International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 2, February- 2013

A Novel Survey on Load Balancing in Cloud Computing

Aarti Khetan
Amity University
Noida, India

Vivek Bhushan
Amity University
Noida, India

Subhash Chand Gupta
Amity University
Noida, India

Abstract- The availability of Virtual Machines (VMs) in cloud is one major concern of cloud computing. Cloud Computing is nothing but a collection of computing resources and services pooled together and is provided to the users on pay-as-needed basis. Sharing of the group of resources may initiate a problem of availability of these resources causing a situation of deadlock. One way to avoid deadlocks is to distribute the workload of all the VMs among themselves. This is called load balancing. The goal of balancing the load of virtual machines is to reduce energy consumption and provide maximum resource utilization thereby reducing the number of job rejections. The aim of this paper is to discuss the concept of load balancing in cloud computing and how it improves and maintain the performance of cloud systems.

Keywords: Cloud Computing, Load Balancing, Deadlock

I. Introduction

Cloud Computing or the future of next generation computing provides its clients with a virtualized network access to applications and or services. No matter from wherever the client is accessing the service, he is automatically directed to the available resources.

There are situations when our system gets hanged up or it seems to take few decades for pages to come out of printer. All this happens because there is a queue of requests waiting for their turn to access resources which are shared among them. Further these requests cannot be serviced as the resources required by each of these requests are held by another process or request by virtual machines. One cause for all these problems is called deadlock.

Load balancing is a new approach that assists networks and resources by providing a high throughput and least response time [7].

The real world example of load balancing can be a website which has thousands of users at the same time. If not balanced then the users have to face the problem of timeouts, response delays and long processing time.

The solutions involve making use of duplicate servers to make the website available by balancing the network traffic.

The following sections discusses about the concept of load balancing, its needs and goals, types. After that it discusses the conclusion and the references.

II. Load Balancing

Load balancing is the process of improving the performance of the system by shifting of workload among the processors. Workload of a machine means the total processing time it requires to execute all the tasks assigned to the machine [6].

Load balancing is done so that every virtual machine in the cloud system does the same amount of work throughout therefore increasing the throughput and minimizing the response time [4].

Load balancing is one of the important factors to heighten the working performance of the cloud service provider. Balancing the load of virtual machines uniformly means that anyone of the available machine is not idle or partially loaded while others are heavily loaded. One of the crucial issue of cloud computing is to divide the workload dynamically.

The benefits of distributing the workload includes increased resource utilization ratio which further leads to enhancing the overall performance thereby achieving maximum client satisfaction [1].

III. Why Balancing is needed in Cloud

We can balance the load of a machine by dynamically shifting the workload local to the machine to remote nodes or machines which are less utilized. This maximizes the user satisfaction, minimizing response time, increasing resource utilization, reducing the number of job rejections and raising the performance ratio of the system.

Load balancing is also needed for achieving *Green computing* in clouds [5]. The factors responsible for it are:

a) ***Limited Energy Consumption:*** Load balancing can reduce the amount of energy consumption by avoiding over hearting of nodes or virtual machines due to excessive workload [5].

b) ***Reducing Carbon Emission:*** Energy consumption and carbon emission are the two sides of the same coin. Both are directly proportional to each other. Load balancing helps in reducing energy consumption which will automatically reduce carbon emission and thus achieve Green Computing [5].

IV. Goals of Load Balancing

Goals of load balancing as discussed by authors of [8], [9] include:

- Substantial improvement in performance
- Stability maintenance of the system
- Increase flexibility of the system so as to adapt to the modifications.
- Build a fault tolerant system by creating backups.

V. Classification of Load Balancing Algorithms

Based on process origination, load balancing algorithms can be classified as [1], [2], [3] :

a) ***Sender Initiated:*** In this type of load balancing algorithm the client sends request until a receiver is assigned to him to receive his workload i.e. the sender initiates the process.

b) ***Receiver Initiated:*** In this type of load balancing algorithm the receiver sends a request to acknowledge a sender who is ready to share the workload i.e. the receiver initiates the process.

c) ***Symmetric:*** It is a combination of both sender and receiver initiated type of load balancing algorithm.

Based on the current state of the system there are two other types of load balancing algorithms [1], [2], [10].

1. Static Load Balancing:

Static load balancing algorithms require aforementioned knowledge about the applications and resources of the system [11], [12], [13]. The decision of shifting the load does not depend on the current state of the system.

The performance of the virtual machines is determined at the time of job arrival. The master processor assigns the workload to other slave processors according to their performance. The assigned work is thus performed by the slave processors and the result is returned to the master processor.

Static load balancing algorithms are not preemptive and therefore each machine has at least one task assigned for itself. Its aims in minimizing the execution time of the task and limit communication overhead and delays.

This algorithm has a drawback that the task is assigned to the processors or machines only after it is created and that task cannot be shifted during its execution to any other machine for balancing the load.

The four different types of Static load balancing techniques are Round Robin algorithm, Central Manager algorithm, Threshold algorithm and randomized algorithm.

2. Dynamic Load Balancing:

In this type of load balancing algorithms e.g., [14]-[20], the current state of the system is used to make any decision for load balancing.

It allows for processes to move from an over utilized machine to an under utilized machine dynamically for faster execution.

This means that it allows for process preemption which is not supported in Static load balancing approach. An important advantage of this approach is that its decision for balancing the load is based on the current state of the system which helps in improving the overall performance of the system by migrating the load dynamically.

Dynamic Load Balancing Policies or Strategies:

The different policies as described in [1], [3] are as follows:

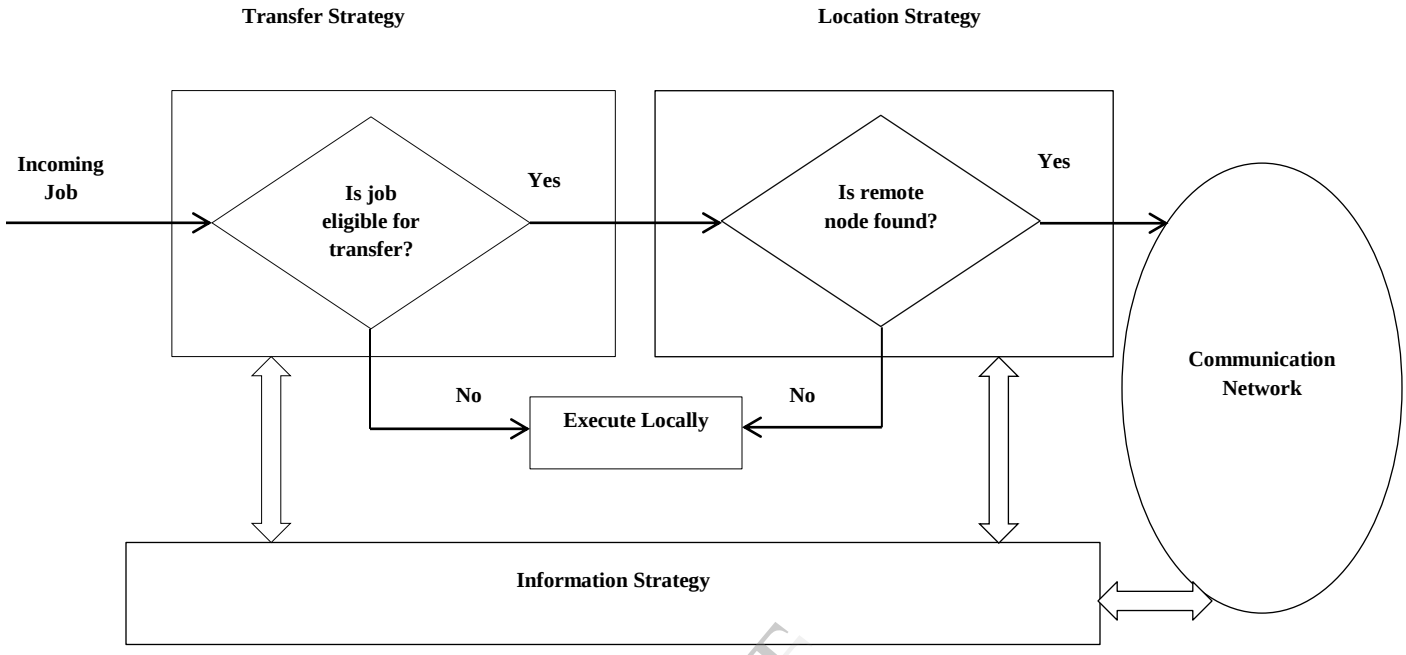

Figure1. Interaction between different components of Dynamic Load Balancing Algorithm [1]

1. ***Location Policy:*** The policy used by a processor or machine for sharing the task transferred by an over loaded machine is termed as Location policy.

2. ***Transfer Policy:*** The policy used for selecting a task or process from a local machine for transfer to a remote machine is termed as Transfer policy.

3. ***Selection Policy:*** The policy used for identifying the processors or machines that take part in load balancing is termed as Selection Policy.

4. ***Information Policy:*** The policy that is accountable for gathering all the information on which the decision of load balancing is based id referred as Information policy.

5. ***Load estimation Policy:*** The policy which is used for deciding the method for approximating the total work load of a processor or machine is termed as Load estimation policy.

6. ***Process Transfer Policy:*** The policy which is used for deciding the execution of a task that is it is to be done locally or remotely is termed as Process Transfer policy.

7. ***Priority Assignment Policy:*** The policy that is used to assign priority for execution of both local and remote processes and tasks is termed as Priority Assignment Policy.

8. ***Migration Limiting Policy:*** The policy that is used to set a limit on the maximum number of times a task can migrate from one machine to another machine.

The two different types of Dynamic load balancing techniques are Local Queue Algorithm and Central Queue algorithm.

VI. Qualitative Metrics for Load Balancing

The different qualitative metrics or parameters that are considered important for load balancing in cloud computing [25] are discussed as follows:

1. ***Throughput:*** The total number of tasks that have completed execution is called throughput. A high throughput is required for better performance of the system.

2. ***Associated Overhead:*** The amount of overhead that is produced by the execution of the load balancing algorithm. Minimum overhead is expected for successful implementation of the algorithm.

3. ***Fault tolerant:*** It is the ability of the algorithm to perform correctly and uniformly even in conditions of failure at any arbitrary node in the system.

4. ***Migration time:*** The time taken in migration or transfer of a task from one machine to any other machine in the system. This time should be minimum for improving the performance of the system.

5. ***Response time:*** It is the minimum time that a distributed system executing a specific load balancing algorithm takes to respond.

6. ***Resource Utilization:*** It is the degree to which the resources of the system are utilized. A good load balancing algorithm provides maximum resource utilization.

7. ***Scalability:*** It determines the ability of the system to accomplish load balancing algorithm with a restricted number of processors or machines.

8. ***Performance:*** It represents the effectiveness of the system after performing load balancing. If all the above parameters are satisfied optimally then it will highly improve the performance of the system.

VII. Conclusion

The purpose of this paper is to focus on one of the major concerns of cloud computing that is Load balancing. The goal of load balancing is to increase client satisfaction and maximize resource utilization and substantially increase the performance of the cloud system thereby reducing the energy consumed and the carbon emission rate. Also the purpose of load balancing is to make every processor or machine perform the same amount of work throughout which helps in increasing the throughput, minimizing the response time and reducing the number of job rejection.

VIII. References

1. **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.**
2. **Ram Prasad Padhey, P. Goutam Prasad Rao, " Load Balancing in Cloud Computing Systems", Department of Computer Science and Engineering, National Institute of Technology, May 2011.**
3. **Abhijit A Rajguru, S.S. Apte, "A Comparative Performance Analysis of Load Balancing Algorithms In Distributed Systems Using Qualitative Parameters", International Journal of Recent Technology and Engineering, Vol. 1, Issue 3, August 2012.**
4. **Hisao Kameda, EL-Zoghdy Said Fathyy and Inhwan Ryuz Jie Lix, "A Performance Comparison of Dynamic vs Static Load Balancing Policies in a Mainframe, Personal Computer Network Model", Proceedings Of The 39th IEEE Conference on Decision & Control, 2000.**
5. **Nidhi Jain Kansal, Inderveer Chana, "Cloud Load Balancing Techniques: A Step Towards Green Computing", IJCSI, Vol. 9, Issue 1, January 2012.**
6. **Sandeep Sharma, Sarabjeet Singh, Meenaksshi Sharma, "Performance Analysis of Load Balancing Algorithms", World Academy of Science, Engineering and Technology, 2008.**
7. **R. Shimonski, Windows 2000 And Windows Server 2003, Clustering and Load Balancing Emeryville, McGrow-Hill Professional Publishing, CA, USA, 2003.**
8. **David Escalnte and Andrew J. Korty, "Cloud Services: Policy and Assessment", EDUCAUSE Review, Vol. 46, July/August 2011.**
9. **Parin. V. Patel, Hitesh. D. Patel, Pinal. J. Patel, "A Survey on Load Balancing in Cloud Computing" IJERT, Vol. 1, Issue 9, November 2012.**

10. **R. X. T and X. F. Z, " A Load Balancing Strategy Based on the Combination of Static and Dynamic, in Database Technology and Applications", 2nd International Workshop, 2010.**
11. **H. Stone, "Multiprocessor Scheduling With the Aid of Network Flow Algorithms", IEEE Transactions on Software Engineering, Vol. SE-3, No. 1, January 1977.**
12. **A. N. Tantawi and D. Tawsley, "Optimal Static Load Balancing in Distributed Computer Systems" Journal of the ACM, April 1985.**
13. **S. H. Bokhari, "Dual Processor Scheduling With Dynamic Reassignment", IEEE Transactions on Software Engineering, Vol. SE-5, July 1979.**
14. **J. A. Stankovic and I. S. Sidhu, "An Adaptive Bidding Algorithm for Processes, Cluster and Distributed Groups, In Proc. 4th Int. Conf. Distributed Compu. Sys. 1984.**
15. **J. Stankovic, "Simulations of Three Adaptive, Decentralized Control, Task Scheduling Algorithms, Computer Networks, Vol. 8, June 1984.**
16. **H. S. Stone, "High Performance Computer Architecture," 2nd Edition, Addison Wesley, Reading, MA, 1990.**
17. **S. Zhou, "A Trace Driven Simulation Study of Dynamic Load Balancing", IEEE Transactions on Software Engineering, Vol. SE-14, September 1988.**
18. **A. Barak And A. Shiloh, "A Distributed Load Balancing Policy For a Multicomputer", Software Practice and Experience, Vol. 15, September 1985.**
19. **K. Goswani, M. Devarakonda and R. Iyer, "Prediction Based Dynamic Load-Sharing Heuristics", IEEE Transactions On Parallel And Distributed Computing, Vol.4, June 1993.**
20. **Y. Wang And R. Morris, "Load Sharing in Distributed Systems", IEEE Trans. Comput., Vol. C-34, March 1985.**
21. **Nidhi Jain Kansal and Inderveer Chana, "Existing Load Balancing Techniques in Cloud Computing: A Systematic Review, Journal of Information Systems and Communication, Vol. 3, Issue 1.**