Load Balancing in Cloud Computing

Minakshi Berwal[1], Dr. Chander Kant[2]
[1]Research Scholar, Department of Computer Science and Application, K.U., Kurukshetra, INDIA
[2]Asst. Professor, Department of Computer Science and Application, K.U., Kurukshetra, INDIA
29minakshi@gmail.com, ckverma@rediffmail.com

Abstract: Cloud computing as a new Internet service concept has become popular to provide different types of services to user. With the advancement of the Cloud, there are so many new possibilities opening up on how applications can be built and how various services can be offered to the end user through Virtualization, on the internet. With the recent introduction of technology, load balancing or resource control in cloud computing is the most challenging issue. Load balancing is a methodology that tends to distribute work load across a number of computers, or other means over the network links to achieve optimal resource utilization, minimum response time, to maximize throughput, and avoid overload. This paper defines cloud computing with its services and also discussed load balancing.

Keywords: Cloud Computing, Load Balancing, Load Balancing Algorithm.

INTRODUCTION

Cloud computing means that instead of using all the computer hardware and software on our desktop, or anywhere inside our company's network, it's provided for us as a service by another company and accessed by us over the Internet. The physical location where the hardware and software is located and how it works doesn't matter to us. The domain of cloud computing is still surfaced by many issues [1]. The major focus of the paper will be to study cloud computing and analyse the research issues in load balancing protocols or understanding its requirement in cloud platform.

Load balancing is usually mechanized to implement failover of the continuance of a service on the failure of 1 or additional parts of it. The components are monitored frequently and when any one becomes nonresponsive, the load balancer that has been used over the network become active and will not sends traffic to it. Resource consumption and energy conservation don't seem to be continuously attentiveness once discussing cloud computing; but with correct load balancing in place of resource consumption are often unbroken to a minimum.

Load Balancing not only serves to keep cost low and enterprise greener, it also helps to put less stress on the circuits of each individual design making them more potentially last longer [1].

1. LITERATURE REVIEW

Due to the recent emergence of cloud computing research in this area is in the preliminary stage. A resource allocation mechanism with pre-empt able task execution which increases the utilization of clouds [2] proposed. Cloud provide various services [3] to the used through various component that it’s constitute.

Cloud is the concept that evolve distribution of resources in virtual environment so that each user can access resources, helps utilization of resources [1] [4]. Load balancing in cloud computing system [4] discussed on basic concepts of Cloud Computing and Load balancing and studied some existing load balancing algorithms [5], which can be applied to clouds. In Load Balancing technique concept of virtual network [6] and fuzzy concept [9] play a major role as virtual network is the key concept to distribute resources on various node [10] so that each node has equivalent access to resources. For even distribution of resources there must be some means to analyse availability [12] of resources as well. For implementation analysis of load balancing technique various simulator are used. Different algorithms different perspective so that simulator [14] can be used to analyse the appropriate algorithm that can be applied in a particular environment. Simulator analyse the efficiency of a technique for execution in real life application [13] [16].

2. CLOUD COMPUTING

Cloud computing is an innovative technology that facilitates the networked computers to share the pooled resources on demand in pay per use model. Cloud computing is highly scalable, dynamic [1] and easily configurable above that it can handle multitenant request simultaneously [3].Cloud computing can also be viewed as alternative way of describing IT (information technology) "outsourcing"; others use it to mean any computing service provided over the Internet or a similar network; and some define it as any bought- in computer service you use that sits outside your firewall [4] [10]. The US National Institute of Standards and Technology (NIST) define cloud computing as "a model for user convenience, on demand network access contribute the computing resources (e.g. networks, storage, applications, servers, and services) that can be rapidly implemented with minimal management effort or service provider interference [8]. Cloud computing can also be defined as it is a new service, which are the collection of technologies and a means of supporting the use of large scale Internet services for the remote applications with good quality of service (QoS) levels [2][17]. Cloud computing is the dynamic provisioning of information technology capabilities (hardware, software, or services) from third parties over a network.

2.1 Component and Services

Any cloud computing system comprises of three major components such as clients (end users), datacenter and distributed servers. These components are shown in Figure 1.

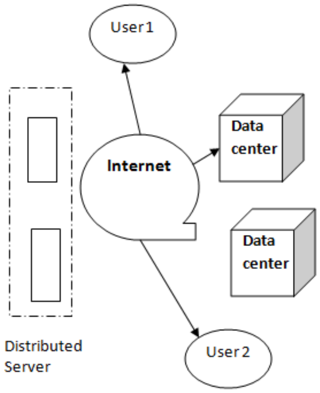

Figure 1: Cloud Components

Brief discussion of explicit role and purpose of each component is presented in the following:

a) Client: End users interact with the clouds to manage information related to the cloud. Clients can be categorized as:-

- Thin: A thin client is a bare bones computer that allows users to access files, programs, and functionality that is hosted on another computer. The server drives the main operating system, information and programs to the thin client when a user logs on.
- Thick: A thick client (also called heavy, rich or fat client) is a computer (client) in client–server architecture that typically provides rich functionality independent of the central server.

b) Datacenter: Datacenter is collection of servers hosting different applications. An end user (say user1) connects to the datacenter to subscribe different applications.

c) Distributed Servers: Distributed servers are the part of a cloud which is available throughout the internet hosting different applications [11].

2.2 Types of Cloud Computing

The cloud computing technology can be viewed from two different perspectives: Capability and Accessibility [13].

2.2.1 Based on Type of Capability: According to this categorizations, cloud system provides three different types of services as follows:

a) Cloud Software as a Service (SaaS): The capability provided to the user is to access the provider's requests running on a cloud infrastructure. The applications are available from various client devices through a thin client interface such as a web browser (e.g., web-based email) as shown in figure 2.

b) Cloud Platform as a Service (PaaS): The capability provided to the consumer is to deploy onto the cloud infrastructure consumer- created or acquired requests created using programming languages and tools supported by the provider

c) Cloud Infrastructure as a Service (IaaS): The capability provided to the user is to provision processing, networks, storage, and other fundamental computing resources where the user is able to deploy and run arbitrary software, which can include operating[2][7] systems and applications.

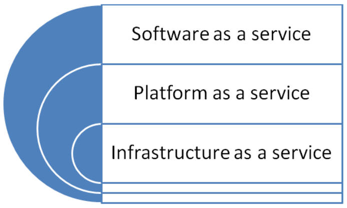

Figure 2.Cloud computing services

2.2.2 Based on Accessibility Type

On the basis of accessibility also clouds are categorized.

a) Public Cloud: This cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.

b) Private Cloud: The cloud infrastructure is operated solely for a single organization. It may be managed by any organization or a third party.

c) Community Cloud: The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns.

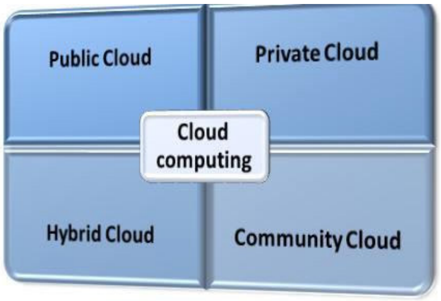

Figure 3 Clouds categorized on basis of accessibility

d) Hybrid Cloud: This cloud infrastructure comprises two or more clouds (private, public or community) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability [7].

2.3 Advantages and Barriers of Cloud Computing

Progress of Cloud Computing is massive with respect to personal uses and business uses. Among numerous advantages or benefits, few are discussed below:

a) Scalability
b) Virtualization
c) Mobility
d) Low Infrastructure Costs
e) Increased Storage

Thus, the cloud computing provides several advantages with form of elasticity, availability and expandability on-demand. Still it has some constraints as discussed [3] in the following:

a) Latency
b) Platform or Language constraints
c) Resource Control
d) Load Balancing

Load Balancing is prime issue as data increasing day by day and maintaining data is of prime concern.

3. LOAD BALANCING

The increase in web traffic and different application in the web world is increasing day by day where millions of data are created every second. Load balancing has become a very prevalent research field due to need of balancing the load on this heavy traffic. Cloud computing use is a concept that use virtual machine instead of physical device to host, store and link the different nodes for their specific purpose. The load balancing is needed on CPU load, memory capacity and network. Load Balancing is done in such a way that the entire load is distributed among various nodes in a distributive system. If there is a failure of any node or host system in the network, it will lead to isolation of web resource in the web world. Load balancing in such situation should be able to provide availability and scalability. Many authors agree with the definition of Cloud Computing as it consists of clusters of distributed computers (Clouds) providing on-demand resources or services over a network with the scale and reliability of a data centre [8].

Load balancing is a process of reassigning the total load to the individual nodes of the computing environment, [9] this facilitates the network and resources and further improving the system performance. The important parts of this process are estimation and comparison of the stability, load and performance of the system, internodes traffic optimization. To construct load balancing mechanism many techniques and strategies are used [5].The load need to be distributed over the resources in cloud-based architecture, consequently each resources does almost the equal amount of task at any point of time. The basic goal is to design some techniques to balance requests to provide the solutions [11].

Cloud vendors are based on automatic load balancing services, which allow users to rise the number of CPUs or memories for their resources to scale with increased demands. This service provided is elective and depends on the clients business needs. Hence load balancing serves two important needs, mainly to promote availability of Cloud resources and secondarily to promote performance [12].

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

a) Cost effectiveness: The major 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 topology or size. Thus the algorithm must be scalable and flexible enough to allow such changes to be handled easily.

c) Priority: It require prioritization of the resources or jobs need to be done on beforehand 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 [11].

4.1 LOAD BALANCING ALGORITHMS

The brief reviews of existing load balancing algorithms are presented in the following:

4.1.1 Round Robin Load Balancer

It is one of the simplest scheduling techniques that utilize the principle of time slices. The time is divided into multiple slices and each node is given a particular time interval i.e. it employs the principle of time scheduling. Each node is given a time slice and in this time slice the node will perform its operations [4].

This algorithm works on random selection of the virtual machines. The datacenter controller assigns the requests to a list of virtual machines on a rotating basis.

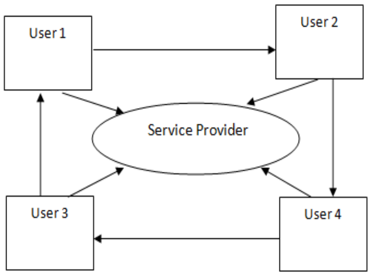

Figure 4: RR Algorithm

The first request is allocated to a virtual machine picked randomly from the group and then the Data Center controller assigns the requests in a circular order to each virtual machine as shown in figure 4. Once the virtual machine is assigned the request, the virtual machine is moved to the end of the list.

4.1.2 Weighted Round Robin

Another way to define round robin algorithm is a better allocation concept known as Weighted Round Robin. Allocation in which one can assign a weight to each virtual machine so that if one virtual machine is capable of handling twice as much load as the other, the more powerful server gets a weight of 2. In this cases, the DataCenter Controller will assign two requests to the powerful virtual machine for each request assigned to a weaker one. The key issue in this allocation is this that it does not consider the advanced load balancing requirement such as processing times for each individual requests [10].

4.1.2 Equally Spread Current Execution Algorithm

Equally spread current execution algorithm process handle with priorities. ESCE algorithm distribute the load randomly by checking the size and transfer the load to that virtual machine which is lightly loaded or handle that task easy and take less time to accomplish the task and maximize throughput. ESCE algorithm a spread spectrum technique in which the load balancer spread the load of the job in hand into multiple virtual machines [6].

It is spread spectrum technique in which the load balancer spread the load of the job in hand into multiple virtual machines. The load balancer helps to sustain a queue of the jobs that need to use and are currently using the services of the virtual machine. The balancer continuously scans this queue and the list of the virtual machines. If there is a virtual machine available that can handle request of the node/client, the virtual machine is allocated to that request. If however there is a virtual machine that is free and there is another virtual machine that needs to be freed of the load, then the balancer allocates tasks of that virtual machine to the free one so as to reduce the overhead of the former virtual machine.

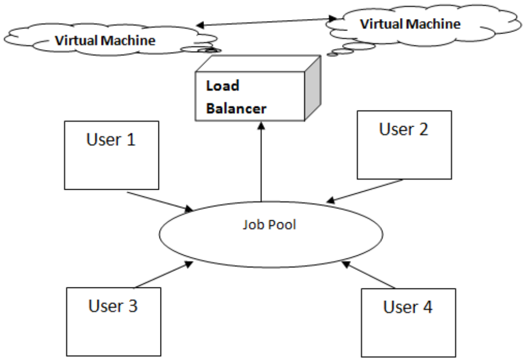

Figure 5: ESCE Algorithm

The jobs are submitted to the virtual machine manager as shown in figure 5, the load balancer also maintains an array of the jobs, their resources requested and the size. The balancer selects the job that matches the criteria for execution at the present time [14].

4.1.3 Active Monitoring Load Balancer Active virtual machine Load Balancer maintains information about each virtual machine and the number of requests currently allocated to which virtual machine. When a request to allocate a new virtual machine arrives, it identifies the least loaded virtual machine. If there are more than one virtual machine, the first identified is selected. The virtual machine which is active, load balancer returns the virtual machine id to the DataCenter Controller the datacenter Controller sends the request to the virtual machine identified by that id DataCenter Controller warns the Active virtual machine Load Balancer of the new allocation [16].

D. Throttled Load Balancer

Throttled algorithm is completely based on virtual machine. In this algorithm a client first request the load balancer to check the right virtual machine which access that load easily and perform the operations which is given by the client or user. In this algorithm the user first requests the load balancer to find a suitable virtual machine to perform the required operation [10] as it is shown in figure 6.The process first starts by maintaining an array of all the virtual machines each row is individually indexed to speed up the lookup process and, if a match is found on the basis of availability of the machine and size then the load balancer accepts the request of the client and allocates that virtual machine to the client.

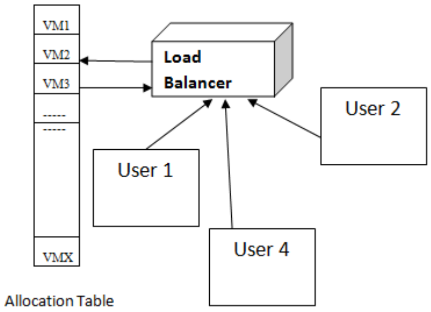

Figure 6: Throttled

If there is no such virtual machine available that matches the certain criteria then the load balancer returns -1 and the request is queued [14].

4. SIMULATORS

The main aim of simulator is to test the implementation work in the absence of the required environment. Thus in the cloud environment two simulator are used CloudSim and Vcloud.

4.1 CloudSim

CloudSim is a new generalized and extensible simulation framework that enables seamless modeling, simulation, experimentation of emerging. The simulation framework has the following novel features: (i) support for modeling and instantiation of large scale Cloud computing infrastructure, (ii)self-contained platform for modeling datacenters, service brokers, scheduling, and allocations policies; (iii) availability of virtualization engine, this availability helps in establishment and supervision of independent, multiple and co-hosted virtualized services on a datacenter node. [15]

5. CONCLUSION

As the aim of cloud computing is to provide services to the users on demand. The issue of disclosing the availability of virtual machines to the client will improve the performance level of the cloud computing. For allocation of efficient virtual machines on demand all have to decide efficient load balancing algorithm. Whenever the user came to know about any free available virtual machines they can choose whether they want to use service from that cloud or not. Thus while caring out these issue we can able to have a better service from the cloud computing.

REFERENCE

[1] Suriya Begum, Dr. Prashanth C.S.R,"Review of Load Balancing in Cloud Computing", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 2, 2013.

[2] M.Sudha, M.Monica," Enhanced Security Framework to Ensure Data Security in Cloud Computing Using Cryptography", ACSA, vol. 1, no. 1, pp. 32-37, 2012.

[3] Bhasker Prasad Rimal, Eummi Choi, Lan Lump"A Taxonomy and Survey of Cloud Computing System"5th International Joint Conference on INC, IMS and IDC, IEEE Explore 25-27Aug 2009, pp. 44-51,2014

[4] Ajith Singh. N, M. Hemalat,"An Approach on Semi-Distributed Load Balancing Algorithm for Cloud Computing System", International Journal of Computer Applications (0975 –8887) Volume 56– No.12, 2012.

[5] Nigni Jain Kansal, Inderveer Chana,"Existing load balancing techiques in cloud computing: A systematic review", Journal of Information Systems and Communication, Vol. 3, Issue 1, 87–91, 2012.

[6] Priyanka Gupta, Ashok Verma,"Concept of VPN on Cloud Computing for Elasticity by Simple Load Balancing Technique" International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 5, 2012.

[7] Gurudatt Kulkarni & Jayant Gambhir, Tejswini Patil, Amruta Dongare,"A Security Aspects in Cloud Computing", 978-1-4673-2008-5, 2012.

[8] Mell, P., Grance, T., & Grance, T. (n.d.). "The NIST Definition of Cloud Computing Recommendations of the National Institute of Standards and Technology".

[9] Marin Marinov," Intuitionistic fuzzy load balancing in cloud computing", 8th Int. Workshop on IFSs, Banská Bystrica, Vol. 18, No. 4, 19–25,2012

[10] Rudra Koteswaramma," Client-Side Load Balancing and Resource Monitoring in Cloud", International Journal of Engineering Research and Applications, Vol. 2, Issue 6, pp.167-171,2012.

[11] Soumya Ray and Ajanta De Sarkar ," Execution Analysis Of Load Balancing Algorithms In Cloud Computing Environment", International Journal on Cloud Computing: Services and Architecture (IJCCSA), Vol.2, No.5, 2012. [12] Zenon Chaczko, Venkatesh Mahadevan, Shahrzad Aslanzadeh, Christopher Mcdermid,"Availabity and Load Balancing in Cloud Computing" International Conference on Computer and Software Modeling IPCSIT vol.14 IACSIT Press, 2011.

[13] Carnegie Mellon, Grace Lewis, "Basics about Cloud Computing", Software Engineering Institute, 2010.

[14] Tanveer Ahmed, Yogendra Singh, "Analytic Study Of Load Balancing Techniques Using Tool Cloud Analyst",International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 2, pp.1027-1030,2012.

[15] MR.Manan D. Shah, MR. Amit A. Kariyani, MR.Dipak L. Agrawal," Allocation Of Virtual Machines In Cloud Computing Using Load Balancing Algorithm" IRACST-(IJCSITS),Vol. 3, No.1,2013.

[16] Raj, G.,"Comparative Analysis of Load Balancing Algorithms in Cloud Computing", 1(3), 120–124, 2012.

[17] Sharma, M., & Sharma, P. , "Efficient Load Balancing Algorithm in VIRTUAL MACHINE Cloud Environment", 8491,439–441.-43, 2012.