"Load Balancing Under Bursty Environment For Cloud Computing."

Naimesh D. Naik
M.Tech(Information Technology)
U.V Patel Engg. College,
GANPAT University.

Ashilkumar R. Patel
M.E (Information Technology)
PIET,IT Dept.
Gujarat Technical University.

ABSTRACT

The cloud computing systems uses distributed resources to deliver a service to end users using several technologies in combination. Over utilization of these resources is responsible for lengthy response time and under utilization of these resources is responsible for wastage of the available resources. Burstiness in user demands also degrades the performance of the cloud computing system. Major challenge for cloud computing system is to satisfy the peak user demands with the most effective utilization of available resources. Current load balancing algorithm does not consider the current resource utilization and burstiness in user demands. This paper presents a dynamic load balancing algorithm which maintains the state of all virtual machine (VM) resources, and based on CPU, memory and storage space utilization, selects the less utilized VM resource To handle the request. Based on the predicted information of burstiness, this algorithm selects the best VM resource on the-fly to handle the request. This load balancing algorithm improves the performance by selecting the best sever under both bursty and non-bursty workloads.

1. INTRODUCTION

Cloud computing is an on demand service in which shared resources, information, software packages and other resources are provided according to the clients requirement at specific time. Its a term which is generally used in case of Internet. The whole Internet can be viewed as a cloud. Capital and operational costs can be cut using cloud computing[2].

In case of Cloud computing services can be used from diverse and widespread resources, rather than remote servers or local machines. There is no standard definition of Cloud computing. Generally it consists of a bunch of distributed servers known as masters, providing demanded services and resources to different clients known as clients in a network with scalability and reliability of datacenter. The distributed computers provide on-demand services. Services may be of software resources (e.g. Software as a Service,SaaS) or physical resources (e.g. Platform as a Service, PaaS) or hardware/infrastructure(e.g. Hardware as a Service, HaaS or Infrastructure as a Service, IaaS). Amazon EC2(Amazon Elastic Compute Cloud) is an example of cloud computing services [2].

The computing power of any distributed system canbe realized by allowing its nodes, to work cooperatively so that large loads are allocated among them in a fair and effective manner. Any strategy for load distribution among node is called load balancing. An effective load balancing policy ensures optimal use of the distributed resources where no resource is under or over utilized.

Cloud computing environment provides the users for accessing the shared pool of distributed resources. Cloud is a pay-go model where the consumers pay for there sources utilized instantly, which necessitates having highly available resources to service the requests on demand. Hence, the management of resources becomes a complex job from the business perspective of the cloud service provider[1].
There are many different kinds of load balancing algorithms available for cloud computing system, which can be categorized mainly into two groups:

❖ **Static algorithms:**

○ Static algorithms divide the traffic equivalently between servers. By this approach the traffic on the servers will be disdained easily and consequently it will make the situation more imperfectly [4].

❖ **Dynamic algorithms:**

○ In dynamic algorithms decisions on load balancing are based on current state of the system. No prior knowledge is needed for load balancing [3]. So it is better than static approach. Dynamic load balancing can be done in two ways

- **Distributed dynamic load balancing :**

In the distributed one, the dynamic load balancing algorithm is executed by all nodes present in the system and the task of load balancing is shared among them. A benefit, of this is that even if one or more nodes in the system fail, it will not cause the total load balancing process to halt, it instead would affect the system performance to some extent[3].

- **Non-distributed dynamic load balancing :**

In the non-distributed one, the dynamic load balancing algorithm is executed by a single node of the system and the task of load balancing is dependent only on that node. In this approach if the load balancing node fails, it will cause the total load balancing process to halt. Resource allocation in cloud computing can be done at two different levels. First, when an application is uploaded to the cloud, the load balancer assigns the requested instances to physical computers, attempting to balance the computational load of multiple applications across physical computers[5].

Major problem with the current load balancing algorithm is they does not consider the current utilization of VM resources. These algorithms divide the upcoming request equally without considering the available memory and storage space and current CPU utilization of the VM resource. These applications are dependent on other applications. These applications are executed either in parallel or sequentially. Cloud users try to access the multiple instances of different applications during a short time period. This will cause a significant arrival peak.

This will increase the competition between these applications to access the available resources and hence the load unbalancing for the cloud system. Current algorithms do not consider the bursty workloads and hence it will decrease the system performance.

Our proposed algorithm considers the current VM resource utilization and bursty workloads for distributing the load to each VM instances. We expect that using the proposed algorithm cloud service provider can meet the service level agreements (SLA) without purchasing additional resources. Our proposed algorithm also ensures that none of VM resources is over utilized when another one is underutilized. This will increase the system performance and provide faster response time. This will also increase the economic profit of an organization as all the resources are better utilized so there is no need for extra resources for handling the request.

1.1 Cloud Components

A Cloud system consists of 3 major components such as clients, datacenter, and distributed servers. Each element has a definite purpose and plays a specific role.

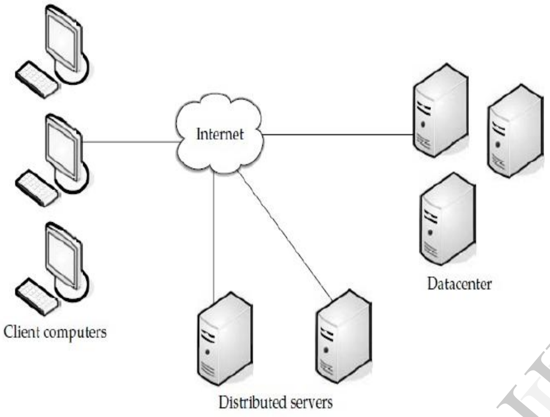


Figure 1: Three components make up a cloud computing solution(adopted from [1]).

1.1.1 Clients

End users interact with the clients to manage information related to the cloud. Clients generally fall into three categories as given in [1]:

- **Mobile:** Windows Mobile Smartphone, smartphones, like a Blackberry, or an iPhone.
- **Thin:** They don't do any computation work. They only display the information. Servers do all the works for them. Thin clients don't have any internal memory.
- **Thick:** These use different browsers like IE or Mozilla Firefox or Google Chrome to connect to the Internet cloud.

Now-a-days thin clients are more popular as compared to other clients because of their low price, security, low consumption of power, less noise, easily replaceable and repairable etc.

1.1.2 Datacenter

Datacenter is nothing but a collection of servers hosting different applications. A end user connects to the datacenter to subscribe different applications. A datacenter may existat a large distance from the clients.

Now-a-days a concept called virtualization is used to install a software that allow multiple instances of virtual server applications.

1.1.3 Distributed Server

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.

1.2 Type of Clouds

Based on the domain or environment in which clouds are used, clouds can be divided Into 3 categories:

- ❖ **Public Clouds**
- ❖ **Private Clouds**
- ❖ **Hybrid Clouds** (combination of both private and public clouds)

1.3 Virtualization

It is a very useful concept in context of cloud systems. Virtualization means "something which isn't real", but gives all the facilities of a real. It is the software implementation of a computer which will execute different programs like a real machine.

Virtualization is related to cloud, because using virtualization an end user can use different services of a cloud. The remote datacenter will provide different services in a fully or partial virtualized manner.

Two types of virtualization are found in case of clouds as given in [1] :

- **Full virtualization**
- **Paravirtualization**

1.3.1 Full Virtualization

In case of full virtualization a complete installation of one machine is done on the another machine. It will result in a virtual machine which will have all the software that are present in the actual server.

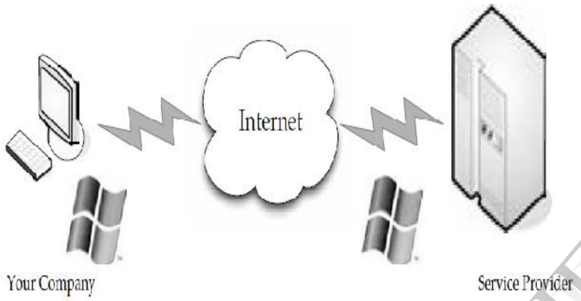


Figure 2: Full Virtualization (adopted from [1]).

Here the remote datacenter delivers the services in a fully virtualized manner. Full virtualization has been successful for several purposes as pointed out in [1]:

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

1.3.2 Paravirtualization

In Paravirtualization, the hardware allows multiple operating systems to run on singlemachine by efficient use of system resources such as memory and processor. e.g. VMware software. Here all the services are not fully available, rather the services are provided partially.

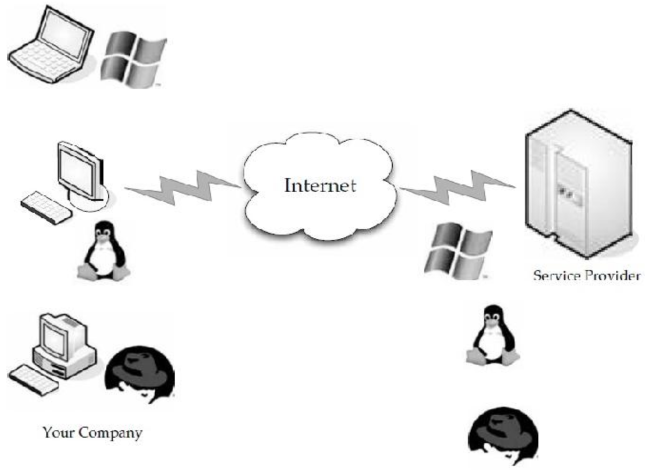


Figure 3: Paravirtualization (adopted from [1]).

Paravirtualization has the following advantages as given in [1]:

- 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 hardwares can be moved or replaced or repaired easily, capacity management is simple and easier.

1.4 Services provided by Cloud computing

Service means different types of applications provided by different servers across the cloud. It is generally given as ”as a service”. Services in a cloud are of 3 types as given in[1] :

I. **Software as a Service (SaaS)**
II. **Platform as a Service (PaaS)**

III. Hardware as a Service (HaaS) or Infrastructure as a Service (IaaS)

1.4.1 Software as a Service (SaaS)

In SaaS, the user uses different software applications from different servers throughthe Internet. The user uses the software as it is without any change and do not need to makelots of changes or doesn't require integration to other systems. The provider does all theupgrades and patching while keeping the infrastructure running [2].

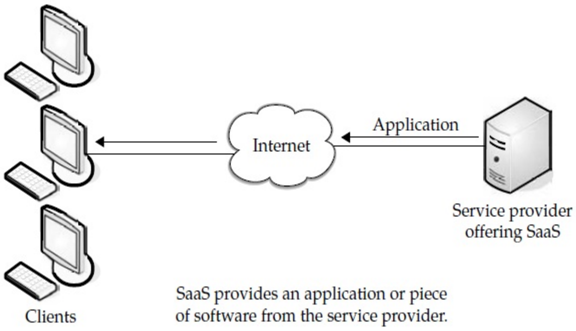

Application
Internet
Service provider offering SaaS
Clients
SaaS provides an application or piece of software from the service provider.

Figure 4: Software as a service (SaaS) (adopted from [1])

The client will have to pay for the time he uses the software. The software that does a simple task without any need to interact with other systems makes it an ideal candidate for Software as a Service. Customer who isn't inclined to perform software development but needs high-powered applications can also be benefitted from SaaS. Some of these applications include (taken from [1]):

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

Benefits:

The biggest benefit of SaaS is costing less money than buying the whole application. The service provider generally offers cheaper and more reliable applications as compared to the organization [1]. Some other benefits include (given in [1]): Familiarity with the Internet, Better marketing, Smaller staff, reliability of the Internet, data Security, More bandwidth etc.

Obstacles:

- SaaS isn't of any help when the organization has a very specific computational need that doesn't match to the SaaS services.
- While making the contract with a new vendor, there may be a problem. Because the old vendor may charge the moving fee. Thus it will increase the unnecessary costs.
- SaaS faces challenges from the availability of cheaper hardwares and open source applications.

1.4.2 Platform as a Service (PaaS)

PaaS provides all the resources that are required for building applications and services completely from the Internet, without downloading or installing a software [1].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.

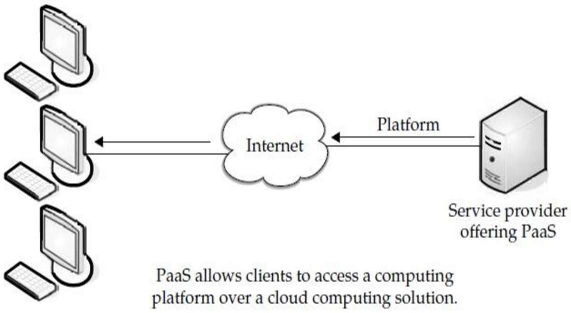

Platform
Internet
Service provider offering PaaS
PaaS allows clients to access a computing platform over a cloud computing solution.

Figure 5: Platform as a service (PaaS) (adopted from [1])

Downfall (taken from [1]):

➢ lack of portability among different providers.

➢ if the service provider is out of business, the user's applications, data will be lost.

1.4.3 Hardware as a Service (HaaS)

It is also known as Infrastructure as a Service (IaaS). It offers the hardware as a service

To a organization so that it can put anything into the hardware according to its will [1].

HaaS allows the user to "rent" resources (taken from [1]) as:

- Server space
- Network equipment
- Memory
- CPU cycles
- Storage space

2.1 Problem Statement:-

Load balancing is a computer networking method to distribute workload across multiple computers or a computer cluster, network links, central processing units, disk drives, or other resources, to achieve optimal resource utilization, maximize throughput, minimize response time, and avoid overload. There are basically two types of algorithm currently available in the market are like. Static algorithm, which divides the traffic equally, is announced as round robin algorithm. However, there were lots of problems appeared in this algorithm. Therefore, another approach is Dynamic which is weighted round robin was defined to improve the critical challenges associated with round robin. In this algorithm each servers have been assigned a weight and according to the highest weight they received more connections. In the situation that all the weights are equal, servers will receive balanced traffic [4]. Problem with this kind of algorithm is that these algorithms are not able to handle bursty workloads. Even these algorithms do not consider the current situation of each node of the system. So that is how we will overcome the issue of balancing load under bursty environment in cloud computing.

2.2 Objectives:-

- To do a literature review for the current existing approaches that are available for load balancing in cloud computing.
- To provide the proposed algorithm by studying the existing algorithms of load balancing in cloud computing.
- And to make the algorithm that uses the proper approach of balancing the load under the bursty environment will be made up of the given two approaches.
- So Now we are providing two algorithm as our objective to provide better solution to make the algorithm that is worth the effort and give desired result which are as follows:-

In algorithm (LB3M) the strategy is to calculate the average completion time of each task for all nodes, respectively and find the task that has the maximum average completion time. Further find the unassigned node that has the minimum completion time less than the maximum average completion time for the task selected in early stage of the algorithm. Then this task is dispatched to the selected node for computation [6]. This strategy achieves better performance than the static algorithms but it also does not consider the current resource utilization of the cloud system.

In another algorithm (ARA) the strategy is to predict the changes in user demands and shifts between the schemes that are greedy i.e. select the best server and random i.e. select the random server based on the predicted workloads. This scheme will improve the performance by making a smart site selection but the problem is that is does not consider the current utilization of available resources [5].

Rashmi K. S., et al [1]A load balancing algorithm has been proposed to avoid deadlocks among the Virtual Machines(VMs) while processing the requests received from the users by VM migration.

The deadlock avoidance enhances the number of jobs to be serviced by cloud service provider and thereby improving working performance and the business of the cloud service provider.

Ram Prasad P., et al [2]”Cloud computing” is a term, which involves virtualization, distributed computing networking, software and web services.
A cloud consists of several elements such as clients, datacenter and distributed servers. It includes fault tolerance, high availability, scalability , flexibility, reduced overhead for users, reduced cost of ownership, on demand services etc.

Mishra,Ratanet al [3]An ant colony optimization has been proposed to initiate the service load distribution under cloud computing architecture.
The pheromone update mechanism has been proved as a efficient and effective tool to balance the load. This modification supports to minimize the make span of the cloud computing based services and portability of servicing the request also has been converged using the ant colony optimization technique. This technique does not consider the fault tolerance issues.

Zenonc.,et al[4]Various models and rules can be applied to load balancers, however these should be based on the scenario the load balancer will be applied for. The network structure or topology should be taken into account when creating the logical rules for the load balancer.

This is due to the pricing of transfer between regions, availability zones and cloud vendors, which all constitute different pricing strategies.

Jianzhe T., et al [5]New static ARA algorithm tunes the load balancer by adjusting the trade-off between randomness and greediness in the selection of sites.
New online ARA algorithm that predicts the beginning and the end of workload bursts and automatically adjusts the load balancer to compensate. They show that the online algorithm gives good results under a variety of system settings. This approach is more robust than the static algorithm.

Che-LunH.,et al [6]In this paper, Authors Have proposed an efficient scheduling algorithm, LB3M, for the cloud computing network to assign tasks to computing nodes according to their resource capability.
Similarly, LB3M can achieve better load balancing and performance than other algorithms, such as MM and LBMM from the case study.

H. El Bakkali,et al [7]In this paper, Authors surveyed the state-of-the-art of load balancing in cloud computing system. We establish the state of the art load balancing in the cloud computing system, giving a definition of this term, its classification and examples of its implementation in classical distributed systems and in the cloud computing system key technologies as well as research directions and cases study of search.

2.3 Proposed Algorithm:-

```
1. Initialize Variables Request;

2. Initialize Variables of avail Resource;
3. Initialize Pheromone on the trail Basis;
4. While (Value of Timer<$T_i$)
        do
        Construct Solutions for Requests;
5. $R_{new}=\min\{f_{obj}(P_k)|k=1,2,\ldots K\}$;
6. If $R_{new} < R$
```

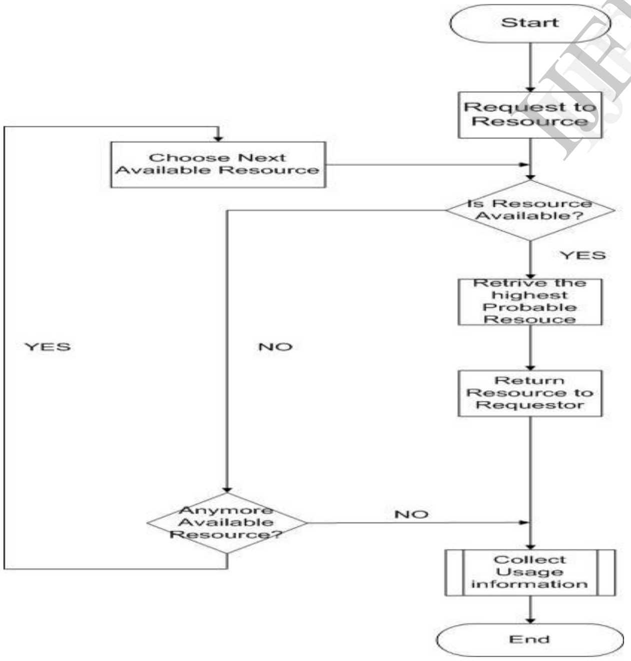
```
        then $R = R_{new}$;
7. Pheromone Update;
        End
8. Allocation ofResource to Request;
        End
```

Figure:6 Pseudo code for distributing load dynamically in cloud

As shown in the above figure is the Algorithm for allocation of Resource to the requests made by requester according to their Probability (OR Pheromone).

2.4 Experimental Setup

Experimental setup of the Load balancing in the cloud computing requires the allocation of the requests that has been made by the requester to the resource. Load Balancer balances the mechanism of allocating the proper resource to the proper request to maintain the balance.


Start
Request to Resource
Choose Next Available Resource
Is Resource Available?
YES
Retrive the highest Probable Resouce
YES
NO
Return Resource to Requestor
Anymore Available Resource?
NO
Collect Usage information
End

There are variety of Algorithms for Load Balancing for Cloud Computing. Different Algorithm Uses different strategy to balance the load by allocating the request to resource which is free at that time of period. The flow of request made by the user and their allocation to the desired resource is shown in the below figure.

In the above architecture shown in figure the cloud user would be accessing his services from the cloud controller server. If user makes a request for a cloud service,the request will first go to the cloud controller server. This request will be transferred to the load balancer.

A monitoring agent would be continuously monitoring the CPU usage, memory and storage space usage and expected load and current load data for each virtual instances. All the data are transferred to the load balancer by monitoring agent. Basedon the data of each virtual instances the request is transfer the appropriate node controller server where virtual machines are running and from where different instances are provided to different users. Finally the request is transfer the virtual instance that is selected by the load balancer.

For implementing the real load balancer we uses the Open source OSes like Ubuntu Flavor and to Create Cloud we require the Software like Eucalyptus etc.

- **<u>Private Cloud using Ubuntu 10.04 server Edition:-</u>**

➢ First of all we have to install server to do Load balancing in cloud Environment.

Setup Server 1:-

1. Insert ubuntu 10.04 server edition cd.
2. select “Install ubuntu enterprise cloud‟
3. Configure the network: select “configure network manually‟

address 192.168.0.221
gateway 192.168.0.1
netmask 255.255.255.0
nameserver 121.242.xxx.xxx

4. Host name for this system: cc
5. Cloud controller address : leave it blank

6. Cloud installation mode:Select following Cloud controller, Walrus storage service, Cluster controller, Storage controller.
7. Partition disks select "Guided-use entries disk and set up LVM"
8. Full name for new user and username for account :cladmin
9. Select no automatic updates
10. Eucalyptus cluster name : cluster1
11. Pool of IP addresses that can be dynamically assigned as
public IP"s of virtual machines: 192.168.0.70-- 192.168.0.80
12. Install grub boot loader to master boot loader: yes
Also install KVM on server1 which helps to install images andbundle them.

```
$ sudo apt-get install qemu-kvm.
```

Setup Server 2:-
1. Insert ubuntu 10.04 server edition cd
2. select "Install ubuntu enterprise cloud"
3. Configure the network: select "configure network manually"
address 192.168.0.222
gateway 192.168.0.221 (IP of cloud controller)
netmask 255.255.255.0
nameserver 121.242.xxx.xxx
Here Cloud controller is detected automatically
4. Host name for this system: nc
5. Cloud installation mode: Node controller
6. Partition disks select „Guided-use entire disk and set up LVM"
7. Full name for new user and username for account :cladmin
8. select no automatic updates
9. Install grub boot loader to master boot loader: yes.

Exchange of Public SSH Keys
On node controller set a temporary password

```
$ sudopasswd eucalyptus
```

On cloud controller

```
$ sudo -u eucalyptus ssh-copy-id -i
/var/lib/eucalyptus/.ssh/id_rsa.pub
eucalyptus@192.168.0.222
```

On node controller remove temporary password

```
$ sudopasswd -d eucalyptus
```

Get credentials
Open web browser and enter following url:
https://192.168.0.221:8443/#login
and click on credentials tab

CONCLUSION

As cloud computing is a new area for research and development, developing a dynamic load balancing algorithm is a major challenge for cloud service provider. This algorithm will ensure the optimum utilization of cloud resources. This algorithm will provide faster response time and it will improve the system performance in the case of changing user demands. This will help the cloud service provider to meet the service level agreements. This algorithm will cut the economic cost for an organization because less resources will be required than static algorithmsto handle the user requests.

REFERENCES

[1.] Rashmi K. S,Suma V.,Vaidehi M., "Enhanced Load Balancing Approach to Avoid Deadlocks in Cloud" in Special Issue of International Journal of Computer Applications (0975 – 8887) on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, June 2012 Page no. [31-35].

[2.] Padhy Ram Prasad, & P. Gautam Prasad Rao. "Load Balancing in Cloud Comuting System" at Department of Computer Science and Engineering National Institute of Technology, Rourkela-769 008, Orissa, India May, 2011 Page no. [1-45].

[3.] Mishra Ratan&JaiswalAnant.“Ant colony Optimization : A Solution of Load balancing in Cloud”in International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.2, April 2012 Page no. [33-50].

[4.] ChaczkoZenon,MahadevanVenkatesh, AslanzadehShahrzad, &Mcdermid Christopher,“Availability and Load Balancing in Cloud Computing” in2011 International Conference on Computer and Software Modeling IPCSIT vol.14 (2011) © (2011)IACSIT Press, Singapore (2011) Page no. [134-140].

[5.] Tai Jianzhe, ZhangJuemin, Li Jun, MeleisWaleed, &MiNingfang“ArA: Adaptive resource allocation for cloud computing environments under bursty workloads”.In the 30th IEEE International Performance Computing and Communications Conference, Page no. [1–8].

[6.] Che-Lun Hung, Hsiao-hsi Wang and Yu-Chen Hu “Efficient Load Balancing Algorithm for Cloud Computing Network” at{clhung, hhwang, ychu}@pu.edu.tw in their Case study Page no. [251-253].

[7.] H. Bakkali EL, A. Khiyaita , M. Zbakh , Dafir EL Kettani (2012 IEEE). “Load Balancing Cloud Computing : State of Art” in volume no:978-1-4673-1053-6/12/$31.00 ©2012 IEEE Page no. [106-109].

[8.] http://www.akashsharma.me/private-cloud-setup-usingeucalyptus-and-xen/ for working with private cloud.

[9.] http://cloudcomputing.sys-con.com/node/2261725 for more information of using virtual OS installation guide.