International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(3), pp.130-134
DOI: http://dx.doi.org/10.21172/1.83.017
e-ISSN:2278-621X

AN INSIGHT TO LOAD BALANCING FOR CLOUD COMPUTING TECHNOLOGY

Karishma Gulati[1], Pronika Chawla[2] and Ms Ekta Malhotra[3]

Abstract- Cloud Computing is one of the most useful emerged technologies which is in trend in IT environment since last decade. It is a standard that has brought a revolution in the area of Distributed Computing. Clients use services as and when required at their remote locations and pay according to the consumption. For better Service-on - demand, load needs to be distributed among various available resources at different Data Centers. So, load balancing is an important concern in cloud computing. Effective Load balancing schemes ensures efficient utilization of resources. This paper presents the issues related to load balancing including the algorithms and metrics along with the challenges faced while balancing the load over a cloud as per specification in Service Level Agreement(SLA).

Keywords – Cloud Computing, Distributed Computing, Data Centers, Service-on-demand, Load Balancing, Resource Utilization and Service Level Agreement(SLA).

I. INTRODUCTION

Today is an era of computation. Computation involves processing of data for performing any computer based operation at faster pace. Cloud Computing is an aggregation of two words, having their specific meaning in the field of technology. They are:' Cloud' and 'Computing'. Cloud [1] is a *pool* of available resources which can be utilized over internet. Technically, Computing refers to data processing tasks; storage allocation, memory access, creating a buffer storage and carrying out other applications. So ,cloud computing can be summarized as the internet based computing providing different services on-demand such as storage, servers and applications to various organizations through internet. This 'Pay-as-you-go' technology saves substantial cost of purchasing and maintaining the infrastructure as compared to the traditional “own and use” technique.

Effective usage of the resources on-demand calls for their availability to the end-users. For ensuring their availability to users on dynamic basis, resources need to be scheduled. In cloud computing environment, the computing is done on the basis of certain criteria specified in SLA[1]. The resources are made available to the users on-demand basis, for which they are required to pay according to the consumption. Computing resources (e.g. servers, software, storage, network, applications and services) need to be provisioned rapidly with minimum efforts ad cost for achieving maximum utilization.

II. CLOUD STAKEHOLDERS

Different people have different perspectives to get associated with Cloud computing environment. Such people are divided into three main categories: End users, Cloud Developers and Cloud Providers.

[1] *Department of Computer Science and Engineering, Manav Rachna International University, Faridabad, Haryana, India*
[2] *Department of Computer Science and Engineering, Manav Rachna International University, Faridabad, Haryana, India*
[3] *Department of Computer Science and Engineering, Manav Rachna International University, Faridabad, Haryana, India*

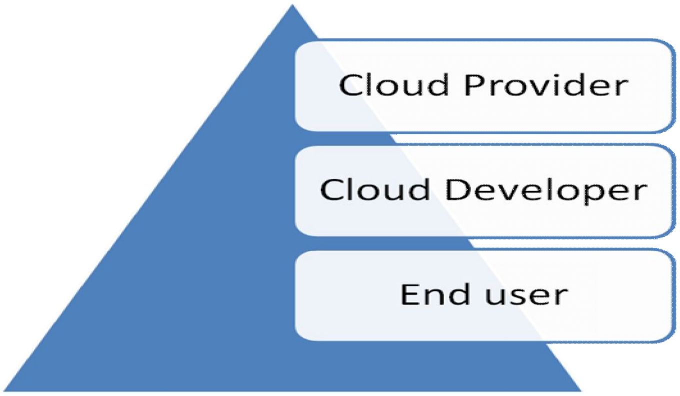

Figure 1: Cloud Stakeholders

2.1 End users

End users are actually the consumers of clouds. They can use various resources remotely (Infrastructure/ Software/Platform/Services) offered by the cloud. Such users must agree to the Service Level Agreement (SLA) specified by the Cloud Provider, before using the cloud services. They use the services on pay-as-you-go basis and pay only for the extent to which the services are consumed. Flexibility while using services can be achieved by incorporating utility computing. So, an end user can access any service, software or infrastructure at minimal cost and with less applicable efforts. Security, Privacy, High Availability, Reduced Cost and Ease-of-use are the major concerns of End users of cloud.

2.2. Cloud Providers

Cloud providers are the enterprises that provide either public or private or hybrid cloud. They build clouds to be used by end users. Various enterprises or business organizations provide Private clouds to other organizations for their domain specific use. They may be used by their employees for storing and managing large volume of data. Security of Private clouds is the main concern of Cloud Providers. Some of the commonly used private clouds are Open Stack, VMware and Cloud Stack. Public clouds are available to be used by an organization or individuals as and when required from anywhere. Confidentiality is the major security issue in using public cloud. Amazon web services, Google Compute Engine, Microsoft Azure, HP cloud are some of the public clouds available in market. A hybrid cloud is a combination of public and private cloud [2].Optimization of hybrid cloud is a challenge to achieve as the needs of users keep on changing. Cloud providers must use “resource provisioning”. There are two main activities involved in Resource provisioning: managing large pool of resources that make up cloud and providing those resources to the end users. Managing Resources , Outsourcing ,Resource Utilization ,Energy Efficiency,-Metering Resources, Providing Resources ,Cost Efficiency ,Meeting end user requirements ,Utility Computing are some of the requirements of Cloud Providers.

2.3. Cloud Developers

Cloud developers lies between cloud providers and end users [2]. They take into consideration both the perspectives of the end user and the cloud provider. They must know all the technical details related to cloud to satisfy the needs of both end users and Cloud Providers. A cloud developer acts as a bridge between the end user of the cloud and the cloud provider. Elasticity/ Scalability, Agility and Adaptability, Virtualization, Reliability, Availability, Data Management and Programmability are the issues related to cloud developers.

III. CLOUD VIRTUALIZATION

Virtualization in context of cloud computing is actually creating an environment where with the use of hardware and software a perception is created that one or more entity exist, although it does not exist actually. In other words, it can be concluded that it is not real but gives the image of a real entity[4].

Virtual environment when implied on a cloud gives the image of running multiple operating systems on a single physical system and enables its users to share the underlying hardware resources easily and remotely at any location. Cloud computing operates fully on the virtualization concept. It improves the efficiency of data centers and activates virtual machine to single physical server.

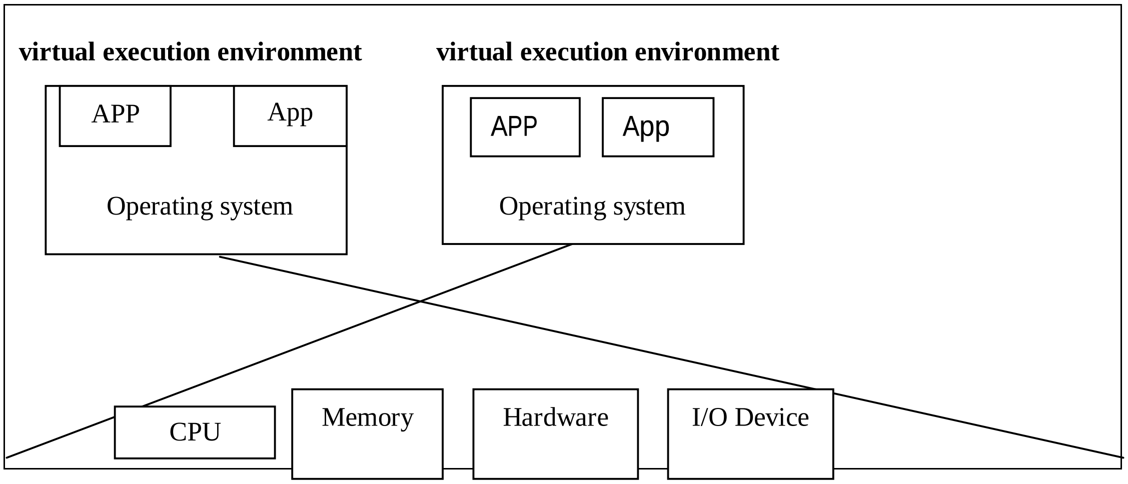

Figure 2: Cloud Virtualization Environment

IV. LOAD BALANCING

Load balancing in cloud computing deals with the problem of effective utilization of resources so as to reduce the response and at the same time there should not be any underutilized resource.

It is the process of reassigning task among the present nodes in the environment such that it ensures no node in the system is over loaded or sits idle for any instant of time. An efficient load balancing algorithm will make sure that every node in the system does more or less same volume of work.[2]

The main concern in load balancing technique is to map the job and allocate to the free node in order to achieve the user demands and to also propagate maximum resource utilization. To achieve the balanced load became one of the crucial concerns in cloud computing since we cannot predict the number of requests that are issued at each second in cloud environment. The unpredictability is due to the ever changing behavior of the cloud.

4.1 Need of load balancing in cloud computing:

Load Balancing majorly focuses on two tasks; one is to assure maximum resource utilization and another one is task scheduling in the distributed environment. Scheduling defines the way in which different nodes are provisioned. Resource provisioning defines which resource will be available to meet user requirements whereas task scheduling defines the manner in which the allocated resource is available to the end user (i.e. whether the resource is fully available until task completion or is available on sharing basis)[3].\

The efficient task and resource scheduling mechanisms will ensure:

1) There should not be bottlenecks under the resources, so that they are easily available on demand
2) Regardless of load is heavy or low; resources must be utilized efficiently
3) Cost of using resources is reduced.
4) It increases the throughput eventually

V. LOAD BALANCING METRICS

There are some qualitative metrics for improving the load balancing mechanism in cloud computing. These matrices are useful in designing algorithms related to load balancing and resource scheduling. Various metrics are as follows:

1) Response Time: It is the time taken by particular load balancing technique to respond. Response time should be minimum[6]

2) Resource Utilization: It is used to check and verify the utilization of resources. Optimum resources should be

utilized for better load balancing[5].

3) Performance: It is used to check the overall efficiency of the system. Performance of the system should be improved when all the parameters are improved and at a reasonable cost.

4) Scalability: It is the ability of an algorithm to be effectively utilized for an increasing number of nodes in a cloud. Scalability should be improved so that when increasing number of end-users use the available resources, it would be effectively utilized.

5) Throughput: It is the total number of tasks that have been completed in a given time frame. Throughput should be high for better performance [8].

6) Fault Tolerant: It is the ability to perform load balancing algorithm uniformly without node failure. For better load balancing, we should have good fault tolerance approach [8].

7) Overhead Associated: It should be minimum for better load balancing. It is basically the amount of overhead involved during the implementation of load balancing algorithm [6].

VI. ALGORITHMS

There are various available load balancing algorithms. Some of them are:

1. Round–Robin: It works on random selection of virtual machine. It is related to time slice mechanism. According to its name, the work is done in round –robin manner. In this, each node has to given a fixed time slot and has to wait for its turn. The time is allocated to each node in which nodes have to perform their task. It randomly choose first node and jobs are allocated to other node in round robin fashion. The main advantages of this algorithm are that it yields no starvation and gives a faster response [7]. But, some processes have different processing time, so at any time some nodes may be heavily loaded while others remain idle.

2. Min –Min Algorithm: In this, initially there is no assigned task. First, find the minimum completion time for all tasks. Then, among these minimum value is selected and task is allotted to that particular node. The execution time of all other tasks is updated on that machine and task gets discarded from the list of tasks. Then, again the same process is repeated until all the tasks are assigned on the resources [3]. The advantage of this algorithm is that it is a simple and fast algorithm.

The main drawback is it assigns the smaller task first and that one is executed first but larger task will be in waiting stage. It is fewer faults tolerant and less scalable

3. Opportunistic Load Balancing: It is one of the static load balancing algorithms. It does not consider the present workload of the virtual machine. It basically deals with the unexecuted tasks faster and in random order to current node, where every task is randomly assigned to the node. So, the tasks are processed in a slow order and the current execution time of the node is not calculated. This algorithm does not provide a good result.

4. Max-Min Load Balancing: It is also one of the static load balancing algorithms [3]. It is basically same as Min-Min algorithm, where the maximum time jobs are selected. In this, once the minimum time jobs are completed, the tasks which are in the queue assigned to the processor? The execution time of every task is updated to the processor. So, it performed in a correct order and time of every task is calculated in advance.

5. Throttled Load Balancing (TLB): It is one of the dynamic load balancing algorithms. In this algorithm load balancer maintain an index table related to virtual machine with their state (available, busy). When a request from client/server comes to data centre for a virtual machine to perform the recommended job. The data centre sends this query to load balancer. Load Balancer scan the index table until the first available virtual machine is found. The id of that particular virtual machine sent to data centre, now the data centre respond to the client/server with the id as requested [11]. Further, the data centre acknowledges the load balancer to new allocation and data centre update the index table accordingly. If all the virtual machines are busy at that time then load balancer return -1 to data centre.

VII. CHALLENGES

1) Automated Service Provisioning: The main feature of cloud computing is its elasticity which allows it to allocate and release its resources automatically. The Problem faced is how to release its resources by maintaining the same performance and using optimum resources[9]

2) Energy Management: The main challenge of load balancing is to save energy. Global economy will be achieved by using a set of global resources supplied by the resource providers rather than by using local resources[5]. So, the challenge faced is to find out better performance with minimum resource consumption.

3) Stored Data Management: To Manage the stored data at individual's site poses a major challenge for cloud computing. Then, how to achieve optimum storage capability in faster way.

4) Virtual Machine Migration: In the virtualization the overall machine can be treated as a single file or set of files. So, to balance a load on physical machine we require virtual machine between them. To avoid the congestion how can we dynamically distribute the load in virtual machines?[9]

VIII. CONCLUSION

In this paper, we have done an in-depth study of load balancing mechanism. We have discussed its significance in cloud computing technology along with the commonly used algorithms for load balancing. We have reviewed various metrics which must be taken into account while designing any load balancing algorithm. Cloud computing is a widely accepted and popular technology used by every industry now a day. We can conclude that the central issue in cloud computing technology is to distribute excess dynamic local workload evenly to all the available nodes in the whole cloud. So, we have also discussed the challenges associated while dealing with balancing excess workload over a cloud. In future, we would like to discuss the solution to these challenges faced so far.

REFRENCES

[1] Katyal, Mayanka, and Atul Mishra. "A comparative study of load balancing algorithms in cloud computing environment." *arXiv preprint arXiv:1403.6918* (2014).

[2] Kaur, Sukhvir, and Supriya Kinger. "Review on Load Balancing Techniques in Cloud Computing Environment." *International Journal of Science and Research (IJSR), Paper ID* 2014812: 2499-2504.

[3] Gopinath, PP Geethu, and Shriram K. Vasudevan. "An in-depth analysis and study of Load balancing techniques in the cloud computing environment." *Procedia Computer Science* 50 (2015): 427-432.

[4] Desai, Tushar, and Jignesh Prajapati. "A survey of various load balancing techniques and challenges in cloud computing." *International Journal of Scientific & Technology Research* 2.11 (2013): 158-161.

[5] Zhang, Qi, Lu Cheng, and Raouf Boutaba. "Cloud computing: state-of-the-art and research challenges." *Journal of internet services and applications* 1.1 (2010): 7-18.

[6] Chaczko, Zenon, et al. "Availability and load balancing in cloud computing." *International Conference on Computer and Software Modeling, Singapore*. Vol. 14. 2011.

[7] Radojević, Branko, and Mario Žagar. "Analysis of issues with load balancing algorithms in hosted (cloud) environments." *MIPRO, 2011 Proceedings of the 34th International Convention*. IEEE, 2011.

[8] Singh, Aarti, Dimple Juneja, and Manisha Malhotra. "Autonomous agent based load balancing algorithm in cloud computing." *Procedia Computer Science* 45 (2015): 832-841.

[9] Sidhu, Amandeep Kaur, and Supriya Kinger. "Analysis of load balancing techniques in cloud computing." *International Journal of Computers & Technology* 4.2 (2013): 737-741.

[10] Al Nuaimi, K., Mohamed, N., Al Nuaimi, M., & Al-Jaroodi, J. (2012, December). A survey of load balancing in cloud computing: Challenges and algorithms. In *Network Cloud Computing and Applications (NCCA), 2012 Second Symposium on* (pp. 137-142). IEEE.

[11] Chaudhary, Divya, and Rajender Singh Chhillar. "A new load balancing technique for virtual machine cloud computing environment." *International Journal of Computer Applications* 69.23 (2013).

[12] Foster, Ian, et al. "Cloud computing and grid computing 360-degree compared." *Grid Computing Environments Workshop, 2008. GCE'08*. Ieee, 2008.

[13] Hassan, Mohammad Mehedi, and Eui-Nam Huh. "Overview of cloud computing and motivation of the work." *Dynamic Cloud Collaboration Platform*. Springer New York, 2013. 1-12.

[14] Patel, Rajni, and Deepak Dahiya. "Aggregation of cloud providers: A review of opportunities and challenges." *Computing, Communication & Automation (ICCCA), 2015 International Conference on*. IEEE, 2015.