ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

An Approach to Load Balancing In Cloud Computing

Radha Ramani Malladi

Visiting Faculty, Martins Academy, Bangalore, India

ABSTRACT: Cloud computing is a structured model that defines computing services, in which data as well as resources are retrieved from cloud service provider via internet through some well-formed, web-based tool and application. Cloud Computing is nothing but a collection of computing resources and services pooled together and is provided to the users. Sharing of the group of resources may initiate a problem of availability of these resources causing a situation of deadlock. One way to avoid deadlocks is to distribute the workload of all the VMs among themselves. This is called load balancing. As the numbers of users are increasing on the cloud, the load balancing has become the challenge for the cloud provider.
Load balancing is a main challenge in cloud environment. It helps to distribute the dynamic workload across multiple nodes to ensure that no single node is overloaded. It helps in proper utilization of resources .It also improves the performance of the system. Load balancing is the process of finding overloaded nodes and then transferring the extra load to other nodes.

KEYWORDS: Cloud Computing; Load Balancing; Deadlock; Scheduling; Resource Allocation; Existing Load Balancing Algorithms;

I. INTRODUCTION

Cloud computing provides online resources and online storage to the user's .It provides access to the resources and all the data at a lower cost to them. In Cloud computing the cloud provider outsources all the resources to their client. There are many existing issues in cloud computing. The main problem is load balancing in cloud computing. Load balancing helps to distribute all loads between all the nodes. It also ensures that every computing resource is distributed efficiently and fairly. It provides high satisfaction to the users. Load balancing is a relatively new technique that provides high resource utilization and better response time. Sometimes our system gets hanged up or it seems to take few decades for pages to come out of printer. All this happens because there is a queue of requests waiting for their turn to access resources which are shared among them. But these requests cannot be serviced as the resources required by each of these requests are held by another process or request by virtual machines. One cause for all these problems is called deadlock. Load balancing is a new approach that assists networks and resources by providing a high throughput and least response time.

1. Cloud computing consist of several characteristics:

• On demand service- Cloud computing provide services to users on their demand .Users can access the services as they want.

• Broad Network Access- In cloud computing capabilities are available over the network .All the capabilities are accessed through different mechanisms.

• Resource Pooling- Different models are used to pooled the resources which provide by the providers to their consumers. All the resources dynamically assigned and reassigned according to consumer demand.

• Rapid Elasticity- Quantity of resources is increase at any time according to the customer's requirements.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

• Measured Service- In cloud computing resource usage can be monitored, controlled for both the provider and consumer of the all service.

2. Challenges in Cloud Computing

There are many challenges in cloud computing:-

1. Security
2. Efficient load balancing
3. Performance Monitoring
4. Consistent and Robust Service abstractions
5. Resource Scheduling
6. Scale and QoS management
7. Requires a fast speed Internet connection.

II. CLOUD COMPUTING MODEL

Fig: 1 shows Cloud computing model which consist services of cloud and different deployment models as:

A. Services of Cloud Computing:

Service means different types of applications provided by different servers across the cloud. There are many services that are provided to the users over cloud.

1) **Software as a Service (SaaS)**: Saas provided all the application to the consumer which are provided by the providers. Applications are running on a cloud infrastructure. Interfaces (web browser)are used access the applications. The consumer does not control the internal function.

That Customers who are not able to developed software, but they need high level applications can also betake advantages from SaaS. There are some of applications of software of services:-

• Customer resource management (CRM)
• Video conferencing
• IT service management
• Accounting
• Web analytics
• Web content management

Advantages:

1) The main advantage of SaaS is costing less money than buying the whole application.
2) It provides reliable and cheaper applications.
3) More bandwidth.
4) Need less staff.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

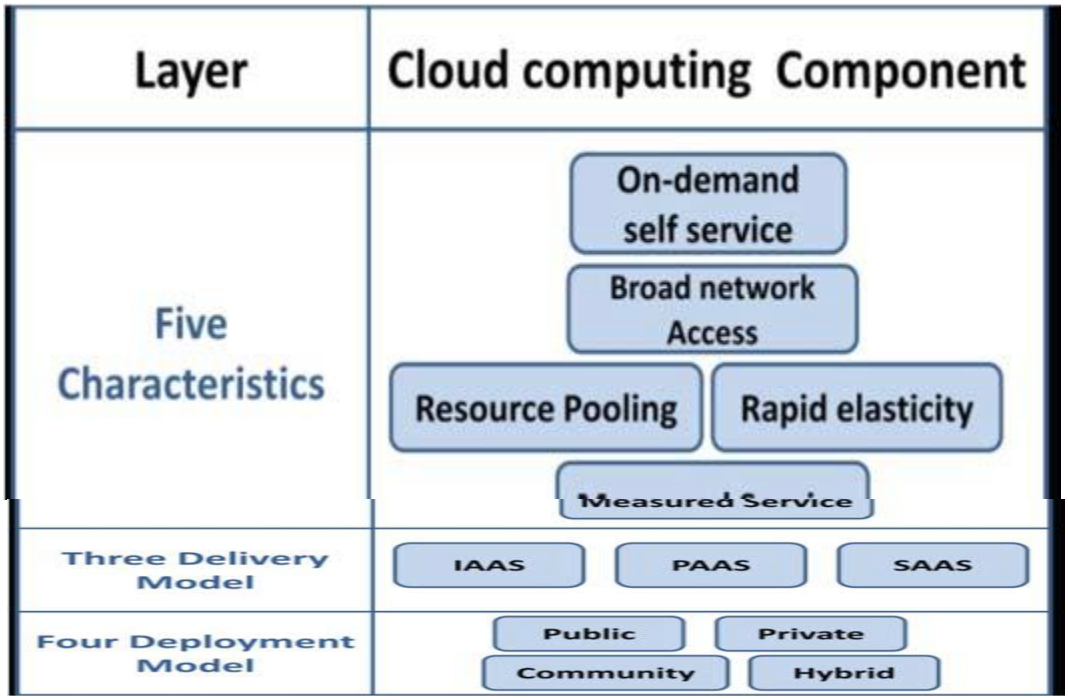


Fig -1 (Model of Cloud Computing)

2) **Platform as a Service (PaaS)**: PaaS provides all the resources to the customers that are required for building applications. It provides all the services on the internet .User not need to download and install the software. Consumers deploy all the application onto the cloud infrastructure .There is different tools and programming languages are provided to the uses to develop the applications. The consumer does not control network, servers, operating systems, or storage. Consumer controls all applications which they deploy.

Disadvantages

- There is very less portability among different providers.

3) **Infrastructure as a Service (IaaS)**: In this service consumer does not manage or control the underlying cloud infrastructure. In infrastructure as a service consumer able to control operating systems, storage, and all applications which they deployed. There is a limited control of customer on the networking components. Infrastructure Providers control storing and processing capacity.

Virtualization is used assign and dynamically resizes these resources to build systems as demanded by customers. Consumers deploy the software stacks that run their services. Provider provide network, services as on demand services. User use these services directly .It can be used to avoid buying, housing , and managing the basic hardware and software infrastructure components, scales up and down quickly to meet demand.

B.Layers of Services

All the services have number of layers. Which manage by the users and providers. Fig: 2 represents the different layers: -

Cloud Deployment Models:

1. **Public Cloud**: The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization .Anyone can use public cloud as they want without restriction.

2.Private Cloud: The cloud infrastructure is used by a single organization. Private cloud is only managed by the organization or a third party. General Public will not able to use the private cloud directly.

3. **Community Cloud**: The cloud infrastructure is shared by many organizations

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

Community cloud supports a Specific community that has shared concerns. Ex:-security requirements, policy, compliance considerations. It may be managed by the organizations or a third party.

4. **Hybrid Cloud**: Hybrid cloud is a combination of two or more clouds (private, community, or public). That remains unique entities but is bound together by standardized technology that enables data and application portability. Ex:- cloud bursting for load-balancing between clouds.

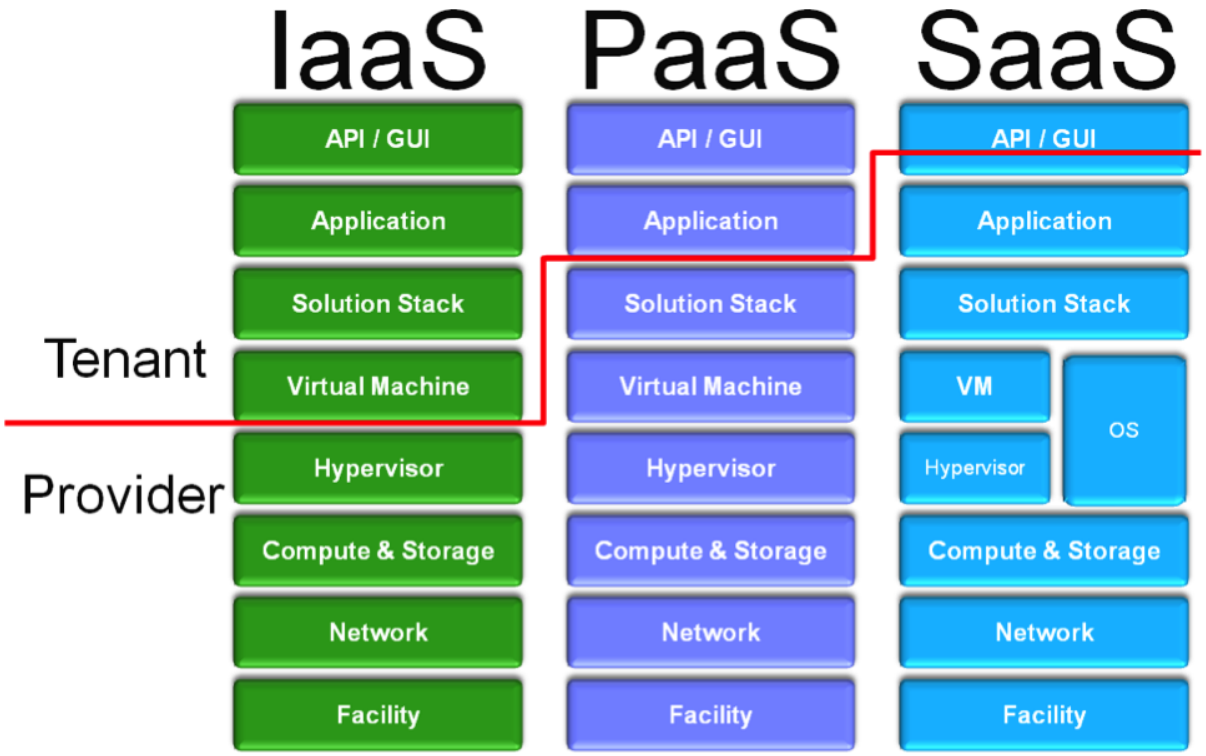


Fig.2 (Layers of services)

III. LOAD BALANCING

Load balancing is the process of improving the performance of the system by shifting of workload among the processors. Workload of a machine means the total processing time it requires to execute all the tasks assigned to the machine. Balancing the load of virtual machines uniformly means that anyone of the available machine is not idle or partially loaded while others are heavily loaded. Load balancing is one of the important factors to heighten the working performance of the cloud service provider. The benefits of distributing the workload includes increased resource utilization ratio which further leads to enhancing the overall performance thereby achieving maximum client satisfaction.

In cloud computing, if users are increasing load will also be increased, the increase in the number of users will lead to poor performance in terms of resource usage, if the cloud provider is not configured with any good mechanism for load balancing and also the capacity of cloud servers would not be utilized properly. This will confiscate or seize the performance of heavy loaded node. If some good load balancing technique is implemented, it will equally divide the load (here term equally defines low load on heavy loaded node and more load on node with less load now) and thereby we can maximize resource utilization. One of the crucial issue of cloud computing is to divide the workload dynamically.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

Load balancing is used to distributing a larger processing load to smaller processing nodes for enhancing the overall performance of system. In cloud computing environment load balancing is required to distribute the dynamic local workload evenly between all the nodes.

A. *Load balancing classification:*

Fig.3 represents different load balancing algorithms. This is mainly divided into two categories: static load balancing algorithm and dynamic load balancing algorithm:

1) **Static approach**: - This approach is mainly defined in the design or implementation of the system. Static load balancing algorithms divide the traffic equivalently between all servers.

2) **Dynamic approach**:- This approach considered only the current state of the system during load balancing decisions. Dynamic approach is more suitable for widely distributed systems such as cloud computing. Dynamic load balancing approaches have two types .They are distributed approach and non-distributed(centralized) approach. It is defined as following:

a) **Centralized approach**: - In centralized approach, only a single node is responsible for managing and distribution within the whole system. Other all nodes are not responsible for this.
b) **Distributed approach**: - In distributed approach, each node independently builds its own load vector. Vector is collecting the load information of other nodes. All decisions are made locally using local load vectors. Distributed approach is more suitable for widely distributed systems such as cloud computing.

B. *Metrics for Load Balancing*:

1. Throughput: - It is used to calculate the all tasks whose execution has been completed. The performance of any system is improved if throughput is high.
2. Fault Tolerance: -It means recovery from failure. The load balancing should be a good fault tolerant technique.
3. Migration time: -It is the time to migrate the jobs or resources from one node to other nodes. It should be minimized in order to enhance the performance of the system.
4. Response Time: - It is the amount of time that is taken by a particular load balancing algorithm to response a task in a system. This parameter should be minimized for better performance of a system.
5. Scalability: - It is the ability of an algorithm to perform Load balancing for any finite number of nodes of a system. This metric should be improved for a good system.

C. *Policies of load balancing algorithm*

There are many policies are used in load balancing algorithms:

• Information policy: It defined that what information is required and how this information is collected. This is also defined that when this information is collected
• Triggering policy: This policy defined that time period when the load balancing operation is starting to manage the load.

• Resource type policy: This policy defined the all types of resources which are available during the load balancing.

• Location policy: This uses all the results of the resource type policy. It is used to find a partner for a server or receiver.

• Selection policy: This policy is used to find out the task which transfers from overloaded node to free node.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

D. Major goals of load balancing algorithms

1. Cost effectiveness: Load balancing help in provide better system performance at lower cost.

2. Scalability and flexibility: The system for which load balancing algorithms are implemented may be change in size after some time. So the algorithm must handle these types' situations. So algorithm must be flexible and scalable.

3. Priority: Prioritization of the resources or jobs needs to be done. So higher priority jobs get better chance to execute.

Goals of Load Balancing

Goals of load balancing include:

- Substantial improvement in performance
- Stability maintenance of the system
- Increase flexibility of the system so as to adapt to the modifications.
- Build a fault tolerant system by creating backups.

Classification of Load Balancing Algorithm

Based on process orientation they are classified as:

a) Sender Initiated: In this sender initiates the process; the client sends request until a receiver is assigned to him to receive his workload
b) Receiver Initiated: The receiver initiates the process; the receiver sends a request to acknowledge a sender who is ready to share the workload
c) Symmetric: It is a combination of both sender and receiver initiated type of load balancing algorithm.

Based on the current state of the system they are classified as:

1. Static Load Balancing

In the static load balancing algorithm the decision of shifting the load does not depend on the current state of the system. It requires knowledge about the applications and resources of the system. The performance of the virtual machines is determined at the time of job arrival. The master processor assigns the workload to other slave processors according to their performance. The assigned work is thus performed by the slave processors and the result is returned to the master processor.

Static load balancing algorithms are not pre-emptive and therefore each machine has at least one task assigned for itself. Its aims in minimizing the execution time of the task and limit communication overhead and delays. This algorithm has a drawback that the task is assigned to the processors or machines only after it is created and that task cannot be shifted during its execution to any other machine for balancing the load. The four different types of Static load balancing techniques are Round Robin algorithm, Central Manager Algorithm, Threshold algorithm and randomized algorithm.

2. Dynamic Load Balancing

In this type of load balancing algorithms the current state of the system is used to make any decision for load balancing, thus the shifting of the load is depend on the current state of the system. It allows for processes to move from an over utilized machine to an under-utilized machine dynamically for faster execution.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

This means that it allows for process pre-emption which is not supported in Static load balancing approach. An important advantage of this approach is that its decision for balancing the load is based on the current state of the system which helps in improving the overall performance of the system by migrating the load dynamically.

Traditional Computing V/S Cloud Computing Environment

There are many similarities as well as differences between traditional scheduling algorithms and the scheduling of VM resources in cloud computing environment.
First of all the major difference between cloud computing environment and traditional computing environment is the target of scheduling. In traditional computing environment, it mainly schedules process or task so the granularity and the transferred data is small; whereas in cloud computing environment, the scheduling target is VM resources so the granularity is large and the transferred data is large as well.
Secondly, in cloud computing environment, compared with the deployment time of VMs, the time of scheduling algorithm can almost be neglected.

Need of Load Balancing

We can balance the load of a machine by dynamically shifting the workload local to the machine to remote nodes or machines which are less utilized. This maximizes the user satisfaction, minimizing response time, increasing resource utilization, reducing the number of job rejections and raising the performance ratio of the system.
Load balancing is also needed for achieving Green computing in clouds [5]. The factors responsible for it are:

1. Limited Energy Consumption: Load balancing can reduce the amount of energy consumption by avoiding over hearting of nodes or virtual machines due to excessive workload.
2. Reducing Carbon Emission: Energy consumption and carbon emission are the two sides of the same coin. Both are directly proportional to each other. Load balancing helps in reducing energy consumption which will automatically reduce carbon emission and thus achieve Green Computing.

IV. LOAD BALANCING ALGORITHMS

The paper describes about three load balancing algorithms which are Round robin algorithm, equally spread current execution load and Throttled Load balancing.

Round Robin: Round robin use the time slicing mechanism. The name of the algorithm suggests that it works in the round manner where each node is allotted with a time slice and has to wait for their turn. The time is divided and interval is allotted to each node. Each node is allotted with a time slice in which they have to perform their task. The complicity of this algorithm is less compared to the other two algorithms. An open source simulation performed the algorithm software know as cloud analyst, this algorithm is the default algorithm used in the simulation. This algorithm simply allots the job in round robin fashion which doesn't consider the load on different machines.

Equally spread current execution load: This algorithm requires a load balancer which monitors the jobs which are asked for execution. The task of load balancer is to queue up the jobs and hand over them to different virtual machines. The balancer looks over the queue frequently for new jobs and then allots them to the list of free virtual server. The balance also maintains the list of task allotted to virtual servers, which helps them to identify that which virtual machines are free and need to be allotted with new jobs. The experimental work for this algorithm is performed using the cloud analyst simulation. The name suggests about this algorithm that it work on equally spreading the execution load on different virtual machine.

Throttled Load balancing: The Throttled algorithm works by finding the appropriate virtual machine for assigning a particular job. The job manager is having a list of all virtual machines, using this indexed list, it allot the desire job to the appropriate machine. If the job is well suited for a particular machine than that job is, assign to the appropriate machine. If no virtual machines are available to accept jobs then the job manager waits for the client request and takes the job in queue for fast processing.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

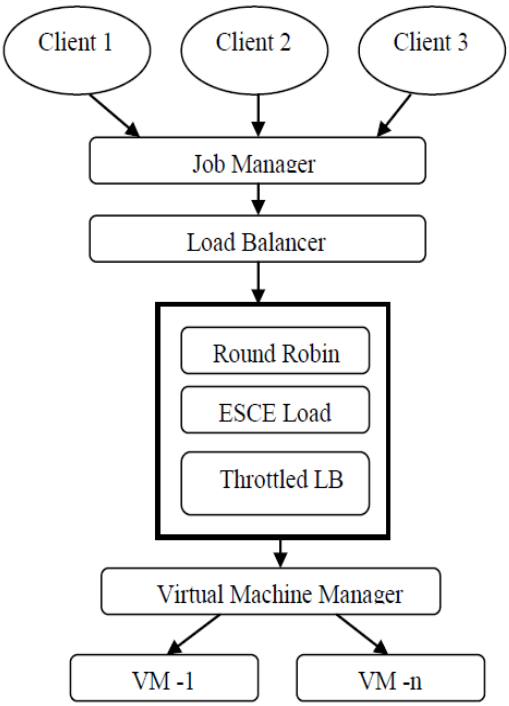
(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

ARA(Adaptive Resource Allocation):In ARA algorithm for adaptive resource allocation in cloud systems, which attempts to counteract the deleterious effect of burrstones by allowing some randomness in the decision making process and thus improve overall system performance and availability. The problem with this strategy is that it only considers the Poisson arrival streams as well as the exponentially distributed service time and the fixed number of choice.

The following figure shows the diagrammatical representation of the algorithm used for load balancing in cloud computing environment .The figure also shows the three algorithms which are studded in this paper using the cloud analyst simulation tool, this tool is based on cloud sim , the cloud sim provides a GUI inter face which helps to perform the experimental work.

The below figure shows the diagrammatical representation of the algorithm used for load balancing in cloud computing environment. The figure also shows the three algorithms which are studded in this paper using the cloud analyst simulation tool, this tool is based on cloud sim , the cloud sim provides a GUI inter face which helps to perform the experimental work.



V. CONCLUSION

Cloud computing mainly deals with software, data access and storage services that may not require end-user knowledge of the physical location and configuration of the system that is delivering the services. In the cloud storage, load balancing is a key issue.

The goal of load balancing is to increase client satisfaction and maximize resource utilization and substantially increase the performance of the cloud system and minimizing the response time and reducing the number of job rejection thereby reducing the energy consumed and the carbon emission rate.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

A few existing algorithms can maintain load balancing and provide better Strategies through efficient scheduling and resource allocation techniques as well. This paper presents a concept of Cloud Computing along with load balancing. Main thing is considered in this is load balancing algorithm. There are many above mentioned algorithms in cloud computing which consist many factors like scalability, better resource utilization, high performance, better response time.

REFERENCES

[1] http://www.ancoris.com/solutions/cloudcomputing.html, "Cloud Computing |Google Cloud -Ancoris."
[2] http://www.personal.kent.edu/~rmuhamma/ OpSystems/Myos/prioritySchedule.htm, "Priority Scheduling - Operating Systems Notes."
[3] A. Sidhu, S. Kinger, "Analysis of load balancing techniques in cloud computing", INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY 4 (2) (2013) pages737–741.
[4] http://www.qualitytesting.info/group/cloudcomputing/forum/topics/software-as-a-s, "Software as a Service (SAAS) - Quality Testing"
[5] Sahu, Yatendra and Pateriya, RK, "Cloud Computing Overview with Load Balancing Techniques", International Journal of Computer Applications, 2013,vol. 65, Sahu2013
[6] Chaudhari, Anand and Kapadia, Anushka, " Load Balancing Algorithm for Azure Virtualization with Specialized VM", 2013,algorithms,vol 1,pages 2, Chaudhari
[7] Nayandeep Sran,Navdeep Kaur , "Comparative Analysis of Existing Load Balancing Techniques in Cloud Computing ",vol 2,jan 2013
[8] Abhijit A Rajguru, S.S. Apte, "A Comparative Performance Analysis of Load Balancing Algorithms In Distributed Systems Using Qualitative Parameters", International Journal of Recent Technology and Engineering, Vol. 1, Issue 3, August 2012.
[9] Nidhi Jain Kansal, Inderveer Chana, "Cloud Load Balancing Techniques: A Step Towards Green Computing", IJCSI, Vol. 9, Issue 1, January 2012.
[10] Parin. V. Patel, Hitesh. D. Patel, Pinal. J. Patel, "A Survey on Load Balancing in Cloud Computing" IJERT, Vol. 1, Issue 9, November 2012.