LOAD BALANCING IN CLOUD COMPUTING

Ayush Agrawal[1], Devesh Katiyar[2], Gaurav Goel[3]

[1]Ayush Agrawal Department of Computer Science, Dr. Shakuntala Misra National Rehabilitation University, Lucknow, India
[2,3]Assistant Professor, Faculty of Computer & Information Technology, Dr. Shakuntala Misra National Rehabilitation University, Lucknow, India

---

Abstract - Cloud computing offers us the way to share data and provides many resources to users. There is an important issue in cloud computing which is Load Balancing. With remarkable gain in users and their need of requests on the cloud computing platform, with lots of usage of resources became a severe concern. Load balancing gives user gratification and resource utilization ratio by guaranteeing an efficient and impartial allocation of all type of resources. Load balancing method is used to distribute tasks from high loaded resources to low loaded or with in the constant resources There are many algorithms which gives the user more satisfying experience on the cloud services. In this paper, we used different type of algorithms to solve the issues in load balancing.

Keywords: Cloud Computing, Load balancing, Load balancing Algorithms.

1. Introduction

Cloud Computing has become the important requirement for the IT companies. It has moved hardware resources and software services that are on the internet rather than the resources which are present at customer end. The user only has to pay for that service only which they use on cloud. Because of its satisfying services many organizations are seeing forward to use it. Due to the high demand of services provided on cloud allows the user to use the technology, continuing deployment, and the development of the many organizations. The Services in Cloud allows many organizations to increase their investments that is related with costly data storages and applications and minimizing these expenses which are required to use the services on cloud. Today there are many cloud services providers which are available like- Cloud Stack, OpenStack, EMC2, AWS Amazon, Google Cloud, Open Nebula etc.

I. Cloud computing has different types of features:

- On request service- The user can request for any services on cloud and can assess anytime.
- Broad Network Access- There are many services present on the cloud which can access over internet. The cloud services are generally get through local networks or on any standard devices.
- Fast Elasticity- The services on this allows less usage of running of workloads which required large number of servers but only for a less period of time.
- Resource Sharing- There are different models by which users can share their resources which are provided by the service provider. All the resources are dynamically allocated and reallocated based on the user's request.

II. There are different types of challenges in Cloud Services:

1. Interoperability
2. Lack of Experience
3. Performance Monitoring
4. Cost Management
5. High Dependence On Internet

Load balancing is one of the important issue in cloud computing. Load balancing involves the process of distribution of different load at various nodes to improve job response time and system utilization. It also helps in the situation when any node is overloaded while other are less loaded or become idle. Load balancing makes sure that all the systems or nodes performs uniformly and complete similar amount of work simultaneously. Since the demand and users for cloud services escalates, the need of load balancing rises proportionally. It helps the users to incur high resource utilization and user satisfaction. Load Balancing is responsible to map all the work set for the cloud to free the resources and make them available to enhance the response time and provide better utilization of the resources. Multiple servers or multiple resources capable of fulfilling user demands are required to reach load balancing. Even if one or more component fails to provide service, load balancing

enables to keep providing serviced by splitting the load on the other available resources, it helps supplying the requests of users without fail. It makes sure every component/resource is lay out equally to provide better responses. It reduces response time, provides scalability and restricts blockages. The figure below explains the development of load balancing in Cloud Computing.

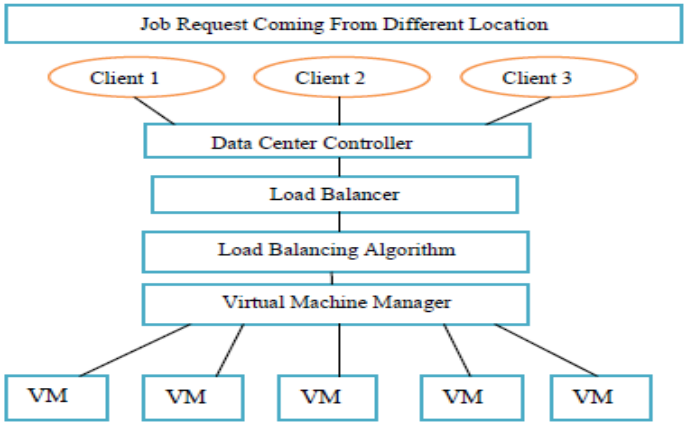

Figure 1 Load Balancing in Cloud Computing

2. Load balancing classification:

Load balancing is basically classified into two categories: static and dynamic load balancing:

1) **Static algorithm: -** This approach is mostly defined in the design or implementation of the system. Static load balancing algorithms split the traffic equally between all the servers.

2) **Dynamic algorithm:** - This approach measured only the current state of the system during load balancing results. A dynamic algorithm rearranges the processes among processors during execution time, Dynamic approach is more appropriate for widely distributed systems such as cloud computing. Major disadvantage of Dynamic algorithms is the run-time overhead due to the transmission of load information among processor and decision-making for the variety of processes and communication stays associated with the task rearrangement itself. Dynamic load balancing algorithms can be central or distributed, depending on whether the accountability for task of global dynamic scheduling should actually reside in the single processor (centralized) or the work involved in making conclusions should be physically distributed among processors.

Dynamic load balancing algorithm have two kinds.

Which are distributed approach and non-distributed (centralized) approach. It is defined as following:

a) **Centralized approach**: - In centralized approach, only a solo node is responsible for working and distribution within the entire system. Further all nodes are not responsible for this.

b) **Distributed approach:** - In distributed approach, each node individualistically builds its individual load vector. Vector assembling the load data of other nodes. All conclusions are made nearby using local load vectors. Distributed approach is more appropriate for generally distributed systems such as cloud computing.

3. Main goals of load balancing algorithms

1. Cost effectiveness: Load balancing helps in offer improved system performance at very lower cost.

2. Scalability and flexibility: The system for which load balancing algorithms are executed may change their size after some time. So these type of algorithm must handle these type's conditions. So that the algorithm can be scalable and flexible.

3. Priority: Arrangement of the resources or jobs needs to be done. So that higher significance jobs get well chance to execute.

4. Load Balancing Algorithms:

a) Round Robin Algorithm:

Round Robin algorithm uses the method of time slice mechanism. In the Process of this Type of mechanism time is spread into several slices and specific node is given a specific time interval or time quantum and because of this quantum the node will perform its processes. The resources for this type of service provider are delivered to the client on the basis of this time quantum. This algorithm just assigns the jobs in round robin technique which doesn't affect the load on different machines. As of result, at any moment some node may have heavy load and other node have no request.

In Round Robin Algorithm the time quantum has a very significant role for scheduling, because if the process of time quantum is very large then Round Robin Scheduling Algorithm became same as of the FCFS Scheduling. If the time quantum is very small, then the method of Round Robin Algorithm is called as Processor Sharing Algorithm and quantity of context switches becomes very high.

b) Equally Spread Current Execution Algorithm (ESCE):

In the Process of spread spectrum method, the load balancer makes effort to reserve the same load on all types of virtual machines connected with the data centre. Load balancer keeps an index table of Virtual machines along with the number of requests currently allocated to the Virtual Machine. If any type of request arises from the data centre to allocate the new VM, it tests the index table for minimum loaded VM. If their case arises where more than one VM is found than first known VM is selected for considered for handling the request of the client/node and the load balancer also returns the VM Id to the respective data centre controller. The data centre transfers the request to the VM recognized by that id. Now the data centre reviews the index table by increasing the share count of identified VM. When the allocated task is accomplished by the VM a request is shifted towards the data centre which is further reported by the load balancer. The figure 2 shows the Function of ESCE algorithm.

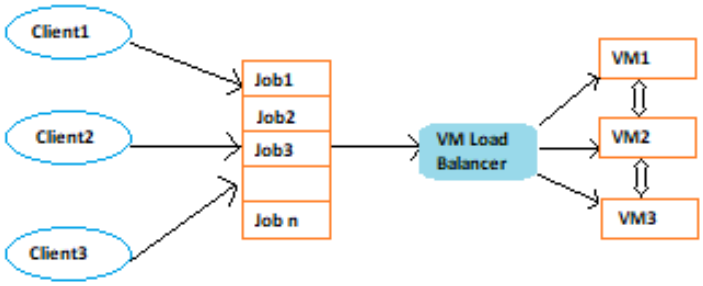

Figure 2 ESCE Algorithm

c) Min-Min Algorithm:

These type of algorithm starts with a task set which are originally not allocated to any of the nodes. At first, the least completion time is considered for all the available nodes. Then, the task which has the least expected completion time is selected and allocated to the node with minimum execution time. Now from the task set the task is removed. This process is continual until all types of the tasks have been allocated to the same nodes. Hence the algorithm become better if the larger task is smaller than the small task.

d) Max-Min Algorithm:

This type of max-min algorithm is just like min-min algorithm. Max-Min algorithm starts with the set of all the submitted tasks in the task-set which are originally unassigned to some node. At start, the least completion time for all types of available tasks is assessed. Then the process with has the extreme execution time is then used and allotted to the resource with minimum response time.

This algorithm overtakes the Min-Min algorithm where short process is in large in numbers as compared to long ones.

e) Throttled Algorithm:

This types of algorithm works by discovering the suitable virtual machine for allocation of a particular task. These type of algorithm the load balancer keeps an index table of virtual machines along with their states. The customer first makes a demand to Data Centre to find an appropriate virtual machine to perform required particular operation. The Data Centre obtains the request from customer for the distribution of Virtual Machine. Then, Data Centre inquiries the load balancer for distribution of Virtual Machine. The load balancer finds the index table from above until the first accessible VM is found or index table is searched fully.

If VM starts, then the VM Id is forwarded to the Data Centre. Then the Data Centre connects the demand to the VM recognised by the Id. Later, the Data Centre recognizes the load balancer of the latest allocation and then the data Centre studies the index table thoroughly.

During dealing out with the request of the customer, if VM is not available, then the load balancer gives value -1 to the data Centre. Then the Data Centre lines the request until the next accessibility of Virtual Machine. When the VM completes the processing request, it give results to the data Centre and recognizes load balancer for the de-allocation of VM. Then the Load balancer informs the allocation table by reducing the allocation for VM by 1.

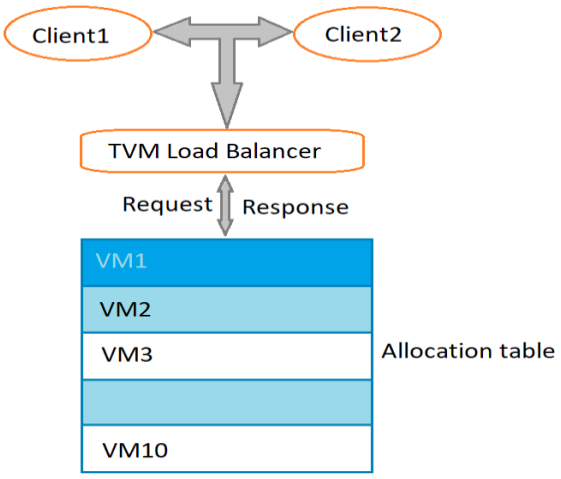

Figure 3 Throttled Algorithm

f) Ant Colony Optimization Algorithm:

This type of algorithms behaves similar like the behaviour of real ants. It is basically based on the capacity of ants to search an ideal path from shell to the source. In Ant Colony Optimization algorithm, when the demand is instigated, the ant starts its movement. Ants instigate from the root node and starts moving from one node to another node and also check whether the node is or under loaded or overloaded. When ants travel towards the network, they inform the pheromone table which saves the data of each node's operation.

g) Honeybee Foraging Behaviour:

This type of algorithm is a nature inspired Algorithm for self-organization. Honeybee attains over-all load balancing via local server actions. The performance of the system is improved with the increase in system range. The main drawback is that throughput will not increase with the increase in size of the system. When the various population of service types is required then this type of algorithm is appropriate.

5. Performance Measurement for Load Balancing:

1. Throughput: - It is used to analyse all tasks whose implementation has been completed. The enactment of any system is enhanced when throughput is max.

2. Fault Tolerance: -It means regaining from failure. The load balancing should be a better fault tolerant technique.

3. Response Time: - It's the amount of time that is engaged by a definite load balancing algorithm to response a task within a system. This limitation must be reduced for better performing of a system.

4.Overhead: It occurs because of more time usage in traveling from one machine to another. It should be minimized for efficient working of an algorithm.

5.Resource utilization: It is used to see the utilization of resources in a system. Resources must be utilized optimally by a load balancing algorithm.

6. Scalability: - It is the capability of an algorithm to execute Load balancing for any finite number of nodes of a system. This metric must be improved for the system.

7.Performance: It is the efficiency of system during the load balancing. Performance must be improved by educing response time, by increasing throughput and at a reasonable cost.

Table 1- Comparison of different load balancing algorithms based on Metric Enrollment

Parameters Algorithms	Throughput	Fault Tolerance	Response Time	Overhead	Resource Utilization	Scalability	Performance
Round Robin	YES	NO	YES	YES	YES	YES	YES
ESCE	NO	NO	NO	NO	YES	YES	NO
Min	YES	NO	YES	YES	YES	NO	YES
Max	YES	NO	YES	YES	YES	NO	YES
ALO	YES	NO	NO	YES	YES	NO	NO
Honey Bee	YES	NO	NO	NO	YES	NO	NO
Throttled	NO	YES	YES	NO	YES	YES	YES

6. CONCLUSION:

In this paper, we analysed different algorithms in cloud computing for load balancing. Cloud computing has generally been implemented by the industry, through there are many types of existing issues like Server Consolidated, Energy Management, Load balancing, Virtual machine Migration, etc. The main issue in all of these is load balancing, that is necessary to allocate the excess dynamic local workload equally to all the nodes in the entire cloud to get higher customer fulfilment and resource consumption proportion. In this paper, we have analysed and equated different dynamic and static load balancing algorithms in cloud computing such as, Max-Min, Ant Colony Optimization Algorithm, round robin, Honeybee, Min-Min, Throttled Algorithm etc. considering the features like overhead, fault tolerance, throughput, scalability etc.

7. REFERENCES

[1] R. Shimon ski, Windows 2000 And Windows Server 2003, Clustering and Load Balancing Emeryville, McGraw-Hill Professional Publishing, CA, USA, 2003.

[2] R. Mata-Toledo, and P. Gupta, "Green data centre: how green can we perform", Journal of Technology Research, Academic and Business Research Institute, Vol. 2, No. 1, May 2010.

[3] Ali M Alakeel, "A Guide to Dynamic Load Balancing in Distributed Computer Systems", International Journal of Computer Science and Network Security, Vol. 10 No. 6, June 2010.

[4] Parin. V. Patel, Hitesh. D. Patel, Pinal. J. Patel, "A Survey on Load Balancing in Cloud Computing" IJERT, Vol. 1, Issue 9, November 2012.

[5] Sahu, Yatendra and Pateriya, RK, "Cloud Computing Overview with Load Balancing Techniques", International Journal of Computer Applications, 2013,vol. 65, Sahu2013.

[6] S. K. Garg, C. S. Yeob, A. Anandasivamc, and R. Buyya, "Environment-conscious scheduling of HPC applications ondistributed Cloud-oriented data centers", Journal of Parallel and Distributed Computing, Elsevier, Vol. 70, No. 6, May 2010, pages 1-18.

[7] O. Elzeki, M. Reshad, M. Elsoud, "Improved max-min algorithm in cloud computing, International Journal of Computer Applications" vol 50 (12) (2012)pages 22–27..