Survey paper on various load balancing algorithms in cloud computing

Vinay Darji, Jayna Shah, Rutvik Mehta

Abstract— Cloud computing is emerging trend in Information Technology community. Cloud resources are delivered to cloud users based on the requirements. Because of the services provided by cloud, it is becoming more popular among internet users. Hence, the number of cloud users is increasing day by day. Because of this, load on the cloud server needs to be managed for optimum resource utilization. This paper focuses on various load balancing techniques with different parameters. This study focuses on parameters like throughput, overhead, response time, resource utilization and performance. This paper concludes that how these parameters can be used to meet service level agreements.

Index Terms— Cloud computing, load balancing, min-min algorithm, resource utilization

1 INTRODUCTION

Cloud computing is a concept used to describe a variety of computing concepts that involve a large number of computers connected through a real-time communication network such as the Internet. In science, cloud computing is a synonym for distributed computing over a network, and means the ability to run a program or application on many connected computers at the same time. The phrase also more commonly refers to network-based services, which appear to be provided by real server hardware, and are in fact served up by virtual hardware, simulated by software running on one or more real machines. Such virtual servers do not physically exist and can therefore be moved around and scaled up (or down) on the fly without affecting the end user.[1]

The number of cloud users is increasing day by day. Because of this, there are several issues that need to be solved. Some of them are as follows:

Privacy

The increased use of cloud computing services such as Gmail and Google Docs has pressed the issue of privacy concerns of cloud computing services to the utmost importance. The provider of such services lie in a position such that with the greater use of cloud computing services has given access to a plethora of data. This access has the immense risk of data being disclosed either accidentally or deliberately. Privacy advocates have criticized the cloud model for giving hosting companies' greater ease to control—and thus, to monitor at will—communication between host company and end user, and access user data (with or without permission). [2]

Legal

As with other changes in the landscape of computing, certain legal issues arise with cloud computing, including trademark infringement, security concerns and sharing of proprietary data resources. [3]

The Electronic Frontier Foundation has criticized the United States government during the Megaupload seizure process for considering that people lose property rights by storing data on a cloud computing service.

These legal issues are not confined to the time period in which the cloud based application is actively being used. There must also be consideration for what happens when the provider-customer relationship ends. In most cases, this event will be addressed before an application is deployed to the cloud. However, in the case of provider insolvencies or bankruptcy the state of the data may become blurred. [4]

Vendor lock-in

Because cloud computing is still relatively new, standards are still being developed. Many cloud platforms and services are proprietary, meaning that they are built on the specific standards, tools and protocols developed by a particular vendor for its particular cloud offering. This can make migrating off a proprietary cloud platform prohibitively complicated and expensive.

Three types of vendor lock-in can occur with cloud computing:

Platform lock-in: cloud services tend to be built on one of several possible virtualization platforms, for example VMWare. Migrating from a cloud provider using one platform to a cloud provider using a different platform could be very complicated.

Data lock-in: since the cloud is still new, standards of ownership, i.e. who actually owns the data once it lives on a cloud platform, are not yet developed, which could make it complicated if cloud computing users ever decide to move data off of a cloud vendor's platform.

Tools lock-in: if tools built to manage a cloud environment are not compatible with different kinds of both virtual and physical infrastructure, those tools will only be able to manage data or apps that live in the vendor's particular cloud environment.

Open standards

Most cloud vendors provide APIs to use their services. These APIs are typically well-documented (often under a Creative Commons license) but also unique to their implementation and hence they are not interoperable. Some vendors have adopted others' APIs and there are a number of open standards under development, with a view to delivering interoperability and portability.

Security

- *Vinay Darji is currently pursuing masters degree program in computer engineering in SVIT, Vasad,Gujarat,India PH-01123456789. E-mail:vinaydarji049@gmail.com*

As cloud computing is achieving increased popularity, concerns are being voiced about the security issues introduced through adoption of this new model. The effectiveness and efficiency of traditional protection mechanisms are being reconsidered as the characteristics of this innovative deployment model can differ widely from those of traditional architectures. An alternative perspective on the topic of cloud security is that this is but another, although quite broad, case of "applied security" and that similar security principles that apply in shared multi-user mainframe security models apply with cloud security. [5]

The relative security of cloud computing services is a contentious issue that may be delaying its adoption. Physical control of the Private Cloud equipment is more secure than having the equipment off site and under someone else's control. Physical control and the ability to visually inspect data links and access ports is required in order to ensure data links are not compromised. Issues barring the adoption of cloud computing are due in large part to the private and public sectors' unease surrounding the external management of security-based services. It is the very nature of cloud computing-based services, private or public, that promote external management of provided services. This delivers great incentive to cloud computing service providers to prioritize building and maintaining strong management of secure services. Security issues have been categorized into sensitive data access, data segregation, privacy, bug exploitation, recovery, accountability, malicious insiders, management console security, account control, and multi-tenancy issues. Solutions to various cloud security issues vary, from cryptography, particularly public key infrastructure (PKI), to use of multiple cloud providers, standardization of APIs, and improving virtual machine support and legal support.

Cloud computing offers many benefits, but is vulnerable to threats. As cloud computing uses increase, it is likely that more criminals find new ways to exploit system vulnerabilities. Many underlying challenges and risks in cloud computing increase the threat of data compromise. To mitigate the threat, cloud computing stakeholders should invest heavily in risk assessment to ensure that the system encrypts to protect data, establishes trusted foundation to secure the platform and infrastructure, and builds higher assurance into auditing to strengthen compliance. Security concerns must be addressed to maintain trust in cloud computing technology.[6]

Data breach is a big concern in cloud computing. A compromised server could significantly harm the users as well as cloud providers. A variety of information could be stolen. These include credit card and social security numbers, addresses, and personal messages.

IT governance

The introduction of cloud computing requires an appropriate IT governance model to ensure a secured computing environment and to comply with all relevant organizational information technology policies. As such, organizations need a set of capabilities that are essential when effectively implementing and managing cloud services, including demand management, relationship management, data security management, application lifecycle management, risk and compliance management. A danger lies with the explosion of companies joining the growth in cloud computing by becoming providers. However, many of the infrastructural and logistical concerns regarding the operation of cloud computing businesses are still unknown. This over-saturation may have ramifications for the industry as whole.

Consumer end storage

Storage-as-a-service provided by cloud service provider reduces the cost required for purchasing high storage capacity disks and magnetic tapes. cheaper low storage devices are becoming more popular that streams the data from cloud storage devices. Unregulated usage is beneficial for IT and tech moguls like Amazon, the anonymous nature of the cost of consumption of cloud usage makes it difficult for business to evaluate and incorporate it into their business plans.

Performance interference and noisy neighbors

Due to its multi-tenant nature and resource sharing, Cloud computing must also deal with the "noisy neighbor" effect. This effect in essence indicates that in a shared infrastructure, the activity of a virtual machine on a neighboring core on the same physical host may lead to increased performance degradation of the VMs in the same physical host, due to issues such as e.g. cache contamination. Due to the fact that the neighboring VMs may be activated or deactivated at arbitrary times, the result is an increased variation in the actual performance of Cloud resources. This effect is dependent on the nature of the applications that are running inside the VMs. Other factors such as scheduling parameters and the careful selection can be used for optimized assignment and hence reduce the effect of this phenomena. This has great impact on choice of cloud providers on cost and performance using traditional benchmarks for service and application performance.

Load balancing

One of the major issue is load balancing. Load balancing [7] is a computer networking method for distributing workloads across multiple computing resources, such as computers, a computer cluster, network links, central processing units or disk drives. Load balancing aims to optimize resource use, maximize throughput, minimize response time, and avoid overload of any one of the resources. Using multiple components with load balancing instead of a single component may increase reliability through redundancy.

2 LOAD BALANCING TYPES

Load balancing can be defined as a process of assigning and reassigning the overall load of the system to the individual nodes of the system for effective resource utilization. There are many different kinds of load balancing algorithms available for cloud computing system, which can be categorized mainly into two groups:

1. **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 [8].

2. Dynamic algorithms

In dynamic algorithms current state of the system is responsible for the decisions that need to be taken for load balancing. No prior knowledge is needed for load balancing [8]. 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 [9].

- **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.[9]

3 CLOUD COMPUTING ARCHITECTURE

This section describes the architectural, business and various operation models of cloud computing.

3.1 A layered model of cloud computing

Generally speaking, the architecture of a cloud computing environment [10] can be divided into 4 layers: the hardware/ datacenter layer, the infrastructure layer, the platform layer and the application layer, as shown in Fig.1. [10]. We describe each of them in detail:

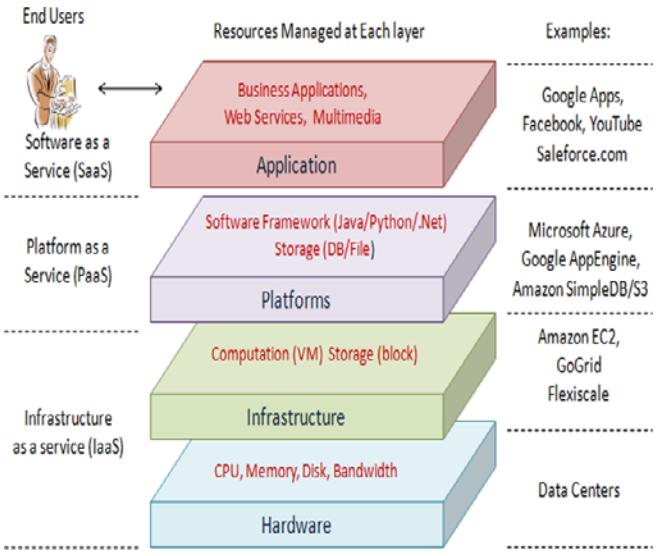


Fig. 1 Cloud computing architecture [10]

The hardware layer: This layer is responsible for managing the physical resources of the cloud, including physical servers, routers, switches, power and cooling systems. In practice, the hardware layer is typically implemented in data centers. A data center usually contains thousands of servers that are organized in racks and interconnected through switches, routers or other fabrics. Typical issues at hardware layer include hardware configuration, fault tolerance, traffic management, power and cooling resource management.

The infrastructure layer: Also known as the virtualization layer, the infrastructure layer creates a pool of storage and computing resources by partitioning the physical resources using virtualization technologies such as Xen, KVM and VMware. The infrastructure layer is an essential component of cloud computing, since many key features, such as dynamic resource assignment, are only made available through virtualization technologies.

The platform layer: Built on top of the infrastructure layer, the platform layer consists of operating systems and application frameworks. The purpose of the platform layer is to minimize the burden of deploying applications directly into VM containers. For example, Google App Engine operates at the platform layer to provide API support for implementing storage, database and business logic of typical web applications.

The application layer: At the highest level of the hierarchy, the application layer consists of the actual cloud applications. Different from traditional applications, cloud applications can leverage the automatic-scaling feature to achieve better performance, availability and lower operating cost. Compared to traditional service hosting environments such as dedicated server farms, the architecture of cloud computing is more modular. Each layer is loosely coupled with the layers above and below, allowing each layer to evolve separately. This is similar to the design of the OSI model for network protocols. The architectural modularity allows cloud computing to support a wide range of application requirements while reducing management and maintenance overhead.

4 LAYERED ARCHITECTURE OF LOAD BALANCING IN CLOUD COMPUTING

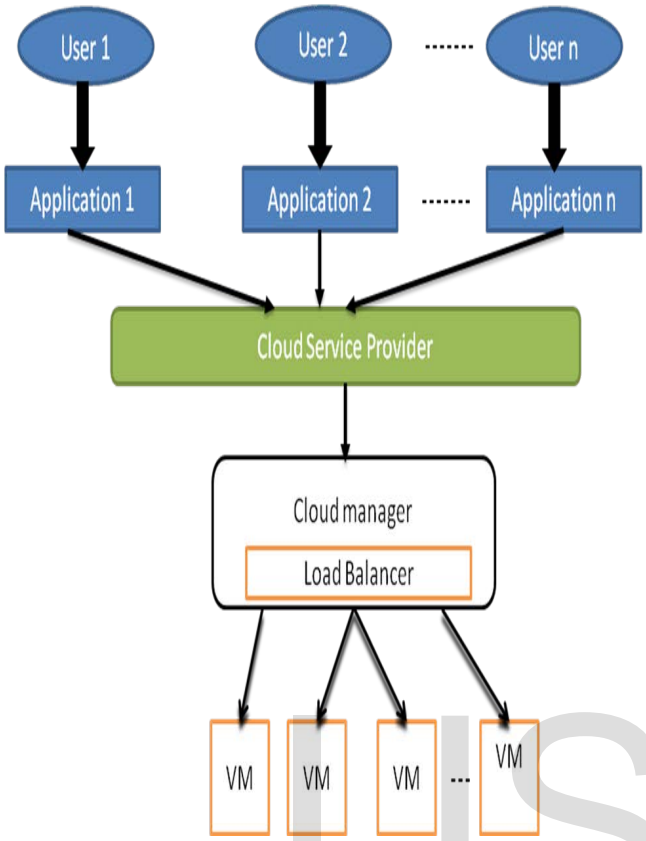


Fig. 2 Cloud Architecture for Load Balancing

In this architecture, various users submit their diverse applications to the cloud service provider through a communication channel. The Cloud Manager in the cloud service provider's datacenter is the prime entity to distribute the execution load among all the VMs by keeping track of the status of the VM. The Cloud Manager maintains a data structure containing the VM ID, Job ID of the jobs that has to be allocated to the corresponding VM and VM Status to keep track of the load distribution. The VM Status represents the percentage of utilization. The Cloud Manager allocates the resources and distributes the load as per the data structure. The Cloud Manager analyzes the VM status routinely to distribute the execution load evenly. In course of processing, if any VM is overloaded then the jobs are migrated to the VM which is underutilized by tracking the data structure. If there are more than one available VM then the assignment is based on the least hop time. On completion of the execution, the Cloud Manager automatically updates the data structure.[11]

Metrics For Load Balancing In Clouds

Various metrics considered in existing load balancing techniques in cloud computing are discussed below:

- **Throughput** is used to calculate the no. of tasks whose execution has been completed. It should be high to improve the performance of the system.[12]
- **Response Time** is the amount of time taken to respond by a particular load balancing algorithm in a distributed system. This parameter should be minimized.[12]
- **Resource Utilization** is used to check the utilization of re-sources. It should be optimized for an efficient load balancing.[13]
- **Performance** is used to check the efficiency of the system. This has to be improved at a reasonable cost, e.g., reduce task response time while keeping acceptable delays.[12]
- **Scalability** is the ability of an algorithm to perform load balancing for a system with any finite number of nodes. This metric should be improved.[12]
- **Overhead** Associated determines the amount of overhead involved while implementing a load-balancing algorithm. It is composed of overhead due to movement of tasks, inter-processor and inter-process communication. This should be minimized so that a load balancing technique can work efficiently.[12]

5. LOAD BALANCING TECHNIQUES

a) A scheduling strategy on load balancing of virtual machine resources

The procedure overview: According to historical data and current state of the system and through genetic algorithm, this strategy computes ahead the influence it will have on the system after the deployment of the needed VM resources and then chooses the least-affective solution, through which it achieves the best load balancing and reduces or avoids dynamic migration. This strategy solves the problem of load imbalance and high migration cost by traditional algorithms after scheduling. Experimental results prove that this method is able to realize load balancing and reasonable resources utilization both when system load is stable and variant.[14]

b) 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 algorithm, 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 as-

signed to a weaker one. The major issue in this allocation is this that it does not consider the advanced load balancing requirements such as processing times for each individual requests.[15]

c) Throttled algorithm

In this algorithm the client first requests the load balancer to find a suitable Virtual Machine to perform the required operation. The process first starts by maintaining a list of all the VMs each row is individually indexed to speed up the lookup process. If a match is found on the basis of size and availability of the machine, then the load balancer accepts the request of the client and allocates that VM to the client. If, however there is no VM available that matches the criteria then the load balancer returns -1 and the request is queued.[15]

d) Current execution spread spectrum

It is spread spectrum technique in which the load balancer spread the load of the job in hand into multiple virtual machines. The load balancer maintains a queue of the jobs that need to use and are currently using the services of the virtual machine. The balancer then continuously scans this queue and the list of virtual machines. If there is a VM available that can handle request of the node/client, the VM is allocated to that request. If however there is a VM that is free and there is another VM that needs to be freed of the load, then the balancer distributes some of the tasks of that VM to the free one so as to reduce the overhead of the former VM. Figure2. better explains the working of the ESCE algorithm. The jobs are submitted to the VM manager, the load also maintains a list of the jobs, their size and the resources requested. The balancer selects the job that matches the criteria for execution at the present time. Though there algorithm offers better results as shown in further section, it however requires a lot of computational overhead.[15]

e) CARTON

CARTON[14][16] is used for cloud control that unifies the use of LB and DRL. LB (Load Balancing) is used to equally distribute the jobs to different servers so that the associated costs can be minimized and DRL (Distributed Rate Limiting) is used to make sure that the resources are distributed in a way to keep a fair resource allocation. DRL also adapts to server capacities for the dynamic workloads so that performance levels at all servers are equal. With very low computation and communication overhead, this algorithm is simple and easy to implement.

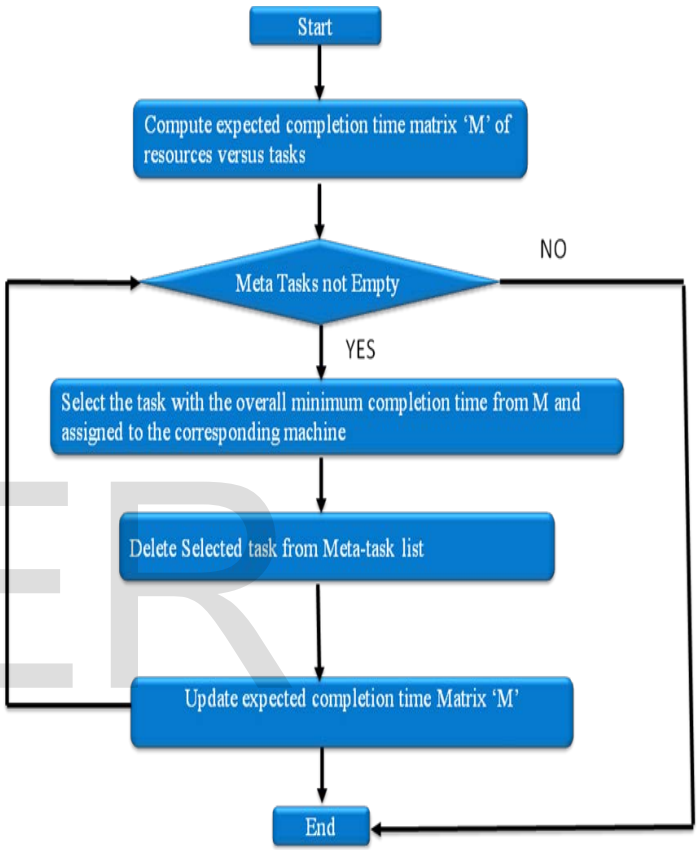
f) CLBVM

Central Load Balancing Policy[14][17] is used for Virtual Machines (CLBVM) that balances the load evenly in a distributed virtual machine/cloud computing environment. This policy improves the overall performance of the system but does not consider the systems that are fault-tolerant.

g) Task Scheduling based on LB

It is a two-level task scheduling mechanism based on load balancing to meet dynamic requirements of users and obtain high resource utilization. It achieves load balancing by first map-ping tasks to virtual machines and then virtual machines to host resources thereby improving the task response time, resource utilization and overall performance of the cloud computing environment.[14][18]

h) min-min algorithm



The Min-min heuristic begins with the set U of all unmapped tasks. Then, the set of minimum completion times, M, for each ti U, is found. Next, the task with the overall minimum completion time from M is selected and assigned to the corresponding machine (hence the name Min-min). Last, the newly mapped task is removed from U, and the process repeats until all tasks are mapped (i.e., U is empty). Min-min is based on the minimum completion time, as is MCT. However, Min-min considers all unmapped tasks during each mapping decision and MCT only considers one task at a time. Min-min maps the tasks in the order that changes the machine availability status by the least amount that any assignment could. Let ti be the first task mapped by Min-min onto an empty system. The machine that finishes ti the earliest, say mj, is also the machine that executes ti the fastest. For every task that Min-min maps after ti, the Min-min heuristic changes the availability status of mj by the

least possible amount for every assignment. Therefore, the percentage of tasks assigned to their first choice (on the basis of execution time) is likely to be higher for Min-min than for Max-min. The expectation is that a smaller makespan can be obtained if more tasks are assigned to the machines that complete them the earliest and also execute them the fastest.

6. COMPARISON OF LOAD BALANCING TECHNIQUES

Metrics\Techniques	a	b	c	d	e	f	g	H
Throughput	Yes	No	No	Yes	No	No	Yes	Yes
Response Time	No	No	Yes	Yes	No	No	Yes	Yes
Resource Utilization	Yes	No	Yes	Yes	No	No	Yes	Yes
Performance	No	No	No	Yes	No	Yes	Yes	Yes
Scalability	No	Yes	No	No	Yes	Yes	No	Yes
Overhead	Yes	No	Yes	Yes	No	No	Yes	Yes

7. CONCLUSION

Cloud computing is emerging trend in Information Technology community. Cloud resources are delivered to cloud users based on the requirements. Because of the services provided by cloud, it is becoming more popular among internet users. Hence, the number of cloud users is increasing day by day. Because of this, issues of cloud computing are needed to be solved and load balancing is one of the important issue among them. In this paper, we examined several load balancing techniques considering number of parameters. We conclude that min-min algorithm considers parameters like makespan which makes it a very good technique for load balancing on cloud and is the best among the techniques.

8. REFERENCES

[1] Martin Randles, David Lamb, A. Taleb-Bendiab, "A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing," IEEE 24th International Conference on Advanced Information Networking and Applications Workshops, pp551-556, 2010

[2] J. F. Yang and Z. B. Chen, "Cloud Computing Research and Security Issues," International Conference on Com- putational Intelligence and Software Engineering (CiSE), Wuhan, 10-12 December 2010, pp. 1-3.

[3] W. Kim, S. D. Kim, E. Lee and S. Y. Lee, "Adoption Issues for Cloud Computing," Proceedings of the 11th International Conference on Information Integration and Web-Based Applications & Services (iiWAS'09), 14-16 December 2009, Kuala, pp. 3-6.

[4] Z. Mahmood, "Data Location and Security Issues in Cloud Computing," International Conference on Emer- ging Intelligent Data and Web Technologies (EIDWT), Tirana, 7-9 September 2011, pp. 49-54.

[5] Kuyoro S. O., Ibikunle F., Awodele O., Cloud Computing Security Issues and Challenges, International Journal of Computer Networks(IJCN), Volume (3) : Issue (5) : 2011

[6] J. C. Roberts and W. Al-Hamdani, "Who Can You Trust in the Cloud? A Review of Security Issues within Cloud Computing," Proceedings of the 2011 Information Secu-rity Curriculum Development Conference, Kennesaw, 7-9 October 2011, pp. 15-19.

[7] K. Ramana, A. Subramanyam and A. Ananda Rao, Comparative Analysis of Distributed Web Server System Load Balancing Algorithms Using Qualitative Parameters, VSRD-IJCSIT, Vol. 1 (8), 2011, 592-600

[8] Chaczko, Z., Mahadevan, V., Aslanzadeh, S., & Mcdermid, C., "Availability of Load Balancing in Cloud Computing", International Conference on Computer and Software Modeling, 2011.

[9] A.Khiyaita, M.Zbakh, H. El Bakkali, Dafir El Kettani, "Load Balancing Cloud Computing : State of Art", IEEE, 2012.

[10] Qi Zhang, Lu Cheng , Raouf Boutaba, " Cloud computing: state-of-the-art and research challenges", Springer, 20 April 2010

[11] Rashmi. K. S., Suma. V, Vaidehi. M, "Enhanced Load Balancing Approach to Avoid Deadlocks in Cloud", International Journal of Computer Applications, June 2012.

[12] Nidhi Jain Kansal, Inderveer Chana, " Cloud Load Balancing Techniques : A Step Towards Green Computing", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012

[13] Nidhi Jain Kansal, Inderveer Chana, "Existing load balancing techniques in cloud computing: a systematic review", Journal of Information Systems and Communication, February 2012

[14] Jinhua Hu, Jianhua Gu, Guofei Sun, Tianhai Zhao, "A Scheduling Strategy on Load Balancing of Virtual Machine sourcein Cloud Computing Environment", 3rd International symposium on Parallel Architectures, Algorithms and Programming, IEEE 2010.

[15] Tanveer Ahmed, Yogendra Singh, "Analytic Study Of Load Balancing Techniques Using Tool Cloud Analyst", International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 2, Mar-Apr 2012.

[16] Stanojevic R. and Shorten R. (2009) IEEE ICC, 1-6.

[17] Bhadani A. and Chaudhary S. (2010) 3rd Annual ACM Banga-lore Conference.

[18] Fang Y., Wang F. and Ge J. (2010) Lecture Notes in Comput-er Science, 6318, 271-277.