International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.4., 2015
(July-Aug)

REVIEW ARTICLE

ISSN: 2321-7758

CLOUD SEGREGATING FOR SHARED CLOUDS USING FREIGHT STABILIZATION MODEL FOR PERFORMANCE ENHANCEMENT

P. NAGA SUSMITHA[1], M. SRI BALA[2]

[1] M.tech Student, Dept. Of CSE, Lakireddy Balireddy College of Engineering, Mylavaram (Krishna DT), Andhra Pradesh, India.

[2] Sr. Assistant Professor, Dept. Of CSE, Lakireddy Balireddy College of Engineering, Mylavaram (Krishna DT). Andhra Pradesh. India.

P. NAGA SUSMITHA

ABSTRACT

Cloud computing is an enhancing technology in the field of computer science. It is a distributed computing network and has the ability to run a program or application on many connected computers at the same time. It is an efficient and scalable network, but maintaining the stability of processing so many applications is a very complex problem. Load balancing is the best solution for this problem. It 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. Load Balancing Model Based on Cloud Partitioning for the Public Cloud environment has an important impact on the performance of network load. Excellent load balancing makes cloud computing more efficient and also improves user satisfaction. This paper announces a better approach of load balance model for the public cloud depends on the cloud segregating concept with a switch mechanism to select different approaches for different circumstances. The algorithm applies the game theory to the load balancing strategy to improve the efficiency in the public cloud environment which ultimately helps to improve the different performance parameters like throughput, response time, latency etc. for the clouds.

Keywords--- Cloud Computing, Dynamic Load Balancing Model, Public Cloud, Cloud Partitioning, Cloud Status, Game Theory.

I. INTRODUCTION

Cloud computing is an attracting technology in the field of computer science. In Gartner's report, it says that the cloud will bring changes to the IT industry. The term "cloud" originates from the world of telecommunications when providers began using virtual private network (VPN) services for data communications. The definition of cloud computing provided by National Institute of Standards and Technology (NIST) says that: "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. So through this cloud computing there is no need to store the data on desktops, portables etc. You can store the data on servers and you can access the data through internet. Cloud computing provides better utilization of distributed resources over a large data and they can access remotely through the internet.

A load balancing is a method of dividing computing loads among numerous hardware. Load balancing is a process of reassigning the total load to the individual nodes of the collective system to make resource utilization effective and to improve the response time of the job, simultaneously removing a

condition in which some of the nodes are overloaded while some others are under loaded. A load balancing algorithm used for balancing purpose which is dynamic in nature does not consider the previous state or behaviour of the system, that is, it depends on the present behaviour of the system. Also load balancing is a relatively new technique that facilitates networks and resources by providing a maximum throughput with minimum response time. Proper load balancing can help in utilizing the available resources optimally. 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. Also load balancers may have a variety of special features.

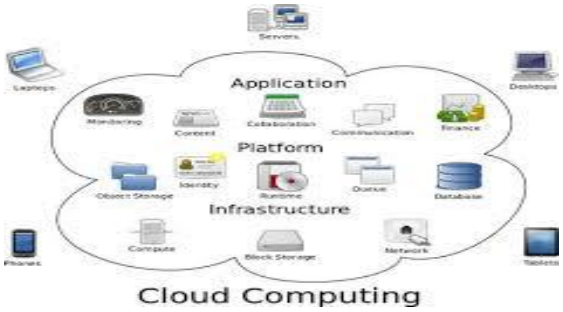

Fig 1. Applications based on cloud computing

Load balancing and provisioning in cloud computing systems is really a challenge job. For solving such problem always a distributed and dynamic solution is required. In this paper we introduce a dynamic strategy to balance workload among nodes. This scheme provides more flexibility and performance in the system. Virtualization is very useful concept in context of cloud systems. Virtualization means "something which isn't real", but gives all the facilities of a real. It is the software implementation of a computer which will execute different programs like a real machine. Virtualization is related to cloud, because using virtualization an end user can use different services of cloud. In this paper we reviewed a dynamic strategy to balance workload among nodes with the help of cloud partitioning and game theory concept. In this work various nodes are used with required computing resources situated in different geographic location.

Goals of Load Balancing

- To improve the performance substantially.
- To have a backup plan in case the system fails even partially.
- To maintain the system stability.
- To accommodate future modification in the system.
- Enhance the performance of the cloud.
- Reduce Traffic in cloud.
- Increasing reliability and throughput.

II. EXISTED ISSUES

The jobs arrival pattern is not predictable and the capacities of each node in the cloud differ, for the load balancing problem, and workload control is difficult to improve system performance and maintain stability. Cloud computing is efficient and scalable but maintaining the stability of processing so many jobs in the cloud computing environment is a very complex problem with load balancing receiving much attention for researchers.

DISADVANTAGES:

Workload control is vital to improve system performance and maintain stability. Cloud computing environment is a very complex issue with load balancing receiving. The job arrival design is not predictable and the capabilities of each node in the cloud differ for load balancing problem.

III. PROPOSED ISSUES

(1) Cloud division rules: Cloud division is not a simple problem. Thus, the framework needs a detailed cloud division methodology. For example, nodes in a cluster may be far from other nodes or there will be some clusters in the same geographic area that are still far apart. The division rule should simply be based on the geographic location.

(2) How to set the refresh period for data statistics analysis, the main controller and the cloud partition balancers need to refresh the information at a fixed period. If the period is too short, the high frequency will influence the system performance. If the period is too much long, the information will be too old to make good decision. Thus, tests and statistical tools are needed to set reasonable refresh periods.

(3) A load status evaluation: A good algorithm is needed to set Load degree high and Load

degree low, and the evaluation mechanism needs to be more comprehensive.

ADVANTAGES

When the environment is huge and compound these divisions streamline the load balancing. The role that loads balancing plays in refining the presentation and maintaining stability.

IV. BACKGROUND

Cloud computing is Internet based computing, whereby shared resources, software and information are provided to computers and other devices on-demand, like a public utility. Infrastructure as a Service is a single tenant cloud layer where the Cloud computing vendor's dedicated resources are only shared with contracted clients at a pay-per-use fee. This greatly minimizes the need for huge initial investment in computing hardware such as servers, networking devices and processing power. Software as a Service also operates on the virtualized and pay-per-use costing model whereby software applications are leased out to contracted organizations by specialized SaaS vendors. This is traditionally accessed remotely using a web browser via the Internet. Platform as a service cloud layer works like IaaS but it provides an additional level of —rented functionality. Clients using PaaS services transfer even more costs from capital investment to operational expenses but must acknowledge the additional constraints and possibly some degree of lock-in posed by the additional functionality layers.

Cloud computing is the problem of load balancing. Further, while balancing the load, certain types of information such as the number of jobs waiting in queue, job arrival rate, CPU processing rate, and so forth at each processor, as well as at neighbouring processors, may be exchanged among the processors for improving the overall performance. Good load balance will improve the performance of the entire cloud. However, there is no common method that can adapt to all possible different situations. Various methods have been developed in improving existing solutions to resolve new problems. Each particular method has advantage in a particular area but not in all situations. Therefore, the current model integrates several methods and switches between the load balance method based on the system status.

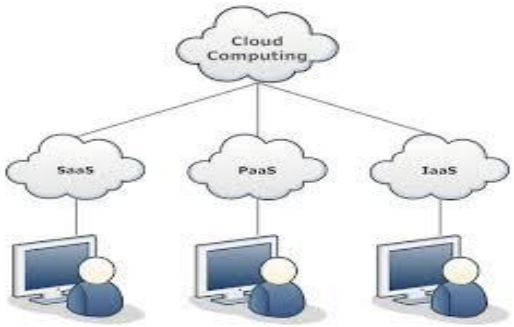

Fig 2. Three Services of Cloud Computing

V. LITERATURE SURVEY

Load balancing in cloud computing was described by B.Adler[7][9] in his white paper "Load balancing in the cloud: Tools, tips and techniques", in which he introduced the tools and techniques commonly used for load balancing in the cloud. Z Chaczko, V. Mahadevan, S. Aslanzadeh, and C. Mcdermid, in their paper "Availability and load balancing in cloud computing,2011" described the role that load balancing plays in improving the performance and maintaining stability[1][10]. Nishant et al. [7][10] used the ant colony optimization method in nodes load balancing. Randles et al.[8][10] gave a compared analysis of some algorithms in cloud computing by checking the performance time and cost. They concluded that the ESCE algorithm and throttled algorithm are better than the Round Robin algorithm in terms of performance time and cost. The Round Robin algorithm is the simplest algorithm that uses the concept of time quantum or slices which play a very important role for scheduling, because if time quantum is very large then Round Robin Scheduling Algorithm is same as the FCFS Scheduling. So for simplicity we use the RR algorithm for our work.

VI. RELATED WORK

In 2013, Xu, Gaochao et al [1] presented A load balancing model based on cloud partitioning for the public cloud. The load balancing model is aimed at the public cloud which has numerous nodes with distributed computing resources in many different geographic locations. Thus, this model divides the public cloud into several cloud partitions. When the environment is very large and complex, these divisions simplify the load balancing. The cloud has a main controller that chooses the suitable partitions

for arriving jobs while the balancer for each cloud partition chooses the best load balancing strategy. The load balance solution is done by the main controller and the balancers. The main controller first assigns jobs to the suitable cloud partition and then communicates with the balancers in each partition to refresh this status information. Since the main controller deals with information for each partition, smaller data sets will lead to the higher processing rates. The balancers in each partition gather the status information from every node and then choose the right strategy to distribute the jobs.

Zhu, Yan et al suggested "Efficient provable data possession for hybrid clouds. They focused on the construction of PDP scheme for hybrid clouds, supporting privacy protection and dynamic scalability. They first provide an effective construction of Cooperative Provable Data Possession (CPDP) using Homomorphic Verifiable Responses (HVR) and Hash Index Hierarchy (HIH). This construction uses homomorphic property, such that the responses of the client's challenge computed from multiple CSPs can be combined into a single response as the final result of hybrid clouds. By using this mechanism, the clients can be convinced of data possession without knowing what machines or in which geographical locations their files reside. More importantly, a new hash index hierarchy is proposed for the clients to seamlessly store and manage the resources in hybrid clouds. Their experimental results also validate the effectiveness of our construction.

There have been many studies of load balancing for the cloud environment. Load balancing in cloud computing was described in a white paper written by Adler who introduced the tools and techniques commonly used for load balancing in the cloud. However, load balancing in the cloud is still a new problem that needs new architectures to adapt to many changes. Chaczko et al. described the role that load balancing plays in improving the performance and maintaining stability. There are many load balancing algorithms, such as Round Robin, Equally Spread Current Execution Algorithm, and Ant Colony algorithm. Nishant et al. used the ant colony optimization method in nodes load balancing. Randles et al. gave a compared analysis of some algorithms in cloud computing by checking the performance time and cost. They concluded that the ESCE algorithm and throttled algorithm are better than the Round Robin algorithm. Some of the classical load balancing methods are similar to the allocation method in the operating system, for example, the Round Robin algorithm and the First Come First Served (FCFS) rules. The Round Robin algorithm is used here because it is fairly simple.

VII. SYSTEM ARCHITECTURE

There are several cloud computing categories with this work focused on a public cloud. A public cloud is based on the standard cloud computing model, with service provided by a service provider. A large public cloud will include many nodes and the nodes in different geographical locations. Cloud partitioning is used to manage this large cloud. A cloud partition is a subarea of the public cloud with divisions based on the geographic locations. The architecture is shown in Figure 3. When job i arrives at the system, the main controller (Admin) decides to which partition the job should be assigned. If this is the last updated job, then the job is assigned to Partition1. If it is an upcoming job, then it is assigned to Partition2. If it's a currently running job then it is assigned to Partition3. If it is an out dated job then it is assigned to Partition4. The Best Partition Searching algorithm is shown in Algorithm 1.

ALGORITHM 1: BEST PARTITION SEARCHING

```
begin
while job do
searchBestPartition (job);
if Update(job) then
Send Job to Partition1;
else if EndDate>CurrentDate then
Send Job to Partition2;
else if ArrivalDate<=CurrentDate&&EndDate
=CurrentDate then
Send Job to Partition3;
else if EndDate<CurrentDate then
Send Job to Partition4;
end if
end while
end
```

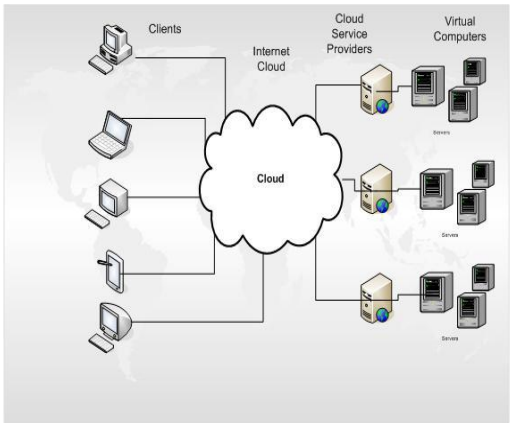

Fig 3. Typical Cloud partitioning

The load balancing strategy is based on the cloud partitioning concept. After creating the cloud partitions, the load balancing then starts, when a job arrives at the system, with the main controller deciding which cloud partition should receive the job. The partition load balancer then decides how to assign the jobs to the nodes. When the load status of a cloud partition is normal, this partitioning can be accomplished locally. If the cloud partition load status is not normal, this job should be transferred to another partition.

LOAD BALANCING:

"The load balancing technique used to make sure that none of the node is in idle state while other nodes are being utilized". In order to balance the load among multiple nodes you can distribute the load to another node which has lightly loaded. Thus distributing the load during runtime is known as Dynamic Load Balancing technique. Load balancing algorithm can be divided into two categories as 1) Static and 2) Dynamic. In static load balancing algorithm, all the information about the system is known in advance, and the load balancing strategy has been made by load balancing algorithm at compile time. This load balancing strategy will be kept constantly during runtime of the system. In contrast, dynamic algorithm is implemented at running time, and the load balancing strategies change according to the real statement of the system. Though, the dynamic algorithm has better adaptability, it is sensitive to the accuracy of the load information or statement of system. Many researchers have proposed several algorithms for load balancing. In cloud computing when a computation is requested by any system it is distributed to all the slaves existing in that cloud. So the way in which the distribution is being done must get the response from all the slaves at the same time so that there should not be any waiting for any particular computing device to reply before further processing could happen. But in the real time clouds heterogeneous computing devices exists and any process's execution time on the slave is required to be estimated. So the main feature that is must in any load balancer is the asymmetric load distribution.

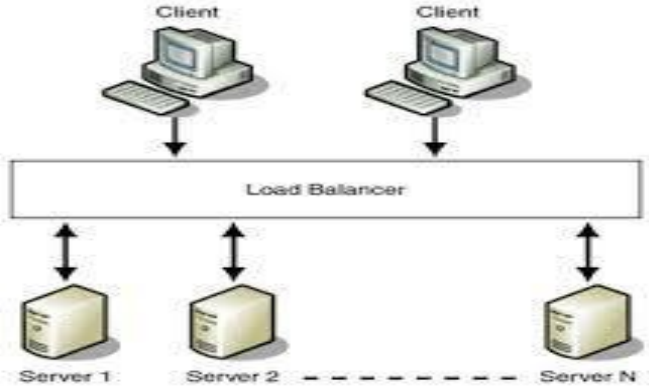

Fig 4. Simple View of Load Balancer

CLOUD PARTITIONING MODEL:

The load balancing strategy is based on the cloud partitioning concept as shown in fig5.After creating the partitions, the load balancing then starts: when a job arrives at the system, then the main balancer decides which cloud partition should receive the job. The partition load balancer then decides how to assign the jobs to the nodes. When the load status of a cloud partition is normal, this partitioning can be accomplished locally. If the cloud partition load status is not normal, this job should be transferred to another partition.

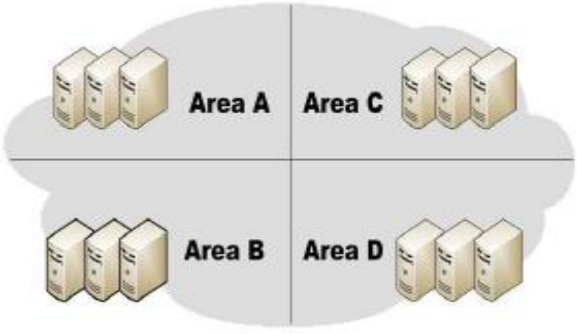

Fig 5. Typical Cloud Partitioning

MAIN CONTROLLER AND BALANCER

The load balance solution is done by the main controller and the balancers. The main controller first assigns jobs to the suitable cloud partition and then communicates with the balancers in each partition to refresh this status information. Since the main controller deals with information for each partition, smaller data sets will lead to the higher processing rates. The balancers in each partition gather the status information from every node and then choose the right strategy to distribute the jobs. The relationship between the balancers and the main controller is shown in Fig.6.

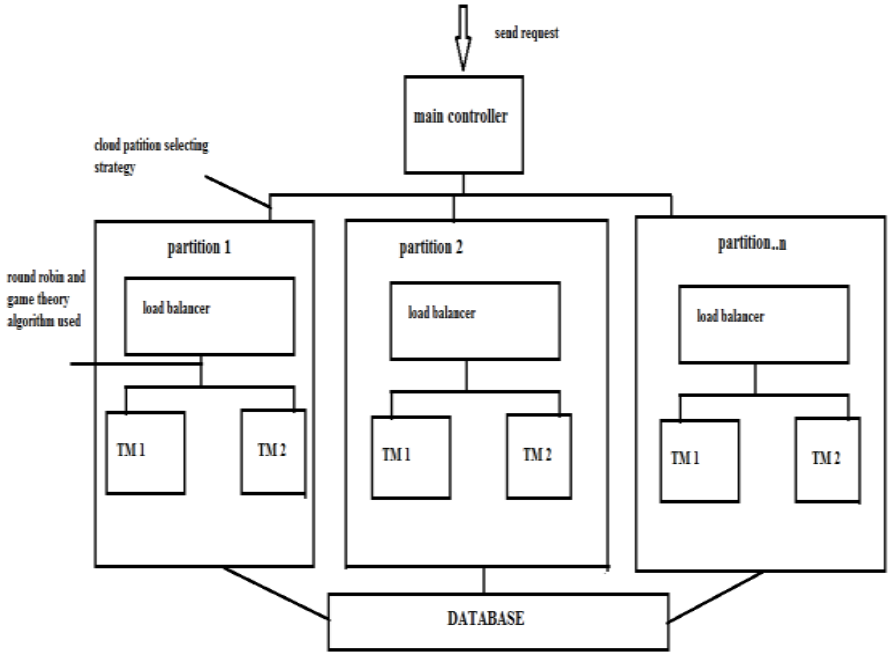

Fig 6. Relation between Balancer and Main Controller

ASSIGNING JOBS TO THE CLOUD PARTITION:

When a job arrives at the public cloud, the first step is to choose the right partition. The cloud partition status can be divided into three types:

(1) Idle: When the percentage of idle nodes exceeds α, change to idle status.

(2) Normal: When the percentage of the normal nodes exceeds β, change to normal load status.

(3) Overload: When the percentage of the overloaded nodes exceeds γ , change to overloaded status. The parameters α, β, and γ are set by the cloud Partition balancers.

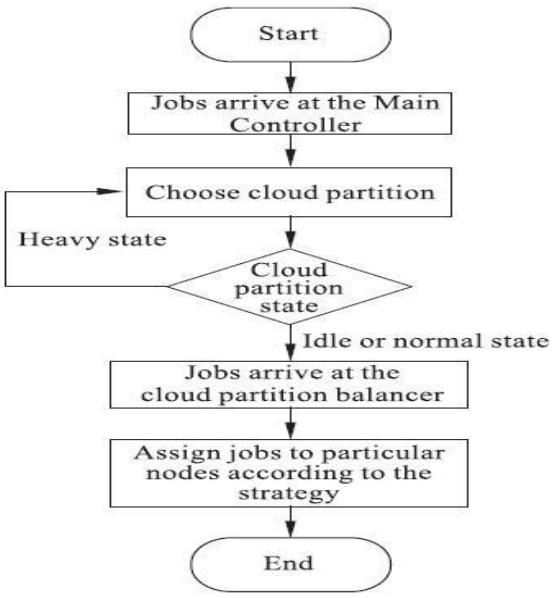

Fig 7. Job Assignment Strategy

CALCULATION OF LOAD DEGREE (LD) FOR A NODE:

The Load Degree (LD) of a node in any cloud partition is calculated from the following equation

$$LD\ (N) = \sum_{i=1}^{m} Xi * Fi$$

Here, N=Current Node, Fi are the parameter either static or dynamic where Fi(1<=i<=m), m represents the total number of parameter . Xi are weights that may differ for different kinds of job for all (1<=i<=n).

POSSIBLE LOAD STATUS OF NODE:

According to the calculation of load degree for the node three load status of the node are defined as follows

IDLE: When LD (N)=0

NORMAL: 0<LD (N) <=High_LD

OVERLOADED: High_LD<=LD (N)

Any cloud partition having the status=HEAVY is not selected by the Load Balancer Manager and likewise any node having the Load Degree (LD) =OVERLAODED is not eligible for the processing. Only cloud partition having IDLE or NORMAL load status and Node having IDLE or NORMAL load degree are selected for scheduling and load balancing.

VIII. CLOUD PARTITION LOAD BALANCING STRATEGY

Here we are going to discuss some load balancing technique for both the partition having either load status=idle or load status=normal. In this section mainly we will discuss about the load balancing technique for the cloud partition having load status=normal using game theory. When the cloud partition is idle, few jobs are arriving and thus the cloud partition has the ability to process jobs as quickly as possible so a simple load balancing method such as "The Round Robin algorithm based on the load degree evaluation" will be used here for its simplicity. When the cloud partition is normal, job arrival pattern much faster than in the idle state and the situation is far more complex, so a different strategy is used for the load balancing as each user wants his jobs completed in the shortest time. The current model uses the game theory approach for non-cooperative games.

There are many simple load balancing algorithm methods such as the First Come First Served (FCFS), Round Robin algorithm.

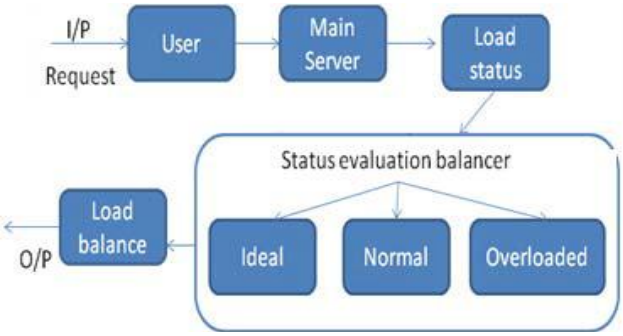

Fig 8. System Model

FIRST COME FIRST SERVE ALGORITHM

Step 1- Main Controller (Admin) maintains an index table of job requests.

Step 2- The job requests are stored in the table on the basis of their arrival time.

Step 3- The Main Controller (Admin) scans the index table from top to bottom.

Step 4- The first job request according to the arrival time is allocated the grant by the Main Controller (Admin).

Step 5- The HR receives the response to the request sent and then posts jobs by providing details about the interview.

In this way all the jobs are processed in the first come first serve basis.

This evaluation of each node's load status is very significant. The first task is to describe the load degree of each nodes. The node load degree is associated to various static parameters and dynamic parameters. The static parameters contain the number of CPU's, the CPU processing speeds, the memory size, etc. Dynamic parameters are the memory utilization ratio, the CPU utilization ratio, the network bandwidth, etc. The cloud partition balancer collects load information from each node to estimate the cloud partition status.

IX. CONCLUSION AND FUTURE SCOPE

The overall goal of this project is to balance the load on clouds. Balancing load on the cloud will improve the performance of cloud services substantially. It will prevent overloading of servers, which would otherwise degrade the performance. Load balancing is the utmost essential issue in the system to allocate load in well-organized manner. It also confirms that each computing resource is dispersed efficiently and objectively. Public cloud is made up of several nodes situated in deferent

geographic location. Cloud partitioning is a method to make partitions of huge public cloud is some segment of cloud. The object of study in game theory is the game, which is a formal model of an interactive situation. It typically involves several players; a game with only one player is usually called a decision problem. Thus with cloud partitioning concept it is possible to provide good load balancing and hence improving the overall performance of cloud environment and user satisfaction.

In future study we will try to find other load balance strategy because other load balance strategies may provide better results, so tests are needed to compare different strategies. Many tests are needed to guarantee system availability and efficiency. Also we will address the development of game theoretic models for load balancing in the context of uncertainty as well as game theoretic models for dynamic load balancing in future.

X. ACKNOWLEDGMENT

I wish to thank all the people who gave me an unending support right from stage the idea was conceived. I would like to thank my Professor for accepting me to work under her valuable guidance. She closely supervises the work over the past few months and advised many innovative ideas, helpful suggestion, valuable advice and support. And finally, the authors would like to thank to GaochaoXu, Junjie Pang, and Xiaodong Fu for their exclusive information.

XI. REFERENCES

[1] Xu, Gaochao, Junjie Pang, and Xiaodong Fu. "A load balancing model based on cloud partitioning for the public cloud."IEEE Tsinghua Science and Technology, Vol. 18, no.1, pp. 34-39, 2013.

[2] B. Adler, Load balancing in the cloud: Tools, tips and techniques, http://www.rightscale.com /infocenter /whitepapers/Load-Balancing-in-the-Cloud.pdf, 2012.

[3] P. Mell and T. Grance, The NIST definition of cloud computing,http://csrc.nist.gov/ publications/nistpubs/800-145/SP800-145.pdf, 2012.

[4] Microsoft Academic Research, Cloud computing, http://libra.msra.cn/Keyword/6051/ cloud-computing?query= cloud%20computing, 2012.

[5] Vinay Kumar Kaushik, Hemant Kumar Sharma, Dinesh Gopalani, L*oad Balancing In Cloud Computing Using High Level Fragmentation Of Dataset*, International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.

[6] Google Trends, Cloud computing, http://www.google. com/trends/explore#q=cloud%20 computing, 2012.

[7] Z.Chaczko, V.Mahadevan, S.Aslanzadeh and C. Mcdermid, *Availability and load balancing in cloud computing*, presented at the 2011 International Conference on Computer and Software Modelling, Singapore, 2011.

[8] M. Randles, D. Lamb, and A. Taleb-Bendiab, *A comparative study into distributed load balancing algorithms for cloud computing,* in Proc. IEEE 24th International Conference on Advanced Information Networking and Applications, Perth, Australia, 2010, pp. 551-556.

[9] Tejinder Sharma, Vijay Kumar Banga, *Efficient and Enhanced Algorithm in Cloud Computing* ,International Journal of Soft Computing and Engineering (IJSCE)ISSN: 2231-2307, Volume-3, Issue-1, March 2013.

[10] S. Aote and M. U. Kharat, *A game-theoretic model for dynamic load balancing in distributed systems*, in Proc. The International Conference on Advances in Computing, Communication and Control (ICAC3 '09),New York, USA, 2009, pp. 235-238.