LOAD BALANCING ALGORITHMS FOR THE CLOUD COMPUTING ENVIRONMENT: A REVIEW

[1]HARDI S. SANGHAVI, [2]DR. TEJAS P. PATALIA

[1]Research Scholar, Master in Computer Engineering(ME-CE), V.V.P. Engineering College, Rajkot-360001
[2]Head & Associate Professor, Computer Engineering Department, V.V.P. Engineering College, Rajkot-360001

hardee11@gmail.com ,Pataliatejas@rediffmail.com

ABSTRACT:
Innovations are necessary to ride the inevitable tide of change. Most of enterprises are striving to reduce their computing cost through the means of virtualization. This demand of reducing the computing cost has led to the innovation of Cloud Computing. Cloud Computing offers better computing through improved utilization and reduced administration and infrastructure costs and also in cost efficient and pay-as per use mode. Cloud Service providers must fulfil the user demands in an efficient for which they need to have proper utilization of the available resources. Resources must be selected as per the requirement of the user. Thus by analyzing current scenario of the research work done in the area of Cloud Computing we have come to conclusion that the most common problem found here is of the Load Balancing. The purpose of applying the technique of Load Balancing is to distribute the available resources efficiently and fairly. As the number of users increases, Load Balancing becomes a serious issue for the service provider.

***KEY WORDS** : Cloud Computing, Load Balancing, Load Balancing Algorithms, Throttled Load Balancer, Dynamic Round Robin Load Balancer.*

1. Introduction

Recently Cloud Computing has become one of the popular techniques used by both industry and educational institutions in order to provide flexible way to store and access the data files. Cloud Computing can be defined as "structural model that defines computing services where resources as well as data are retrieved from cloud service provider via Internet through some well-formed web-based tools and application."

Cloud Computing is a service oriented design that reduces the cost of access to gather the information of the clients offer greater flexibility and demand based services and so on[2]. Cloud Computing services are provided by the Cloud Service Providers. The client does not need to buy additional hardware and software. They have to just send their requirements to the service provider and pay for the same.

The economic appeal of Cloud Computing is often described as "converting capital expenses to operating expenses", generally known as "pay as you go". Hours purchased via Cloud Computing can be distributed non-uniformly in time. In the Networking community, this way of selling bandwidth is already known as "usage-based pricing". The service provider is sole responsible for the maintenance and up gradation of the services they are providing to their clients.

As the use of internet is increasing day-by-day, the usage of cloud computing services also increases. To handle the incoming requests of the users, Load Balancing is one the most important issue to be handling by the service providers. There are many algorithms such as Round Robin, Throttled, Active Monitoring etc. are widely used for executing the client's request in the minimum response time.

2. Definition: Cloud Computing

Although many formal definitions of Cloud Computing are used due to its characteristics and usage in the industrial as well as academics but the one provided by the U.S. NIST (National Institute of Standards and Technology) appears to have included all the key elements related to the field of Cloud Computing:

"Cloud Computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction."[1][2]

The above definition includes the architecture of cloud, security essentials and deployment strategies. It is composed of five essential characteristics, three service models, and four deployment models.

The cloud computing system consists of mainly three components: Clients, Datacenter and Distributed Servers. [1][3]

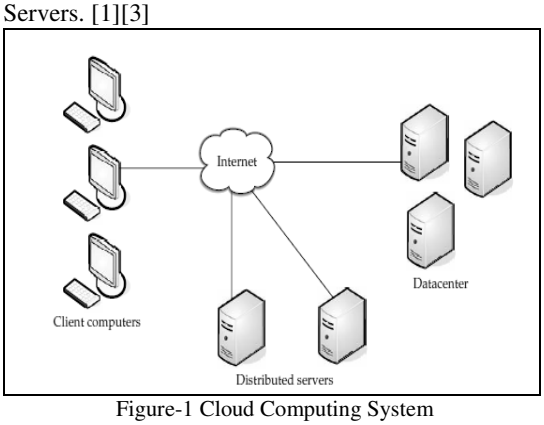

Figure-1 Cloud Computing System

Each component has a specific purpose and plays specific role in delivering a functional cloud-based application.

Clients: Clients are the simple computing devices such as PC, laptop, mobiles etc.. The end users interact with the clients to maintain the information related to cloud. Generally they can be categorized in to three groups, which are as follows:

- **Mobile Clients:** It includes smart phones like a Blackberry, Windows Mobile Smartphone or an iPhone.

- **Thin Clients:** They are the computers without any internal hard disk. The computing work for such clients is done by the server. Thin clients just display the result of the computations done by the server to its end user.

- **Thick Clients:** They are the regular computers having their own internal hard disk. Thick clients make use of the web browsers such as Internet Explorer or Firefox to get connected to the cloud. Thin clients are increasingly used nowadays due to its lower hardware cost, lower IT costs, less power consumption and ease of use.

Datacenter: The datacenter is the collection of the servers where the prescribed application of the end user is housed. It can be a large room in the basement of the building or a room full of servers anywhere in the world that can be access through Internet. By using the concept of virtualization, multiple instances of virtual machines can be installed on a single physical server.

Distributed Servers: Distributed servers are the parts of a cloud which are present throughout the Internet hosting different applications. But while using the application from the cloud, the user will feel that he is using this application from its own machine.

The essential characteristics that define the overall working of cloud computing technology are as below: [2][3][4]

On-demand self service
On-demand self service allows user to access cloud computing services as per the requirement without any interaction with the service provider. The user can schedule usage of cloud services as required. Thus, it provides cost savings to both the users and cloud service provider.

Broad Network access
High bandwidth network communication provides access to a large pool of IT resources. Thus provides efficient and effective replacement to in-house data centers. Most of the organizations use either three-tier architecture or two-tier architecture in order to connect to the computing platforms such as laptops, palm tops, PDAs.

Location Independent Resource Pooling
The cloud service provider pools together the available computing resources to serve multiple users in an efficient and effective manner. But these resources can be located anywhere around the world physically and assigned as virtual components whenever they are required. The computing resources are dynamically assigned and reassigned to the users according to their requirements. The user does not have knowledge or control over the exact location of the provided resources.

Measured service
The amount of the cloud services used can be continuously observed and recorded. Also the user of the services can be billed according to his/her usage. Resource usage thus being observed, controlled and reported provides transparency to both the user of the utilized resource and the service provider

Rapid elasticity
Capabilities of the computing resources can be scaled-up or scaled-down rapidly as per the requirement of the user. For the consumers, the capabilities available appear to be unlimited and can be purchased in any quantity at any time.

3. Virtualization
Virtualization is one of the most important concept in context of cloud computing. In general, virtualization means “something which is not physically present”, but provides all the properties of the real. In Cloud Computing, Virtualization means the software

implementation of a computer which will execute different programs like a real machine. An end user can use different services of a cloud with the help of the concept of virtualization. The remote datacenter will provide different services in a fully or partial virtualized manner.

There are two types virtualization found in case of cloud computing, as given in [1][5]:

- Full Virtualization
- Paravirtualization

Full Virtualization

The complete installation of one machine is done on another machine in case of full virtualization. Thus it will result in to a virtual machine which will have all the software already present in the real server.

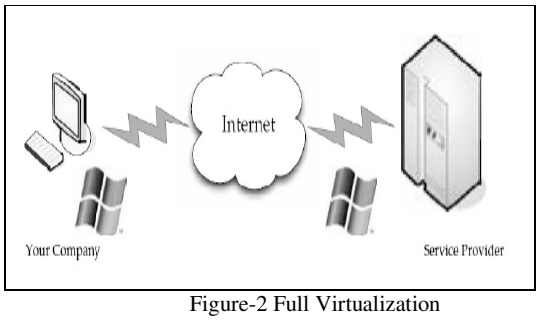

Figure-2 Full Virtualization

This kind of virtualization allows not only run unique applications but also different operating systems. The remote datacenter delivers the services in a fully virtualized manner. Full virtualization has been successful for several purposes:

- Sharing a computer system among multiple users
- Isolating users from each other and from the control program
- Emulating hardware on another machine

Paravirtualization

In case of paravitualization, the hardware allows multiple operating systems to run on single machine by efficient use of system resources such as memory and processor. e.g. VMware software. Here the services are provided partially.

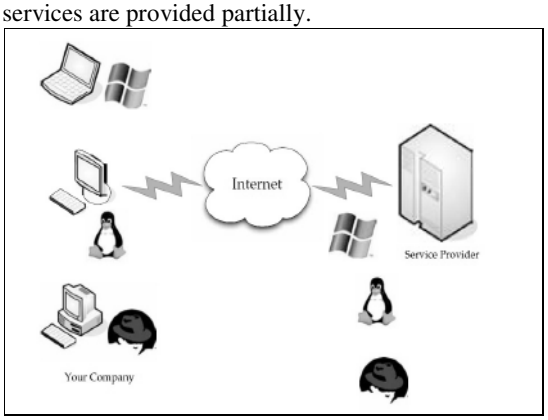

Figure-3 Paravirtualization

Paravirtualization has the following advantages:

- **Disaster recovery:** In the event of a system failure, guest instances are moved to hardware until the machine is repaired or replaced.
- **Migration:** As the hardware can be replaced easily, hence migrating or moving the different parts of a new machine is faster and easier.
- **Capacity management:** In a virtualized environment, it is easier and faster to add more hard drive capacity and processing power. As the system parts or hardware can be moved or replaced or repaired easily, capacity management is simple and easier.

Virtual Machines

A "Virtual Machine" is a software implementation of a machine that executes programs like a physical device. They are divided in to two major groups based on their use and degree of correspondence to any real machine.

- **System Virtual Machine**

This type of virtual machine provides a complete system platform which supports the execution of a complete operating system (OS). These usually emulate an existing architecture, and are built with the purpose of either providing a platform to run programs where the real hardware is not available for use or of having multiple instances of virtual machines leading to more efficient use of computing resources.

- **Process Virtual Machine**

It is also known as Language Virtual Machine and is designed to run a single program which means that it supports a single process. Such VMs are usually closely suited to one or more programming languages and built with the purpose of providing program portability and flexibility.

4. Deployment Models

Depending on the environment in which the clouds are used, clouds can be broadly divided in to four categories:

- Public Cloud
- Community Cloud
- Private Cloud
- Hybrid Cloud

Public Cloud

The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services. Thus public cloud provides "flexibility" of how to use the resources.

Community Cloud

Community cloud shares infrastructure between several organizations from a specific community with

common concerns, whether managed internally or by a third party and either hosted internally or externally.

Private Cloud
Here the cloud infrastructure is operated solely for an organization. It may be managed by the organization or a third party and may exist on premise or off premise. Thus private cloud provides "control" over the infrastructure being used and who can use them.

Hybrid Cloud
Hybrid cloud is a composition of two or more clouds (public, private or community) that remain unique entities but are bound together by standardized technology that enables data and application portability.

5. SERVICE MODELS

In Cloud Computing, the term "service" is the concept of being able to use re-usable, fine-grained components across a vendor's network[2][3][6]. The following three service models are used to categories the cloud services:

- Software as a Service (SaaS)
- Platform as a Service (PaaS)
- Infrastructure as a Service (IaaS)

Software as a Service (SaaS)
The end-user can use the applications hosted by the cloud service provider on the cloud infrastructure. The applications are accessible through thin clients interface such as web browser. The end-user have no control over the underlying cloud infrastructure. The cloud service provider is sole responsible for the maintenance and up gradation of the applications they have hosted.

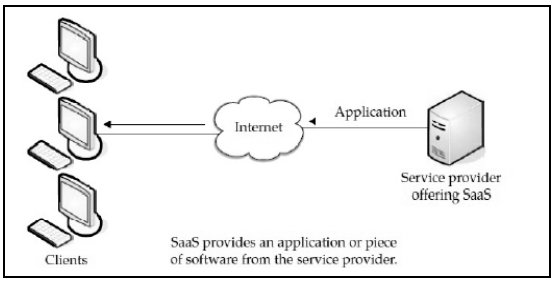

Figure-4 Software as a Service

The client has to pay for the time he/she uses the software. Also the end user does not have to purchase the software or to install it. The limitation of using these kinds of services is there is minimal customization available to the end-user. Also there are security and privacy risks for the data being exchanged over the network. Examples of SaaS are: SalesForce.com, Google Mail, Google Docs and so forth[3][6].

Platform as a Service (PaaS)
PaaS is mainly used as a development platform which supports the entire Software Development Life Cycle (SDLC). PaaS services are software design, development, testing, deployment, and hosting. Other services can be team collaboration, database integration, web service integration, data security, storage and versioning etc. Here the end-users are allowed to develop cloud services and applications directly on the PaaS cloud. The main difference between SaaS and PaaS is that SAAS only hosts the completed cloud applications whereas PaaS offers a development platform that hosts both completed and in-progress cloud applications.[3][6]

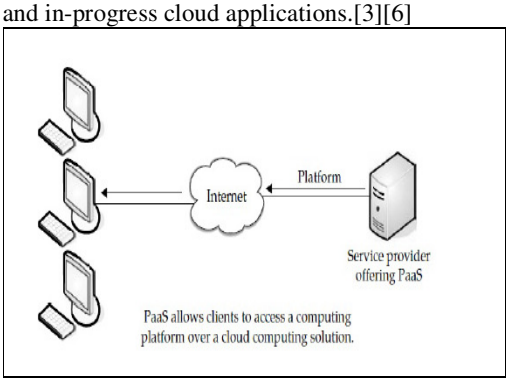

Figure-5 Platform as a Service

This requires PaaS, in addition to supporting application hosting environment, to possess development infrastructure along with programming languages, tools etc. the major issue faced by the end-user is up gradation of the softwares being used in PaaS. Also there is an issue of compatibility of different versions of the applications hosted. Examples of PaaS are: Google AppEngine, Microsoft Azure.

Infrastructure as a Service (IaaS)
IaaS is also known as Hardware as a Service. In IaaS, the end-user can directly access the IT infrastructure such as processing, storage, networks, and other fundamental computing resources provided. In general words, the end- users uses the hardware of the cloud service provider on rent. The concept of Virtualization is most efficiently use in IaaS to provide efficient service to its users. The IaaS cloud can be scaled-up or scaled-down according to the requirement of the end-user. Different cloud service providers have different pricing models according to the quality of the service they provide to their end-users.

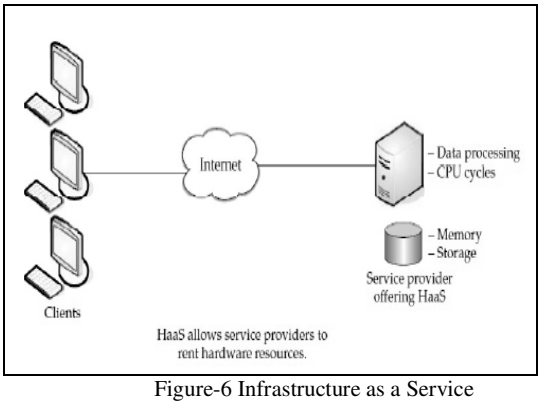

Figure-6 Infrastructure as a Service

For the efficient and effective usage of the available infrastructure, establishment of the efficient load balancing algorithm is compulsory. Examples of IaaS are: Amazon EC2, Rackspace.

6. LOAD BALANCING

Load balancing is the process of improving the performance of distributed and parallel computing with the help of distribution of load among the processors or nodes. As the use of the web increasing day by day, with this there is need to increase the requirement for load balancing. The introduction of E-Commerce has led many businesses to carry out the most of their day-to-day business online. As a result of the increase in demand of the web, web sites providers want to ensure the availability of access for their users and make sure that their requests are processed as quickly as possible.

Traditional algorithms were used on different web servers to distribute the load on different systems but these algorithms do not always give expected performance with the large-scale and distinct structure of service-oriented data centre. To overcome the above mentioned shortcomings, the load balancing techniques were introduced and have been widely studied by the researchers and implemented by the industry. The goal of performing load balancing is to achieve optimal resource utilization, maximum throughput, maximum response time, and avoiding overload.

7. ISSUES AND PARAMETERS OF LOAD BALANCING

There are various issues which need to be handled while dealing with load balancing in the cloud computing environment [4][7]. Each load balancing algorithm is designed to achieve some specific goal. Some algorithms are designed to achieve minimum response time for the client's request, some aims to achieve effective and efficient utilization of available resources.

Some major issues which must be considered while designing any load balancing algorithms are as follows:

- **Geographical distribution of nodes**

The geographical location of different nodes matters a lot while observing the real time cloud computing systems, especially in case of large-scaled applications. A well distributed system of nodes in cloud computing environment helps to reduce fault tolerance.

- **Dynamic versus Static behavior of algorithm**

The state or behavior of the system needs to be considered before designing any load balancing algorithm. Depending on the state of the system the load balancing algorithms can be broadly divided in to two groups they are, static algorithms and dynamic algorithms.

- **Complexity of algorithm**

The overall performance of the system is mainly affected by the complexity of the load balancing algorithm. Many a times a complex algorithm provides better throughput and resource utilization while less complex algorithm may give poor performance in terms of fault tolerance and migration time. Thus depending on the system requirements care should be taken to decide a better or suitable load balancing algorithm.

- **Traffic analysis over different geographical locations**

For any load balancing algorithm, it is very important to analyze the traffic flow in real-time scenarios over different geographic regions, and then balance the overall workload accordingly. All regions over the globe have a different time zone and have certain peak hours during which the network load is supposed to be at its peak. Therefore, load balancer must be capable of handling the traffic in peak hours in every location so as to achieve maximum resource utilization and throughput.

8. METRICS OF LOAD BALANCING ALGORITHMS

In cloud computing, the load balancing is required to distribute the workload evenly across all the nodes. It helps to achieve higher user satisfaction by ensuring fair and efficient allocation of computing resources. The following metrics helps to measure the overall performance of the load balancing algorithm: [6][7]

- **Throughput**

Throughput is the term used to calculate the total number of tasks whose execution has been already completed.

- **Overhead Associated**

Overhead Associated determines the amount of overhead included while implementing a load balancing algorithm which includes overhead occurred due to movement of tasks, inter-process communication.

- **Fault Tolerance**

Fault Tolerance is the ability of an algorithm to perform uniform load balancing in case of link failure.

- **Migration Time**

Migration Time is the time required while migrating the jobs or resources from one node to another. It should be minimized to enhance the overall performance of the cloud computing system.

- **Response Time**

Response Time is the amount of time taken to respond by a load balancing algorithm.

- **Resource Utilization**

Resource Utilization measures the overall utilization of the computing resources using a particular load balancing algorithm.

- **Scalability**

Scalability is the ability to scale-up or scale-down according to the requirements of the system.

- **Performance**

Performance measures the efficiency of the system which must be improved at a reasonable cost.

9. EXISTING LOAD BALANCING ALGORITHMS

Now in this section we will discuss the existing load balancing algorithms which are widely used now days due to its characteristics. The various load balancing algorithms need to be studied are as follows:

- **Random Scheduling Algorithm**

This method distributes the load across the available servers by selecting anyone using method of random number generation and sending current connection to it [6]. The drawback of this technique is that at pick hours of working there are chances of getting wrong output and there is no guarantee of load being distributed equally.

- **Round Robin Algorithm**

In this, the DataCenter controller assigns the requests to a list of VMs on a rotating basis. The first request is allocated to a VM- picked randomly from the group and then the DataCenter controller assigns the subsequent requests in a circular order. Once the VM is assigned the request, the VM is moved to the end of the list.

In this RRLB; there is a better allocation concept known as Weighted Round Robin Allocation in which one can assign a weight 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 DataCenter Controller will assign two requests to the powerful VM for each request assigned to a weaker one.

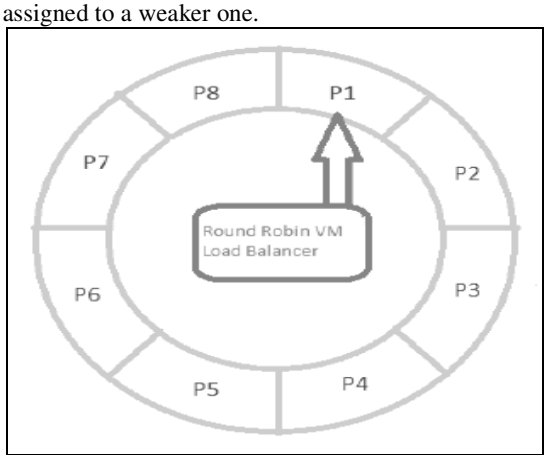

Figure-7 Round Robin Load Balancer

The basic purpose of this algorithm is the system builds a standard circular queue and walks through it, sending one request to each machine before getting to the start of the queue and repeating it again. Each request is allowed to be executed for the fix time quantum. If still working of the server is left then it is queued and above mentioned procedure is repeated again until the process is complete.

The major issue in this allocation is that it will not check whether the server is heavily loaded or not, just assign the server to the requested client. Thus many a times it happens that some servers are heavily loaded while some are lightly loaded. Also RRLB works well in homogeneous environment.

- **Dynamic Round Robin Algorithm**

Dynamic Round Robin is a self-adaptive algorithm. This algorithm works in the same manner as Round Robin algorithm works. The only difference between them is that slicing of time among the servers is done on the basis of their current load of that server. For this, continuous monitoring of the VMs is required resulting in to some overload on the DataCenter controller. But this overload helps to improve the efficiency of the system. This kind of technique works well in heterogeneous environment.

- **Active Monitoring Load Balancing Algorithm**

In Active Monitoring Load Balancing Algorithm, information about each VM is maintained along with the number of requests currently allocated to it.

When a new request arrives the least loaded VM is searched out from the available information. If there are more than one then the first identified VM is selected. After the selection of VM, its id is send back to the DataCenter controller to forward the request to the VM whose id has been identified.

At times, multiple tasks are assigned to a single VM which increases the complexity of the algorithm.

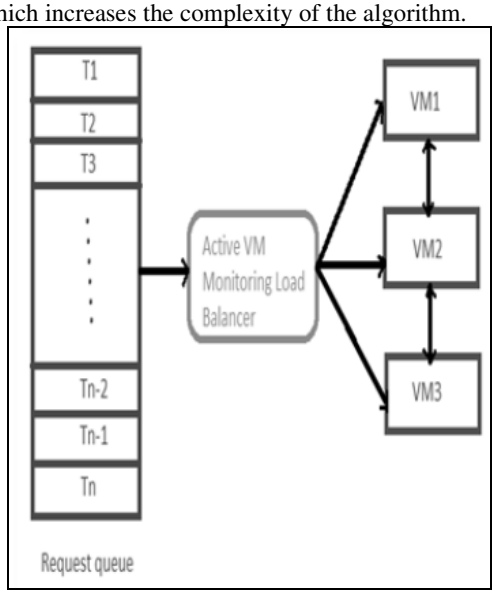

Figure-8 Active Monitoring Algorithm

- **Throttled Algorithm**

Throttled algorithm is totally based on virtual machine [9].

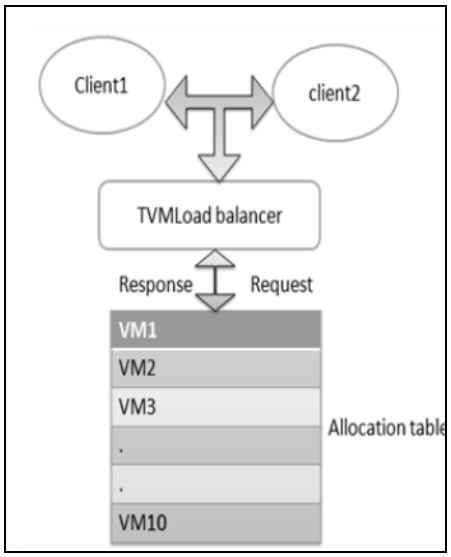

Figure-9 Throttled Algorithm

In this algorithm, when the client sends request to DataCenter controller, a suitable VM is found which can execute the request in an efficient manner using specific technique. The current state of the VM whether it is available or busy for each VM is maintained. While doing so, the current load of the VM and processing time required by the request is not considered.

- **Modified Throttled Algorithm**

In the paper [9], the authors have combined the two traditional load balancing algorithms, Throttled algorithm and Round Robin algorithm, in order to improve the working efficiency of the Throttled algorithm and named the improved version of the algorithm as Modified Throttled algorithm.

10. CONCLUSION AND FUTURE WORK

Recently Cloud Computing has become one of the popular techniques used by both industry and educational institutions in order to provide flexible way to store and access the data files. Cloud Computing can be defined as "structural model that defines computing services where resources as well as data are retrieved from cloud service provider via Internet through some well-formed web-based tools and application."

Cloud Computing is a service oriented design that reduces the cost of access to gather the information of the clients offer greater flexibility and demand based services and so on. Cloud Computing services are provided by the Cloud Service Providers. The client does not need to buy additional hardware and software. They have to just send their requirements to the service provider and pay for the same.

The economic appeal of Cloud Computing is often described as "converting capital expenses to operating expenses", generally known as "pay as you go". Hours purchased via Cloud Computing can be distributed non-uniformly in time. In the Networking community, this way of selling bandwidth is already known as "usage-based pricing". The service provider is sole responsible for the maintenance and up gradation of the services they are providing to their clients.

As the use of internet is increasing day-by-day, the usage of cloud computing services also increases. To handle the incoming requests of the users, Load Balancing is one the most important issue to be handling by the service providers. There are many algorithms such as Round Robin, Throttled, Active Monitoring etc. are widely used for executing the client's request in the minimum response time.

It needs to distribute load for better performance and computation to satisfy client for their requested services. It also requires maintains the information an index table of virtual machines and also the state of VMs. There provides some features such as performance guarantee, algorithm overload and robustness, Load balance and cost movement. Virtual

Machine request multiple type of resource, the resource manager allocates it in optimal manner.
The main aim of this research can be said load balancing in efficient way such that all resource has fully utilize and get high performance with low response time. We also get maximum throughput with high availability. It also needs client satisfaction and efficient services. So, select algorithm such that allocating resources according to load status information are update locally and globally. There is need to schedule requested resources which is helpful for utilization of resources. It also need such a way distribute load that provides scalability of server.

As a future scope, the present work need to be focused on changing the data structures used for maintaining the index table and also by incorporating the paradigms of parallel and high performance computing the response time and utilization of VMs may be further optimized.

11. REFERENCES:

1. Anthony T. Velte, Toby J. Velte and Robert Elsenpeter, "Cloud Computing- A Practical Approach," Tata McGraw Hill Publications, pp. 6-8.

2. Lee Badger, Tim Grance, Robert Patt-Corner, Jeff Voas , " DRAFT Cloud Computing Synopsis and Recommendations ", Special publication 800-146, May 2011

3. Tharam Dillon, Chen Wu, Elizabeth Chang, "Cloud Computing: Issues and Challenges", 24th IEEE International Conference on Advanced Information Networking and Applications, IEEE, 2010

4. N. S. Raghava and Deepti Singh, "Comparative Study on Load Balancing Techniques in Cloud Computing", Open Journal of Mobile Computing and Cloud Computing, Scientific Online

5. Stuti Dave and Prashant Mehta,"Role of Virtual Machine live Migration in Cloud Load Balancing", IJARCCE, 2012

6. Nayandeep Sran and Navdeep Kaur, "Comparative Analysis of Existing Load Balancing Techniques in Cloud COmputing", International Journal of Engineering Science Invention, January 2013, pp. 60-63

7. Ms. Nitika, "Comparative Analysis of Load Balancing Algorithms in Cloud Computing", International Journal of Engineering and Science,2012

8. Jitendra Bhatti, Tirth Patel, Harshal Trivedi and Vishrut Majumdar, "HTV Dynamic Load Balancing Algorithm for Virtual Machines instances in Cloud", International Symposium on Cloud and Services Computing,IEEE, 2012

9. Shridhar G. Domanal and G.Ram Mohana Reddy, "Load Balancing in Cloud Computing using Modified Throttled Algorithm", IEEE, 2013

10 Bhathiya, Wickremasinghe."Cloud Analyst: A Cloud Sim-based Visual Modeller for Analysing Cloud Computing Environments and Applications"

11 Ali M. Alakeel,"A Guide to Dynamic Load Balancing in Distributed Computer Systems", IJCSNS International Journal of Computer Science and Network Security,VOL.10 No.6, June 2010.

12 Soumya Ray, 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, October 2012

13 Ken W. Batcher and Robert A. Walker, "Dynamic Round Robin Task Scheduling To Reduce Cache Misses For Embedded Systems", EDAA, 2008.

14 Ching Chi Lin, Pangfeng Liu and Jan-Jan Wu, "Energy-Aware Virtual Machine Dynamic Provision and Scheduling for Cloud Computing", 4th International Conference on Cloud Computing,IEEE, 2011

15 http://en.wikipedia.org/wiki/Cloud_computing

16 http://en.wikipedia.org/wiki/Virtual_machine