ISSN 2277-3061

LOAD BALANCING ISSUES AND ITS SOLUTION IN CLOUD COMPUTING: A REVIEW

Settu Bharti [1], Naseeb Singh [2]
[1] Research Scholar, Department of Computer Science Engineering, AIET, Faridkot
settu.bharti@yahoo.com
[2] Assistant Professor, Department of Computer Science Engineering, AIET, Faridkot
naseebdhillon@hotmail.com

ABSTRACT

Cloud computing is an emerging paradigm in the computer industry where the computing is moved to a cloud of computers. Cloud computing is a way to increase the capacity or add capabilities dynamically without investing in new infrastructure, training new personnel, or licensing new software. This paper is focused on the load balancing issues of cloud computing and techniques to overcome the waiting time and turnaround time. Load Balancing is done with the help of load balancers where each incoming request is redirected and is transparent to client who makes the request. Based on predetermined parameters, such as availability or current load, the load balancer uses various scheduling algorithm to determine which server should handle and forwards the request on to the selected server.

Keywords

Cloud Computing, Load Balacing, Virtual Machine, Data Center, Data Center Broker.

Council for Innovative Research
Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol.14, No.3

www.ijctonline.com, editorijctonline@gmail.com

INTRODUCTION

Internet has been a driving force towards the various technologies that have been developed since its inception. Arguably, one of the most discussed among all of them is Cloud Computing. Over the last few years, Cloud computing paradigm has witnessed an enormous shift towards its adoption and it has become a trend in the information technology space as it promises significant cost reductions and new business potential to its users and providers [1]. Cloud computing can be defined as "Cloud is a parallel and distributed computing system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements (SLA) established through negotiation between the service provider and consumers".

Load Balancing[2] is an emerging computer paradigm where data and services placed massively in the cloud and which can be accessed from any connected devices over the internet. It is known as provider of dynamic services using very large scalable and virtualized resources over the internet. Load Balancing is a computer networking method to distribute workload across multiple computer clusters, network links or other resources to achieve optimal resource utilization, maximize throughput, minimize response time and avoid overload. It is a mechanism that distributes the dynamic local work load[3] evenly across all the nodes in the whole cloud to avoid a situation where some nodes are heavily loaded while others are idle. Its goal is to improve the overall performance and resource utility of the system.

A. CLOUD COMPUTING: AN OVERVIEW

Cloud computing[4] involves distributed computing over a network, where a program or application may run on many connected computers at the same time. The cloud makes it possible for you to access your information from anywhere at any time. While a traditional computer setup requires you to be in the same location as your data storage device, the cloud takes away that step. The cloud removes the need for you to be in the same physical location as the hardware that stores your data.

B. SERVICE MODELS

The cloud service providers three different services based on different capabilities such as SaaS (Software as a Service), PaaS (Platform as a Service), IaaS (Infrastructure as a Service) [5].

1. Software as a Service (SaaS): Software as a Service consists of software running on the provider's cloud infrastructure, delivered to (multiple) clients (on demand) via a thin client (e.g. browser) over the Internet.

2. Platform as a Service (PaaS): This gives a developer the flexibility to develop applications on the provider's platform. Entirely virtualized platform that includes one or more servers, operating systems and specific applications.

3. Infrastructure as a Service (IaaS): The service provider owns the equipment and is responsible for housing, running and maintaining it a service .

C. DEPLOYMENT MODELS

Depending on infrastructure ownership, there are four deployment models of cloud computing [6].

1. The Public Cloud: Which describes cloud computing in the traditional mainstream sense; resources are dynamically provisioned on a self-service basis over the Internet. It is usually owned by a large organization (e.g. Amazon, Google AppEngine)

2. The Private Cloud: It defers from the traditional data enter in its predominant use of virtualization. The private cloud is more appealing to enterprises especially in mission and safety critical organizations.

3. The Community Cloud: Thus refers to a cloud infrastructure shared by several organizations within a specific community. A typical example is the Open Cirrus Cloud Computing Testbed.

4. The Hybrid Cloud: It comprises of a combination of any two (or all) of the three models discussed above.

LOAD BALANCING

Load Balancing [5] is done with the help of load balancers where each incoming request is redirected and is transparent to client who makes the request. Based on predetermined parameters, such as availability or current load, the load balancer uses various scheduling algorithm to determine which server should handle and forwards the request on to the selected server. To make the final determination, the load balancer retrieves information about the candidate server's health and current workload in order to verify its ability to respond to that request. Load balancing solutions can be divided into software-based load balancers and hardware-based load balancers. Hardware-based load balancers are specialized boxes that include Application Specific Integrated Circuits (ASICs) [28] customized for a specific use.

One of the most commonly used applications of load balancing is to provide a single Internet service from multiple servers, sometimes known as a server farm. Commonly load-balanced systems include popular web sites, large Internet Relay Chat networks, high-bandwidth File Transfer Protocol sites, Network News Transfer Protocol (NNTP) [13] servers and Domain Name System (DNS) servers. Lately, some load balancers have evolved to support databases; these are called database load balancers.

For Internet services, the load balancer is usually a software program that is listening on the port where external clients connect to access services. The load balancer forwards requests to one of the "backend" servers, which usually replies to the load balancer. This allows the load balancer to reply to the client without the client ever knowing about the internal separation of functions. It also prevents clients from contacting back-end servers directly, which may have security benefits by hiding the structure of the internal network and preventing attacks on the kernel's network stack or unrelated services running on other ports.

Some load balancers provide a mechanism for doing something special in the event that all backend servers are unavailable. This might include forwarding to a backup load balancer, or displaying a message regarding the outage. Load balancing gives the IT team a chance to achieve a significantly higher fault tolerance. It can automatically provide the amount of capacity needed to respond to any increase or decrease of application traffic.

It is also important that the load balancer itself does not become a single point of failure. Usually load balancers are implemented in high-availability pairs which may also replicate session persistence data if required by the specific application.

• O.M.elzeki, et.al,(2012): discusses in Improved Max-Min Algorithm in Cloud Computing that focuses on the cloud computing which further deals with the allocation of the tasks to the resources while observing different parameters like Waiting time, Average waiting time, Turn Around time, Processing cost. So, an algorithm named as Max-Min in improved manner from load balancing has been shown to overcome such kinds of problems. The algorithm calculates the expected completion time of the submitted tasks on each resource. Then the task with the overall maximum expected execution time is assigned to a resource that has the minimum overall completion time.

• Amandeep Kaur Sidhu, (April-2013) discussed in Analysis of load balancing techniques in cloud computing that aims to share of data, calculations and resources transparently over a scalable network of nodes.

• Gytis Vilutis* et al,(2012) discussed that it is complicated to determine the quantity of resources in order to satisfy work load with peaks. Some projects are lost because of under provisions of cloud resources which leads to postponed work and that can reduce the probability of projects not to be executed. The author discussed two problems: to deploy maximum quantity of servers wishing to satisfy all its users requirement and to keep minimum quantity of servers in full usage even the users load is at minimum level.

• Upendra Bhoi* et al,(April2013) Discussed that in enhanced Max-Min Task Scheduling Algorithm in cloud computing helps in supplying a high performance computing based on protocols which allowed shared computation and storage over long distances. It depends upon expected execution time instead of completion time. Max-Min algorithm assign task with maximum execution time to resource produces minimum completion time while Enhanced Max-min assign task with average execution time to resource produces minimum execution time.

• Klaitham Al Nuaimi* et al, (2012) discuss about the overall approach to enhance the performance of cloud. Cloud provides a flexible and easy way to keep and retrieve data and files. Especially for making large data sets and files. In Load Balancing algorithm are classified as static and dynamic algorithm. Static algorithm is for stable and homogenous environments whereas dynamic are more flexible and can adapt to various changes by providing better results.

• Tushar Desai* et al,(Nov 2013) discusses about the imerging technology i.e a new standard of large scale distributed computing and parallel computing. It provides shared resources, information or other resources as per clients requirements at specific times.For better management of available good load balancimg techniques are required. And through beter load balancing in cloud , performance is increased and user gets better services. So in this author has discussed many different load balancing techniques used to solve the issue in cloud computing environment.

• Haozheng Ren* et al, (2012) explains that The load balancing algorithm is an important means to achieve efficient utilization of resources. This paper presents a dynamic load balancing algorithm based on virtual machine migration under cloud computing environment .It Minimizes the allocation time of user requests and Maximize system throughput.

METRICS FOR LOAD BALANCING IN CLOUD

Various metrics considered in existing load balancing techniques in cloud computing are discussed below-

- **Scalability** is the ability of an algorithm to perform load balancing for a system with any finite number of nodes. This metric should be improved.
- **Resource Utilization** is used to check the utilization of re-sources. It should be optimized for an efficient load balancing.
- **Performance** is used to check the efficiency of the system. This has to be improved at a reasonable cost, e.g., reduce task response time while keeping acceptable delays.
- **Response Time** is the amount of time taken to respond by a particular load balancing algorithm in a distributed system. This parameter should be minimized.
- **Overhead Associated** determines the amount of overhead involved while implementing a load-balancing algorithm. It is composed of overhead due to movement of tasks, inter-processor and interprocess communication. This should be minimized so that a load balancing technique can work efficiently.

LOAD BALANCING ALGORITHMS

In order to balance the requests of the resources it is important to recognize a few major goals of load balancing algorithms:

a) **Cost effectiveness**: primary aim is to achieve an overall improvement in system performance at a reasonable cost.

b) **Scalability and flexibility**: the distributed system in which the algorithm is implemented may change in size or topology. So the algorithm must be scalable and flexible enough to allow such changes to be handled easily.

c) **Priority**: prioritization of the resources or jobs need to be done on before hand through the algorithm itself for better service to the important or high prioritized jobs in spite of equal service provision for all the jobs regardless of their origin.

Following load balancing algorithms are currently prevalent in clouds:-

Round Robin: In this algorithm [7], the processes are divided between all processors. Each process is assigned to the processor in a round robin order. The process allocation order is maintained locally independent of the allocations from remote processors. Though the work load distributions between processors are equal but the job processing time for different processes are not same. So at any point of time some nodes may be heavily loaded and others remain idle. Thi s algorithm is mostly used in web servers where http requests are of similar nature and distributed equally.

Connection Mechanism: Load balancing algorithm [8] can also be based on least connection mechanism which is a part of dynamic scheduling algorithm. It needs to count the number of connections for each server dynamically to estimate the load. The load balancer records the connection number of each server. The number of connection increases when a new connection is dispatched to it, and decreases the number when connection finishes or timeout happens.

Randomized: Randomized algorithm is of type static in nature. In this algorithm [7] a process can be handled by a particular node n with a probability p. The process allocation order is maintained for each processor independent of allocation from remote processor. This algorithm works well in case of processes are of equal loaded. However, problem arises when loads are of different computational complexities. Randomized algorithm does not maintain determini stic approach. It works well when Round Robin algorithm generates overhead for process queue.

Equally Spread Current Execution Algorithm: Equally spread current execution algorithm [9] process handle with priorities. it distribute the load randomly by checking the size and transfer the load to that virtual machine which is light ly loaded or handle that task easy and take less time , and give maximize throughput. It is spread spectrum technique in which the load balancer spread the load of the job in hand into multiple virtual machines.

Throttled Load Balancing Algorithm: Throttled algorithm [9] is completely based on virtual machine. In this client first requesting the load balancer to check the right virtual machine which access that load easily and perform the operations which is give by the client or user. In this algorithm the client first requests the load balancer to find a suitable Virtual Machine to perform the required operation.

A Task Scheduling Algorithm Based on Load Balancing: Y. Fang et al. [10] discussed a two-level task scheduling mechanism based on load balancing to meet dynamic requirements of users and obtain high resource utilization. It achieves load balancing by first mapping tasks to virtual machines and then virtual machines to host resources thereby improving the task response time, resource utilization and overall performance of the cloud computing environment.

Min-Min Algorithm: It begins with a set of all unassigned tasks. First of all, minimum completion 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 resources. Then according to that minimum time, the task is scheduled on the corresponding machine. Then the execution time for all other tasks is updated on that machine by adding the execution time of the assigned task to the execution times of other tasks on that machine and assigned task is removed from the list of the tasks that are to be assigned to the machines. Then again the same procedure is followed until all the tasks are assigned on the resources. But this approach has a major drawback that it can lead to starvation [12].

Max-Min Algorithm: Max-Min is almost same as 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 any resources. Then according to that maximum time, the task is scheduled on the corresponding machine. Then the execution time for all other tasks is updated on that machine by adding the execution time of the assigned task to the execution times of other tasks on that machine and assigned task is removed from the list of the tasks that are to be assigned to the machines[12].

PROBLEM DESCRIPTION

Although cloud computing has been widely adopted. Research in cloud computing is still in its early stages, and some scientific challenges remain unsolved by the scientific community, particularly load balancing challenges [15].

• **Automated service provisioning:** A key feature of cloud computing is elasticity, resources can be allocated or released automatically. How then can we use or release the resources of the cloud, by keeping the same performance as traditional systems and using optimal resources?

• **Virtual Machines Migration**: With virtualization, an entire machine can be seen as a file or set of files, to unload a physical machine heavily loaded, it is possible to move a virtual machine between physical machines.

• In the paper titled “Improved Max-Min Algorithm in cloud computing“, the author is trying out to allocate the task with maximum execution time to the resource with minimum completion time.

• In this approach, if we are having more no of tasks(lets say 10,000), then the average turn-around time of the tasks will be very high which will decrease the efficiency of the entire system.

• And if the average turnarounds time will be high then the processing cost as well as waiting time will also be increased.

• Thus Load balancing is improving the performance by balancing the load among the resources like network links, CPU, disk and even on cloud and other storage devices.

OBJECTIVES

- In Max-Min algorithm, in cloud computing describes the solving of large tasks first and delay in small tasks. So the main objective is to improve the Max-Min Algorithm in cloud computing. Max-Min strategy resolves the priority system and selects the task with the maximum completion time and assigns it to the resource on which achieve minimum execution time.
- To improve execution time over the completion time of the task.
- To improve the Turn Around Time.
- Supplying high performance computing based on protocols which allow shared computation and storage over long distances.

METHODOLOGY

- All tasks will sorted according to their minimum execution length.
- Now we will calculate the expected completion time of each task on all resources.
- Expected Completion time of task on a resource can be calculated as: CT(i,j)=ET(i,j)+ r(j), where ET(I,j) is the expected execution time of task t(i) on machine m(j) and r(j) is the ready time of m(j) i.e. the time when m(j) becomes ready to execute t(i).
- Now we will find minimum expected completion time of each task in MT(meta task table) and the resource that will obtain it.(tasks are collected into a set called meta task(MT)).
- Now we will arrange the resources in the descending order of MIPS(million instruction per second).
- Finally we arrange our tasks into the group.
- No. of groups = No. of Tasks/ Number of resources.
- So, the choice of cloudlet in the group.

  Task: Size = more/max.

  Task: Size = less/ min.
- T1,T2,T3,T4,T5,T6 are tasks and R1,R2,R3 are resources.
- 6/3= 2, 12/3 =4

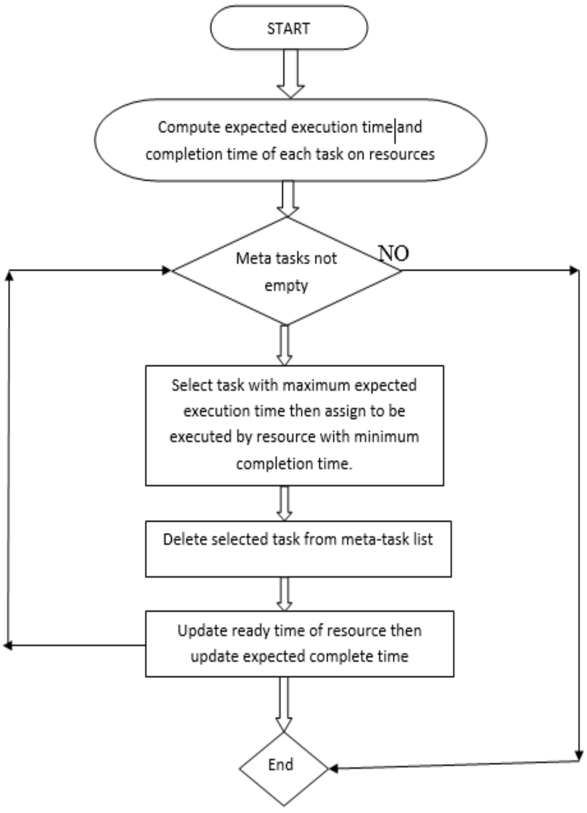
START
Compute expected execution time and completion time of each task on resources
Meta tasks not empty
NO
Select task with maximum expected execution time then assign to be executed by resource with minimum completion time.
Delete selected task from meta-task list
Update ready time of resource then update expected complete time
End

Fig 1. Flowchart of the Proposed Methodology

CONCLUSION

This paper is based on cloud computing technology which has a very vast potential and is still unexplored. The capabilities of cloud computing are endless. Cloud computing provides everything to the user as a service which includes platform as a service, application as a service, infrastructure as a service.

One of the major issues of cloud computing is load balancing because overloading of a system may lead to poor performance which can make the technology unsuccessful. So there is always a requirement of efficient load balancing algorithm for efficient utilization of resources. Our paper focuses on the various load balancing algorithms and their applicability in cloud computing environment.

REFERENCES

[1] O.M. Elzeki . “Improved Max-Min Algorithm in Cloud Computing”. International Journal of Computer Applications(0975-8887) Volume 50-No.12,july 2012.

[2] “A technical support seminar on cloud computing technology” by Prashant Gupta.

[3] Amandeep Kaur Sidhu. “Analysis of load balancing techniques in cloud computing”. International Journal of computers & technology volume 4 No. 2, March-April, 2013, ISSN 2277-3061.

[4] Ektemal Al-Rayis. “Performance Analysis of load balancing Architectures in Cloud computing” 2013 European Modeling Symposium. 978-1-4799-2578-0/13$31.00@2013 IEEE.

[5] Haozheng Ren. “The load balancing Algorithm in cloud computing Environment” 2nd International Conference on computer science and network technology 2012.

[6] Tushar Desai. “A survey of various load balancing techniques and challenges in cloud computing” International Journals of scientific and technology research volume 2. Issue11,Nov2013.

[7] Upendra Bhoi. “Enhanced max-min Task scheduling Algorithm in cloud computing”. International Journal of Application or Innovation in Engineering & management(IJAIEM), April 2013.

[8] Klaithem Al Nuaimi, “A survey of load balancing in cloud computing challenges and algorithm”. 2012 IEEE second symposium on network cloud computing and applications.

[9] Gytis Vilutis, “Model of load balancing and scheduling in cloud computing”. Proceedings of the ITI 2012 34th Int.Conf. on Information Technology Interfaces, June 25-28,Cavat,Croatia.