Load Balancing Tactics in Cloud Computing: A Systematic Study

Ramandeep Kaur[#1], Navtej Singh Ghumman[#2]
#1 M.Tech Student, Department of Computer Science and Engineering, SBS State Technical Campus, Ferozepur (City), India.

#2 Assistant Professor, Department of Computer Science and Engineering, SBS State Technical Campus, Ferozepur (City), India.

ABSTRACT
Cloud computing has recently emerged as new paradigm in field of technology. Cloud computing is attractive to business owners and IT people. It is still in its infancy and many issues are to be addressed. This paper covers the cloud computing basics and discusses load balancing in cloud computing environment as the one of the major challenges of cloud computing. It also discusses the various existing load balancing algorithms.

Key words – Cloud computing, Load balancing, existing load balancing algorithms.

1. INTRODUCTION

Cloud computing is a buzzword that offers different services to different people. It is also called as anywhere anytime computing. Users can access its service being at any place across the internet. It acts as online storage of user's application data and many more. Cloud computing means that you are using you are just sitting at your desktop and software and the hardware you are using is provided to you by some another company and accessing it over the Internet and the most important factor us that you don't care[1,2,3].Cloud computing is implemented using virtualization approach.

Cloud Computing is being used in our everyday life. You start your day with checking your mail which is a cloud computing service you are using. It provides various advantages and advantages as mentioned below:

1.1 Some advantages and disadvantages of cloud computing:

Advantages:

- Scalable and High Performance
- Environment friendly
- Cost -efficient approach
- Backup and Recovery
- Location independence
- Quick deployment and ease of installation

Disadvantages:

- Lack of security of private data
- Technical difficulties and downtime
- Dependency and Vendor lock-in

1.2 Characteristics of Cloud Computing [5,6].

1) Dynamic computing infrastructure: Cloud computing requires a dynamic computing infrastructure. There should be high levels of redundancy so that it has high availability and mostly it should be extended without architecture rework.

2) IT service-centric approach: Cloud computing is a IT service-centric rather than server-centric model. In most cases users prefer to use business service or application and they would prefer easily and quickly access to their dedicated application.

3) Self-service based usage model: This requires certain level of user self-service. This would help administrative staff to focus on more strategic, high-valued responsibilities.

4) Minimally or self-managed platform: It is preferable to have a self-managed technology. Cloud computing enable self-management through software automation with the following capabilities such as:

- A provisioning engine for deploying services and tearing them down recovering resources for high levels of reuse.
- Mechanisms for scheduling and reserving resource capacity.
- Capabilities for configuring, managing, and reporting to ensure resources can be allocated and reallocated to multiple groups of users.
- Tools for controlling access to resources and policies for how resources can be used or operations can be performed.

5) *Consumption-based billing:* Cloud computing is pay as you use model. User is charged based on consumption-based model. Cloud computing must provide mechanisms to capture usage information thus helping chargeback reporting.

1.3 Challenges in Cloud Computing

There are many challenges in cloud computing:-

1) Security
2) Efficient load balancing
3) Real Benefits / Business Outcome
4) Consistent and Robust Service abstractions
5) Resource Scheduling
6) Fast Internet speed.
7) Datacenter energy consumption
8) Service quality

2. CLOUD COMPUTING ARCHITECTURE

Cloud computing architecture explains the basic framework of a cloud computing environment.

2.1 Three components make up Cloud Computing

Fig. 1 shows different cloud components.

1) *Clients:* End-users interact with the clients to manage the information on the cloud. Clients are categorized as [6]:
   - *Mobile:* Windows Mobile, Smartphone.
   - *Thin:* They don't have any processing power. They only display the information. Servers do all the work for them. Thin clients don't have any internal memory.
   - *Thick:* Different browsers like IE, Mozilla Firefox or Google Chrome are used as thick clients. But nowadays thin clients are preferable.
2) *Datacenters:* A datacenter is collection of different servers hosting different applications. End-user connects to datacenter to subscribe to different applications.
3) *Distributed servers:* The servers which are distributed over the network are distributed servers. The user is accessing the application from some remote server but he feels as if he is using application from his own machine.

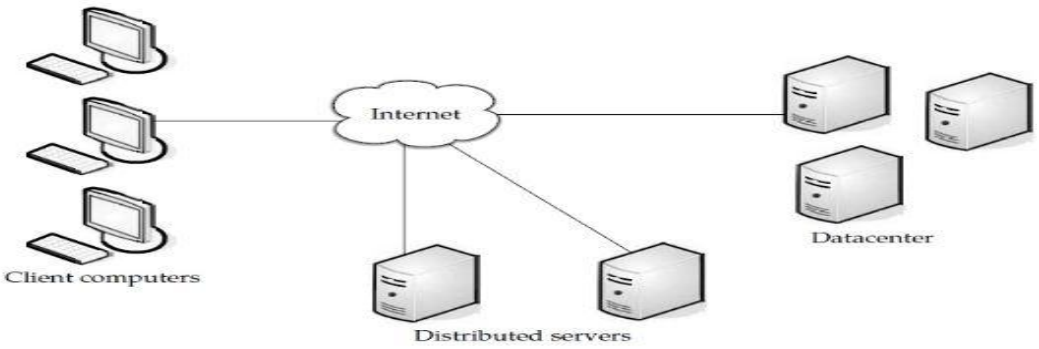

Fig. 1 Components make up cloud computing solution [7]

2.2 Layers of Abstraction

Service means different types of applications that are provided to end-user by cloud computing environment [8]. There are various services that are provided to user across the cloud. Fig. 2 shows the layers of abstraction in cloud computing.

1) *Software as a Service (SaaS):* Software as a service is a way of delivering a service to a user such that user need not install or purchase any software rather he can access it over the Internet. The provider manages the software including its security, availability and performance [9]. Various types of applications offered by SaaS are:
   - Customer resource management (CRM)
   - Video conferencing
   - IT service management
   - Accounting
   - Web analytics
   - Web content management
2) *Platform as a Service (PaaS):* Platform as a service provides all the resources which are required for building applications. Here the software development platform is virtualized. This allows the user to rent the virtualized servers and associated services for running existing applications. User need not have to control operating system, servers or storage.
3) *Infrastructure as a Service (IaaS):* Infrastructure as a service is also known as Hardware as a Service (HaaS). It provides the user with hardware where he can install his own operating system and software and use it. HaaS allows you to rent resources such as :
   - Server Space
   - Network Equipment
   - Memory
   - CPU cycles
   - Storage space

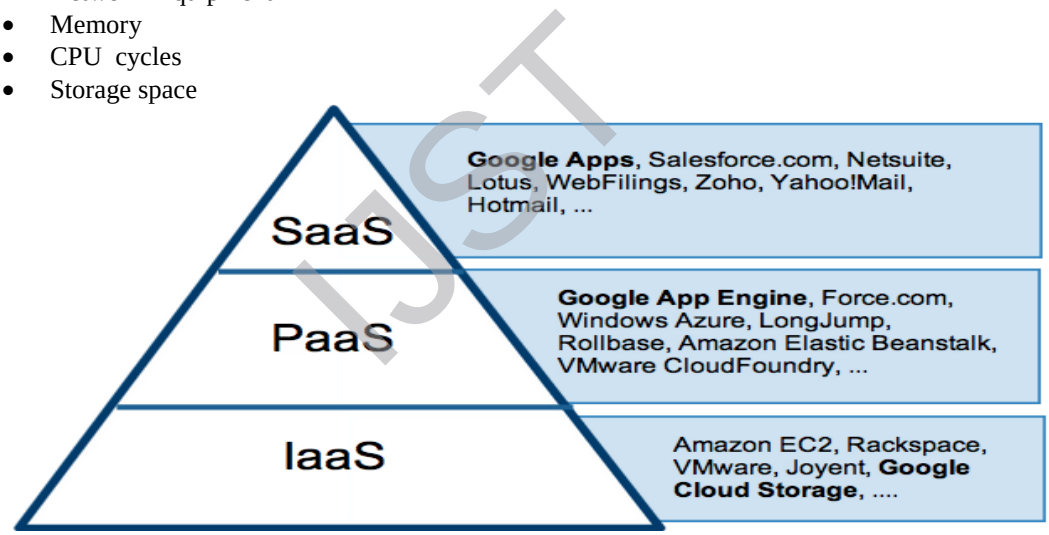

Fig. 2 Abstractions of cloud computing [10]

2.3 Cloud Computing Deployment Models

There are several different cloud deployment models for implementing cloud technology [11,12]. Fig. 3 shows the types of deployment models.

1) *Public Cloud:* As the name suggests, public cloud applications, storage and resources are made available to the general public and is managed by third party. This type of cloud is the most familiar and a cheaper one. It provides the services which may be free or pay-as-you-use model. We use it every day from Hotmail to Google Docs to Flicker.
2) *Private Cloud:* In this model, cloud infrastructure is solely owned by a one or more organization. The organization is only responsible for its management or it may pass on this to some third party. Any authorized person is not allowed to access the services. Its services are dedicated to particular set of users thus adds to the security of resources and applications.
3) *Community Cloud:* The cloud infrastructure is owned, managed and operated by several organizations in a community with different security and reliability requirements. It may exist on or off-premises.
4) *Hybrid Cloud:* So as the name suggests, hybrid cloud is composition of at least one private cloud and at least one public cloud. It allows to take scalability and cost-effectiveness that public cloud offers and without exposing data and applications to third-party [13].

There is rarely a clear-cut solution as which model is best suited. It all depends upon organization's needs. Every organization wants to cut down capital expenditure and control operating costs.

2.4 Concept of Virtualization

Virtualization is something which is not 'real' but it behaves as real. It is software implementation of a machine which executes different programs as an actual machine would execute [14].Fig 4 shows concept of virtualization.

Virtualization concept is employed in clouds. Virtualization is of two types in clouds:

1) *Para Virtualization:* Here services are not fully available, rather services are partially available. In this type of virtualization, hardware allows running multiple operating systems on a same machine. For example, Windows OS and Linux OS can be installed on same machine using Para Virtualization. It offers following advantages:
   - Capacity Management
   - Migration
   - Recovery from crash failure

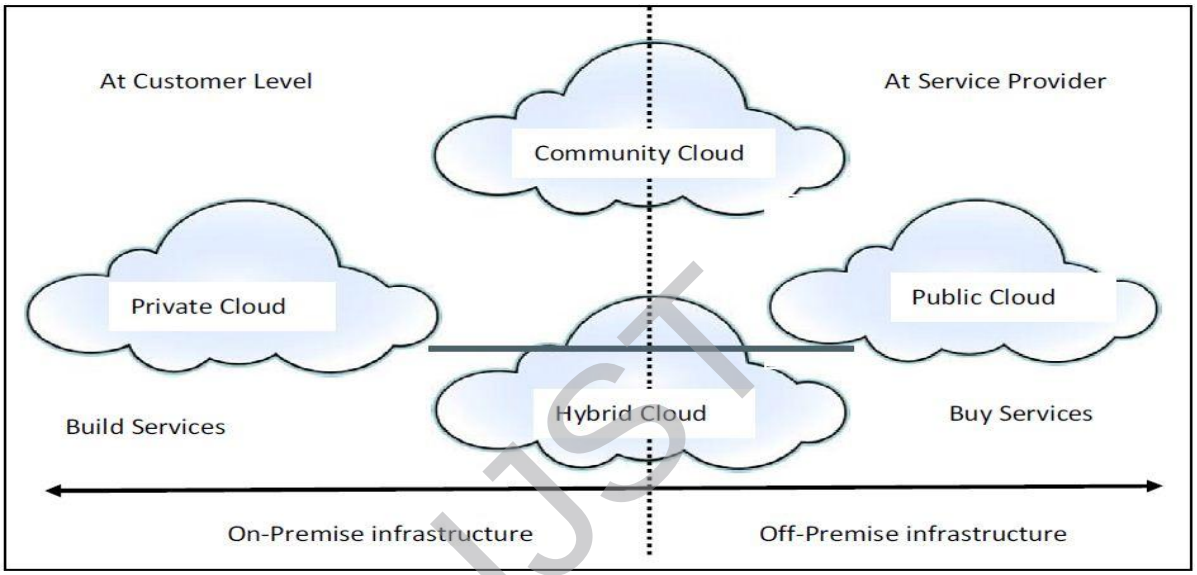

Fig. 3 Cloud computing deployment models

2) *Full Virtualization:* Here the virtual machine provides full functionality of an actual machine. In this case, one machine is completely installed on another machine. It has been successful in emulating hardware on another machine and sharing of system among multiple users.

Fig. 4 Cloud hosting and virtualization solutions [15]

3. LOAD BALANCING

Load Balancing is technique of assigning the load to the different nodes in such a manner that there is no node that is over-loaded or under-loaded. It helps in effective resource utilization and improves response time of the job. There are various types of algorithms available for load balancing.

3.1 Goals of Load Balancing Algorithms

Some of the goals of load balancing algorithms are [16]:

- To have a backup plan in case system fails.
- To improve performance substantially.
- To allow future modification in system.
- To maintain stability of system.

3.2 Classification of load balancing algorithms

Load balancing algorithms have been classified based on who initiated the process and on current state of system.

1) Classification based on who initiated the process: They are of three types based on this classification:
   - *Sender-initiated:* If the sender initiates the algorithm conveying the other nodes that it is over-loaded.
   - *Receiver-initiated:* If the receiver initiates the algorithm conveying the other nodes that it is under-loaded and thus asks for the load.
   - *Symmetric:* If either sender or receiver initiates the algorithm.
2) Classification based on current state of system: There are of two types based on this classification:
   - *Static:* The algorithm does depend upon current state of system. Thus prior knowledge of the system is not needed.
   - *Dynamic:* The algorithm depends upon the current state of the system. It varies as the state of system changes.

3.3 Policies in dynamic load balancing

There are various policies available for dynamic load balancing [17].

1) *Transfer policy:* It selects the job to be transferred from local node to some node at remote place.
2) *Selection policy:* It specifies the processes participating in load exchange.
3) *Location policy:* The selection of destination node for the transferred task is the location strategy.
4) *Information policy:* It is responsible for collecting information about the nodes.
5) *Load estimation policy:* It tells how to estimate the load of any particular node.
6) *Process transfer policy:* It decides whether to execute the process locally or remotely.
7) *Priority transfer policy:* It determines the priority of local and remote processes at any given time.
8) *Migration Limiting policy*: It defines the upper limit for process migration.

3.4 Qualitative parameters for load balancing algorithms

Some of the parameters are listed below as [17]:

1) *Nature:* It is related with the nature of the load balancing algorithms whether they are static or dynamic.
2) *Reliability:* It is related with the reliability of load balancing algorithms. Static load balancing algorithms are less reliable as compared to dynamic load balancing algorithms.
3) *Adaptability:* It checks whether load balancing algorithm adapts to changing situations. Static load balancing algorithm fails in varying conditions. Dynamic load balancing algorithms are adaptive to almost every situation.
4) *Predictability:* It is related with deterministic or non-deterministic nature of the outcome of the algorithm. As the average execution time and workload assignment is fixed in case of static approach, so static load balancing algorithm is predictable and dynamic load balancing algorithm is unpredictable.
5) *Waiting Time*: It is the total amount of time spent waiting in ready queue.
6) *Turnaround Time*: the time interval between submission of the task and its completion.
7) *Throughput:* The amount of data transferred from one node to another in a given period of time successfully.

4. EXISTING LOAD BALANCING ALGORITHMS

There are several existing load balancing algorithms available which are listed below as:

1) *Index Name server (INS):* This algorithm [18] computes the optimum selection point which depends upon several parameters. These include Hash code of block of data to be downloaded, server's location holding target block of data, maximum bandwidth required for downloading, and path parameter.
2) *Downloading algorithm from FTP server (DDFTP):* This algorithm [19] performs load balancing by dividing the file of size n into n/2 divisions. This minimizes the node communication, thereby reducing the network overhead which further eliminates the need for run-time supervision of nodes.
3) *Honeybee Foraging Algorithm:* This algorithm [20] is influenced by behavior of honeybee for searching and obtaining food. Here the web servers are allocated dynamically to the users as demand keeps on varying depending upon their requirement. All servers are grouped under virtual servers and each has its own virtual service queue. Each server processing request has its own queue to compute profit which is equal to quality that honeybee posses in their waggle dance.
4) *Active Clustering:* This algorithm [21] exploits the similarity of the jobs and linking the similar ones in a group. The performance of the system is improved through high availability of resources. The performance degrades by system diversity.
5) *Biased Random Sampling:* This algorithm [22] is based on constructing a virtual graph which depicts connectivity of each node with each and every node of a system. This algorithm offers a highly reliable and scalable approach.
6) *CARTON:* The proposed algorithm [23] integrates the concept of LB (Load Balancing) and DRL (Distributed Rate Limiting). LB takes care of equal distribution of load to servers. DRL makes sure that resources are distributed in way to maintain fair resource allocation.
7) *ACCLB:* This proposed algorithm [24] is based on Ant Colony and Complex Network Theory. This algorithm exploits the scale-free and small-world quality of network. This is an excellent fault-tolerant, scalable and adaptive approach.
8) *Min-min Algorithm:* This algorithm [25] computes the minimum completion time for unscheduled jobs and then assigns the job to the node with a minimum completion time.
9) *Max-min Algorithm:* This algorithm [26] works similar to min-min algorithm. The basic difference is that it gives more priority to larger tasks in contrast to max-min algorithm. Therefore, larger jobs (jobs with high completion time) are assigned first and shorter jobs keep on waiting for their scheduling.
10) *Two-phase Load balancing Algorithm (OLB + LBMM):* This algorithm [27] uses the combined approach of OLB (Opportunistic Load Balancing) and LBMM (Load Balance Min Min).This combined approach provides effective resource utilization and improve performance.
11) *Power Aware Load Balancing (PALB):* This algorithm [28] preserves the state of all calculated nodes and based on their utilization percentages decides which of the calculated nodes would be operating. This could be implemented to the cluster controller of the cloud that is power aware and use job scheduler to simulate requests from users for virtual machine instances.
12) *LBVS:* Load Balancing Virtual Storage strategy proposed by Liu et al., [29] that provide a large scale net data storage model and Storage as a Service model based on Cloud Storage. Storage virtualization is achieved using a three-layered architecture and load balancing is achieved using two load balancing modules.
13) *Exponential Smooth Forecast based on Weighted Least Connection (ESWLC):* This algorithm [30] assigns jobs to particular node based upon the number of connections that exist for that node. This algorithm suffers from various drawbacks so it does not take into account the processing speed, storage capacity and bandwidth. Therefore, ESWLC was proposed to take into account these parameters.
14) *Join Idle Queue:* Y. Lua et al. [22] proposed this algorithm was proposed by for distributed scalable web services. This allows for large scale load balancing for distributed load balancing.
15) *Load Balancing for Real-time, Location-based Event Processing on Cloud Systems:* This algorithm [31] was proposed for distribution of workload for location-based event processing on cloud systems. Sungmin Yi et al., introduced various event processing techniques for range queries such as:
    - Round Robin Data Distribution.
    - Round Robin Query Distribution.
    - Data/Query Distribution through Space Partitioning.
    - Skew Aware Distribution.
16) *Compare and Balance:* This approach [32] employs a compare and balance to reach an equilibrium state and manage unbalanced system's load. Here, the node randomly select any node and compare the load with itself.
17) *Vector Dot:* This is novel based algorithm proposed by A. Singh et al. [33].This algorithm uses the dot-product to distinguished nodes based on item requirements and helps in removing overheads storage nodes, servers and switches.

5. CONCLUSIONS

Cloud Computing is an attractive technology which relieves end-user from the burden installing and maintaining software, problems of system failure and many more benefits which we had just gone through. Still there is much more to be explored yet. Cloud computing also has various issues/challenges which are still unaddressed. This paper discusses about cloud computing basics, its characteristics, service models and types, load balancing in brief, which is one of the major challenges of cloud computing and also various existing load balancing tactics.

6. REFERENCES

[1] http://www.explainthatstuff.com/cloud-computing-introduction.html. [Accessed (July 3, 2014)], "Woodford, Chris. (2009) Cloud computing."
[2] Mell, Peter and Grance, Tim, "The NIST definition of cloud computing", National Institute of Standards and Technology, 2009,vol 53, p.50, Mell2009.
[3] http://reliscore.com/blog/cloud-computing-the-very-basics, "Cloud Computing - The Very Basics."
[4] http://www.saga.rs/en/products-solutions/virtualization/key-characteristics-of-cloud-computing.html, "Key Characteristics of Cloud Computing."
[5] http://www.zdnet.com/news/the-five-defining-characteristics-of-cloud-computing/287001, "The five defining characteristics of cloud computing."
[6] Siddharth Sonawane, Prathmesh Arnikar, Ankita Fale, Sagar Aghav, Shikha Pachouly, "Load Balancing in Cloud Computing", JOURNAL OF INFORMATION, KNOWLEDGE AND RESEARCH IN COMPUTER ENGINEERING, vol 3(1), pp.573-576.
[7] Anthony T.Velte, Robert Elsenpeter ,Toby J.Velte, , A Book: "Cloud Computing A Practical Approach", TATA McGraw-HILL Edition 2010.
[8] http://www.qualitytesting.info/group/cloudcomputing/forum/topics/software-as-a-s, "Software as a Service (SAAS)- Quality Testing."
[9] http://www.wikinvest.com/concept/Software_as_a_Service, "Software_as_a_Service."
[10] http://www.gartner.com/technology/summits/na/applications/, "Gartner Application Architecture, Development & Integration Summit 8 - 10 December 2014 | Las Vegas, NV."
[11] http://www.davidakka.com/uncategorized/cloud-models-what%E2%80%99s-right-for-you/, "Cloud models-what's right for you."
[12] Marinos, Alexandros, and Gerard Briscoe, "Community cloud computing."Cloud Computing. Springer Berlin Heidelberg, 2009, pp.472-484.
[13] http://searchcloudcomputing.techtarget.com/definition/hybrid-cloud, "hybrid cloud."
[14] Hashizume, et al, "An analysis of security issues for cloud computing" Journal of Internet Services and Applications 2013 pp.4-5.
[15] http://www.rdanet.com/solutions/cloud-and-virtualization, "Cloud Hosting and Virtualization Solutions."
[16] Ali M. Alakeel, "A Guide to Dynamic Load Balancing in Distributed Computer Systems", International Journal of Computer Science and Network Security, vol 10 (6), pp. 153-160, June 2010.
[17] AA. Rajguru , S.S. Apte, "A Comparative Performance Analysis of Load Balancing Algorithms in Distributed System using Qualitative Parameters". International Journal of Recent Technology and Engineering (IJRTE) vol 1(3), August 2012.
[18] Tin-Yu Wu, et al, "Dynamic load balancing mechanism based on cloud storage" Computing, Communications and Applications Conference (ComComAp), 2012, pp. 102-106.
[19] International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (Cyber C), IEEE, pp.447-454, October 2011
[20] Al-Jaroodi, N.Mohamed, "DDFTP: Dual-Direction FTP," CCGRID '11 Proceedings of the 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp.504-513.
[21] Anandharajan TRV, Dr. M.A. Bhagyaveni, "Co-operative Energy Aware Load-Balancing technique for an Efficient Computational Cloud", IJCSI International Journal of Computer Science Issues, vol. 8(2), pp.571-576, March 2011.
[22] M. Randles, D. Lamb, A. Taleb-Bendiab, "A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing",IEEE International Conference on Advanced Information Networking and Applications Workshops,April 2010, pp.551-556.
[23] Stanojevic, R.,Shorten, R., "Load Balancing vs. Distributed Rate Limiting: An Unifying Framework for Cloud Control," *Communications, 2009. ICC '09. IEEE International Conference on* , vol., no., pp.1,6, 14-18 June 2009.
[24] Z.Zhang, X.Zhang, "A load balancing mechanism based on ant colony and complex network theory in open cloud computing federation," *Industrial Mechatronics and Automation (ICIMA), 2010 2nd International Conference on* , vol.2, no., pp.240,243, 30-31 May 2010.

[25] Stanojevic R. and Shorten R., IEEE ICC, pp.1-6, 2009.
[26] T. Kokilavani and George D I Amalarethinam, "Load Balanced Min-Min Algorithm for Static Meta-Task Scheduling in Grid Computing" International Journal of Computer Applications (0975 – 8887) Volume 20,No.2, April 2011.
[27] S. Wang, K. Yan, W. Liao, and S. Wang, "Towards a Load Balancing in a Three-level Cloud Computing Network", *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, vol 1, pp.108-113.
[28] Priya, S. Mohana, and B. Subramani. "A New Approach For Load Balancing In Cloud Computing." *International Journal Of Engineering And Computer Science (IJECS-2013) Vol* 2 (2013), pp.1636-1640.
[29] Jeffrey M. Galloway, Karl L.Smith, Susan S. Vrbsky, "Power Aware Load Balancing for Cloud Computing" ,Proceedings of the World Congress on Engineering and Computer Science 2011,vol 1,WCECS 2011.
[30] Lee, R. and B. Jeng, "Load-balancing tactics in cloud." *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2011 International Conference on , pp.447-454* .
[31] Sotomayor, B., RS. Montero IM. Llorente, and I. Foster, "Virtual infrastructure management in private and hybrid clouds," *Internet Computing, IEEE* , vol.13, no.5, pp.14,22, Sept.-Oct. 2009.
[32] J Laha, R Satpathy, Kaustuva Dev, "Load Balancing Techniques :Major challenges in Cloud Computing-A Systematic Review", International Journal of Computer Science and Network, vol 3(1),February 2014, pp.1-8
[33] A. Singh, M. Korupolu, and D. Mohapatra, "Server-storage virtualization: integration and load balancing in data centers", Proceedings of the ACM/IEEE conference on Supercomputing (SC), Article 53 ,12 pages November 2008.