International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 3 Issue 4, April 2014

Analysis of Load Balancing Algorithms in Cloud Computing and Study of Game Theory

Shilpa S, Prof. Shubhada Kulkarni, Prof.Sharada Kulkarni

Abstract: Cloud Computing is a mechanism that offers several services such as public storage, application, hardware, software, processing etc., to a shared pool of users. Still being in its nascent stage and with ever increasing customer base and amount of data to store; proper access control, data lineage, cost and stability are still some issues in Cloud Computing, with Load Balancing being the biggest challenge. This article compares the various algorithms associated with load balancing in cloud computing and explains the application of game theory assuming the nodes and jobs in the cloud as players of the game. Game theory when associated with load balancing can help improve the efficiency of cloud computing by reducing the response time and utilizing the resources optimally.

Index Terms: Cloud Computing, Game Theory, Load Balancing

1. INTRODUCTION

Cloud computing is a service delivery framework allowing users to share resources. Hardware, software or application service can be enabled on demand of the consumers. Cloud Computing adjusts to the changing demands of the users and also spares them from investing huge amounts in purchasing of hardware, software or even in infrastructure. [1] Cloud computing services offer the pay-as-you-use model, hence it is a cost effective model making IT management much responsive and user friendly to the changing needs. [2] Cloud computing infrastructure overcomes the physical barriers in the system and automats the management of the resources providing users in the cloud with the computational services and data storage facilities. The user base can range from an individual to infinity. [3]

Cloud Computing has some interesting characteristics provisioned by the user to his best interest. Some of them are listed below:

Abstraction and virtualization are the fundamental concepts of cloud computing. It abstracts the details of system implementation from the users and creates virtual platforms to map the workloads to the nodes in the cloud. [4]

Computing applications can be set as per the user's requirement so as to provide the computing services automatically.

Cloud computing is a broad network and capabilities provided on it can be accessed through standard mechanisms.

Several physical and virtual resources serve multiple users across several locations. These resources expand or contract their availability depending on the demand of the users. Resources may be memory, storage, bandwidth etc.,

Cloud Computing is transparent as its usage can be monitored and controlled. The cloud usage and operation can be controlled and accordingly optimizes its resources by supplementing the service to the cloud. [5]

A cloud can be public or private and may be shared by several or selected users. It can also be a combination of the two referred to as a Hybrid cloud. The public cloud may be partitioned based on the geographical locations to manage the unlimited service calls. Resources are globally distributed and have varying capacity to be hosted across the cloud. The cloud computing environment is large, complex and dynamic integrating numerous computing resources.

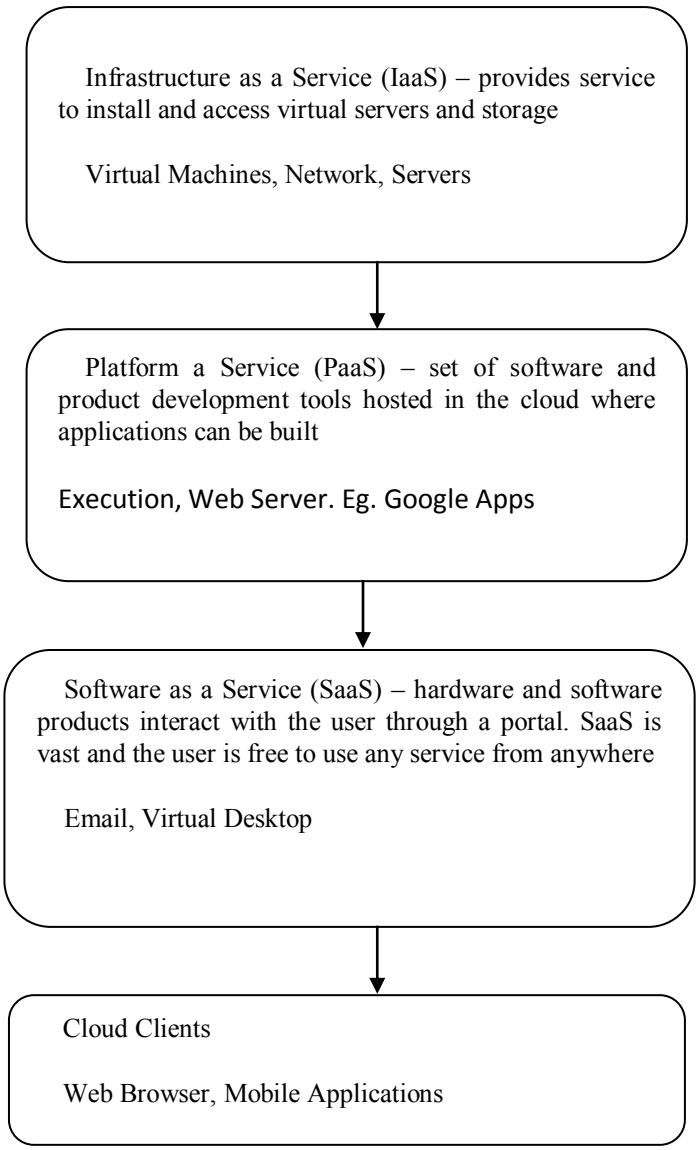

Fig. 1.Layers of cloud computing

I.2 Challenges in Cloud Computing

Though cloud computing is a brilliant concept of sharing work load but has certain challenges associated with it. The users may not be able to evenly assign their control over the data or application available at the cloud cluster. Availability and usage of resource affects the overall performance of the cloud. With the growing demand, there is an ever growing user base with each user expecting unlimited availability of resource or service in the cloud for him. Cloud computing has to meet the expectations of all these users efficiently. Since data and disseminated resources are stored in the open environment, data storage increases and becomes an issue in handling the users and their jobs in the cloud. Maintaining the load information in this vast system and dispersing it to the right resource involves hefty cost. As cloud based services become more numerous and dynamic, resource provisioning becomes more challenging.

Security, stability, work load balance, network bandwidth, response time, transfer delay, data transfer cost etc are some of issues in Cloud Computing with Load Balancing being one of the major challenges.

The cyber (servers, routers) and physical (fibre etc.,) components account for initial provisioning and operation of the cloud and are also prone to attacks. A game theoretic approach allows provisioning and operation of the cloud infrastructure. [6]

I.3 Related Work

Several studies have been conducted to understand Cloud Computing, its effects – favourable and adverse. Meenakshi, Sidhu & Kinger gave a comparative analysis of load balancing algorithms basis various parameters depicting their efficiency and effectiveness. Grousa,

Penmatsa, Chronopoulos, in detail have explained the Game theory for load balancing and job allocation in Cloud Computing or Computing Grids. Sim Kwan Mong has explained the genetic approach of Ant Colonization and how the same can be applied to distribute the load in Cloud Computing. Wu et al modelled a controlled game for cooperative behaviour evolving utility from the energy efficiency of the cloud. [7] Kone et al also worked towards the resource allocation problem in virtual machines and suggested he pay as you use system where resources and virtual machines are selfishly allocated. [8] Cloud computing being an ever changing process would involve an on-going research work and no method can be declared perfect for addressing the issues in cloud computing.

I.4 Problem Definition:

Load arrives randomly in a cloud computing environment and can choke the server's bandwidth by overloading some nodes while its other nodes are idle. This problem persists and grows with the increasing customer base in the cloud. The solution to this problem may be equal distribution of work load among all the nodes in the cloud by efficient scheduling and resource allocation thereby improving the overall performance of the system, reducing the response time and total cost of the system.

To analyse the comparison of various load balancing algorithms based on the necessary qualitative metrics

2. LOAD BALANCING

Load balancing is a general method to suitably distribute work load across multiple customer networks. With all the storage and application services getting virtual load balancing aims to be self adaptive and organize cloud services to the increasing computing traffic and load so as to achieve maximum utilization of resources by deputing the load to each resource and node equally. Load balancing algorithms are optimized if the resources are utilized efficiently.

This method if applied accurately helps to improve the efficiency of the entire network by reducing the response time and optimizing the resource utilization. One of the most critical issues faced by Cloud computing is Load Balancing. With the no. of users increasing and the storage and other services to be provided to this growing lot, cloud computing is facing an issue of allocating and completing the job. Some nodes have been overloaded while others are idle. There has to be a mechanism that understands the flow of jobs being received and the current status of each node to assign the jobs accordingly to the cloud nodes. Load Balancing helps to distribute the jobs uniformly across the nodes of the cloud, hence improving the overall efficiency of the cloud. [5]

Every node is assigned a job without keeping any node idle or overloading it, hence increasing the throughput of the cloud. Load balancing can be classified basis on how the received jobs are delegated to nodes (Static Load Balancing) OR on the status of the other nodes/total system (Dynamic Load Balancing). Most of the load balancing algorithms are strategized and developed following these two concepts. [9]

In Load Balancing, traffic in the cloud is diverted and data is received or sent without bottlenecks and delays. The jobs may be prioritized and then sent to the nodes for processing. The nodes are equally loaded with work or are selected depending on their workload status or availability. Mapping resources to the nodes is not easy as the cloud is under constant change. Load balancing model divides the cloud in several cloud partitions and directs the user's request despite his location to the right resource rather than choking a single node with all job requests to process. The figure below –

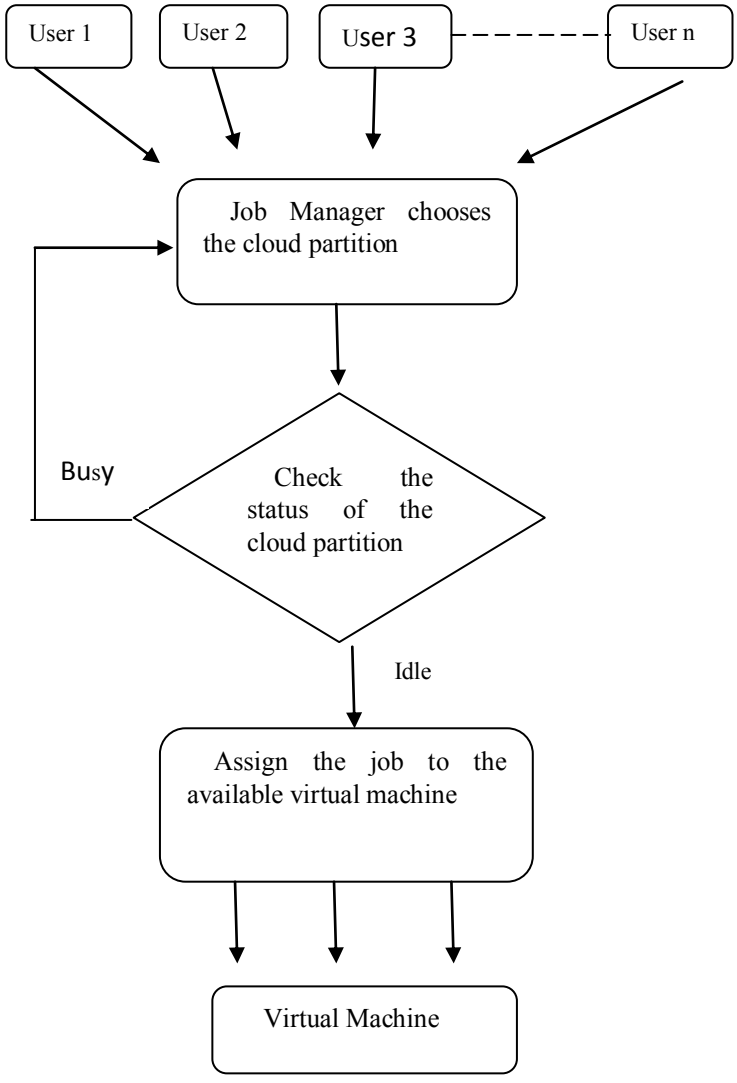

Fig .2. Execution flow of a load balancing strategy.

Load Balancing is a fine method to improve the performance of the cloud but there is no common method

to follow and achieve the desired result. Rather there are several methods and each method has its own advantages and disadvantages. Provisioning of services must coordinate completion time of these tasks providing a unified response to the users. There are many algorithms associated with load balancing in cloud computing. Some of them are explained and compared in the coming section.

Some concerns associated with load balancing to be kept in mind when strategizing and building a load balancing algorithm are:

As cloud computing is elastic, resources are allocated automatically. It is important to detect the right use or release of these resources in the cloud to keep the performance of the cloud intact.

Virtual machines are moved between physical machines to unload a heavily loaded physical machine. The distribution of load when the virtual machine is moving has to be rightly sensed to avoid bottle necks.

The entire concept of cloud computing is based on the idea to have lesser individual resources. Rather, it endorses to have global resources owned by few providers and open for public utility. The accurate use of these data centres is important for effective use of load balancing.

Management and Distribution of data to the cloud for optimal storage of data while maintaining the fast access and processing of jobs is crucial.

Small data centres may be more beneficial in terms of cost and energy consumption as compared to large data centres. Cloud computing being on a global scale may find it difficult to ensure adequate response time with optimal distribution of resources. [5]

2.1 Load Balancing Algorithms

Load Balancing algorithms are used to improve the overall performance of the cloud. Load balancing or job allocation is the main research concepts for resource management. Cost, scalability, flexibility and its executing flow are some major factors that decide the effectiveness and efficiency of an algorithm.

The centralized approach has a single point of success or failure as all information is in custody of only a single node and all jobs are allocated by this node simply. Decentralized system being more robust than the centralized approach delivers only partial information to its nodes to help them reach a sub optimal decision. Depending on the stage of the system at which load balancing would be implemented, the load balancing algorithms can be classified as static or dynamic.

Static Load Balancing algorithms assign work to the nodes as and when the job arrives to the cloud. Hence some job is assigned to each node but the job cannot be shifted during execution. In this stage, the information of the entire system is known and the strategy for balancing the load is made during the compilation time. Though the jobs are distributed equally to each node, users may still have a long waiting time while other jobs are being processed, keeping the users dissatisfied.

On the contrary, dynamic load balancing algorithm is implemented at the run time and strategies for balancing load change depending on the real time status of the system. [6] Dynamic Load Balancing examines the current state of the system and shifts jobs accordingly from an overloaded machine to an idle or a normal machine to improve the execution speed, involving participation of all the nodes. These nodes may work collaboratively or independently towards reducing the response time or increasing the overall efficiency of the cloud. For a small size system, non distributed load balancing algorithm can be applied where a single node in the system or in each cluster of nodes is allocated to take care of load balancing. Dynamic algorithms have better adaptability but are sensitive to the accuracy of the load information or the real time statement of the system, thereby causing mistakes while allotting the jobs. The processors or nodes are treated unfair in a non cooperative game as the users are given the priority of choosing their job allocation strategies serving their selfish interests.

Hybrid Load Balancing algorithms provide solution to the issues faced by the static and dynamic algorithms. A game theoretic load balancing algorithm is one such approach towards allocating the right job to the right resource in the cloud.

There are several load balancing algorithms structured to address the load balancing issues in cloud computing. Some of the Load Balancing algorithms used in Cloud Computing are as described below:

Round Robin – It is a random sampling algorithm wherein jobs are assigned to nodes in a round robin circular order. Each node is distributed equal work but since the processing time of each job is not the same it results in under loading and over loading of nodes.

Equally Spread Current Execution (ECSE) – A spread spectrum technique handles the jobs with priorities. It analyzes the job's size and allocates it to the node which is comparatively idle or can handle the task easily in lesser time.

Throttled Load Balancing – it is a sender initiated request where the user requests the load balancer to check its job and assign the appropriate virtual machine to perform its job with maximum throughput.

Task Scheduling Algorithm – two level task scheduling is executed to meet user's requirements and improving the resource utilization. It maps the jobs to the virtual machines in the cloud and then maps these virtual machines to the host resources.

Load Balancing Ant Colony Optimization (LBACO) – This is based on Genetic theory. Ants travel through various routes to find food. They leave behind trails of their path by depositing pheromone for other ants and finally find the shortest path to travel to their destination. [10] This is a method to construct solutions based on the characteristics of the past solutions. This algorithm is

used for mobile networks.

2.2 Game Theory approach for Load Balancing

This algorithm like LBACO is also based on Genetic behaviour and is inspired by human networking where local information at each node is utilized for the distribution of jobs. Load balancing in cloud computing is viewed as a game with the nodes and jobs as players in the game. The main intention is to overcome the load balancing problem by using Game Theory. A game theoretic approach is an improvised strategy for load balancing through efficient job scheduling and resource allocation techniques. All nodes should be processing equal amount of work at any time basis the size, processing time and importance of the job.

Motivation for Game theory approach for Load Balancing in Cloud Computing

Given that there is large no. of jobs in the cloud; Load Balancing formulates the allocation of these jobs to computing resources optimizing the resource utilization and response time. Every user has a different requirement and selfishly wants all the resources available for them while the computing resource has to substantially meet all these requirements. This selfish behaviour in the cloud cannot be overcome using conventional algorithms for load balancing.

For static load balancing, a cooperative game is modelled taking into account the collected system information. The decision makers cooperate in making decisions to reach an optimal solution which is in concurrence and agreement with all players. The load balancing problem is formulated as a cooperative game among "n" heterogeneous computers and the communication networking subsystem. Nash Bargaining Solution (NBS) provides the Pareto Optimal solution for cooperative load balancing game. [11]

For dynamic or non cooperative games, the current state of the system is taken into account along with the static load balancing scheme to try and reach the Nash Equilibrium. Since there are several interactions involved, each player has a strategy as the best response to the other player's strategy. Nash equilibrium is attained based on these choices of strategies. Each player makes the most optimized decision for himself and the optimal solution is achieved when no player can benefit by changing his decision. It is referred as Wardrop equilibrium when there are infinite no. of decision makers, whereas Nash equilibrium is obtained for finite no of decision makers. The expected response time majorly remains low and may increase when the system utilization is very high. [12]

A global approach to the game is when there is one decision maker supervising all the jobs and status of nodes and that decision maker is responsible for optimizing the resources in minimum response time. [13]. Tree and star networking systems are examples of Global approach.

Load Balancing effects cloud computing and improves the performance by redistributing the load among the processors. Jobs are transferred from one node to another through the network involving some delay (queuing delay +processing delay) as it has to determine the destination node through remote processing. [11]

A distributed system model has n no. of users and m no of computing resources. Nash Equilibrium can be defined for a distributed system model as a strategy "s" for every user "u" as

$$Su = ArgMin\, Dj(S1, S2..Su ....Sn) \quad (1)$$

The Nash equilibrium is achieved when no user can decrease its average expected response time by unilaterally changing its strategy. A cloud which has a normal rate should dispatch jobs immediately when he receives to the nodes that would process them. These processors maintain a waiting queue therefore, the

$$\text{Total Response time} = \text{Processing time} + \text{Waiting time} + \text{Transfer time} \quad (2)$$

$$\text{Relative Processing Rate} = \frac{\text{Job's Processing rate}}{\text{Lowest Processing rate in the clust}} \quad (3)$$

$$\text{Job Generation rate} = \frac{\text{User's job generation rate}}{\text{Total job generation rate of all users}} \quad (4)$$

Having learnt the different ways the players can turn the game in a system, one of the issues that the system might face is in deciding the period it is to be refreshed. The refresh period cannot be too long or too short or else it would invite the inconsistency problem with erroneous load strategy and mistaken nodes choice. If it is long, the information may get obsolete and the system would not be able to make the right decision. Likewise, if the period is short, the frequency would be very high impacting the performance of cloud computing. The refresh period also has to be accurately planned to avoid inconsistency and errors.

Comparison of different algorithms for load balancing

The table below compares the different load balancing algorithms briefed above basis the necessary qualitative metrics in cloud computing.

Table I

Algorithm / Parameters	Round Robin	ESCE	Throttled	Task Scheduling	Load Balancing Ant Colony Optimization
Static/Dynamic	Static	Dynamic	Dynamic	Dynamic	Dynamic
Throughput	Low	Average	Average	Good	High
Response time	Low	Average	Average	Good	Good
Scalability	Low	Average	High	High	High
Priority	Low	High	High	High	High
Fault tolerance	Low	Low	Average	High	High
Overhead	High	Average	Average	Low	Low
Cost (Virtual Machine Usage/hour)	High	Average	Average	Average	Average
Power Consumption (Idle nodes off)	High	High	High	Average	Average
Complexity	Low	Low	Low	High	High
Fairness	Low	Average	Average	Average	Average
Performance	Low	Average	Average	Good	Good

Game theory is a better approach in many aspects and involves low cost for maintaining the structure. We look forward to creating a load balancing algorithm that works as a non cooperative game among users and a cooperative game among processors. The load balancing algorithm would be effective if the average job arrival rate from users is less than the average processing rate of the jobs in the cloud. Therefore, it is important to correctly allocate the distribution of jobs from the users to the processors. This is possible if we can accurately determine the time a job takes to travel from the user to the cloud which is dependent on the to the average size of the job, distance between user and processor, bandwidth available within the cloud etc., [6]

The considered factors are briefed below: [14]

Throughput – It is the amount of work that all the nodes can process in a given time period.

Response time – the elapsed time between the demand placed and the beginning of a response after completion of the job. Simply put, it is the time taken to respond by a particular load balancing algorithm. This parameter should be minimized.

Scalability – Ability of a computer application (hardware/software/service) to continue its function effectively even when its size, topography etc is changed.

Priority – Preference of tasks based on factors like cost, time, size etc.,. [15]

Fault tolerance - System designed such that it can tolerate and continue functioning despite any failure.

Overhead – refers to the processing time required by the system for installation, operation or any transaction.

Cost – it is the cost involved in configuring the system and in processing the jobs demanded by the users, better defined as the usage of the machine per hour.

Power Consumption – Information technology consumes tremendous power and involves high energy costs. Efficient power management is vital for the success of IT environment such as Cloud computing, Grid computing etc.,

Complexity – Making the entire system difficult. With increasing users associating with the cloud and its properties, the complexity of the system increases.

Fairness – indicates that each user has the equal response time and all get their jobs completed within approximately the same time.

Performance – it is the speed and accuracy at which the

jobs are completed and is measured against the preset standards. In simple words, it is the total efficiency of the system. This can be improved by reducing the task response time and waiting time maintaining a reasonable cost of the system.

3. CONCLUSION

Cloud Computing is one of the fastest growing IT fascism being widely accepted though in its nascent stage. Load Balancing has been one of the major issues in Cloud Computing. This paper has given a brief of load balancing algorithms and explains the Game theory concept for balancing load in a cloud in a better way. Cloud computing is a dynamic mechanism involving changes every day, hence the algorithm has to evolve with the on-going transformation.

A cooperative game among nodes or processors ensure minimal execution time whereas a non cooperative game between the users allow users to choose their best benefit strategy providing optimum job response time.

This is a conceptual piece of work and some more work is to be done leading to its execution and implementation. Amendments in the game-theoretic approach to load balancing would be made as per the ongoing changes in the technology of cloud computing and the problems that arise accordingly.

REFERENCES CITED

[1] "Cloud Computing", 7Th IEEE International Conference on Cloud Computing, Alaska, 2014.

[2] Ram Prasad Padhy, P Goutam Prasad, "Load Balancing in Cloud Computing System," National Institute of Technology at Rourkela, May 2011.

[3] Guiyi Wei, Athanasio V, Vasilakos, Yao Zheng, Naixue Xiong, "A game theoritic method for fair resource allocation for cloud computing services," J SuperComputers, pp. 252-269, 2010.

[4] Meenakshi, "Comparative Study of Load Balancing Algorithms in Cloud Computing Environment," IJERT, vol. 2, no. 10, pp. 2628 - 2633, 2013.

[5] Amandeep Kaur Sidhu &. Supriya Kinger, "Analysis of Load Balancing Techniques in Cloud Computing," International Journal of Computers & Technology, vol. 4, no. 2, pp. 737-741, March - April 2013.

[6] N.S.V Rao, S W Poole, Fei Hi, Jun Zhuang, Mac Y T, Yau, "IEEE Explore," 2012. [Online]. Available: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6167441&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6167441.

[7] D Wu, Y Cai, l Zhou et al , "Cooperative Strategies for energy aware ad hoc networks: a correlated equilibrium game-theoritic approach," in IEEE, 2013.

[8] Z Kong, C Z Xu & M Guo , "Mechanism design for stochastic virtual resource allocation in non cooperative cloud systems," in International Conference on Cloud Computing, Washington DC, USA, July 2011.

[9] Doddini Probhuling L, "Load Balancing Algorithms in Cloud Computing," International Journal of Advanced Computer and Mathematical Sciences, vol. 4, no. 3, pp. 229- 233, 2013.

[10] Kwnag Mong Sim, "Ant colony optimization for routing and load-balancing: survey and new directions," IEEE, vol. 33, no. 5, pp. 560-572, Sept 2003.

[11] A. T. Chronopoulos, "Game Theory Based Load Balanced Job Allocation," San Antonio. [Online]. Available: http://graal.ens-lyon.fr/~lmarchal/aussois/slides/chronopoulos.pdf.

[12] Satish Penmatsa, "Game Theory BASED JOB ALLOCATION/LOAD BALANCING IN," San Antonio, 2007.

[13] Daniel Grosu &. A. Chronopoulos, "A Game-Theoritc Model and Algorithm for Load Balacing in Distributed Systems," in 16th International Parallel and Distributed Processing Symposium (IPDPS 2002), Fort Lauderdale, 15-19 April 2002.

[14] Shanti Swaroop Moharana, Rajadeep D Ramesh & Digamber Powar, "Ananlysis of Load Balancers in Cloud Computing," International Journal of Computer Science and Engineering, vol. 2, no. 2, pp. 101-108, 2013.

[15] K C Gouda, Radhika T V & Akshatha M, "Priority based resource allocation model for cloud computing," International Journal of Science, Engineering and Technology Research, vol. 2, no. 1, Jan 2013.

Shilpa S is currently pursuing MTech in Computer science & engg. from Gogte Institute of Technology – Belgaum.,She received B.E in Computer science & engg. From Maratha Mandal's Engg College, Belgaum.Her current research interest include load balancing in cloud computing and game theory.

Prof. Shubhada.Kulkarni is working as an Asst. Prof at Gogte Institute of Technology – Belgaum., She received Masters Degree in Computer Science & Engineering from Viswesvaraya Technological University –Belgaum and has a rich teaching experience of 18 years.

Prof. Sharada.Kulkarni is working as an Asst. Prof at Gogte Institute of Technology – Belgaum., She received Masters Degree in Computer Science & Engineering from Viswesvaraya Technological University – Belgaum and has a rich teaching experience of 18 years.

