Cloud solutions are simple, don't require long term contracts and are easier to scale up and down as per the demand. Prefect planning and migration services are needed to ensure a successful implementation. Both Public and Private Clouds can be deployed together to leverage the best of both.

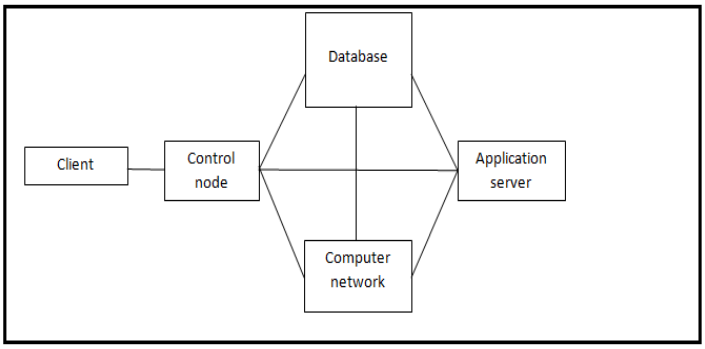

Fig 1. Working of Cloud with Server

2.1 Virtualization

Virtualization can be used for various system layers, such as hardware and operating system-level virtualization, higher end virtual machines. For virtual machine operating systems hypervisors are designed and it can be tightly coupled with operating system. Fig 2 explains this more. Virtual machine monitors can be implemented by multiplexing the virtual hardware onto the physical hardware safely and efficiently. Virtual CPUs on Physical CPUs and VM's Physical Memory on Machine's Memory. Virtualization layer enhance the virtual device, multiplexes and

ISSN 2278-7763

drives the physical device. The applications running on virtual infrastructure provides a layer of abstraction between computing.

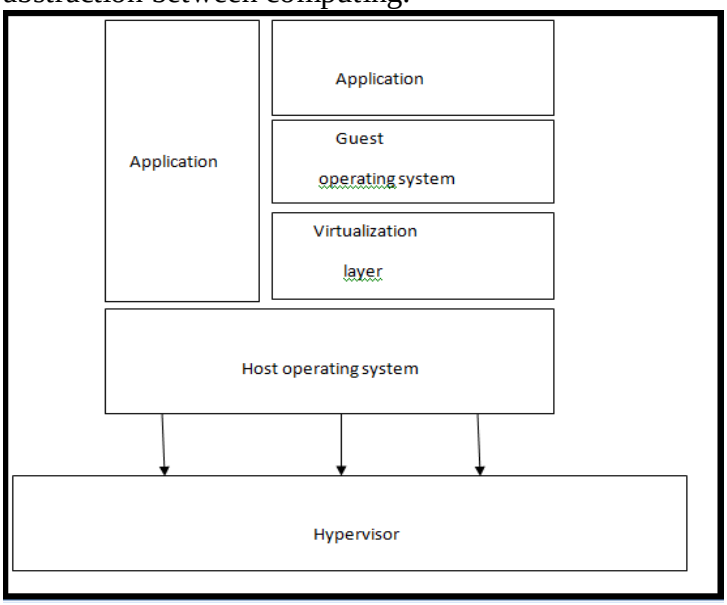

Fig 2: Hypervisor architecture

3 OVERVIEW OF LOAD BALANCING

Load balancing is the important concept in network. The load balancer accepts multiple requests from the client and distributing each of them across multiple computers or network devices based on how busy the computer or network device is. Load balancing helps to prevent a server or network device from getting overwhelmed with requests and helps to distribute the work. For example the client can send application request to the server at that time the server over loaded in another process the current process is wait for some time till the serve is idle. Here the client can wait. To avoid this first we check the utilization of the server and process the client request. The CPU utilization can properly do by load balancing algorithm. The load balancing algorithm which is dynamic in nature does not consider the previous state or behaviour of the system, that is, it depends on the present behaviour of the system.

3.1 GOALS OF LOAD BALANCING

- To improve the performance substantially
- To have a backup plan in case the system fails even partially
- To maintain the system stability
- To accommodate future modification in the system.

3.2 LOAD BALANCING CHALLENGES IN THE CLOUD COMPUTING

ISSN 2278-7763

Although cloud computing has been widely adopted. Research in cloud computing is still in its early stages, and some scientific challenges remain unsolved by the scientific community, particularly load balancing challenges :

- Automated service provisioning: A key feature of cloud computing is elasticity, resources can be allocated or released automatically. How then can we use or release the resources of the cloud, by keeping the same performance as traditional systems and using optimal resources?

- Virtual Machines Migration: With virtualization, an entire machine can be seen as a file or set of files, to unload a physical machine heavily loaded, it is possible to move a virtual machine between physical machines. The main objective is to distribute the load in a datacenter or set of datacenters. How then can we dynamically distribute the load when moving the virtual machine to avoid bottlenecks in Cloud computing systems?

- Energy Management: The benefits that advocate the adoption of the cloud is the economy of scale. Energy saving is a key point that allows a global economy where a set of global resources will be supported by reduced providers rather that each one has its own resources. How then can we use a part of datacenter while keeping acceptable performance?

- Stored data management: In the last decade data stored across the network has an exponential increase even for companies by outsourcing their data storage or for individuals, the management of data storage or for individuals, the management of data storage becomes a major challenge for cloud computing. How can we distribute the data to the cloud for optimum storage of data while maintaining fast access?

- Emergence of small data centers for cloud computing: Small datacenters can be more beneficial, cheaper and less energy consumer than large datacenter. Small providers can deliver cloud computing services leading to geo-diversity computing. Load balancing will become a problem on a global scale to ensure an adequate response time with an optimal distribution of resources.

4 LOAD BALANCING ALGORITHMS FOR CLOUD COMPUTING

Distribute workload of multiple network links to achieve maximum throughput, minimize response time and to avoid overloading. We use three algorithms to distribute the load. And check the performance time and cost

4.1 Round Robin Algorithm (RR)

It is the simplest algorithm that uses the concept of time quantum or slices Here the time is divided into multiple slices and each node is given a particular time quantum or time interval and in this quantum the node will perform its operations. The resources of the service provider are provided to the client on the basis of this time quantum. In Round Robin Scheduling the time quantum play a very important role for scheduling, because if time quantum is very large then

ISSN 2278-7763

Round Robin Scheduling Algorithm is same as the FCFS Scheduling. If the time quantum is extremely too small then Round Robin Scheduling is called as Processor Sharing Algorithm and number of context switches is very high. It selects the load on random basis and leads to the situation where some nodes are heavily loaded and some are lightly loaded. Though the algorithm is very simple but there is an additional load on the scheduler to decide the size of quantum and it has longer average waiting time, higher context switches, higher turnaround time and low throughput.

4.2 Equally Spread Current Execution Algorithm (ESCE)

In spread spectrum technique load balancer makes effort to preserve equal load to all the virtual machines connected with the data centre. Load balancer maintains an index table of Virtual machines as well as number of requests currently assigned to the Virtual Machine (VM). If the request comes from the data centre to allocate the new VM, it scans the index table for least loaded VM. In case there are more than one VM is found than first identified VM is selected for handling the request of the client/node, the load balancer also returns the VM id to the data centre controller. The data centre communicates the request to the VM identified by that id. The data centre revises the index table by increasing the allocation count of identified VM. When VM completes the assigned task, a request is communicated to data centre which is further notified by the load balancer. The load balancer again revises the index table by decreasing the allocation count for identified VM by one but there is an additional computation overhead to scan the queue again and again.

4.3 Throttled Load Balancing Algorithm (TLB)

In this algorithm the load balancer maintains an index table of virtual machines as well as their states (Available or Busy). The client/server first makes a request to data centre to find a suitable virtual machine (VM) to perform the recommended job. The data centre queries the load balancer for allocation of the VM. The load balancer scans the index table from top until the first available VM is found or the index table is scanned fully. If the VM is found, the load data centre. The data centre communicates the request to the VM identified by the id. Further, the data centre acknowledges the load balancer of the new allocation and the data centre revises the index table accordingly. While processing the request of client, if appropriate VM is not found, the load balancer returns -1 to the data centre. The data centre queues the request with it. When the VM completes the allocated task, a request is acknowledged to data centre, which is further apprised to load balancer to de- allocate the same VM whose id is already communicated.

4.4 Biased Random Sampling

A distributed and scalable load balancing approach that uses random sampling of the system domain to achieve self-organization thus balancing the load across all nodes of the system. Here a virtual graph is constructed, with the connectivity of each node (a server is treated as a node) representing the load on the server. Each server is symbolized as a node in the graph, with each in degree directed to the free resources of the server. The load balancing scheme used here is fully decentralized, thus making it apt for large network systems like that in a cloud. The performance is degraded with an increase in population diversity.

ISSN 2278-7763

4.5 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.

4.6 Max-Min Algorithm

Max-Min is almost same as the min-min algorithm except the following: after finding out minimum execution times, the maximum value is selected which is the maximum time among all the tasks on any resources. Then according to that maximum 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.

4.7 Token Routing

The main objective of the algorithm is to minimize the system cost by moving the tokens around the system. But in a scalable cloud system agents cannot have the enough information of distributing the work load due to communication bottleneck. So the workload distribution among the agents is not fixed. The drawback of the token routing algorithm can be removed with the help of heuristic approach of token based load balancing. This algorithm provides the fast and efficient routing decision. In this algorithm agent does not need to have an idea of the complete knowledge of their global state and neighbors working load. To make their decision where to pass the token they actually build their own knowledge base. This knowledge base is actually derived from the previously received tokens. So in this approach no communication overhead is generated.

4.8 Dynamic Load balancing algorithm

In a distributed system, dynamic load balancing can be done in two different ways: Distributed and Non-distributed. In the distributed one, the dynamic load balancing algorithm is executed by all nodes present in the system and the task of load balancing is shared among them. The interaction among nodes to achieve load balancing can take two forms: Cooperative and Non-cooperative. In the first one, the nodes work side-by-side to achieve a common objective, for example, to improve the overall response time, etc. In the second form, each node works independently toward a goal local to it, for example, to improve the response time of a local task. Dynamic load balancing algorithms of distributed nature, usually generate more messages than the non-distributed ones because, each of the nodes in the system needs to interact with every other node. A benefit, of this is that even if one or more nodes in the system fail, it will not cause

ISSN 2278-7763

the total load balancing process to halt, it instead would effect the system performance to some extent. Distributed dynamic load balancing can introduce immense stress on a system in which each node needs to interchange status information with every other node in the system. It is more advantageous when most of the nodes act individually with very few interactions with others.

In non-distributed type, either one node or a group of nodes do the task of load balancing. Non-distributed dynamic load balancing algorithms can take two forms: centralized and semi-distributed. In the first form, the load balancing algorithm is executed only by a single node in the whole system: the central node. This node is solely responsible for load balancing of the whole system. The other nodes interact only with the central node. In semi-distributed form, nodes of the system are partitioned into clusters, where the load balancing in each cluster is of centralized form. A central node is elected in each cluster by appropriate election technique which takes care of load balancing within that cluster. Hence, the load balancing of the whole system is done via the central nodes of each cluster. Centralized dynamic load balancing takes fewer messages to reach a decision, as the number of overall interactions in the system decreases drastically as compared to the semi distributed case. However, centralized algorithms can cause a bottleneck in the system at the central node and also the load balancing process is rendered useless once the central node crashes. Therefore, this algorithm is most suited for networks with small size.

5 EVALUATION OF LOAD BALANCING ALGORITHMS

In Sec.4 we have discussed about several cloud computing algorithms. The metrics on which the existing load balancing techniques have been measured are discussed below:

Throughput : This metric is used to estimate the total number of tasks, whose execution has been completed successfully. High throughput is necessary for overall system performance.
Overhead : Overhead associated with any load balancing algorithm indicates the extra cost involved in implementing the algorithm. It should be as low as possible.
Fault Tolerance : It measures the capability of an algorithm to perform uniform load balancing in case of any failure. A good load balancing algorithm must be highly fault tolerable.
Migration Time : It is defined as, the total time required in migrating the jobs or resources from one node to another. It should be minimized.
Response Time : It can be measured as, the time interval between sending a request and receiving its response. It should be minimized to boost the overall performance.
Resource Utilization : It is used to ensure the proper utilization of all those resources, which comprised the whole system. This factor must be optimized to have an efficient load balancing algorithm.
Scalability : It is the ability of an algorithm to perform uniform load balancing in a system with the increase in the number of nodes, according to the requirements. Algorithm with higher scalability is preferred.
Performance : It is used to check, how efficient the system is. This has to be improved at a reasonable cost, e.g., reducing the response time though keeping the acceptable delays.

ISSN 2278-7763

Table 1 Comparision of Load Balancing Algorithms

Metrics/ Techniques	Throughput	Overhead	Fault tolerance	Migration time	Response time	Resource Utilization	Scalability	Performance
Round Robin [22]	YES	YES	NO	NO	YES	YES	YES	YES
Dynamic Round Robin [26]	YES	YES	YES	YES	NO	YES	NO	NO
PALB [8]	YES	YES	YES	YES	YES	YES	NO	NO
Active Monitoring [23]	YES	YES	NO	YES	YES	YES	YES	NO
FAMLB [13]	YES	YES	YES	YES	NO	YES	YES	YES
Min-Min [6]	YES	YES	NO	NO	YES	YES	NO	YES
Max-Min [27]	YES	YES	NO	NO	YES	YES	NO	YES
OLB+LBMM [4]	NO	NO	NO	NO	NO	YES	NO	YES
Throttled [23]	NO	NO	YES	YES	YES	YES	YES	YES
Honeybee Foraging [28]	NO	NO	NO	NO	NO	YES	NO	NO
Active Clustering [28]	NO	YES	NO	YES	NO	YES	NO	NO
Biased Random Sampling [28]	NO	YES	NO	NO	NO	YES	NO	YES

6 CONCLUSION AND FUTURE SCOPE OF WORK

This paper explains the concept of load balancing, types of load balancing algorithms, general idea about dynamic load balancing algorithms and the different policies that can be used in it. Cloud Computing has widely been adopted by the industry, though there are many existing issues like Load Balancing, Virtual Machine Migration, Server Consolidation, Energy Management, etc. which have not been fully addressed. Central to these issues is the issue of load balancing, that is required to distribute the excess dynamic local workload evenly to all the nodes in the whole Cloud to achieve a high user satisfaction and resource utilization ratio. It also ensures that every computing resource is distributed efficiently and fairly. This paper presents a concept of Cloud Computing along with research challenges in load balancing. Cloud Computing is a vast concept and load balancing plays a very important role in case of Clouds. There is a huge scope of improvement in this area. In next level, we are going to compare all the load balancing algorithms, determine the performance of algorithm and decided which algorithm is best for reducing the overhead in the cloud computing environment.

References

[1] A.Khiyaita, M. Zbakh, H. El Bakkali and Dafir El Kettani, “Load Balancing Cloud Computing: State of Art” , 9778-1-4673-1053-6/12/$31.00, 2012 IEEE.

[2] A.M. Alakeel, “A Guide to dynamic Load balancing in Distributed Computer Systems”, International Journal of Computer Science and Network Security (IJCSNS), Vol. 10, No. 6, June 2010, pages 153-160.

ISSN 2278-7763

[3] G. Pallis, "Cloud Computing: The New Frontier of Internet Computing", IEEE Journal of Internet Computing, Vol. 14, No. 5, September/October 2010, pages 70-73.

[4] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing", EECS Department, University of California, Berkeley, Technical Report No., UCB/EECS-2009-28, pages 1-23, February 2009

[5] M. Nelson, B.-H. Lim, and G. Hutchins, "Fast Transparent Migration for Virtual Machines," Proc. USENIX Ann. Technical Conf., 2005.

[6] M. D. Dikaiakos, G. Pallis, D. Katsa, P. Mehra, and A. Vakali, "Cloud Computing: Distributed Internet Computing for IT and Scientific Research", IEEE Journal of Internet Computing, Vol. 13, No. 5, September/October 2009, pages 10-13.

[7] "Multicores in Cloud Computing: Research Challenges for Applications", Lizhe Wang†, Jie Tao‡, Gregor von Laszewski†, Holger Marten – Journal of computers, vol. 5,no. 6, June 2010.

[8] Mladen A. Vouk, Cloud Computing Issues, Research and Implementations, Proceedings of the ITI 2008 30th Int. Conf. on Information Technology Interfaces, 2008, June 23-26.

[9] Wayne Jansen Timothy Grance" Guidelines on Security and Privacy in Public Cloud Computing" NIST Draft Special Publication 800-144.

[10] Zhong Xu, Rong Huang,(2009)"Performance Study of Load Balanacing Algorithms in Distributed Web Server Systems", CS213 Parallel and Distributed Processing Project Report.

