A Survey Of Various Load Balancing Techniques And Challenges In Cloud Computing

Tushar Desai, Jignesh Prajapati

Abstract: Cloud computing is emerging technology which is a new standard of large scale distributed computing and parallel computing. It provides shared resources, information, software packages and other resources as per client requirements at specific time. As cloud computing is growing rapidly and more users are attracted towards utility computing, better and fast service needs to be provided. For better management of available good load balancing techniques are required. So that load balancing in cloud becoming more interested area of research. And through better load balancing in cloud, performance is increased and user gets better services. Here in this paper we have discussed many different load balancing techniques used to solve the issue in cloud computing environment.

Index Terms: Cloud Computing, Cloud Service Model, Load Balancing, Optimization, Priority, Task Scheduling, Utility Computing, Virtualization.

1 INTRODUCTION

A Cloud refers to a distinct IT environment that is designed for the purpose of remotely provisioning scalable and measured IT resources [1]. It is a type of computing in which resources are shared rather than owning personal devises or local personal servers which can be used to handle applications on system. The word cloud in cloud computing is used as a metaphor for internet so we can define a cloud computing as the internet based computing in which the different services like storage, servers and application are provided to organizations computers and device using internet[2]. So as compared to traditional "own and use" technique if we use cloud computing, the purchasing and maintenance cost of infrastructure is eliminated. It allows the users to use resources according to the arrival of their needs in real time. Thus, we can say that cloud computing enables the user to have convenient and on-demand access of shared pool of computing resource such as storage, network, application and services, etc.. On pay per use basis.

1.1 CLOUD COMPUTING ARCHITECTURE

Cloud computing is growing in the real time environment and the information about the cloud and the services it provide and its deployment models are discussed. Figure 1 illustrating the three basic service layers that constitute the cloud computing. It provides three basic services that are Software as a Service, Platform as a Service and Infrastructure as a Service [2]. The rest of the paper is organized as follows. In section 2 we discussed virtualization of cloud. In section 3, load balancing and it's necessitate is discussed. In section 4 various load balancing techniques are discussed. And in section 5 Challenges of load balancing in cloud computing are explained.

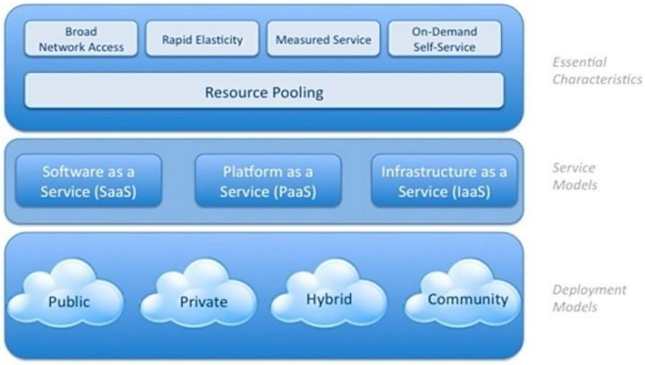

Figure 1: cloud computing architecture

2 CLOUD VIRTUALIZATION

In context of cloud computing the virtualization is very worthwhile concept. Virtualization is like "something that is not real" but provides all the facilities that are of real world. This is a software implementation of computer on which different programs can be executed as in the real machine. Virtualization is a part of cloud computing, because different services of cloud can be used by user. All these different services are provided to end user by remote data centers with full virtualization or partial virtualization manner [4]. There are two types of virtualization which are available and it is described below.

2.1 FULL VIRTUALIZATION

In full virtualization the entire installation of one system is done on other system. Due to this all the software that are present in actual server will also available in virtual system and also sharing of computer system among multiple users and emulating hardware located on different systems are possible.

2.2 PARA VIRTUALIZATION

In this type of virtualization, multiple operating systems are allowed to run on a single system by using system resources like memory and the processor (VMware software). Here complete services are not fully available, but partial services

- *Tushar Desai is currently pursuing masters degree program in Computer science and engineering in Gujarat Technical University, India, PH-09033218544. E-mail: tushar_it2510@yahoo.in*
- *Jignesh Prajapati is currently working as Assistant Professor in Computer Science & Engineering Department in Parul Institute of Technology, Gujarat, India. PH-09033462993. E-mail: jig4physics@gmail.com*

are provided. Disaster recovery, migration and capacity management are some salient features of Para virtualization.

3 LOAD BALANCING

Load balancing is one of the main issues related to cloud computing. The load can be a memory, CPU capacity, network or delay load. It is always required to share work load among the various nodes of the distributed system to improve the resource utilization and for better performance of the system. This can help to avoid the situation where nodes are either heavily loaded or under loaded in the network. Load balancing is the process of ensuring the evenly distribution of work load on the pool of system node or processor so that without disturbing, the running task is completed. The goals of load balancing [16] are to:

- Improve the performance
- Maintain system stability
- Build fault tolerance system
- Accommodate future modification.

There are mainly two types of load balancing algorithms:

3.1 STATIC ALGORITHM

In static algorithm the traffic is divided evenly among the servers. This algorithm requires a prior knowledge of system resources, so that the decision of shifting of the load does not depend on the current state of system. Static algorithm is proper in the system which has low variation in load.

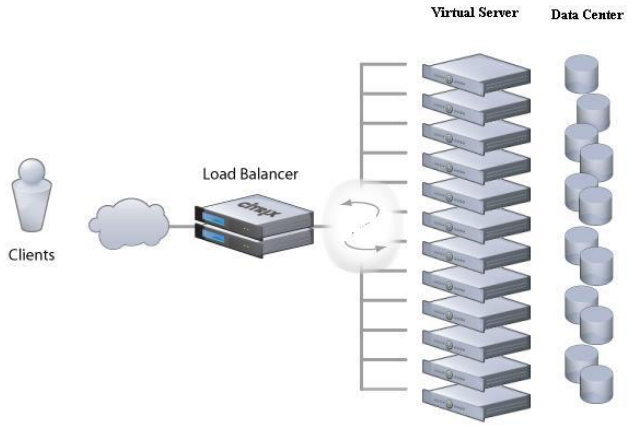

Figure 2: load balancing in cloud computing [3]

3.2 DYNAMIC ALGORITHM

In dynamic algorithm the lightest server in the whole network or system is searched and preferred for balancing a load. For this real time communication with network is needed which can increase the traffic in the system. Here current state of the system is used to make decisions to manage the load.

4 AVAILABLE LOAD BALANCING TECHNIQUES

In this section we discuss existing load balancing techniques in cloud computing. Here we classify load balancing algorithm in two main types that are Static load balancing and Dynamic load balancing. In 2009, B Sotomayor et al [4] introduced a static well-known load balancing technique called Round Robin, in which all processes are divided amid all available processors. The allocation order of processes is maintained locally which is independent of the allocation from the remote processor. In this technique, the request is sent to the node having least number of connections, and because of this at some point of time, some node may be heavily loaded and other remain idle [4]. This problem is solved by CLBDM. In 2010, S C. Wang et al. [5] presented a dynamic load balancing algorithm called load balancing Min-Min (LBMM) technique which is based on three level frameworks. This technique uses Opportunistic Load Balancing algorithm which keep each node busy in the cloud without considering execution time of node. Because of this it causes bottle neck in system. This problem is solved by LBMM three layer architecture. First layer request manager which is responsible for receiving task and assigning it to one service manager to second level. On receiving the request service manager divide it into subtasks. After that service manager will assign subtask to service node to execute task. In 2011, B. Radojevic et al [6] introduced a static load balancing algorithm called CLBDM (Central Load Balancing Decision Model). CLBDM is an enhancement of the Round Robin technique. This is based on session switching at application layer. In round robin, request is sent to the node having least number of connections. RR is enhanced and in CLBDM, the calculation of the connection time between the client and the node is done and if the connection time goes above the threshold then problem is raised. If a problem is arises, then the connection between the client and the node is terminated and the Task is forwarded to the further node using Round Robin law. In 2011, L. Colb et al [7] introduced the Map Reduced based Entity Resolution load balancing technique which is based on large datasets. In this technique, two main tasks are done: Map task and Reduce task which the author has described. For mapping task, the PART method is executed where the request entity is partitioned into parts. And then COMP method is used to compare the parts and finally similar entities are grouped by GROUP method and by using Reduce task. Map task reads the entities in parallel and process them, so that overloading of the task is reduced. In 2011, J Hu et al. [8] introduced a static scheduling strategy of load balancing on virtual machine resource. This technique considers the historical data and also the current state of system. Here, central scheduler and resource monitor is used. The scheduling controller checks the availability of resources to perform a task and assigns the same. Resource availability details are collected by resource monitor. In 2011, J Al-Jaroodi et al. [9] proposed a dynamic load balancing technique named DDFTP (Duel Direction Downloading Algorithm from FTP server). This can also be implemented for load balancing in cloud computing. In DDFTP, file of size m is divided into m/2 partition and each node starts processing the task. For example if one server starts from 0 to incremental order than other will start from m to detrimental order independently from each other. As on downloading two consecutive blocks the task is considered as finished and assigned next task to server. Because of reduction in network communication between client and node network overhead is reduced. In 2012, K. Nishant et al [10] introduced a static load balancing technique called Ant Colony Optimization. In this technique, an ant starts the movement as the request is initiated. This technique uses the Ants behavior to collect information of cloud node to assign task to the particular node. In this technique, once the request is initiated, the ant and the pheromone starts the forward movement in the pathway from

the “head” node. The ant moves in forward direction from an overloaded node looking for next node to check whether it is an overloaded node or not. Now if ant find under loaded node still it move in forward direction in the path. And if it finds the overloaded node then it starts the backward movement to the last under loaded node it found previously. In the algorithm [8] if ant found the target node, ant will commit suicide so that it will prevent unnecessary backward movement. In 2012, T. Yu Wu et al. [11] introduced a dynamic load balancing technique called Index Name Server to minimize the data duplication and redundancy in system. This technique works on integration of de duplication and access point optimization. To calculate optimum selection point some parameter are defined: hash code of data block to be downloaded, position of server having target block of data, transition quality and maximum bandwidth. Another calculation parameter to find weather connection can handle additional node or is at busy level B(a), B(b) or B(c). B(a) denote connection is very busy to handle new connection , B(b) denotes connection is not busy and B(c) denotes connection is limited and additional study needed to know more about connection. In 2012, B. Mondal et al [12] have proposed a load balancing technique called Stochastic Hill Climbing based on soft computing for solving the optimization problem. This technique solves the problem with high probability. It is a simple loop moving in direction of increasing value which is uphill. And this make minor change in to original assignment according to some criteria designed. It contains two main criteria one is candidate generator to set possible successor and the other is evaluation criteria which ranks each valid solution. This leads to improved solution. In 2013, D. Babu et al [13] proposed a Honey Bee Behavior inspired Load Balancing [HBB-LB] technique which helps to achieve even load balancing across virtual machine to maximize throughput. It considers the priority of task waiting in queue for execution in virtual machines. After that work load on VM calculated decides weather the system is overloaded, under loaded or balanced. And based on this VMs are grouped. New according to load on VM the task is scheduled on VMs. Task which is removed earlier. To find the correct low loaded VM for current task, tasks which are removed earlier from over loaded VM are helpful. Forager bee is used as a Scout bee in the next steps.

5 Challenges For Load Balancing

There are some qualitative metrics that can be improved for better load balancing in cloud computing [14][15].

Throughput: It is the total number of tasks that have completed execution for a given scale of time. It is required to have high through put for better performance of the system.

Associated Overhead: It describes the amount of overhead during the implementation of the load balancing algorithm. It is a composition of movement of tasks, inter process communication and inter processor. For load balancing technique to work properly, minimum overhead should be there.

Fault tolerant: We can define it as the ability to perform load balancing by the appropriate algorithm without arbitrary link or node failure. Every load balancing algorithm should have good fault tolerance approach.

Migration time: It is the amount of time for a process to be transferred from one system node to another node for execution. For better performance of the system this time should be always less.

Response time: In Distributed system, it is the time taken by a particular load balancing technique to respond. This time should be minimized for better performance.

Resource Utilization: It is the parameter which gives the information within which extant the resource is utilized. For efficient load balancing in system, optimum resource should be utilized.

Scalability: It is the ability of load balancing algorithm for a system with any finite number of processor and machines. This parameter can be improved for better system performance.

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

6 Conclusion and Future Work

In this paper, we have surveyed various load balancing techniques for cloud computing. The main purpose of load balancing is to satisfy the customer requirement by distributing load dynamically among the nodes and to make maximum resource utilization by reassigning the total load to individual node. This ensures that every resource is distributed efficiently and evenly. So the performance of the system is increased. We have also discussed virtualization of cloud and required qualitative matrix for load balancing.

References

[1]. Basic concept and terminology of cloud computing-http://whatiscloud.com

[2]. L. Wang, J. Tao, M. Kunze,”Scientific Cloud Computing: Early Definition and Experience”, the 10th IEEE International Conference Computing and Communications 2008.

[3]. Load Balancing in Cloud computing, http://community.citrix.com/display/cdn/Load+Balancing

[4]. Sotomayor, B., RS. Montero, IM. Llorente, and I. Foster, "Virtual infrastructure management in private and hybrid clouds," in IEEE Internet Computing, Vol. 13, No. 5, pp: 14-22, 2009.

[5]. Wang, S-C., K-Q. Yan, W-P. Liao and S-S. Wang, "Towards a load balancing in a three-level cloud computing network," in proc. 3rd International Conference on. Computer Science and Information Technology (ICCSIT), IEEE, Vol. 1, pp: 108-113, July 2010.

[6]. Radojevic, B. and M. Zagar, "Analysis of issues with load balancing algorithms in hosted (cloud) environments." In proc. 34th International Convention on MIPRO, IEEE, 2011.

[7]. Kolb, L., A. Thor, and E. Rahm, E, "Load Balancing for MapReduce based Entity Resolution," in proc. 28th International Conference on Data Engineering (ICDE),

IEEE, pp: 618-629, 2012.

[8]. J. Hu, J. Gu, G. Sun, and T. Zhao, “A Scheduling Strategy on Load Balancing of Virtual Machine Resources in Cloud computing Environment”, Third International Symposium on Parallel Architectures, Algorithms and Programming (PAAP), 2010.

[9]. Al-Jaroodi, J. and N. Mohamed. "DDFTP: Dual-Direction FTP," in proc. 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), IEEE, pp:504-503, May 2011.

[10]. Nishant, K. P. Sharma, V. Krishna, C. Gupta, KP. Singh N. Nitin and R. Rastogi, "Load Balancing of Nodes in Cloud Using Ant Colony Optimization." In proc. 14th International Conference on Computer Modeling and Simulation (UKSim), IEEE, pp: 3-8, March 2012.

[11]. T-Y., W-T. Lee, Y-S. Lin, Y-S. Lin, H-L. Chan and J-S. Huang, "Dynamic load balancing mechanism based on cloud storage" in proc. Computing, Communications and Applications Conference (ComComAp), IEEE, pp: 102-106, January 2012.

[12]. Brototi M, K. Dasgupta, P. Dutta, “Load Balancing in Cloud Computing using Stochastic Hill Climbing-A Soft Computing Approach”, in proc. 2nd International Conference on Computer, Communication, Control and Information Technology(C3IT)-2012.

[13]. Dhinesh B. L.D , P. V. Krishna, “Honey bee behavior inspired load balancing of tasks in cloud computing environments”, in proc. Applied Soft Computing, volume 13, Issue 5, May 2013, Pages 2292-2303.

[14]. Foster, I., Y. Zhao, I. Raicu and S. Lu, “Cloud Computing and Grid Computing 360-degree compared,” in proc. Grid omputing Environments Workshop, pp: 99-106, 2008.

[15]. Buyya R., R. Ranjan and RN. Calheiros, “InterCloud: Utilityoriented federation of cloud computing environments for scaling of application services,” in proc. 10th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP), Busan, South Korea, 2010.

[16]. D. Escalnte, Andrew J. Korty, “Cloud Services: Policy and Assessment”, Educause review July/August 2011.