A Survey on Heterogeneous Load Balancing Techniques in Cloud Computing

Mittal Patel
PG Student
Department of Computer Engineering
Kalol Institute of Technology & Research Centre

Chaita Jani
Assistant Professor
Department of Computer Engineering
Kalol Institute of Technology & Research Centre

Abstract

As the IT industry growing in day to day, the necessitate of computing and storage is increasing sharply. Exchange of data over the network is also always increasing. New advance technology, cloud computing is become accepted because of providing above services beneficially. Other imperative technologies such as virtualization and scalability by creating virtual machines in cloud computing. In cloud computing web traffic and service provisioning is growing day by day so load balancing is becomes a big research issue in cloud computing. Cloud Computing is a new propensity emerging in IT environment with in vast requirements of infrastructure and resources. Load Balancing technique for cloud computing is a vital aspect of cloud computing environment. Load balancing is one of the inconvenient issues in distributed environment. Resources of service provider need to balance the load of client request. Different load balancing algorithms have been proposed in order to manage the resources of service provider efficiently and effectively. The purpose of applying the technique of Load Balancing is to allocate the available resources powerfully and quite. As the number of users increases, Load Balancing becomes a serious issue for the service provider in this paper, an overall review of the latest load balancing technique in the Cloud Computing environment is submitted. The ideas of every algorithm are discussed and finally sum up as an overview.

Keywords: Cloud Computing, Load Balancing, Load Balancing Algorithms, Virtualization., virtual machines, cloud Service provider

I. INTRODUCTION

Recently Cloud Computing has grown to be one of the popular techniques used by both industry and educational institutions in order to provide bendable way to store and access the data files. Cloud Computing can be defined as "structural model that defines computing services where resources as well as data are retrieved from cloud service provider via Internet through some well-formed web-based tools and application." Cloud Computing is a service oriented design that reduces the cost of access to gather the information of the clients offer greater flexibility and demand based services and so on[1]. Cloud Computing services are provided by the Cloud Service Providers. The client does not need to buy additional hardware and software. They have to just send their requests to the service provider and pay for the same. The economic application of Cloud Computing is often described as "converting capital expenses to operating expenses", generally known as "pay as you go". Hours purchased via Cloud Computing can be distributed non-uniformly in time. In the Networking community, this way of exchange bandwidth is already known as "usage-based pricing". The service provider is one and only responsible for the maintenance and up gradation of the services they are providing to their clients. Load balancing is the process of reassigning the total loads to the individual nodes of the united system to make the best response time and also good utilization of the resources. Cloud computing is an internet computing in which the load balancing is the one of the challenging task. Various methods are to be used to make a better system by allocating the loads to the nodes in a balancing manner but due to network obstruction, bandwidth usage etc, there were problems are occurred. These problems were solved by some of the existing techniques. A 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 current performance of the system. There were various goals that related to the load balancing such as to improve the performance substantially, to maintain the system stability etc. Depending on the current state of the system, load balancing algorithms can be categorized into two types they are static and dynamic algorithms. In the static algorithm there was prior knowledge of the system is needed and not depend on the current system. In the case of dynamic algorithm it is based on the current system and it is better performance than the static algorithm.

II. INTRODUCTION TO CLOUD COMPUTING

In this section, Cloud Computing is a model for enabling well-situated, on-demand network access to a shared pool of configurable computing resources that can be quickly provisioned and released with nominal management effort or service provider communication." The above definition includes the architecture of cloud, security fundamentals and deployment strategies. It is composed of five essential characteristics, three service models, and four deployment models. The cloud

computing system consists of mainly three components: Clients, Data canter and Distributed Servers. Each component has a specific purpose and plays specific role in delivering a functional cloud-based application.

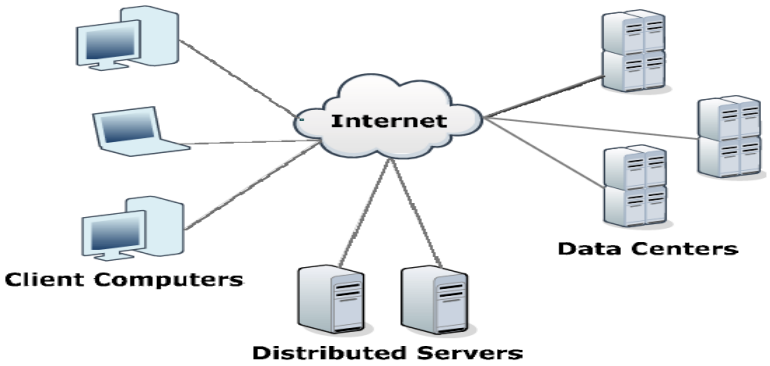

Fig. 1: Cloud Computing System [2]

1) *Clients:*

Clients are the simple computing devices such as PC, laptop, mobiles etc. The end users relate with the clients to maintain the information related to cloud. Generally they can be considered in to three groups, which are as follows:

- Thin Clients
- Mobile Client
- Thick Client

2) *Datacenter:*

The datacenter is the gathering of the servers where the approved application of the end user is housed. It can be a large room in the vault of the building or a room full of servers anywhere in the world that can be access through Internet. By
Using the concept of virtualization, multiple instances of virtual machines can be installed on a single physical server.

3) Distributed Servers:

Distributed servers are the parts of a cloud which are present throughout the Internet hosting diverse applications. But while using the application from the cloud, the user will feel that he is using this application from its own machine.
The essential characteristics that define the overall working of cloud computing technology are as below: [1] [3] [4]

- On-demand self service
- Broad Network access
- Location Independent Resource Pooling
- Measured service
- Rapid elasticity

A. Types of Cloud:

1) *Public Cloud:*

In Public cloud is obtainable for public use instead for a large industry and is owned by an organization selling cloud services. Customers have no visibility and organize Excess where the computing infrastructure is hosted. The computing infrastructure is shared among any organizations.

2) *Private Cloud:*

The computing infrastructure is operated for the exclusive use of an organization. The cloud almost certainly managed by the organization or third party. Private clouds are more expensive and more sheltered when compared to public clouds. Private clouds may be either on or off property.[5] Externally hosted private clouds are also only used by one organization, but are hosted by third party specialize in cloud infrastructure. Externally hosted private clouds are reasonably priced than On-premise private clouds

3) *Hybrid Cloud:*

Combines several clouds (private community of public) where those clouds retain their exclusive identities, but are bound together as a unit. A related term is Cloud convulsive. In Cloud convulsive organization is use their own computing infrastructure for general usage, but access the cloud for high load requirements [5]. This ensures that a unexpected increase in computing requirement is handled gracefully. Hybrid cloud may offer consistent or proprietary access to data and applications, as well as application portability.

4) *Community Cloud:*

A community cloud one where the cloud has been prepared to serve a common occupation or purpose. For example one organization or for several organization, but they share conman concerns such as their mission, security, policies, regulatory compliance needs and so on.

B. Service Model:

1) *Software as a Service (SaaS):*

The end-user can use the applications hosted by the cloud service provider on the cloud infrastructure. The applications are easily reached through slender clients interface such as web browser. The end-user has no organize over the essential cloud infrastructure. [6] The cloud service provider is one and only conscientious for the maintenance and up gradation of the applications they have hosted.

2) *Platform as a Service (PaaS):*

In the PaaS models, cloud providers deliver a computing platform, usually general of programming language execution environment, operating system, and database and web server. Application developers can change and run their software solutions on a cloud platform without the cost and insolubility of buying and overseeing the basic hardware and software layers.[5] The service provider manages the cloud infrastructure, the operating systems and the enabling software

3) *Infrastructure as a service (IaaS):*

Infrastructure as a service (IaaS) involves offering computing. IaaS provides a virtual-machine, virtual storage, virtual infrastructure, raw block storage, and file or object storage, load balancer, IP addresses, firewalls.[5,6] virtual local area networks and software providers supply these resources on-demand from their large pools installed in data centres bundles. The IaaS service provider manages the entire infrastructure.

C. Virtualization:

Virtualization is one of the most significant concept in context of cloud computing. In general, virtualization means something which is not physically present", but provides all the properties of the authentic. In Cloud Computing, Virtualization means the software implementation of a computer which will accomplish different programs like a real machine. An end user can use different services of a cloud with the help of the concept of virtualization. The remote data canter will afford different services in a full or partial virtualized manner.

There are two types virtualization found in case of cloud computing, as give in [7] [8]

- Full Virtualization
- Para virtualization

1) *Full Virtualization:*

The complete installation of one machine is done on another machine in case of full virtualization. Thus it will result in to a virtual machine which will have all the software already present in the real server this kind of virtualization allows not only run unique applications but also different operating systems [7]. Remote data center delivers the services in a fully virtualized manner. Full virtualization has been successful for several purposes.

2) *Para virtualization:*

In case of Para virtualization the hardware allows multiple operating systems to run on single machine by efficient use of system resources such as memory and processor.[8]E.g. VMware software. Here the services are provided partially.

III. LOAD BALANCING IN CLOUD COMPUTING

In a Load balancing is the technique of distributing the load between various resources in any system. Thus load requires to be distributed over the resources in cloud-based architecture, so that each resource does round about the equal amount of work at any point of time. Basic necessity is to provide some techniques to balance requests to provide the solution of fast response for request. Cloud Load Balancers manage online traffic by distributing workloads between numerous servers and resources automatically. They maximize throughput, minimize response time, and avoid overload. There are a number of issues while dealing with load balancing in a cloud computing environment. Every load balancing algorithm must be such as to instate the required target.

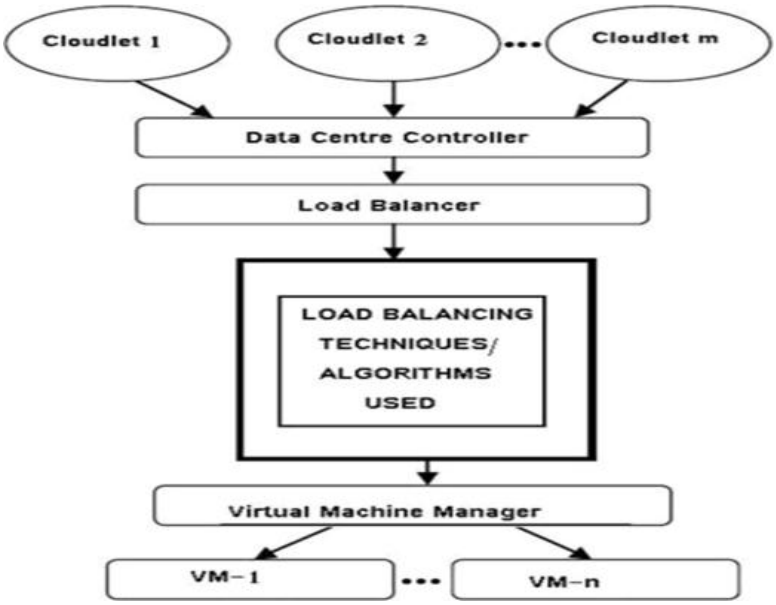

Fig. 2: Load Balancing Algorithms Execution[9]

IV. ISSUE EFFECTING LOAD BALANCING IN CLOUD COMPUTING

There are numerous types of load balancing techniques which are available for cloud computing. These load balancing techniques are: geographical distribution, static and dynamic.

A. The Geographical Distribution:

The geographical distribution of the nodes matters a lot in the cooperative performance of any real time cloud computing systems, exclusively in case of the big scaled applications like Twitter, Facebook etc. Geographical load balancing (GLB) can be defined as a series of decisions about online assignment and/or migration of virtual machines (VMs) or computational tasks to geographically distributed datacenters in order to meet the service level agreements (SLAs) or service deadlines for VMs/tasks and to decrease the operational cost of the cloud system. [10]

B. Static Load Balancing Algorithm:

In static load balancing algorithms, the execution of the processors is determined at the beginning of the execution, it does not depend on current state of the system. The aim of static load balancing is to reduce the overall execution time of a synchronous program while minimizing the communication delays. [10]

C. Dynamic Load Balancing Algorithm:

In dynamic load balancing algorithm, the decisions in load balancing are based on the existing state of the system, no previous knowledge is needed. The major advantage of dynamic load balancing is that if someone node fails, it will not stop the system; it will only involve the performance of the system. Therein are four policies for dynamic load balancers: selection policy, transfer policy, location policy and information policy. The task of load balancing is pooled among distributed nodes. In a distributed system, dynamic load balancing can be done in two different ways: distributed and non-distributed. [10]

V. CHALLENGES FOR LOAD BALANCING

In cloud computing, the load balancing is requisite to distribute the workload consistently across all the nodes. It helps to realize higher user satisfaction by ensuring fair and efficient allocation of computing resources. The following metrics helps to measure the overall performance of the load balancing algorithm. [11][12]

A. Throughput:

Throughput is the term used to approximation the whole number of tasks whose execution has been already completed. High throughput is necessary for overall system performance.

B. Associated Overhead:

It describes the amount of overhead during the implementation of the load balancing algorithm which includes overhead occurred due to movement of tasks, inter-process announcement. It should be as low as possible.

C. Fault Tolerant:
Fault Tolerance is the capacity of an algorithm to perform consistent load balancing in case of link failure. A good load balancing algorithm must be highly fault tolerable.

D. Migration Time:
Migration Time is the time required while migrating the jobs or resources from one node to another. For better performance of the system this time should be always less.

E. Response Time:
It can be measured as, the time interval between sending a request and receiving its response. This time should be minimized for better performance.

F. Resource Utilization:
It is used to ensure the proper utilization of all those resources, which comprised the whole system. For efficient load balancing in system, optimum resource should be utilized.

G. Scalability:
Scalability is the ability to scale-up or scale-down according to the requirements of the system. 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.

H. 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. It is the overall efficiency of the system. If all the parameters are improved then the overall system performance can be improved.

VI. EXISTING LOAD BALANCING ALGORITHM SAND

A. Round Robin Algorithm:
This algorithm uses the round robin scheme for allocating jobs. It selects the first node randomly and then, allocates jobs to all other nodes in a round robin fashion Processors are assigned to each process in a spherical order without any sort of priority and hence there is no starvation. [14]

B. Throttled Algorithm:
Throttled Algorithm is completely based on virtual machine in this algorithm, when the client sends request to Data Centre controller; an appropriate VM is found which can accomplish the request in an efficient manner using specific technique. The current state of the VM whether it is available or busy for each VM is maintained. While doing so, the current load of the VM and processing time required by the request is not considered. Application, paper, the state of the art on ensemble methodologies. Furthermore, we have exposed the positive synergy between sampling techniques and Bagging ensemble learning algorithm. [13]

C. Opportunistic Load Balancing Algorithm:
This algorithm [15] attempts to keep each node busy. consequently it does not consider the present workload of each computer. It provides load balance schedule but it results in very poor make-span.

D. Min Min Load Balancing Algorithm:
The Min Min [16] algorithms begin with a set of all unassigned tasks. First of all, minimum achievement 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 resource. According to that minimum time, the task is scheduled on the equivalent machine. The execution time for all other tasks is updated on that machine and the task is removed from the list.

E. Max Min Load Balancing Algorithm:
Max Min [16] is comparable to 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 the resources. Then according to that maximum time, the task is scheduled on the corresponding machine. The execution time for all other tasks is updated on that machine and assigned task is removed from the list of the tasks that are to be assigned to the machines.

F. Two Phase Load Balancing Algorithm:

The two phase scheduling [17] algorithm combines Opportunistic Load Balancing and Min-Min scheduling algorithms to utilize better executing efficiency and maintain the load balancing of the system. This mutual approach hence helps in an efficient utilization of resources and enhances the work efficiency.

G. Ant Colony Optimization (ACO) Based Load Balancing Algorithm:

ACO [18] is a probabilistic technique for solving computational problems which can be condensed to finding efficient paths through graphs. The aim of ant colony optimization is to search for an optimal path in a graph on the behaviour of ants looking for a path between their colony and source of food.

H. Honey Bee Foraging Load Balancing Algorithm:

This is a inhabitants based search algorithm [19]. A colony of bees can extend itself over a distance and in multiple directions simultaneously to develop a large number of food resources. Flowers with bountiful nectar or pollen grains are visited more often than the ones with fewer grains. The scout bees which have gone in search of food come back and perform a waggle dance indicating the direction in which the food was found. a complete scheduling of operations is specified in the problem. Each solution can be thought as a path from the hive to the food resource.

VII. CONCLUSION AND FUTURE WORK

This paper is based on cloud computing technology which has vast prospective. Each algorithm is observed and their scheduling criteria like average response time, data centre service time and total cost of different data centres is found. Researchers are doing many works to resolve those issues. For Load Balancing issue, the solution is to develop suitable algorithms that balance the load across the partitioned cloud environment. We have also discussed virtualization of cloud.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their precious comments and suggestions that contributed to the expansion of this work.

REFERENCES

[1] Lee Badger, Tim Grance, Robert Patt-Corner, Jeff Voas , " DRAFT Cloud Computing Synopsis and Recommendations ", Special publication 800-146, May 2011
[2] Basic concept and terminology of cloud computing- http://whatiscloud.com.
[3] Tharam Dillon, Chen Wu, Elizabeth Chang, "Cloud Computing: Issues and Challenges", 24th IEEE International Conference on Advanced Information Networking and Applications, IEEE, 2010
[4] Youngmi N. S. Raghava and Deepti Singh, "Comparative Study on Load Balancing Techniques in Cloud Computing", Open Journal of Mobile Computing and Cloud Computing, Scientific Online.
[5] Pengyi Yang, Alok singh,Vikas Kumar Tiwari, Dr. Bhupesh Gour A Survey on Load Balancing in Cloud Computing Using Soft Computing Technique's IJARCCE Vol. 3, Issue 9, September 2014.
[6] HARDI S. SANGHAVI,DR. TEJAS P. PATALIA LOAD BALANCING ALGORITHMS FOR THE CLOUD COMPUTING ENVIRONMENT: A REVIEW NOV 14 TO OCT 15| VOLUME – 03, ISSUE – 02.
[7] Anthony T. Velte, Toby J. Velte and Robert Elsenpeter, "Cloud Computing- A Practical Approach," Tata McGraw Hill Publications, pp. 6-8.
[8] Stuti Dave and Prashant Mehta,"Role of Virtual Machine live Migration in Cloud Load Balancing", IJARCCE, 2012.
[9] N. S. Raghava and Deepti Singh "Comparative Study on Load Balancing Techniques in Cloud Computing" OPEN JOURNAL OF MOBILE COMPUTING AND CLOUD COMPUTING Volume 1, Number 1, August 2014.
[10] Shiny "Load Balancing In Cloud Computing:A Review" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 15, Issue 2 (Nov. - Dec. 2013), PP 22-29.
[11] Nayandeep Sran and Navdeep Kaur,"Comparative Analysis of Existing Load Balancing Techniques in Cloud COmputing", International Journal of Engineering Science Invention, January 2013, pp. 60-63.
[12] Nitika, "Comparative Analysis of Load Balancing Algorithms in Cloud Computing", International Journal of Engineering and Science,2012.
[13] Shridhar G. Domanal and G.Ram Mohana Reddy, "Load Balancing in Cloud Computing using Modified Throttled Algorithm", IEEE, 2013
[14] Shanti Swaroop Moharana, Rajadeepan D Ramesh, Digamber Powar, "Analysis Of Load Balancers in Cloud Computing", International Journal of Computer Science and Engineering 05/2013; 2(2):101-108.
[15] Che-Lun Hung1, Hsiao-hsi Wang2 and Yu-Chen Hu2, "Efficient Load Balancing Algorithm for Cloud Computing Network". IEEE Vol. 9, pp: 70-78, 2012.
[16] T. Kokilavani," Load Balanced Min-Min Algorithm for Static Meta-Task Scheduling in Grid Computing" International Journal of Computer Applications (0975 – 8887) Volume 20– No.2, April 2011.2012.
[17] Wang, K. Yan, W. Liao, and S. Wang, "Towards a Load Balancing in a Three-level Cloud Computing Network", Proceedings of the 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), Chengdu, China, September 2010, pages 108-113.
[18] Ratan Mishra and Anant Jaiswal " Ant Colony Optimization: A Solution of Load balancing in Cloud" International Journal of Web and Semantic Technology.Vol. 3, No. 2. pp. 33-50. April 2012.
[19] Ram Prasad Padhy ,P Goutam Prasad Rao, "Load Balancing in Cloud Computing Systems", National Institute of Technology, Rourkela, India, 2011