A REVIEW ON LOAD BALANCING TECHNIQUE IN THE PUBLIC CLOUD USING PARTITIONING METHOD

[1] G. DAMODAR, [2] D. BARATH KUMAR

[1]M.Tech Student, Department of CSE.
gdyadav509@gmail.com

[2]Assistant Professor, Department of CSE
bandlabarathkumar@gmail.com

ABSTRACT— Cloud computing is a provision of providing networked, on-line, on-demand accommodations pay per use substructure. Several issues as scalability, security, performance etc are discussed so far by many researchers for the cloud computing. Cloud partitioning is an optimal approach for public cloud. In public cloud environment sundry nodes are utilized with required computing resources situated in different geographic locations, so this strategy simplifies the load distribution across the multiple nodes, but fault tolerance and load balancing are most consequential quandaries obtaining high performance in the system. Load balancing is the process of distribution of workload among different nodes or processor. The purport of load balancing is to amend the performance of a cloud environment through a congruous distribution strategy. Game theory is the formal study of conflict and cooperation. Game theoretic concepts apply whenever the actions of several agents are interdependent. The game theoretic algorithms avail to obtain a utilizer optimal load balancing which ultimately amends overall performance of cloud computing. This paper introduces a more preponderant approach for public cloud load distribution utilizing partitioning and game theory concept to increment the performance of the system.

Index Terms— Load Balancing, Overload, Balancers, Public Cloud Infrastructure, Cloud partitioning.

I INTRODUCTION

Cloud Computing is a concept that has many computers interconnected through an authentic time network like internet. Cloud computing betokens distributed computing. Cloud computing enables convenient, on-demand, dynamic and reliable utilization of distributed computing resources. The cloud computing model has five main characteristics on demand accommodation, broad network access, resource pooling, flexibility, quantified accommodation. Cloud computing is efficient and scalable but to maintain the stability of processing many jobs in the cloud computing is a very conundrum. The job advent pattern cannot be prognosticated and the capacities of each node in the cloud differ. Hence for balancing the utilization of internet and cognate resources has incremented widely. Due to this there is tremendous increase in workload. So there is uneven distribution of this workload which results in server overloading and may crash. In such the load, it is crucial to control workloads to amend system performance and maintain stability. The load on every cloud is variable and dependent on sundry factors. To handle this quandary of imbalance of load on clouds and to increment its working time, and the load balancing strategies change according to

efficiency, this paper endeavors to implement "A Model for load balancing by Partitioning the Public Cloud". Good load balancing makes cloud computing more efficient and additionally amends utilizer contentment. This article is aimed at the public cloud which has numerous nodes. A system having main controller, balancers, servers and a client is implemented here. It introduces a switch mechanism to cull different strategies for different situations. This paper divides the public cloud into cloud partitions and applies different strategies to balance the load on cloud. This paper gives a conception for balancing the load on clouds. It avails to evade overloading of servers and amend replication times. The rudimental designs of the system and algorithms to implement it are described in this paper

Goals of Load Balancing

✓ To amend 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.

II LOAD BALANCING

"The load balancing technique used to ascertain that none of the node is in idle state while other nodes are being utilized". In order to balance the lode among multiple nodes you can distribute the load to another node which has lightly loaded. Thus distributing the load during runtime is kenned 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 kenned in advance, and the load balancing strategy has been made by load balancing algorithm at compile time. This load balancing strategy will be kept perpetually during runtime of the system.

In contrast, dynamic algorithm is implemented at running

the authentic verbal expression of the system. Though, the dynamic algorithm has more preponderant adaptability, it is sensitive to the precision of the load information or verbalization 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 subsisting in that cloud. So the way in which the distribution is being done must get the replication from all the slaves at the same time so that there should not be any waiting for any particular computing contrivance to reply afore further processing could transpire. But in the authentic time clouds heterogeneous computing contrivances subsists 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.
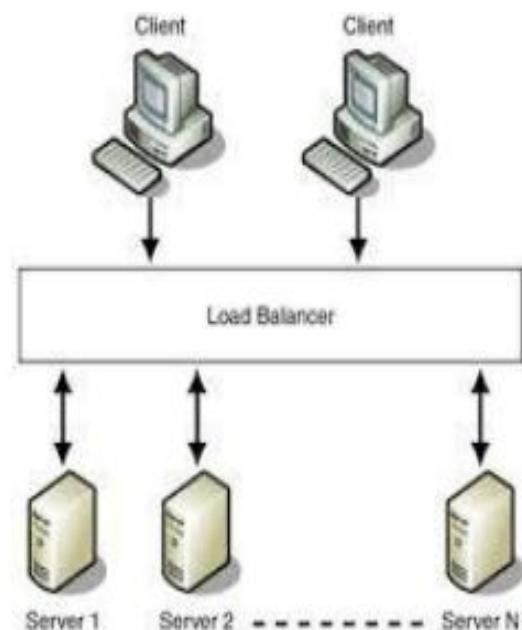

Figure: 1 – Simple view of Load Balancer

A more preponderant ratio of workload is required to be given to those with higher computation capabilities. But sometimes just higher computation power only cannot avail is deciding how much quota of the task is required should be cloud.

assigned to that system.

This assignation of felicitous task to opportune system in heterogeneous computing infrastructure is done by load balancer. Load balancer is withal responsible for 'Priority Activation' which betokens that when the number of slave computing contrivance contrivances drops below a certain point the load balancer must wake some of the slumbering contrivances to maintain the computing performance for the client. Load balancing in cloud computing was described in by Adler. There are many load balancing algorithms, such as Round Robin (RR), Equally Spread Current Execution Algorithm (ESCEA), and Ant Colony algorithm (ACA). Nishant et al. utilized 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 ESCEA algorithm and throttled algorithm are more preponderant than the Round Robin algorithm. Some of the classical loads balancing methods are homogeneous 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 utilized here because it is fairly simple.

III CLOUD PARTITIONING MODEL

A public cloud is one predicated on the standard cloud computing model, in which an accommodation provider makes resources, such as applications and storage, available to the general public over the Internet. Public cloud is composed of several nodes situated in deferent geographic location. Cloud partitioning is a method to make partitions of astronomically immense public cloud is some segment of
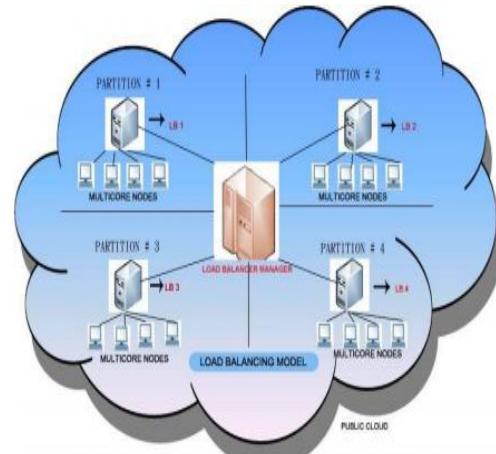

Figure – 2: Cloud partitioning model

A cloud partition has several nodes belongs to a particular area, these subarea of the public cloud predicated on the geographic locations. These subareas are considered to be as cluster of nodes with a load manages. Model of the cloud partitioning is depicted on the following figure. In this model load balancing is implemented in two steps, in first steps public cloud is partitioned into four subarea denominated Partition#1, Partition#2, Partition#3 and Partition#4. Each partition has a Load balancer (LB) associated with multiple multi-core nodes. There is a main controller system which manages the load balancer called Load Balancer Manager (LBM). After partitioning the public cloud into different partitions, load balancing then commences.

A. Working of Load Balancer Manager (LBM):

In this model, Load Balancer Manager (LBM) is responsible for the following task

- Receives the jobs from different end users.
- Choose a categorical partition for the received jobs.
- Check the status of the cloud partition (Status may be in one of these: IDLE, NORMAL, and HEAVY).
- If the partition Status= HEAVY the no allocation eligible for the processing.

- will be done, it signifies all nodes are overloaded already.

If partition Status= IDLE or NORMAL then forward the Jobs to the respective Load balancer (LB). Now the load balancer activated and commences its work.

B. Possible status of Cloud Partition:

Cloud partition can be in one of the following three statuses –

- IDLE – In this, most of the nodes are in idle state.
- NORMAL – In this status, some of the nodes are in idle status while some others are overloaded.
- HEAVY – In this status of the cloud partition, most of the nodes are overloaded.

Any node is available to get job for execution only if it returns to mundane status. Load Balancer Manager checks the status of the partition and dispatches the utilizer requests to the partition which is in IDLE status.

C. Calculation of Load Degree (LD) for a node:

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

LD (N) = $\sum X_i * F_i$

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

Average Load Degree (LD) of the cloud partition will be calculated as n

Avg_ LD= $\sum$ LD (N_i)/n

D. Possible load status of node:

According to the calculation of load degree for the node three loads 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 culled by the Load Balancer Manager and likewise any node having the Load Degree (LD)= OVERLAODED is not

Only cloud partition having IDLE or NORMAL load status and Node having IDLE or NORMAL load degree are culled for scheduling and load balancing.

IV EXPERIMENTAL RESULTS

Go to main controller application then enter the partition size (like how many partitions to be created into the cloud). After click on start partition, Load balancer after creating the partitions.
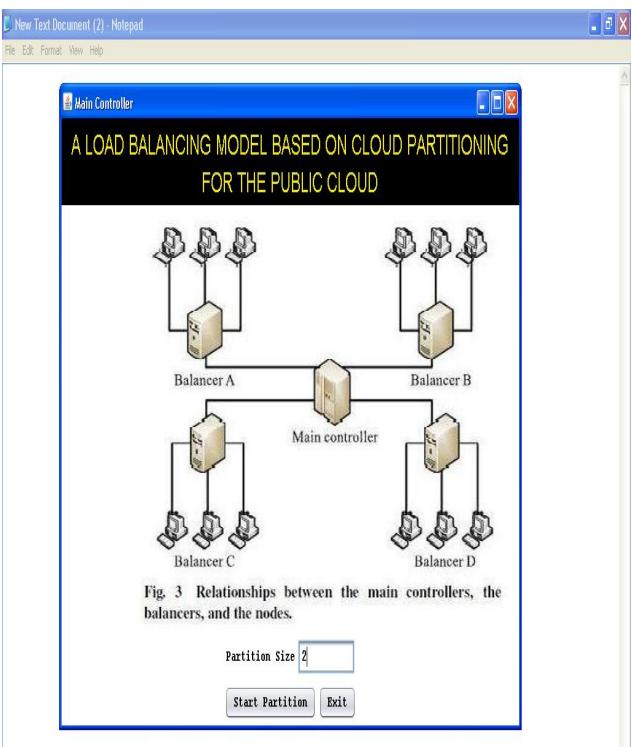
Fig. 3 Relationships between the main controllers, the balancers, and the nodes.

contentment. Cloud partitioning is a method to make
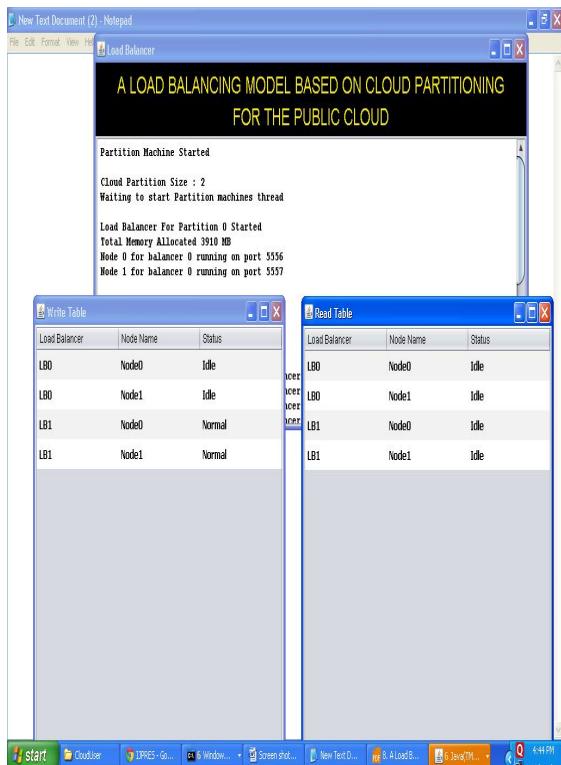
Go to cloud user application, then run the application:

After starting the client application, after 3 seconds it automatically sends 20 requests, First the request will come to load balancer then it identifies the load onto the available partition servers then it map the request to a node which has less load. Updating the read and write tables after receiving and processing the requests. This is the final output.

CONCLUSION

The replication time and data transfer cost is a challenge of every engineer to develop the products that can increase the business performance and high customer gratification in the cloud predicated sector. The several strategies lack efficient scheduling and load balancing resource allocation techniques leading to incremented operational cost and give less customer gratification. Load balancing in the cloud computing environment has a paramount impact on the performance. Good load balancing makes cloud computing more efficient and amends utilizer

partitions of immensely colossal public cloud is some segment of cloud. A public cloud is one predicated on the standard cloud computing model, in which an accommodation provider makes resources, such as applications and storage, available to the general public over the Internet. Public cloud is composed of several nodes situated in deferent geographic location. Cloud partitioning is a method to make partitions of sizably voluminous 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 conventionally called a decision quandary.

FUTURE WORK

This is the matter of fact that the developed system represents a noble scheme of load balancing in public cloud infrastructure, but still it requires to be incorporated with authentic time functional scenario in authentic time cloud computing platform. Since, the developed scheme represents a conceptual framework; it still possesses opportunities for further enhancement in terms of authentic time application predicated system development, implementation of load balancing with certain authentic time cloud framework.

REFERENCES

[1] Yu-lung Lo and Min-Shan Lai, "The Load Balancing of Database Allocation in the Cloud", Proceedings of the International Multi Conference of Engineers and Computer Scientists IMECS 2013, Vol I, , March 13 - 15, 2013, Hong Kong.

[2] Belabbas Yagoubi, Meriem Meddeber, "Distributed Load Balancing Model for Grid Computing", Revue ARIMA, Vol. 12, 2010, pp. 43-60.

[3] Venubabu Kunamneni, "Dynamic Load Balancing for the Cloud", International Journal of Computer Science and Electrical Engineering (IJCSEE), ISSN No. 2315-4209, Vol. 1, Issue 1, 2012.

[4] N. G. Shivaratri, P. Krueger, and M. Singhal, "Load distributing for locally distributed systems", Computer, Vol. 25, No. 12, December. 1992, pp. 33-44.

[5] S. Penmatsa and A. T. Chronopoulos, "Game-theoretic static load balancing for distributed systems, Journal of Parallel and Distributed Computing, Vol. 71, No. 4, April. 2011, pp. 537-555.

[6] D. Grosu, A. T. Chronopoulos, and M. Y. Leung, "Load balancing in distributed systems: An approach using cooperative games", in Proc. 16th IEEE Intl. Parallel and Distributed Processing Symp., Florida, USA, April. 2002, pp. 52-61.

[7] K. Nishant, P. Sharma, V. Krishna, C. Gupta, K. P. Singh, N. Nitin, and R. Rastogi, "Load balancing of nodes in cloud using ant colony optimization", in Proc. 14th International Conference on Computer Modelling and Simulation (UKSim), Cambridge shire, United Kingdom, March. 2012, pp. 28-30.

[8] Adhikari, J., Patil, S, "Double threshold energy aware load balancing in cloud computing" Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013 pp.1 - 6

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

[10] Anthony T. Velte, Toby J.Velte, Robert Elsenpeter, "Cloud Computing A Practical Approach", TATA McGRAWHILL, 2010.[13] Ali M. Alakeel, "A Guide to Dynamic Load Balancing in Distributed Computer Systems", IJCSNS International Journal of Computer Science and Network Security, Vol.10 No.6, June 2010.