ISSN: 2277-3061

Volume: 17 Issue: 1

International Journal of Computers & Technology

LOAD BALANCING IN CLOUD ENVIRONMENT: A REVIEW

Sandeep Kaur [(1)], Parminder Pal Kaur[(2)]

[(1)] Research Scholar, Department of Computer Engineering, BMS College of Engineering, Shri Muktsar Sahib
sandeepk257@gmail.com

[(2)] Assistant Professor, Department of Computer Science & Engineering, Universal Group of Institutes, Lalru
parminder_741@yahoo.com

ABSTRACT

Cloud computing is the means of accessing a shared pool of configurable computing resources (including hardware, software, networks, servers, storage applications and services) that can be rapidly provided, used and released with minimal effort on the part of users or service providers. But it has some of the main concerns like load management and fault tolerance. In this paper we are discussing load balancing approach in cloud computing. Load balancing is helped to distribute the 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 This paper focuses on the load balancing algorithm which distributes the incoming jobs among VMs optimally in cloud data centers. In this paper, we have reviewed several existing load balancing mechanisms and we have tried to address the problems associated with them.

KEYWORDS

Cloud computing, Load balancing, Virtual machine, Host, Datacenter, Datacenter Broker

INTRODUCTION

Cloud is a computing framework which usually denotes storing and accessing data and programs over the internet, instead of your computer's hard drive. Cloud Computing is handled by means of the great prospective paradigm utilized for placement of applications taking place over Internet. This perception also elucidates about the applications which are widen on the way to be manageable over and done with the Internet. Cloud applications utilize huge information centers as well as operational servers which are utilized to host net applications plus services [1]. Cloud computing is a service that is distributed over the internet for data access, computing and cloud storage by creating scalability, elasticity and low cost. Second invention platform for division which suggests [1] a variety of services and applications to the user actually attain them. A cloud computing is one of the rising information technologies used in computation now days. It is green technologies which agree to accessing, computing and storing the assets by offering a variety of services. Cloud computing normally includes models like Infrastructure-as-a-service [1], Platform-as-a-service and Platform-as-a-service. To reduce the computation time and to conquer the storage space issues, most of the organization now a day's make regular use of cloud computing from the established process of calculation. It mainly focuses on allocating data and computations over a scalable information centers of network.Cloud computing [1] is an emerging paradigm in the computer industry where the computing is moved to a cloud of computers. It has become one of the buzz words of the industry. The core concept of cloud computing is, quite simply, that the vast computing resources that we need will reside somewhere out there in the cloud of computers and we'll connect to them and use them as and when needed. Computing can be described as any activity of using and/or developing computer hardware and software. It includes everything that sits in the bottom layer, i.e. everything from raw compute power to storage capabilities. Cloud computing [1] ties together all these entities and delivers them as a single integrated entity under its own sophisticated management.

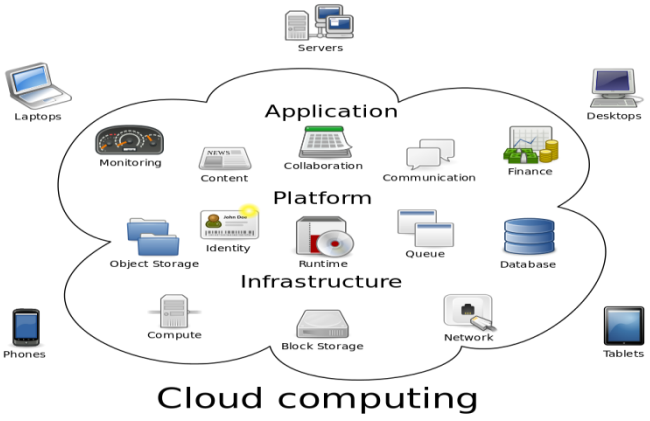

Figure 1. Cloud Computing Model

Cloud is a term used as a metaphor for the wide area networks (like internet) or any such large networked environment. It came partly from the cloud-like symbol used to represent the complexities of the networks in the schematic diagrams. It represents all the complexities of the network which may include everything from cables, routers, servers, data centers

ISSN: 2277-3061

Volume: 17 Issue: 1

International Journal of Computers & Technology

and all such other devices. Computing started off with the mainframe era. There were big mainframes and everyone connected to them via “dumb” terminals. This old model of business computing was frustrating for the people sitting at the dumb terminals because they could do only what they were “authorized” to do. They were dependent on the computer administrators to give them permission or to fix their problems. They had no way of staying up to the latest innovations. The personal computer was a rebellion against the tyranny of centralized computing [4] operations. There was a kind of freedom in the use of personal computers. But this was later replaced by server architectures with enterprise servers and others showing up in the industry. This made sure that the computing was done and it did not eat up any of the resources that one had with him. All the computing was performed at servers. Internet grew in the lap of these servers. With cloud computing we have come a full circle. We come back to the centralized computing infrastructure. But this time it is something which can easily be accessed via the internet and something over which we have all the control. Cloud computing is Internet ("cloud") based development and use of computer technology ("computing"). It is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet. Users need not have knowledge of, expertise in, or control over the technology infrastructure "in the cloud" that supports them.

KEY CHARACTERISTICS

1. On-demand self-service: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service’s provider.

2. Broad network access: Cloud computing provide the users with various capabilities over the network which are accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops etc.)

3. Resource pooling: The provider’s computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. Examples of resources include storage, processing, memory, network bandwidth, and virtual machines

4. Rapid elasticity: Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out, and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.

5. Measured Service: Cloud systems automatically control and optimize resource use by leveraging a metering capability1 at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

TYPES OF CLOUDS

Clouds are divided into 4 categories: -

1. Public cloud computing: It mainly depends on third individual to suggest services by paying them on regular basis according to the procedure. Public Cloud environment is made available to all unrestricted consumers who can subscribe the needed services [3].

2.Private Cloud Computing: The organization itself regulates the services. Usually administrations go for private cloud only in the case of involvement of sensible information. Scaling can be done very professionally by adding hardware and thus the environment can be expanded. The security will be more due to the control of internal structure contained in it and therefore data will be secured.

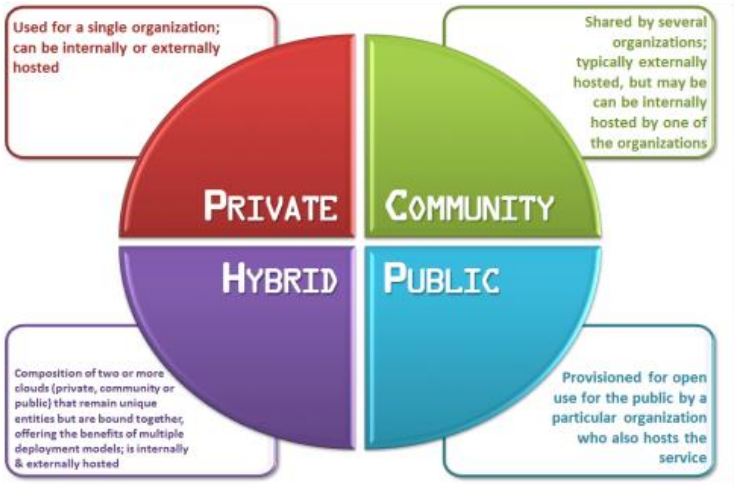

Figure 2. Types of Cloud

ISSN: 2277-3061

Volume: 17 Issue: 1

International Journal of Computers & Technology

3.Hybrid Cloud Computing: It is the mixture of both public & private cloud computing. A less sensible data will be stored in public and all others in Private Cloud.

4) Community Cloud: -When cloud infrastructure construct by many organizations jointly, such cloud model is called as a community cloud. The cloud infrastructure could be hosted by a third-party provider or within one of the organizations in the community

SERVICES OF CLOUD MODEL

There are different types of services are providing by cloud models like: Software as a Service(SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) [6] which are deployed as public cloud, private cloud, community cloud and hybrid clouds.

i.Platform-as-a-service: The platform as services refers to the division on pay to alter the direct readying of application level assets or internet application, software deployment framework and runtime atmosphere. It is a platform wherever package will be developed, deployed and tested. They are resources of entire life cycle where software can be operated on a PAAS. For example: Microsoft Azure, Google app engine, Amazon EC2, IBM Smart cloud etc. In other words, the platform as a service helps the customer to create their own applications. The cloud application supports a set of applications program interface. It acts as an interface between application and hardware.

ii.Infrastructure as a service: The infrastructure as a service is also called hardware as service. They deliver computing capabilities as consistent services and basic storage over the network. Pooled and made available to holder workloads are services, storage schemes, networking tools, data center space etc. Example: storage services provided by Amazon S3, Amazon EBS etc, where users can consume the property and virtual desktop like network, virtualized services, routers and storage etc. are main examples of this model.

iii.Software-as-a-Service: The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through a thin client interface such as a web browser (e.g., web-based email). The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

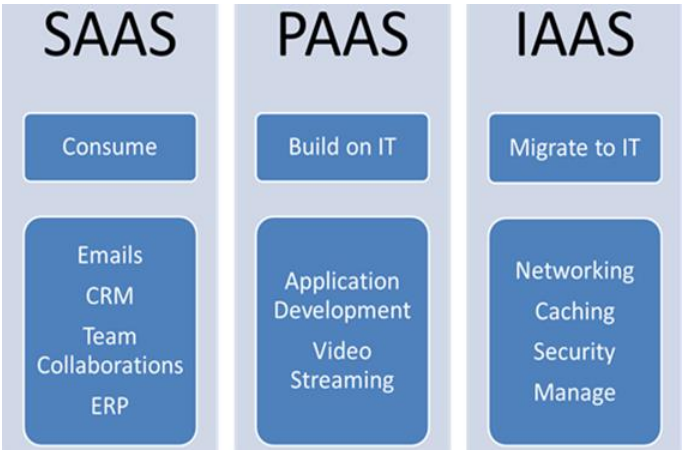

Figure 3. Models of Cloud

LOAD BALANCING

Load balancing is a computer network method for distributing workloads across multiple computing resources, so that time efficiency can be increased and also proper utilization of resources takes place while also avoiding a situation where some of the nodes are heavily loaded while other nodes are idle or doing very little work. Load balancing techniques in clouds, consider various parameters such as performance, response time, scalability, throughput, resource utilization, fault tolerance, and associated overhead.One of the foremost usually used applications of load balancing is to produce quality of service from multiple servers, typically called a server data center. Usually load-balanced systems are properly working inside popular internet sites, big chat networks, high-bandwidth file transfer protocol sites, and domain name System (DNS) servers. It additionally prevents the clients from contacting back-end servers directly, which can have security advantages by hiding the structure of the inner network. Some load balancers give a mechanism for improving the one parameter specially within back end server Load balancing offers the IT team an opportunity to attain a considerably higher fault tolerance. It will mechanically give the capability required to handle any increase or decrease of application traffic. It is additionally necessary that the load balancer itself doesn't become the cause of failure. Sometimes load balancers enforced in high-availability servers can additionally replicate the user's session needed by the application. Load balancing is dividing work load between a set of computers in order to receive the good response time and all the nodes are equally loaded and, in general, all users get served quicker.

GOALS OF LOAD BALANCING

The goals of load balancing are:

• To improve the performance of the system.

• To have a backup of the load or entire server just in case the system fails or even partly fails.

• To maintain the system stability

• To accommodate future modification within the system

LOAD BALANCING CLASSIFICATION

This is chiefly divided into 2 categories: static load balancing mechanism and dynamic load balancing mechanism:

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 preemptive 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 underutilized machine dynamically for faster execution. This means that it allows for process preemption 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.

A) Centralized approach: - In centralized approach, solely one node is liable for managing and distribution among the complete cloud system model. Alternative all nodes aren't liable for handling the requests and providing the response.

b) Distributed approach: - In distributed approach, every node severally builds its own load vector. The work is divided among all the nodes of the server. They aggregate the load information of alternative nodes. Distributed approach is additional appropriate for complicated and very large systems inside the cloud computing.

RELATED WORK

Nguyen KhacChien et al. (2016) has proposed a load balancing algorithm which is used to enhance the performance of the cloud environment based on the method of estimating the end of service time. They have succeeded in enhancing the service time and response time of the user.

Ankit Kumar et al (2016) focuses on the load balancing algorithm which distributes the incoming jobs among VMs optimally in cloud data centers. The proposed algorithm in this research work has been implemented using Cloud Analyst simulator and the performance of the proposed algorithm is compared with the three algorithms which are preexists on the basis of response time. In the cloud computing milieu, the cloud data centers and the users of the cloud-computing are globally situated, therefore it is a big challenge for cloud data centers to efficiently handle the requests which are coming from millions of users and service them in an efficient manner.

S.Yakhchi et al. (2015) discusses that the energy consumption has become a major challenge in cloud computing infrastructures. They proposed a novel power aware load balancing method, named ICAMMT to manage power consumption in cloud computing data centers. We have exploited the Imperialism Competitive Algorithm (ICA) for detecting over utilized hosts and then we migrate one or several virtual machines of these hosts to the other hosts to decrease their utilization. Finally, we consider other hosts as underutilized host and if it is possible, migrate all of their VMs to the other hosts and switch them to the sleep mode.

Surbhi Kapoor et al. (2015) aims at achieving high user satisfaction by minimizing response time of the tasks and improving resource utilization through even and fair allocation of cloud resources. The traditional Throttled load balancing algorithm is a good approach for load balancing in cloud computing as it distributes the incoming jobs evenly among the VMs. But the major drawback is that this algorithm works well for environments with homogeneous VMS, does not considers the resource specific demands of the tasks and has additional overhead of scanning the entire list of VMs every time a task comes. The issues have been addressed by proposing an algorithm Cluster based load balancing which works well in heterogeneous nodes environment, considers resource specific demands of the tasks and reduces scanning overhead by dividing the machines into clusters.

Shikha Garg et al. (2015) aims to distribute workload among multiple cloud systems or nodes to get better resource utilization. It is the prominent means to achieve efficient resource sharing and utilization. Load balancing has become a challenge issue now in cloud computing systems. To meets the user’s huge number of demands, there is a need of

distributed solution because practically it is not always possible or cost efficient to handle one or more idle services. Servers cannot be assigned to particular clients individually. Cloud Computing comprises of a large network and components that are present throughout a wide area. Hence, there is a need of load balancing on its different servers or virtual machines. They have proposed an algorithm that focuses on load balancing to reduce the situation of overload or under load on virtual machines that leads to improve the performance of cloud substantially.

ReenaPanwar et al. (2015) describes that the cloud computing has become essential buzzword in the Information Technology and is a next stage the evolution of Internet, The Load balancing problem of cloud computing is an important problem and critical component adequate operations in cloud computing system and it can also prevent the rapid development of cloud computing. Many clients from all around the world are demanding the various services rapid rate in the recent time. Although various load balancing algorithms have been designed that are efficient in request allocation by the selection of correct virtual machines. A dynamic load management algorithm has been proposed for distribution of the entire incoming request among the virtual machines effectively.

Mohamed Belkhouraf et al. (2015) aims to deliver different services for users, such as infrastructure, platform or software with a reasonable and more and more decreasing cost for the clients. To achieve those goals, some matters have to be addressed, mainly using the available resources in an effective way in order to improve the overall performance, while taking into consideration the security and the availability sides of the cloud. Hence, one of the most studied aspects by researchers is load balancing in cloud computing especially for the big distributed cloud systems that deal with many clients and big amounts of data and requests. The proposed approach mainly ensures a better overall performance with efficient load balancing, the continuous availability and a security aspect.

Lu Kang et al. (2015) improves the weighted least connections scheduling algorithm, and designs the Adaptive Scheduling Algorithm Based on Minimum Traffic (ASAMT). ASAMT conducts the real-time minimum load scheduling to the node service requests and configures the available idle resources in advance to ensure the service QoS requirements. Being adopted for simulation of the traffic scheduling algorithm, OPNET is applied to the cloud computing architecture.

Hiren H. Bhatt et al. (2015) presents a Flexible load sharing algorithm (FLS) which introduce the third function. The third function makes partition the system in to domain. This function is helpful for the selection of other nodes which are present in the same domain. By applying the flexible load sharing to the particular domains in to the distribute system, the performance can be improved when any node is in overloaded situation.

RESEARCH GAP

Cloud computing thus involving distributed technologies to satisfy a variety of applications and user needs. Sharing resources, software, information via internet are the main functions of cloud computing with an objective to reduced capital and operational cost, better performance in terms of response time and data processing time, maintain the system stability and to accommodate future modification in the system .So there are various technical challenges that needs to be addressed like Virtual machine migration, server consolidation, fault tolerance, high availability and scalability but central issue is the load balancing , it is the mechanism of distributing the load among various nodes of a distributed system to improve both resource utilization and job response time while also avoiding a situation where some of the nodes are heavily loaded while other nodes are idle or doing very little work. It also ensures that all the processor in the system or every node in the network does approximately the equal amount of work at any instant of time. In the NBST algorithm, the jobs are present in the queue and we know the length i.e., number of instructions in request. The load balancing algorithm aims at reducing the load over resources. For achieving this, arrange all the virtual machines in order according to their execution speed that is in MIPS (Million instructions per second). After arrangement of machines, sorting of cloudlets is performed on the basis of their length (million instructions). Mid-point is taken of those sorted cloudlets list and sorted virtual machines list and then the divided cloudlet lists are mapped to the corresponding lists of virtual machines.

To make the final determination, the load balancer retrieves information about the candidate server's health and current workload in order to verify its ability to respond to that request. Load balancing solutions can be divided into software-based load balancers and hardware-based load balancers. Hardware-based load balancers are specialized boxes that include Application Specific Integrated Circuits (ASICs) customized for a specific use. They have the ability to handle the high-speed network traffic whereas Software-based load balancers run on standard operating systems and standard hardware components.The currently proposed work will work effectively and efficiently only in the homogeneous environment where all the machines are of same capacity. But as we know that in the cloud computing model, the configurations of the machines will be different from each other as the different users will have different requirements. The tasks of the user are allocated on the basis of availability of virtual machine. There is no checking of capacity of the virtual machine before allocating the request. There can be a scenario where a machine with high configuration is sitting idle and we have assigned the task to the low configuration machine. This may lead to overutilization and under-utilization of resources. There is no checking of the requirement of the user whether user wants to use the machine of high configuration or low configuration. No fault tolerance mechanism has been proposed in the current work.

CLOUD SIM

Cloud service providers charge users depending upon the space or service provided. In R&D, it is not always possible to have the actual cloud infrastructure for performing experiments. For any research scholar, academician or scientist, it is not feasible to hire cloud services every time and then execute their algorithms or implementations. For the purpose of research, development and testing, open source libraries are available, which give the feel of cloud services. Nowadays,

in the research market, cloud simulators are widely used by research scholars and practitioners, without the need to pay any amount to a cloud service provider.

CONCLUSION

In present days cloud computing is one of the greatest platform which provides storage of data in very lower cost and available for all time over the internet. But it has more critical issue like security, load management and fault tolerance. In this paper we are discussing load balancing approaches. Resource scheduling management design on Cloud computing is an important problem. Scheduling model, cost, quality of service, time, and conditions of the request for access to services are factors to be focused. A good task scheduler should adapt its scheduling strategy to the changing environment and load balancing Cloud task scheduling policy. Cloud Computing is high utility software having the ability to change the IT software industry and making the software even more attractive.

REFERENCES

[1] S. Yakhchi, S. Ghafari, M. Yakhchi, M. Fazeli and A. Patooghy, "ICA-MMT: A Load Balancing Method in Cloud Computing Environment," IEEE, 2015.

[2] S. Kapoor and D. C. Dabas, "Cluster Based Load Balancing in Cloud Computing," IEEE, 2015.

[3] S. Garg, R. Kumar and H. Chauhan, "Efficient Utilization of Virtual Machines in Cloud Computing using Synchronized Throttled Load Balancing," 1st International Conference on Next Generation Computing Technologies (NGCT-2015), pp. 77-80, 2015.

[4] R. Panwar and D. B. Mallick, "Load Balancing in Cloud Computing Using Dynamic Load Management Algorithm," IEEE, pp. 773-778, 2015.

[5] M. Belkhouraf, A. Kartit, H. Ouahmane, H. K. Idrissi, Z. Kartit and M. E. Marraki, "A secured load balancing architecture for cloud computing based on multiple clusters," IEEE, 2015.

[6] L. Kang and X. Ting, "Application of Adaptive Load Balancing Algorithm Based on Minimum Traffic in Cloud Computing Architecture," IEEE, 2015.

[7] N. K. Chien, N. H. Son and H. D. Loc, "Load Balancing Algorithm Based on Estimating Finish Time of Services in Cloud Computing," ICACT, pp. 228-233, 2016.

[8] H. H. Bhatt and H. A. Bheda, "Enhance Load Balancing using Flexible Load Sharing in Cloud Computing," IEEE, pp. 72-76, 2015.

[9] S. S. MOHARANA, R. D. RAMESH and D. POWAR, "ANALYSIS OF LOAD BALANCERS IN CLOUD COMPUTING," International Journal of Computer Science and Engineering (IJCSE) , pp. 102-107, 2013.

[10] M. P. V. Patel, H. D. Patel and . P. J. Patel, "A Survey On Load Balancing In Cloud Computing," International Journal of Engineering Research & Technology (IJERT), pp. 1-5, 2012.

[11] R. Kaur and P. Luthra, "LOAD BALANCING IN CLOUD COMPUTING," Int. J. of Network Security, pp. 1-11, 2013.

[12] Kumar Nishant, , P. Sharma, V. Krishna, Nitin and R. Rastogi, "Load Balancing of Nodes in Cloud Using Ant Colony Optimization," IEEE, pp. 3-9, 2012.

[13] Y. Xu, L. Wu, L. Guo,, Z. Chen, L. Yang and Z. Shi, "An Intelligent Load Balancing Algorithm Towards Efficient Cloud Computing," AI for Data Center Management and Cloud Computing: Papers from the 2011 AAAI Workshop (WS-11-08), pp. 27-32, 2011.

[14] A. K. Sidhu and S. Kinger, "Analysis of Load Balancing Techniques in Cloud Computing," International Journal of Computers & Technology Volume 4 No. 2, March-April, 2013, ISSN 2277-3061, pp. 737-741, 2013.

[15] O. M. Elzeki, M. Z. Reshad and M. A. Elsoud, "Improved Max-Min Algorithm in Cloud Computing," International Journal of Computer Applications (0975 – 8887), pp. 22-27, 2012.

[16] B. Kruekaew and W. Kimpan, "Virtual Machine Scheduling Management on Cloud Computing Using Artificial Bee Colony," Proceedings of the International Multi Conference of Engineers and Computer Scientists 2014 Vol I,IMECS 2014, 2014.

[17] R.-S. Chang, J.-S. Chang and P.-S. Lin, "An ant algorithm for balanced job scheduling in grids," Future Generation Computer Systems 25 (2009) 20–27, pp. 21-27, 2009.

[18] Z. Chaczko, V. Mahadevan, S. Aslanzadeh and C. Mcdermid, "Availability and Load Balancing in Cloud Computing," International Conference on Computer and Software Modeling IPCSIT vol.14 (2011) © (2011) IACSIT Press, Singapore, pp. 134-140, 2011.

[19] R. K. S, S. V and V. M, "Enhanced Load Balancing Approach to Avoid Deadlocks in Cloud," Special Issue of International Journal of Computer Applications (0975 – 8887) on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, June 2012, pp. 31-35, 2012.

[20] Kumar Nishant, P. Sharma, V. Krishna, N. and R. Rastogi, "Load Balancing of Nodes in Cloud Using Ant Colony Optimization," IEEE, pp. 3-9, 2012.

[21] Ankit Kumar, Mala Kalra," Load Balancing in Cloud Data Center Using Modified Active Monitoring Load Balancer", IEEE pp. 1-5, 2016.

[22] Saraswathi AT, Kalaashri.Y.RA, Dr.S. Padmavathi, "Dynamic Resource Allocation Scheme in Cloud Computing", ELSEVIER, pp. 30-36, 2015