ISSN 2277-3061
Volume 16 Number 3
INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

ALLOCATION OF HETEROGENOUS CLOUDLETS ON PRIORITY BASIS IN CLOUD ENVIRONMENT

Sumanpreet Kaur [(1)], Mr. Navtej Singh Ghumman [(2)]
[(1)] Research Scholar, Department of Computer Science & Engineering, SBSSTC, Ferozepur, Punjab.
sumanmanes07@gmail.com
[(2)] Assistant Professor, Department of Computer Science & Engineering, SBSSTC, Ferozepur, Punjab.
navtejghumman@yahoo.com

ABSTRACT

Load balancing is one of the main challenges in cloud computing which is required to distribute the dynamic workload across multiple nodes to ensure that no single node is overwhelmed. It helps in optimal utilization of resources and hence in enhancing the performance of the system. In the natural environment, the cloudlets will be processed in the FIFO (First in First Out approach). We propose an improved load balancing algorithm for job scheduling in the Grid environment. Hence, in this research work, various types of leases have been assigned to the cloudlets like cancellable, suspendable and non-preemtable. The leases have been assigned on the basis of cost assigned to them and the requirement specified by the user. The datacenter broker will receive the list of all the virtual machines and will categorize them into two classes i.e. Class A and Class B. Class A will have high end virtual machines and will process the non-preemptable cloudlets. Class B will contain the low end virtual machines and will process the suspendable and cancellable cloudlets. The machines in each class will be further sorted in descending order according to their MIPS. Multiple parameters have been evaluated like waiting time, turnaround time, execution time and processing cost. Further, this research also provides the anticipated results with the implementation of the proposed algorithm. In the cloud storage, load balancing is a key issue. It would consume a lot of cost to maintain load information, since the system is too huge to timely disperse load. The main contributions of the research work are to balance the entire system load while trying to minimize the make span of a given set of jobs. Compared with the other job scheduling algorithms, the improved load balancing algorithm can outperform them according to the experimental results.

Keywords

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

INTRODUCTION

Cloud Computing is the model for convenient on-demand network access, with minimum management efforts for easy and fast network access to resources that are ready to use. It is an upcoming paradigm that offers tremendous advantages in economic aspects, such as reduced time to market, flexible computing capabilities, and limitless computing power. Popularity of cloud computing is increasing day by day in distributed computing environment. There is a growing trend of using cloud environments for storage and data processing needs. To use the full potential of cloud computing, data is transferred, processed, retrieved and stored by external cloud providers. However, data owners are very skeptical to place their data outside their own control sphere. Their main concerns are the confidentiality, integrity, security and methods of mining the data from the cloud. The Greek myths tell of creatures plucked from the surface of the Earth and enshrined as constellations in the night sky. Something similar is happening today in the world of computing. Data and programs are being swept up from desktop PCs and corporate server rooms and installed in “the compute cloud”. In general, there is a shift in the geography of computation. Cloud computing is here. With its new way to deliver services while reducing ownership, improving responsiveness and agility, and especially by allowing the decision makers to focus their attention on the business rather than their IT infrastructure, there is no organization that has not though about moving to the Cloud.

The move to the Cloud is a crucial step for any company, but has to be made with a lot of caution because it could turn against users. Organizations need to clearly understand the benefits and challenges, especially for the most critical applications. There are several concerns but, as shown in an IDC survey about the issues of the Cloud [GEN09], security is the main concern. The question is why security is such a complicated challenge in the decision of moving to the Cloud. The answer is easy: lack of control over their data. 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.

Load balancing is the pre requirements for increasing the cloud performance and for completely utilizing the resources. Load balancing is centralized or decentralized. Load Balancing algorithms are used for implementing. Several load balancing algorithm are introduced like round robin algorithm a mining improvement in the performance. The only differences with this algorithm are in their complicity. The effect of the algorithm depends on the architectural designs of the clouds [4]. Today cloud computing is a set of several data centers which are sliced into virtual servers and located at different geographical location for providing services to clients. The objective of paper is to suggest load balancing for such virtual servers for higher performance rate.

In general, load balancing algorithms follow two major classifications:

- Depending on how the charge is distributed and how processes are allocated to nodes (the system load);
- Depending on the information status of the nodes (System Topology).

RELATED WORK

Surbhi Kapoor (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 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 (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. 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.

Reena Panwar (2015) describes that the cloud computing has become essential buzzword in the Information Technology and is a next stage the evolution of Internet. 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 (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. The proposed approach mainly ensures a better overall performance with efficient load balancing, the continuous availability and a security aspect.

Lu Kang (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 (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.

Shanti Swaroop Moharana (2015) specifies that the Cloud Computing is high utility software having the ability to change the IT software industry and making the software even more attractive. It has also changed the way IT companies used to buy and design hardware. The elasticity of resources without paying a premium for large scale is unprecedented in the history of IT industry. The increase in web traffic and different services are increasing day by day making load balancing a big research topic. Cloud computing is a new technology which uses virtual machine instead of physical machine to host, store and network the different components.

Nguyen Khac Chien (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.

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. Load Balancing is done with the help of load balancers where each incoming request is redirected and is transparent to client who makes the request. Based on predetermined parameters, such as availability or current load, the load balancer uses various scheduling algorithm to determine which server should handle and forwards the request on to the selected server. 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. VM's are categorized only on a single parameter which is MIPS. Multiple parameters like RAM and Bandwidth should also be considered for allocation of cloudlets to VM

- In the existing work, there has been no criteria explained for categorizing the jobs into 3 lease types.
- The existing work is applicable only in the homogeneous environment where all the Vm's are of same capacity.
- No cost has been computed for the jobs of different lease types.

RESEARCH OBJECTIVES

- To study the existing load balancing algorithms.
- To design the improved load balancing algorithm with heterogeneous Virtual Machines.
- To categorize the cloudlets into different lease types by introducing various parameters like cost and priority.
- To increase the profits of cloud service provider.
- To develop the proposed algorithm and compare the performance of proposed algorithm with existing algorithm.

TYPES OF LEASES

There are 3 types of leases available for the cloudlets/jobs assigned by the user:

- ***Cancellable***
- ***Suspend able***
- ***Non-Preempt able***

If the algorithm finds two or more low priority jobs the lease type of the job should be considered. If the lease type is Non-preempt able, then the job is ignored for the candidate set. Priority is given to cancellable lease type than suspend able lease type as the jobs with such lease type can be killed. The job with suspend able lease type should be suspended and resumed. If there are two or more low priority jobs with suspend able lease type then the level of completion of job is considered. The job which has finished only a minimum portion of job is chosen for preemption.

Table 1. Types of Leases

PARAMETERS	CANCELLABLE	SUSPENDABLE	NON-PREEMPTABLE
COST	LOW	MEDIUM	HIGH
PRIORITY	LOW	MEDIUM	HIGH
DEADLINE GUARENTEE	NO	NO	YES
SUSPENSION	YES	YES	NO
JOB KILL	YES	NO	NO

RESEARCH METHODOLOGY

- Open the Cloud Sim Simulator in Netbeans IDE of Java and create the heterogeneous virtual machines of different MIPS.
- The datacenter broker will receive the list of all the virtual machines and will categorize them into two classes:
  - **Class A - High end VM – will process the Non-Preemptable cloudlets**
  - **Class B - Low end VM will process the suspendable and cancellable cloudlets.**
- The machines in each class will be further sorted in descending order according to their MIPS.
- The DCB will group the Cloudlets according to their lease type and sort the non-preemtable cloudlets in ascending order of deadline.
- High priority cloudlets are assigned to machines belonging to class A whereas low priority (suspendable and cancellable) cloudlets are assigned to machines belonging to class B.
- If all the machines in the class A are occupied and any high priority cloudlet arrives, then it will executed by the VM of class B.
- When a high priority job arrives, availability of the VM is checked in class A. If the VM is available in class A, then job is allowed to run on the VM in class A. If the VM is not available, then algorithm finds a free VM in class B.
- Again, if the VM is not available in Class B, then the algorithm find a low priority job taking into account the job's lease type.
- Priority is given to cancellable lease type than suspendable lease type as the jobs with such lease type can be killed.
- If there are two or more low priority jobs with suspendable lease type then the level of completion of job is considered. The job which has finished only a minimum portion of job is chosen for preemption.

- Jobs with suspendable and cancellable lease type will never execute in Class A because there is no guarantee of deadline.
- Calculate the different parameters like waiting time, processing time and cost.
- Repeat the same procedure for all the remaining cloudlets.

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.

EXPERIMENTAL SIMULATION

Table 2. Simulation Environment

Operating system	Windows 8.1
Programming language	Java
Java version	JDK 8.0
Number of virtual machines created	3
MIPS (Million Instructions per Second)	100
Bandwidth	300 MB/sec
RAM	256 MB
Number of CPU	1

Table 3. Results of Existing Work

S.NO	NO. OF CLOUDLETS	CANCELLABLE	SUSPENDABLE	NON PREEMTABLE	TOTAL PROCESSING TIME	TOTAL WAITING TIME
1	5	3	2	0	0.002	0.0011
2	10	4	3	3	0.0041	0.0046
3	40	14	13	13	0.0176	0.103
4	60	20	20	20	0.027	0.2442
5	100	34	33	33	0.0446	0.6703
6	150	50	50	50	0.0675	1.547
7	200	68	66	66	0.0896	2.7342
8	300	100	100	100	0.135	6.2246
9	400	134	133	133	0.1796	11.0413
10	500	168	166	166	0.2246	17.2535
11	700	234	233	233	0.3146	33.8903
12	1000	334	333	333	0.4496	69.3377
13	2000	668	666	666	0.8996	277.8959
14	3000	1000	1000	1000	1.35	625.9821
15	5000	1668	1666	1666	2.2496	1738.4994

In the table 2 and table 3, we have mentioned the results of the various experiments conducted on the existing and proposed algorithm by taking the same configuration of virtual machines and the same properties of the cloudlets.

Table 3. Results of Proposed Work

Sno.	No of Cloudlets	Total Processing Time	Total Waiting Time	Cancellable Cost	Suspendable Cost	Np Cost
1	5	0.00112449	0.000845	12	18	0
2	10	0.001869614	0.002212188	14	30	48
3	40	0.007640135	0.041648617	60	128	192
4	60	0.011940808	0.104276133	84	210	288
5	100	0.019738325	0.283836188	134	330	528
6	150	0.029729929	0.64948866	204	510	768
7	200	0.039636108	1.154722683	270	660	1056
8	300	0.059617367	2.608130365	408	990	1584
9	400	0.078796732	4.600945847	540	1328	2112
10	500	0.099014824	7.249583227	672	1668	2640
11	700	0.138846247	14.26462971	936	2340	3710
12	1000	0.197629721	29.00593174	1334	3330	5328
13	2000	0.395418812	116.2249526	2670	6660	10656
14	3000	0.593291317	261.6056824	4008	9990	15984
15	5000	0.98847098	726.7319212	6672	16668	26640

Response Time

- It is defined as total time taken by a load balancing algorithm to finish the execution of a cloudlet.
- Response Time (RT) = FT – ST

Where,

FT = finish time of execution

ST = start time of execution

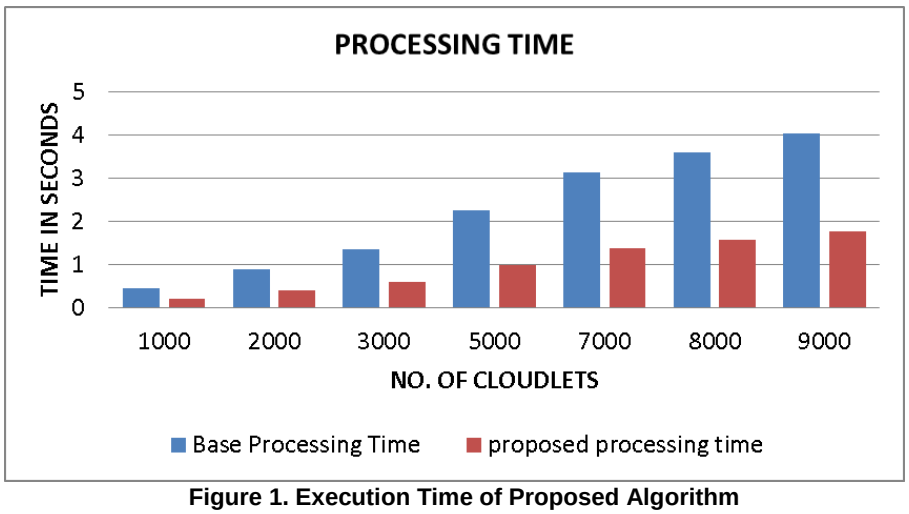

Figure 1. Execution Time of Proposed Algorithm

From the bar chart in figure 1, it is clear that the processing cost has been reduced.

Processing Cost

It is obtained by addition of cost per storage, cost per memory and cost per memory.

Processing Cost = RT * unit_cost.

where, RT = response time

unit_cost = cost per unit time

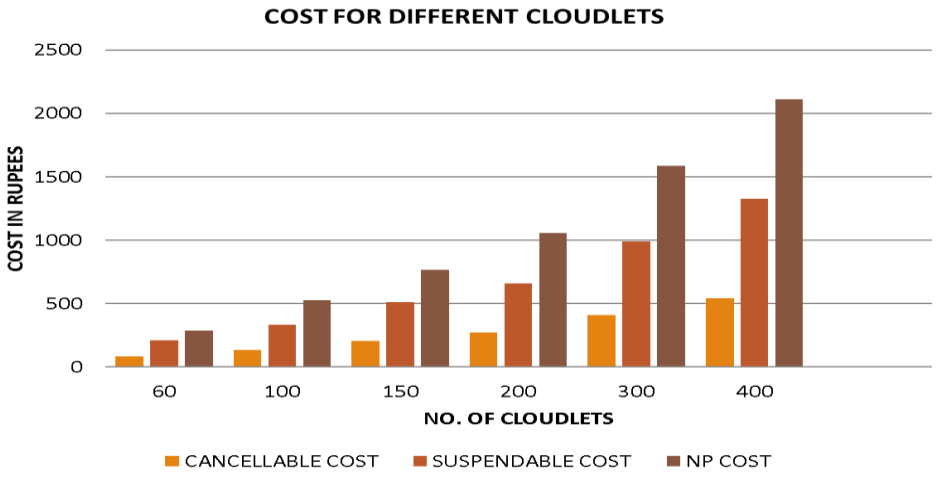

Figure 2. Processing Cost

Figure 2 illustrates cost for different types of cloudlets. The cost of the cancellable is lesser than the suspendable cloudlets and the cost of suspendable cloudlets is lesser than the non- preemtable cloudlets. The cost is associated with the type of the server that the client requested. Figure 3 shows the waiting time for the base work and proposed work. It is clear from the line chart, that the waiting times of the cloudlets have been reduced, thereby decreasing the cost and increasing the overall efficiency of the system.

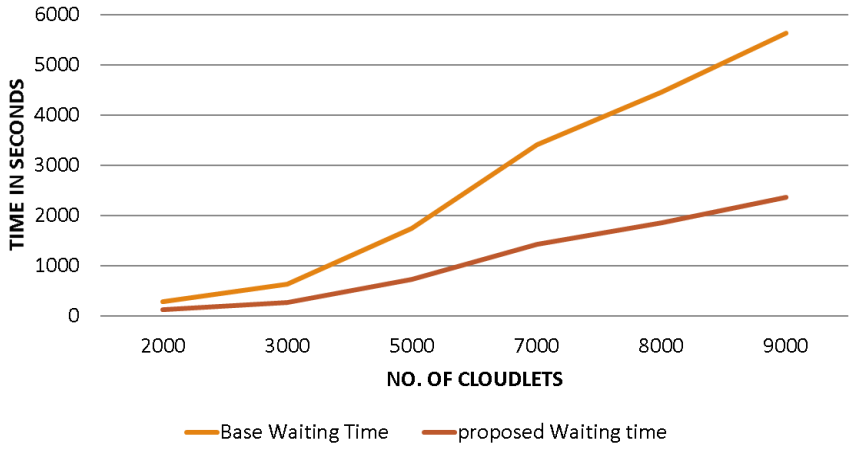

Figure 3. Waiting Time

CONCLUSION AND FUTURE SCOPE

The performance of improved load equalization algorithmic program has been studied in this research work. The request time for the policies applied are same which suggests there's no impact on data centers request time after modifying the algorithm. The processing cost, waiting time, execution time are calculated using various number of experiments. The experiments conducted are compared with previous algorithms. The results indicate that the approaches surpass to previous algorithmic program in terms of execution time, waiting time and processing cost associated with them. The experimental results are obtained by applying the new planned algorithmic program within the Cloud Sim simulator developed in java programming language, shows that the new work has outperformed the present programming algorithms in giant scale distributed systems. To get a much better answer and more precise results, the model ought to be created a lot of realistic by considering problems regarding load equalization like information section, communication price and flow time and results may be tested in real cloud setting. Moreover, fault tolerance, virtual machine migration and the power consumed by the virtual machine are also the considerable factors that can be explored in the future work.

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.