A REVIEW OF LOAD BALANCING TECHNIQUE IN CLOUD COMPUTING

Dr. Gaurav Sharma[1], Er. Manisha Gupta[2.]

[1]Associate Professor, [2]M.tech Student

Department of computer Science & Engg, JMIT, Radaur, (Haryana).

ABSTRACT

Internet is the most important part and revolutionary creation in the field of technology. Cloud computing is related to internet computing. Cloud computing is simply define as "cloud" is the delivery of demand services resource and everything from application to datacenter over the internet on the pay for use basis. Cloud computing has benefitted for service provider and clients. The main task of cloud computing is to provide the satisfactory level of performance to the user. There are many techniques to handle the large services and operations performed in cloud computing. To improve the performance of user utilization and operations it is very important to research some area in cloud computing. One of the important issues in cloud computing is load balancing. In cloud computing load balancing is a technique to distribute the load between two or more data server and to distribute the load on different nodes. It is a method in which workload on resources of a node spreads to respective resources on other node in network without disturbing the running task. In this paper we describe several technique of load balancing like Static load balancing and dynamic load balancing and their further types. The main aim of load balancing algorithm is optimize the resource usage, virtual machine, maximize throughput, and to assign a work to the cloud node so that the response time of the request can be minimized and request processing become effective. Average response time, data center request service time and total cost of different data centre will be the parameters considered for performance.

Keywords: *Cloud computing, Round Robin, Load Balancing and Throttled, cloud analyst, response time and virtualmachine.*

I. INTRODUCTION

Cloud computing is a scattered internet based paradigm, designed for remote sharing and usage of different resources and utility like storage, computational capabilities and applications etc. with high reliability over the large networks. yet, due to dynamic incoming requests, dynamic resource allocation is desire in it. This inherent dynamism in cloud computing requires efficient load balancing mechanisms. Load balancing concerns distribution of resources among the users or requests in rigid manner so that no node is overloaded or sitting idle. Like in, all other internet based scattered computing tasks, load balancing is an important aspect in cloud computing. In the absence of load balancing plan, efficiency of some overloaded nodes can sharply degrade at times, leading to violation of SLA (Service level Agreement). In traditional distributed computing, parallel computing and grid computing surroundings load balancing algorithms are categorized as static, dynamic or mixed scheduling algorithms based on their nature Cloud computing provides a way to use and access multiple server based competition resources via a digital network internet connection using (www) world wide web. Cloud users can access the server resources using a computer, net book, pad computers, smart phone or many other devices. There in cloud computing applications are managed and provided by the cloud server. In cloud computing data is also stored remotely in the cloud configuration. Cloud computing provides a way to delivery of computing resources over the internet. we use cloud computing services to store pictures and video online use webmail or a social networking site. Cloud computing makes IT management easier and more responsive and it is cost effective to the changing needs of the business. Cloud computing allows the business to expand their resources as per requirement, when there is increase in demand of various services. Basically there are three type of services which provided by Cloud computing, (Iaas) Infrastructure as a Service, (Paas) Platform as a Service and (Saas) Software as a Service. All these services are shown in the Fig 1.

A. *Infrastructure as a service (Iaas)* **:-** Iaas earlier called as the Hardware as a Service model (HaaS). This type is the basis of cloud computing. Infrastructure as a Service (IaaS) provides access to various fundamental resources i.e.,

virtual machines, virtual storage, physical machines etc. Resources can be subscribed by the developer on pay as per usage. IaaS is offered in three models as Public, Private and Hybrid.

B. *Platform as a Service (PaaS):-* It helps to provide runtime environment for applications, development tools, etc. It (PaaS) provides the flexibility to develop and test web based applications in cloud without installing the operating system.

C. *Software as a Service (SaaS):-* Software as a Service is a deployment model which allows to use software applications as a service to users and these application are running on a cloud infrastructure. SaaS is also known as "On demand Software" [1] [10].

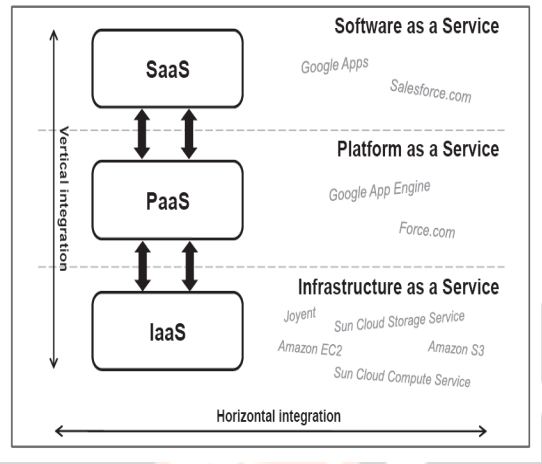

Fig1. Cloud services.

II DEPLOYMENT MODELS

This model defines the types of access to the cloud. There are four types of deployment models which are public cloud, private cloud, community cloud and Hybrid Cloud.

Public cloud:- It is a deployment model which provides system and services which are easily accessible by public. It may be less secure because of its openness, for example email.

Private cloud:- Private cloud provides systems and services which are accessible within an organization. This type of cloud provides more security because of its privacy.

Community cloud:- It provides system and services which are accessible by group of organization. In this type of cloud information is confined to the owner of organization.

Hybrid Cloud:- It is a composition of two or more clouds(public cloud , private cloud and Community Cloud) and uses all the services of all these cloud [2].

Load balancing in cloud computing:- In cloud computing load balancing is one of the major issues. It provides internet service to users from multiple servers [7]. Load balancing distribute workloads over multiple computing resources, such as central processing units, computers or network links. It has a main controller and balancer to gather and analyze the information. Load balancer is used to improve the performance of data center. The load balancer determines web server which serve the request. It may be provided either through hardware or software There are various scheduling algorithm to determine which server handle and forward the request to the selected server.Load balancing is one of the main challenges in cloud computing. It is a structure that distributes the dynamic workload evenly across all the nodes in the all cloud to avoid a situation where some nodes are heavily loaded while others are idle or doing little work.

The load balancing problem can be divided in two sub problems:

1. Admission of new request for VM provisioning and placement of VMs on host.
2. Reallocation/migration of VMs.

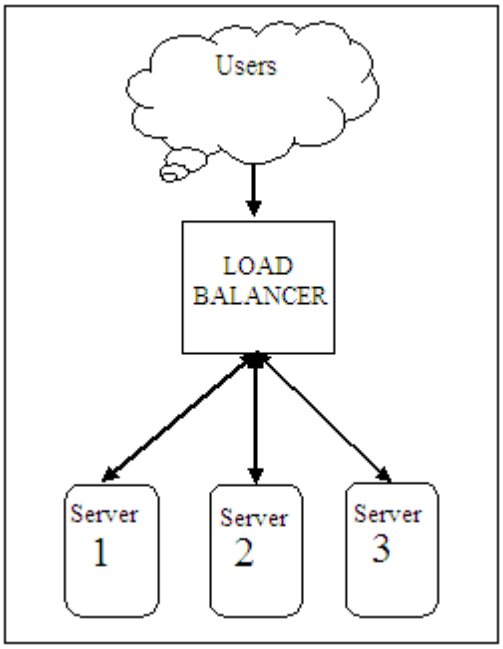

Fig. 2 load balancer in cloud computing.

III LITERATURE SURVEY

The cloud provides the skill that datacenter are geographically distributed across the network and network contains hundred of server. When a user submit a task (**i.e cloudlet**) it is handled by the datacenter controller[11].The Data Center Controller uses a VM Load Balancer to resolve which VM should be assigned the next request for processing. In this paper we explain the number of load balancing algorithm like Static load balancing(Round robin algorithm, Min-Min Scheduling Algorithm, (LBIMM) Max-Min Algorithm, Improved Max-Min Algorithm), Dynamic load balancing (Throttled) algorithm, Equally Spread Current Execution (ESCE) Algorithm **:-**

*Types of Load Balancing Algorithm***:-** The main objectives of load balancing are to improve the performance, maintain stability of system and accommodate future changes of server. Load balancing algorithm can be classified in different ways :

A. *Static Load balancing algorithm* **:-** Static algorithm refers to load balancing algorithm that distribute the work load on a fixed set of rules related to characteristics of the input work load. Static load balancing is easy to design and implement. The aim of static load balancing algorithm is to decrease the running time and minimize the response delay. Round Robin algorithm is one of the static load balancing technique. This technique of the static load balancing is used by the Central Processing Unit (CPU) during execution of process. Round Robin algorithm is the modern generation algorithm of First Come First Serve (FCFS).

1) *Round Robin* **:**– In this algorithm datacenter controller assigns the task to a list of VMs on a rotating basis. The first task is allocated to a VM picked randomly from the group and then the subsequent requests are assigned in a circular order. Once the VM is assigned a request, it is moved to the end of the list. Though the work load distributions between processors are equal but for different processes the job processing time are not same. So at any point of time some nodes may be heavily loaded while others remain idle.

2) *Weighted Round Robin***:**– It is the advanced version of Round Robin in which a weight is assigned to each VM so that if one VM is capable of handling twice as much load as the other, the powerful server gets a weight of 2. In such cases, the Data Center Controller will assign two requests to the powerful VM for each request assigned to a weaker one. The main important issue in this allocation is same as that of Round Robin that is it also does not consider the advanced load balancing requirements such as processing time for each individual requests [5].

3) *Dynamic Round Robin:-* This algorithm mainly works for minimizing the power consumption of physical machine.[4] The two main rules used by this algorithm is as follows:

i) If a virtual machine has finished its execution and there are other virtual machines run on the same physical machine, this kind physical machine will accept no more new virtual machine. Such physical machines are called to be in "retiring" state, i.e. when rest of the virtual machinesfinishes their execution, then this physical machine can shutdown.

ii) The second rule says that if a physical machine is in retiring state for a long time then instead of waiting, all the running virtual machines are migrated to other physical machines. After the successful migration, we can shut down the physical machine. This waiting time threshold is called as "retirement threshold". The algorithm reduces the power consumption cost but Itdoes not scale up for large data centers.

4) *Min-Min Scheduling Algorithm***:-** It starts with a set of tasks. Then the resource which has the minimum completion time for all tasks is found. Next, the task with the small size is selected and assigned to the corresponding resource (thus the name Min-Min). Finally, the task is removed from set and the same procedure is repeated by Min-Min until all tasks are assigned. The method is simple but it does not consider the existing load on a resource before assigning a task. So proper load balance is not achieved.[6]

5) *Improved Min-Min Scheduling Algorithm (IMM)* **:-** It starts by executing Min-Min algorithm at the first step. [6] At the second step is that it select the smallest size task from the heaviest loaded resource and calculates the completion time for that task on all other resources. Then the minimum completion time of that task is compared with the make span time produced by Min-Min. If it is less than make span time then the task is reassigned to the resource that produce it, and the ready time of both resources are updated. The process repeats as for as no other resources can outcome less completion time for the smallest task on the heavy loaded resource than the make span. Thus the overloaded resources are freed and the under loaded or idle resources are more utilized.

6) *Max-Min Algorithm:-*It's structure as Min-Min algorithm. But it grant more priority to the larger tasks. The jobs that have vast execution time or vast completion time are executed first. The major problem is that tiny jobs have to be waiting for lengthy time.[4]

7) *Improved Max-Min Algorithm:-* Max-Min Algorithm extension is Improved Max-Min Algorithm. The Max-Min algorithm selects the task with the utmost finishing point time and allocate it to the resource on which achieve smallest execution time. The basic idea of a better version of Max-Min algorithm assign task with largest execution time to resource produces smallest complete time rather than original Max-Min assign task with largest completion time to resource with minimum execution time. It uses the advantages of Max-Min and also covers its disadvantages [4] [9].

B. *Dynamic load balancing Algorithm:-* Dynamic scheduling (often referred to as dynamic load balancing) is based on the redistribution of processes amidst the processors during execution time. This redistribution is performed by transferring tasks from the weighty loaded processors to the lightly loaded processors with the aim of improving the performance of the application. It is particularly useful when the requirement of process is not known a priori and the primary goal of the system is to maximize the utilization of resources. The major drawback of dynamic load balancing scheme is the run-time overhead due to the assignment of load information among processors, decision-making for the selection of processes and processors for job transfers and the communication jam associated with the task relocation itself. The dynamic load balancing algorithms can be centralized or distributed depending on whether the responsibility for the task of global dynamic scheduling should physically reside in a single processor (centralized) or the work involved in making decisions should be physically distributed among processors . The most important feature of makingdecisions centrally is simplicity . However centralized algorithms suffer from the problem of bottleneck and single point failure. Distributed load balancing algorithms are free from these problems.

1) *Throttled:-*The Throttled Load Balancer (TLB) maintains a record of the state of each virtual machine (busy/idle) [2]. When a request arrives it searches the table and if a match is found on the basis of size and availability of the machine, then the request is accepted otherwise -1 is returned and the request is queued [11]. During allocation of a request the current load on the VM is not considered which can in turn increase the response time of a task.

2) *Modified Throttled* :- Like the Throttled algorithm it also maintains an index table containing list of virtual machines and their states. The first VM is selected in same way as in Throttled. When the next request arrives, the VM at index next to earlier assigned VM is chosen depending on the state of VM and the usual steps are followed, unlikely of the Throttled algorithm, where the index table is decompose from the first index every time the Data Center queries Load Balancer for allocation of VM. It gives better response time compare to the previous one. But in index table the state of some VM may change during the allocation of next request due to deallocation of some tasks. So it is not always beneficial to start searching from the next to already assigned VM[3].

*C. Equally Spread Current Execution (ESCE) Algorithm***:-**This method the load balancer continuously scans the job queue and the list of virtual machines. If there is a VM vacant that can handle the request then the VM is allocated to that request . If there is an overburden VM that needs to be freed of the load, then the load balancer distributes some of its tasks to the VM having minimum load to make every VM equally loaded.[8] The balancer tries to improve the response time and processing time of a job by selecting it whenever there is a match. But it is not fault tolerant and has the problem of single point of failure [4].

D. *Joint Idle Queue (JIQ)* **:-** This algorithm consist of mainly primary and secondary load balancing system, which communicate through a data structure called as I-Queue. Each I-Queue is associate with an Dispatcher. An I-Queue is a list of subset of processor that have reported to be ideal. Each Dispatcher can access all these processor.

1) *Primary Load Balancing:-* This system consist of information of an ideal servers present in the I-Queue and avoid communication overhead from server loads. As the task arrives, the Dispatcher consult the I-Queue, if the I-Queue is nonempty, then Dispatcher remove the first ideal processor from the I-Queue and give the task to this ideal processor. If the I-Queue is empty, the Dispatcher guide the job to a randomly chosen processor.

When a processor become idle, it inform an I-Queue of it's idleness, or joins the I-Queue. For all these algorithms in this class, all idle processor joins only one I-Queue to avoid extra communication to withdraw from I-Queues. The challenge with distributed dispatchers of incoming task and idle processor at the dispatchers. It is possible that a task arrive at an empty I-Queue while there are idle proccessors in other I-Queues. This pose a new load balancing problems in the reverse direction from processor to dispatchers: How can we assign idle processor to I-Queue so that when a task arrives at a dispatcher, there is a high probability that it will find an idle processor in its I-Queue.

2) *Secondary Load Balancing:-* when a processor turn in idle, it chooses one I-Queue based on load balancing algorithm and notify the I-Queue of its idleness, or joins it. We consider here two load balancing algorithm in the reverse direction: **Random and SQ(d).** we calling the algorithm with Random load balancing in the reverse direction JIQ-Random and with SQ(d) load balancing JIQ-SQ(d). an idle processor chooses d random I-Queues are off the critical path, JIQ-Random has the further advantage of having a One-Way communication, without requiring message from the I-Queue[12].

IV REFERENCE

[1] AmandeepKaurSidhu, SupriyaKinger , "Analysis of Load Balancing Techniques in Cloud Computing ", International Journal of Computers &
Technology , Volume 4 No. 2, March-April, 2013, pp. 737-741.

[2] Jasmin James, Dr. BhupendraVerma, "Efficient VM Load Balancing Algorithm for a Cloud Computing Environment ", International Journal on
Computer Science and Engineering (IJCSE) , Vol. 4 No. 09 Sep 2012 , pp.
1658-1663.

[3] Shridhar G. Domanal, G. Ram Mohana Reddy, "Load Balancing in Cloud Computing Using Modified Throttled Algorithm", IEEE,
International conference.CCEM 2013.

[4] M.Aruna , D. Bhanu, R.Punithagowri , "A Survey on Load Balancing Algorithms in Cloud Environment ", International Journal of Computer
Applications, Volume 82 – No 16, November 2013 , pp. 39-43.

[5] Qi Zhang, Lu Cheng, RaoufBoutaba; Cloud computing: sate-of-art and research challenges; Published online: 20th April 2010, Copyright :
The Brazillian Computer Society 2010.

[6] Huankai Chen, Professor Frank Wang, Dr. Na Helian, Gbola Akanmu, "User-Priority Guided Min-Min Scheduling Algorithm For Load Balancing
in Cloud Computing", IEEE, 2013.

[7] Rajesh George Rajan, V.Jeyakrishna (Dec 2013), "A Survey on Load Balancing in Cloud Computing Environment", Vol.2, Issue.12, pp.4726-4728.

[8] Dharmesh Kashyap, JaydeepViradiya(Nov 2014),"A Survey of Various Load Balancing Algorithms in Cloud Computing",Vol.3,Issue.11,pp.115-119.

[9] UpendraBhoi,PurviN.Ramanuj(April 2013) "Enhanced Max-Min Task Scheduling Algorithm in Cloud Computing",Vol.2,Issue.4,

[10] http://en.wikipedia.org/wiki/Cloud_computing.

[11] Subasish Mohapatra, K.Smruti Rekha, Subhadarshini Mohanty , "AComparison of Four Popular Heuristics for Load Balancing of Virtual
Machines in Cloud Computing ", International Journal of Computer
Applications, Volume 68– No.6, April 2013 , pp. 33-38.

[12] V. Gupta, M. Harchol-Balter, K. Sigman, and W. Whitt. Analysis of join-the-shortest-queue routing for web server farms. *Perforance Evaluation*, (64):1062–1081, 2007.