ISSN No. 0976-5697

Volume 5, No. 1, Jan-Feb 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Survey of Load Balancing Techniques in Cloud Environment

Kumughato G[1], Mrs Jeba priya[2]
PG student of Information Technology[1], Assistant Professor[2]
Department of Computer Science & Information Technology
Karunya University, Coimbatore, India
kumughatochophy@karunya.edu.in

Abstract: Cloud computing of late has been the talk in the field of computer technology environment whereby the resources are distributed over the internet. The main advantages has been that of sharing the computing power as well as the storage area over a large network within limit budget and also the provision of being available whenever there is a need and demand but it also comes along with some issues and challenges. Load balancing is one such area and several techniques and methodology has been purposed to achieve a good load balancing in the cloud environment. A good load balancing technique will lead to a better performance and efficiency of the computation in the cloud environment. The resource in the cloud based environment should be utilized and distributed evenly in such a way that it should not lead to the over utilization on one particular resource whereas the other resources that are available for computation should not be kept idly. The overall objective should be that to utilize each and every resources that are available in cloud efficiently without putting an extra load on the other resources. This paper is a survey on the current load balancing techniques that are available in the cloud

Keywords: Load balancing, Cloud computing, resource allocation, para-virtualization, dynamic allocation.

I. INTRODUCTION

Cloud computing is a new form of a distributed system that provides resources in the distributed environment [1]. This new technology provides the requirement of the resources on the go and based on the user need. Cloud computing also provide a user of the resource in the system to pay for the resource as it is being utilized [2] and there is no need for initial investment and budget constraint, it comes with the optionality of pay as you use the resources. Due to this, it has found its way to the business industry and there is a rapid demand for this technology and business and IT establishment are currently moving to the cloud domain. The only back draw is that it should meet the criteria of being available all the time as during the transaction of the business process any failure cannot be tolerated. The other issue is that of load balancing, although the load balancing is achieved to some extent due to the concept of virtualization yet a satisfactory extend of load balancing cannot be done alone by the use the virtualization. Virtualization is a concept in which the system is not present in real world but it will give a sense of its presence to the user and will make the user feel as if it is working in a real world scenario with all the system requirement available for computation while in real sense the system is unavailable but just the extension of the real world making the end user believe that it is working on the real world environment. There are basically two type of virtualization that is done for the cloud service. One is the full virtualization while the other is the para virtualization.

The main difference between the two is that in the full virtualization the installation of one server machine is done over the other server machine and will contain every software that should be available on the real machine. In the case of para virtualization not everything is emulated and we can have many operating system running on a same virtual machine with the condition that the guest OS should be tweaked or modified to be compatible to run on the current virtual machine. Each guest OS will be aware and know the existence of one another and how much is the load it is applying on the real server. Para virtualization is more beneficial over the full virtualization because it allows the processor and memory to be shared among different guest OS with the aide of resource sharing

The provider of cloud computing are offering the availability of their service over the internet to the customer and are done dynamically by providing the resource required such as those relating to software and hardware. Some of the major cloud service provider includes Rackspace, Amazon Web Services, Google Drive, Sky drive etc. The service model of the cloud computing are further dived based on the service that they are providing. The National Institute of Standards and Technology (NIST) Definition of Cloud Computing for the service model are described [3]

a. ***Infrastructure as a service (IaaS):*** As the name suggest in this service model it has got to do more with the providing of the infrastructure to the end user of the system. By the term infrastructure it is related to the hardware component such as processing unit, the network equipment and also the storage space to the end user who are able to perform processing and also able to run the software. In this given model of the service infrastructure, the service provider is solely maintained by the provider of the service and end user will have to pay for only what resource they have consumed. Some of the examples are Window Azure and Amazon EC2
b. ***Platform as a service (Paas):*** This service provide for the platform for the execution of the application user have built. This service provide the necessary tool and programming languages and also support the operating system along with the database and further option of the availability of web server. Some of the common examples of PaaS are Google App Engine, Force.com
c. ***Software as a service (SaaS):*** The cloud service provider can also provide software as s service to the end user or the client. In the SaaS there is no need to worry for the process of installation, licensing, and

updating as well as the maintenance of the software as everything are taken care by the provider of the service. The end user will have to just pay for the subscription of the software which makes things more convenient and smooth even for many IT company as a lot of budget are spend alone in the maintenance of the software. SaaS is usually referred to as a on-demand software. Example of the common SaaS are Google App.

In this survey paper firstly some basic concept and ideas about cloud computing is described and also some idea on virtualization and their types are discussed, also the major service model on the cloud are summarized briefly. In section II the focus is about the load balancing and why there is a need for load balancing in the cloud environment, also the types of load balancing are briefed out. In section III the different available load balancing techniques for cloud computing are summarized out .Section IV describe the various metrics that are considered for the load balancing technique. Finally section V gives the conclusion of the paper.

II. LOAD BALANCING

Load balancing is a methodology in which the work load of one server are distributed evenly among the different system that are available in the networking environment so as to reduce the load on one particular system and to achieve equilibrium among the different system. This will in turn lead to the efficient utilization of the computational resources and also provide the process to be performed more quickly thereby saving time and energy. The main reason behind the use of the load balancing as such is the optimization of the resource that are available over the internet and to reduce the response duration from the server side as well as to achieve an efficient throughput and mainly to prevent the overloading of a particular resources while other may be in idle state without performing any operation.

Load balancing is considered to be one of the main challenging features for the cloud environment [4]. As cloud comes with the feature of elasticity so load balancing for any computation process cannot be ignored. A load balancing in the cloud environment should provide for the instance of the application to be executed dynamically over the cluster of the system without the need for stopping the whole working operation and without even having to change any configuration of the system for that matter. The load balancing in the cloud will provide the mechanisms where the server are distributed evenly for the request that are received and provide for the quicker response without having to wait a longer time. The load balancing will ensure that the work load are distributed evenly among the different resources and that no resources are over-utilized or under-utilized and that a proper ratio utilization are done on the resource that are available over the cloud. It is noteworthy to mention that load balancing will also increase the scalability and provide the mechanism of fail-over by allocating and de-allocating instance request of the application

There are basically two types of load balancing that are available they are the static load balancing and the dynamic load balancing. In the static load balancing, the balancing is done prior and they are done based on the deterministic or the probabilistic nature as such no changes can be made during the execution of the operation. Also in the static load balancing the resource are shared in an equal manner and we cannot determine exactly the time of its execution period. In a dynamic load balancing, the load are distributed dynamically during the execution of the system and there need to be a monitoring system to provide for the communication among the various server and the current load of a particular system. The current states of the resource are monitored and load are changed if necessary in a dynamic manner. The static load balancing are more easier to implement over the dynamic load balancing which are more complex as it is done dynamically

III.LOAD BALANCING TECHNIQUES IN THE CLOUD

Following are some of the load balancing techniques for the cloud computing environment discussed in this section, these are vetctorDot, Dual Direction FTP, CARTON and Join-Idle-Queue

A. Vectordot Load Balancing Technique: The vector Dot load balancing technique proposed in [5] is used in the environment which includes the combination of both server and storage virtualization. The given vectorDot algorithm helps in solving the problem of hierarchy and multidimensional limitation. When a given server or storage disks or even a network switch is overloaded the relocation of the application process are moved to another service system center without hindering the current operation that are being executed. The vectorDot help in this process by deciding where the relocation needed to be done. It handles the need of specific requirement such as processing power, storage capacity and the network bandwidth and where can this requirement be given from another servers without causing a further overload on a new server. The direction of the path can also have an impact on the node load of other system so VectorDot provide a methodology whereby a suitable path from the current loaded server to a new server is achieved by putting a minimum load on other node along the path. The vectorDot will calculate the node load fraction vector as well as its threshold vector, the node can be a server, or a storage or even a switch. The item load fraction vector will collect the requirement it will be needing.

B. The Dual-Direction FTP Technique: Jameela Al-Jarood et al [6] proposed a DDFTP. It is a technique used for the download of large files from the replicated server. This technique divide or partition the file block into equal size at different location where the files are located and are downloaded simultaneously from different server. Load balancing are achieved dynamically by the following manner. Suppose say for example we are going to download a file from two server over the cloud then the file size are first of all calculated and then the file are divided into equal blocks and one server will have to send the first partition block beginning from the front end whereas the second server will also have to send the second partition of the block size from the back end of the file block. This download operation are done simultaneously from the two server. During the operation of download if server one finishes its download operation earlier than that of server two,

then it will help in the operation of sever two so if sever one is free it will help in the load under which server two is currently under operation . In this way load balancing are achieved dynamically, also this technique are done based on the TCP so the file block size transfer are carried out in a manner of first-in-first-out (FIFO) sequence so there is no need to add additional header file to keep track of the file block being transferred. In this technique of load balancing, every server will work according to its own capacity and the server which are having higher bandwidth capacity will compensate those of the server which are connected with the lower bandwidth connection and this are done automatically. Thus this technique provide load balancing for the large file download operation in the cloud environment.

C. Carton Technique: Rade Stanojevic et al. [7] introduce a new framework. This method combines both the load balancing as well as the distributed rate limiting (DRL) to provide an equitable allocation of the resource to the incoming request from the application process. The load balancing are used mainly for reducing the cost involved in the process of execution while the DRL achieves the need of sharing the resource in an equal manner. In this technique the algorithm are achieved in two state. In the first state, a method called Sub gradient is done that will allocate a job to the process such that the cost is reduced. In the second state a DRL algorithm is applied such that the resource on the server are allocated in an equal manner to all the instance of the application.

D. Join-Idle-Queue: Y. Lua et al. [8] proposed a new load balancing algorithm called Join-Idle-Queue technique to achieve a dynamic and scalable web services. In this technique, the processor which are not performing any task will inform the dispatcher of its idealness. There will arise a situation where the ideal processor will have to be careful regarding which dispatcher to inform as again extra load may be given on the ideal processor. To overcome this situation, two load balancing are done namely the primary and secondary load balancing. The primary load balancing make use of the I-queue where the idea processor are aligned in queue. With the arrival of the first task it is given to the first processor in the I-queue. The secondary load balancing chooses the ideal processor from the I-queue in random manner or the shortest queue length.

IV. LOAD BALANCING METRICS FOR CLOUD COMPUTING

The metrics of the load balancing for cloud computing are discussed below:

a. Performance- is used for measuring the effectiveness of the system and how well they are performing under the load

b. Scalability- for the given technique of load balancing it should be scalable to include many other system and still be able to perform well

c. Resource utilization ratio- the load balancing algorithm should provide and make that the resource in the cloud are utilized by sharing the resource equally.

d. Throughput- It is the amount of work or task that are completed in a given amount of time. A good load balancing algorithm should provide for the lesser throughput time

e. Response time- It is the time taken for the system to respond to a particular load balancing technique

f. Overhead- It gives the overhead that are associated with the given load balancing algorithm. The overhead incurred can be due to the migration of the task and the process of passing of communication between the various process. The overhead should be decreased.

V. CONCLUSIONS

This paper had briefly described about the various load balancing techniques that are available for the cloud computing environment and also in the given paper the need for load balancing in the cloud computing are discussed. One of the key component for the success of cloud computing technology lies in a good load balancing technique. With a good load balancing technique the reduction in budget can be achieved, also the resource are utilized in an efficient manner. With the proper resource utilization also comes the factor of efficient energy consumption.

VI. REFERENCES

[1]. M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, “Above the Clouds: A Berkeley View of Cloud Computing”, EECS Department, University of California, Berkeley, Technical Report No., UCB/EECS-2009-28, pages 1-23, February 2009.

[2]. R. W. Lucky, “Cloud computing”, IEEE Journal of Spectrum, Vol. 46, No. 5, May 2009, pages 27-45.

[3]. Peter Mell, Timothy Grance, “The NIST Definition of Cloud Computing”, NIST Special Publication 800-145, September 2011.

[4]. B. P. Rima, E. Choi, and I. Lumb, “A Taxonomy and Survey of Cloud Computing Systems”, Proceedings of 5th IEEE International Joint Conference on INC, IMS and IDC, Seoul, Korea, August 2009, pages 44-51.

[5]. Singh A., Korupolu M. and Mohapatra D. (2008) ACM/IEEE conference on Supercomputing.

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

[7]. Stanojevic R. and Shorten R. (2009) IEEE ICC, 1-6.

[8]. Lua Y., Xiea Q., Kliotb G., Gellerb A., Larusb J. R. and Greenber A. (2011) Int. Journal on Performance evaluation.