Review on Load Balancing Model in Cloud Computing

Varsha S. Salunkhe
Computer Department
MET BKC Adgoan, Nashik Savitribai Phule Pune University, Maharashtra, India.

ABSTRACT
In the cloud computing environment, load balancing plays a vital role in improving performance of the systems. It is beneficial to use load balancing in cloud computing. Server overloading occurs due to tremendous use of internet in day to-days world. This article studies on surveys on load balancing that describes different algorithms for balancing the workload for the cloud using optimal resources for better efficiency and performance. Good load balancing techniques are required for the better management of available resources. This article discusses about the various load balancing algorithms and strategies used in a cloud computing environment.

Keywords
Cloud computing, Load balancing, Public cloud, Cloud Partition

1. INTRODUCTION
Cloud Computing as defined by National Institute of Standard and Technology is a model for facilitating suitable, on - demand network access to shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction[14]. In the world of internet, cloud is formed from logical grouping of systems depending upon their locations. Cloud computing mainly deals with computation, software, data access and storage services that may not require end-user knowledge of the physical location and configuration of the system that is delivering the services [2]. Cloud services are widely used now-a-days. Cloud computing offers infrastructure, platform and software services to the customers under the pay per usage model.

Load balancing is the technique of distributing the entire load equally to all individual nodes in a group of systems. Load balancing schemes avoids overloading of servers and improves response time. Load balancing in computer networks is a technique used to distribute workload across several network links of computers[12]. Load balancing simplifies networks and resources by providing a maximum throughput with minimum time, thus it helps to improve performance by optimized use of available resources and reduces latency and response time. Multiple resources are distributed to multiple servers or nodes so that load is distributed evenly to carry out client's request. Load balancing helps to accomplish a high user satisfaction and better resource utilization. When one or more components of any service fail, load balancing facilitates continuation of the service by implementing fair-over, that is, it helps in provisioning and deprovisioning of instances of applications without any failure. It distributes each computing resource equally and efficiently. There are key issues other than consumption of resources and conservation of energy in cloud computing. However, resource consumption can be kept lower with proper load balancing to reduce cost and make enterprise greener. Scalability, one of the very important features of cloud computing, is also enabled by load balancing. Availability of cloud has also an important impact on load balancing. The network structure or topology should be taken into account when creating the logical rules for the load balancer. [5]

2. RELATED WORK
Many studies and analysis have been performed on load balancing for the cloud environment. There are many load balancing algorithms, such as Round Robin, Particle Swarm Optimization algorithm, Equally Spread Current Execution Algorithm, Self-organized Load Balancing algorithm, Ant Colony algorithm etc. However, load balancing in the cloud is still a challenge that gives rise to new algorithms in order to achieve a better load balancing model. Adler[13] described the load balancing in cloud computing by introducing the tools and techniques commonly used for load balancing in the cloud. The Round Robin algorithm is used because it is fairly simple. To improve the performance of the algorithm variants in round robin algorithm is introduced by Pooja Samal[6]. Some of the existing load balancing algorithms are described as follows:

2.1 Ant Colony Optimization
Main purpose of this approach is to balance the load of nodes efficiently and equally. This algorithm performs task of identification of nodes by the ants and tracing its path consequently in search of different types of nodes. This approach follows the foraging behavior of ants. In original approach, each ant build their own individual result set and later on it builds into a complete solution. However, in this modified approach the ants continuously update a single result set rather than updating their own result set so that the solution set is formed by continuously updating result set.. This algorithm can be used for better working of large networks and better utilization of available resources. The main benefit of this approach lies in its detections of overloaded and underloaded nodes. The other advantage of the approach lies in the fact that the task of each ant is specialized rather than being general [3].

2.2 Round Robin Algorithm
It is the simplest algorithm to distribute load among nodes. A simple Round Robin algorithm requires only information about the names of the nodes. Pooja Samal[6] analyzed variants in Round Robin algorithms for load balancing in cloud computing to avoid overloading and underloading of servers . The scheduling algorithm is based on criteria's like Context-switch, Throughput, CPU utilization, Turnaround time, Waiting time, Response time. This algorithm is easy to implement and shows better response time as compared to other algorithms.

2.3 PSO Algorithm

This algorithm overcomes the drawback of Honey Bee Algorithm where task is assigned in first fit manner. In PSO(Particle Swarm Optimization) algorithm, virtual machines assigns the task in best fit manner i.e. task will check all the virtual machine and assign the task to proper virtual machine which will have least memory wastage in order to achieve QoS. It minimizes the task completion time and task response time and maximizes throughput. Threshold value s used to find the overloaded servers and then load is balanced [7].

2.4 ESCE Algorithm

Equally spread current execution algorithm [12] which assigns a job to each node with priority .It makes use of spread spectrum technique in which it distributes the load over various nodes by checking its load size. Once the available resource (virtual machine) size and the size of the job matches, then immediately the job scheduler allocates appropriate virtual machine or resource to the job in a queue. This algorithm enhances response time and processing time. The equal distribution of jobs is done. After the process has completed its load distribution, there is no such virtual machines left that are underutilized. Due to this advantage, there is a reduction in the cost of virtual machine as well as the cost of data transfer.

2.5 Cooperative Load Balancing

In this approach, system acts as to play cooperative game among themselves. It works on static load balancing problem. The main aim was to derive a fair and optimal allocation scheme using Nash Bargaining Solution (NBS) which provides a Pareto optimal allocation. NBS provides optimality and fairness to allocation. It consider a single class job distributed system that consists of many heterogeneous machines It considers a game in which each computer is a player and it must minimize the expected execution time of jobs that it processes.

2.6 Honey Bee Behavior Inspired Load Balancing Algorithm

Dhinesh et al. [8] proposed an algorithm named honeybee behavior inspired load balancing algorithm. It achieves global load balancing through local server actions Load is balanced across the virtual machines for maximizing the throughput by modelling the foraging behavior of honey bees that includes the method to find and reap food. In bee hives, a class called the scout bees which search for food sources, when they find the food, they come back to the beehive to advertise this news by using a dance called waggle/tremble/vibration dance. Another class of bees is Forager bees. They follow the Scout Bees to the location that they found food and then begin to reap it. After that they return to the beehive and do a tremble or vibration dance to other bees in the hive giving an idea of how much food is left. The tasks removed from the overloaded VMs act as Honey Bees. Current workload of all available VMs can be calculated based on the information received from the data center. It is best suited for the conditions where different types of service is required.

2.7 Self-organized Load Balancing Algorithm

Giuseppe[9] has presented a new technique for load balancing in which highest capacity node act as super peers. These nodes offer different types of services. The goal of the algorithm was to balance the queues of service requests for neighboring peers as uniformly as possible. To build a system that has ability to direct incoming requests for the various hosted services to those nodes that can efficiently fulfill them, one option is to build the system as an overlay network, in which the nodes hosting instances of each of many different types of services can self-organize as "virtual clusters", and efficiently load-balance incoming requests amongst themselves At first level, algorithm find out the capacity of every peer, i.e., the amount of service requests that peer is able to fulfill in a client time unit. This way, super peers are well positioned to effectively balance their neighbors' request queues.

2.8 Active Clustering Load Balancing Algorithm

In Active Clustering algorithm [11], a group of similar nodes operate for distributing loads among themselves. A node initiates the process selecting a different node called the matchmaker node from its neighbors such that it does not match the former one. The matchmaker node then forms a link between one of its neighbors similar to start node. The matchmaker node then disconnects the connection between itself and the start node and this process is followed repeatedly. The performance of the system is improved with high availability of resources, thereby increasing the throughput.

2.9 Load Balancing using Cloud Partitioning

Junjie Pang[1] proposed a concept of better load balancing model to handle the workload on huge public cloud. He introduced a strategy based upon cloud partitioning concept which divides a huge cloud into partitions which are geographically distributed. Based upon the incoming job or workload, the best suitable strategy is selected to get the optimum response time and efficient results.

Initially cloud is partitioned into small divisions forming cluster which provides different types of services. Client requests a job to be executed. This job first enters the load balancer that is, main controller where resources required for the job are identified. Then, load balancer searches for the best suitable partition for the job and checks whether it is available or not. If the status of the partition is normal, then the job is assigned to the particular partition and then the partition balancers choose the appropriate node to process the job. Load status information of each node is updated at partition balancers through which it can select the appropriate node for job execution. Similarly, load status information of partition balancers are updated and maintained at main controller. The node load degree is related to various static parameters and dynamic parameters. The static parameters include the number of CPU's, the CPU processing speeds, the memory size, etc. Dynamic parameters are the memory utilization ratio, the CPU utilization ratio, the network bandwidth, etc.

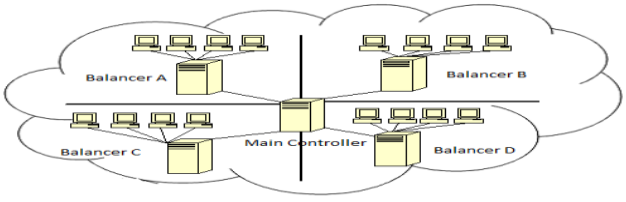

FIGURE 1: Relationship between the main controllers, servers acting as balancers and nodes

Load balancing strategy distributes each computing resource equally and efficiently. Load balancing scheme is introduced for providing more flexibility and performance in the system. Balancing load based upon the incoming job requests ensures that the load is distributed to the nodes where job is best handled. Hence, a switch mechanism for partitioning a cloud is found to be useful to choose different strategies for different situations and service requests.

3. SUMMARY OF VARIOUS LOAD BALANCING ALGORITHMS

Various Surveys has been done on different algorithms which are used for better load balancing in cloud environment. These algorithms offer different features. Some of the features are listed below:

Table 1. Load Balancing Algorithms and Features

Algorithm	Features
Ant Colony Optimization	• Dynamic • Optimal Resource utilization • Ability to handle failover, jobs and machines
Cooperative Load Balancing	• Pareto optimal allocation • Minimizes execution time
PSO Algorithm	• Improves utilization of memory • Minimizes task response time and task completion time
ESCE Algorithm	• Response time is high • Processing time is reduced
Honey Bee Behavior	• Maximizing the throughput • Minimum Waiting Time • Low Overhead
Self-organized Load Balancing algorithm	• Reduced time for client requests
Active Clustering	• Performs better with high resources • Utilizes the increased system • Resources to increase throughput
Load Balancing using Cloud Partitioning	• Efficient resource utilization • Dynamic • High response time

4. CONCLUSION

Load balancing is still a challenge in cloud computing hence it is required to develop a strategy that is more efficient and distributes load more evenly. Various techniques and algorithms are been discussed here for good load balancing. Switch mechanism used in distributing the incoming jobs to appropriate partition helps in better load balancing. Different surveys are analyzed here to discover a better load balancing models among existing ones. This concept can be extended by providing security to the data that is supposed to be balanced.

5. REFERENCES

[1] Gaochao, Xu Junjie Pang, Xiaodong Fu, "A load balancing model based on cloud partitioning for the public clouds," IEEE. Trans On cloud computing year 2013

[2] R. W. Lucky, "Cloud computing", IEEE Journal of Spectrum, Vol. 46, No. 5, May 2009, pages 27-45.

[3] K. Nishant, P. Sharma, V. Krishna, C. Gupta, K. P. Singh, N. Nitin, and R.Rastogi, "Load balancing of nodes in cloud using ant colony optimization", in Proc. 14th International Conference on Computer Modelling and Simulation (UKSim), Cambridgeshire, UK, Mar. 2012, pp. 28-30.

[4] D. Grosu, A. T. Chronopoulos, and M. Y. Leung, "Load balancing in distributed systems: An approach using cooperative games", in Proc. 16th IEEE Intl. Parallel and Distributed Processing Symp., Florida, USA, Apr. 2002, pp. 52-61.

[5] Z. Chaczko, V. Mahadevan, S. Aslanzadeh, and C. Mcdermid, "Availability and load balancing in cloud computing", International Conference on Computer and Software Modeling, Singapore, 2011.

[6] Pooja Samal, Pranati Mishra, "Analysis of variants in Round Robin Algorithms for load ba;ancing in Cloud Computing", IJCSIT, vol.4(3),2013.

[7] Anju Baby, "Load Balancing in Cloud Environment using PSO algorithm "International Journal for applied science and engineering technology,vol.2, Issue IV ,April 2014,ISSN:2321-9653

[8] Dhinesh Babu L.D, P. VenkataKrishna, "Honey bee behavior inspired load balancing of tasks in cloud computing environments", Applied Soft Computing 13 (2013) 2292–2303.

[9] Giuseppe Valetto, Paul Snyder, Daniel J. Dubois, Elisabetta Di Nitto and Nicolo M. Calcavecchia, "A self-organized load balancing algorithm for overlay based decentralized service networks".

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

[11] Karanpreet Kaur, Ashima Narang , Kuldeep Kau, "Load Balancing Techniques of Cloud Computing", IJMCR, Volume 1 issue 3 April 2013 ISSN 2320-7167.

[12] Brain Adler, "Adler, Load balancing in the cloud: Tools, tips and techniques", white papers, www.rightscale.com

[13] The NIST Definition of Cloud Computing, NIST Special Publication 800-145, www.nist.gov/itl/csd/cloud-102511.cfm