International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

A Survey on Load Balancing Challenges In Cloud Environment

S.Hendry Leo Kanickam*[1], S. Arockia Sobana[2]

*[1]Assistant Professor, Department of Information Technology, St.Joseph's College, Trichy, Tamilnadu, India

[2]M.Sc CS, Department of Information Technology, St.Joseph's College, Trichy, Tamilnadu, India

ABSTRACT

Cloud computing helps to share data and provide many resources to users. Users pay only for those resources as much they used. Cloud computing stores the data and distributed resources in the open environment. The amount of data storage increases quickly in open environment. Load balancing is a main challenge in cloud environment. Load balancing is important role to distribute the dynamic workload across multiple nodes to ensure that no single node is overloaded. cloud computing is a structured model that defines computing services, in which data as well as resources are retrieved from cloud service provider via internet through some well formed web-based tool and application. Cloud Computing is nothing but a collection of computing resources and services pooled together and is provide to the users on pay-as-needed basis. 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 and also contains comparison of various existing static load balancers as well as conventional dynamic load balancer also.

Keywords: Load balancing, static, dynamic, ARA

I. INTRODUCTION

Cloud computing is a new technology .It providing online resources and online storage to the users. It provides all the data at a lower cost. In cloud, computing users can access resources all the time through internet. They need to pay only for those resources as much they use .In Cloud computing cloud provider outsourced all the resources to their client. There are many existing issues in cloud computing. The main problem is load balancing in cloud computing. Load balancing helps to distribute all loads between all the nodes. It also ensures that every computing resource is distributing efficiently and fairly. It helps in preventing bottlenecks of the system, which may occur due to load imbalance. It provides high satisfaction to the users. Load balancing is a relatively new technique that provides high resource utilization and better response time.

Load balancing is a new approach that assists networks and resources by providing a high throughput and least response time [5].In cloud platforms, resource allocation (or load balancing) takes place majority at two levels.

At first level: The load balancer assigns the requested instances to physical computers at the time of uploading an application attempting to balance the computational load of multiple applications across physical computers.

At second level: When an application receives multiple incoming requests, each of these requests must be assigned to a specific application instance to balance the computational load across a set of instances of the same application [1].

The following sections discusses about the concept of load balancing, its needs and goals, types and comparison between traditional computing

environment and cloud computing environment and different algorithms.

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. 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. Load balancing is one of the important factors to heighten the working performance of the cloud service provider. The benefits of distributing the workload includes increased resource utilization ratio which further leads to enhancing the overall performance thereby achieving maximum client satisfaction [2].In cloud computing, if users are increasing load will also be increased, the increase in the number of users will lead to poor performance in terms of resource usage, if the cloud provider is not configured with any good mechanism for load balancing and also the capacity of cloud servers would not be utilized properly. This will confiscate or seize the performance of heavy loaded node. If some good load balancing technique is implemented, it will equally divide the load (here term equally defines low load on heavy loaded node and more load on node with less load now) and thereby we can maximize resource utilization. One of the crucial issue of cloud computing is to divide the workload dynamically.

III. CLASSIFICATION OF ALGORITHMS

a) Sender Initiated: In this sender initiates the process; the client sends request until a receiver is assigned to him to receive his workload
b) Receiver Initiated: The receiver initiates the process; the receiver sends a request to acknowledge a sender who is ready to share the workload
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 they are classified :

A. Static Load Balancing

In the static load balancing algorithm the decision of shifting the load does not depend on the current state of the system. It requires knowledge about the applications and resources 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 assign 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.

B. Dynamic Load Balancing

In this type of load balancing algorithms the current state of the system is used to make any decision for load balancing, thus the shifting of the load is depend on the current state of the system. It allows processes to move from an over utilized machine to an underutilized 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 depends on the current state of the system that helps

in improving the overall performance of the system by migrated the load dynamically.

Need of Load Balancing

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:

- ✓ **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.
- ✓ **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.

LOAD BALANCING ALGORITHMS

The paper describes about three load balancing algorithms which are Round robin algorithm, equally spread current execution load and Throttled Load balancing[11].

Round Robin: Round robin use the time slicing mechanism. The name of the algorithm suggests that it works in the round manner where each node is allotted with a time slice and has to wait for their turn. The time is divided and interval is allotted to each node. Each node is allotted with a time slice in which they have to perform their task. The complicity of this algorithm is less compared to the other two algorithms. An open source simulation performed the algorithm software know as cloud analyst, this algorithm is the default algorithm used in the simulation. This algorithm simply allots the job in round robin fashion that does not consider the load on different machines.

Equally spread current execution load: This algorithm requires a load balancer which monitors the jobs which are asked for execution. The task of load balancer is to queue up the jobs and hand over them to different virtual machines. The balancer looks over the queue frequently for new jobs and then allots them to the list of free virtual server. The balance also maintains the list of task allotted to virtual servers, which helps them to identify that which virtual machines are free and need to be allotted with new jobs. The experimental work for this algorithm is performed using the cloud analyst simulation. The name suggests about this algorithm that it work on equally spreading the execution load on different virtual machine.

Throttled Load balancing: The Throttled algorithm work by finding the appropriate virtual machine for assigning a particular job. The job manager is having a list of all virtual machines, using this indexed list, it allot the desire job to the appropriate machine. If the job is well suited for a particular machine than that job is, assign to the appropriate machine. If no virtual machines are available to accept jobs then the job manager waits for the client request and takes the job in queue for fast processing.

ARA(Adaptive Resource Allocation):In ARA algorithm for adaptive resource allocation in cloud systems, which attempts to counteract the deleterious effect of burrstones by allowing some randomness in the decision making process and thus improve overall system performance and availability [1]. The problem with this strategy is that it only considers the Poisson arrival streams as well as the exponentially distributed service time and the fixed number of choice.

QUALITATIVE MATRIX FOR LOAD BALANCING

The different qualitative metrics or parameters that are considered important for load balancing in cloud computing 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: The resources of the system utilized to the degree. 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.

IV. CONCLUSION

As such cloud computing being wide area of research and one of the major topics of research is dynamic load balancing, so the following research will be focusing on algorithm considering mainly two parameters firstly, load on the server and secondly, current performance of server. The goal of load balancing is to increase client satisfaction and maximize resource utilization and substantially increase the performance of the cloud system and minimizing the response time and reducing the number of job rejection thereby reducing the energy consumed and the carbon emission rate.

V. REFERENCES

[1]. JianzheTai,JueminZhang,JunLi,WaleedMeleis and NingfangMi "A R A: Adaptive Resource Allocation for Cloud Computing Environments under Bursty Workloads" 978-1-4673-0012-4/11 ©2011 IEEE.

[2]. 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.

[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]. Nidhi Jain Kansal, InderveerChana, "Cloud Load Balancing Techniques: A Step Towards Green Computing", IJCSI, Vol. 9, Issue 1, January 2012.

[5]. R. Shimonski, Windows 2000 And Windows Server 2003, Clustering and Load Balancing Emeryville, McGrow-Hill Professional Publishing, CA, USA, 2003.

[6]. David Escalnte and Andrew J. Korty, "Cloud Services: Policy and Assessment", EDUCAUSE Review, Vol. 46, July/August 2011.

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

[8]. R. Mata-Toledo, and P. Gupta, "Green data center: how green can we perform", Journal of Technology Research,Academic and Business Research Institute, Vol. 2, No. 1,May 2010, pages 1-8. © 2014 IJEDR | Volume 2, Issue 1 | ISSN: 2321-9939 IJEDR1401163 International Journal of Engineering Development and Research (www.ijedr.org) 912

[9]. 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.

[10]. Jitendra Bhatia, Tirth Patel, HarshalTrivedi, VishrutMajmudar," HTV Dynamic Load Balancing Algorithm for Virtual Machine Instances in Cloud",18,Dec2 012,Pages 15-20 IEEE.

[11]. Dr. Hemant S. Mahalle, Prof. Parag R. Kaveri ,Dr.VinayChavan," Load Balancing On Cloud Data Centres" , International Journal of Advanced Research in Computer Science and Software Engineering,Vol.3,Issue 1,January 2013.

Author's Profile

Mr.S.Hendry Leo Kanickam working as a Assistant Professor in Department of Information Technology ,St. Joseph's College (autonomous) Trichy, India. He received his M.Phil Degree in Bharathidasan University in 2008 and also He is pursuing Ph.D (Computer Science) in Bharathidasan University.

Ms.S.Arockia sobana is studying II M.Sc Computer Science in the Department of Information Technology ,St. Joseph's College (autonomous) Trichy, India