Dynamic Load Balancing Techniques for Improving Performance in Cloud Computing

Srushti Patel
PG Student,
S.P.College of engineering,
Visnagar, 384315, India

Hiren Patel, PhD
Professor, S. P. College of Engineering
Visnagar,384315,India

Nimisha Patel
Research Scholar,Rai University, Ahmedabad
Associate Professor
S. P. College of Engineering
Visnagar,384315,India

ABSTRACT
Cloud Computing is an emerging area in IT sector which enables a wide range of users to access distributed, scalable, virtualized hardware and/or software, applications and platforms are provided over the Internet. Cloud Computing is a shared pool of Configurable computing resources which require the proper distribution of dynamic workload among multiple computers to ensure no single node is underloaded or overloaded. Load Balancing aims to reduce response time of jobs, increase overall performance, reduce communication cost of servers, Resource optimization, maintain cost of VMs, Maximize throughput and avoid overload of any single node. In this paper we discuss the various techniques related to Load Balancing in Cloud Environment and further we propose a modified agent based technique which is used for Balancing a load of the all host and also manage the new arrival jobs to increase the overall performance of system.

Keywords
Cloud Computing, Load Balancing techniques, Dynamic workload Distribution, Resource utilization.

1. INTRODUCTION
The National institute of standard and technology(NIST) that defines the, "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. This cloud model is composed of five essential characteristics(On-demand self-service, Broad network access, Resource pooling, Rapid elasticity, Measured Service), three service models(SAAS, PAAS, IAAS), and four deployment models(Public, Private, Hybrid, Community) [1].

This technology is used for spreading large amount of datasets and files around the world. Handling such type of large amount of datasets that require the some techniques for optimizing the overall performance and user satisfaction[2]. Therefore the load balancing in Cloud Computing is major issues. Load balancing technique is used for improving the overall performance of the system by Distributing the total workload among multiple computing resources or data centers, network links, central processing units, disk drives, on the cloud Server, Processor. This load balancing techniques helps to achieve optimal resource utilization, maximize throughput, minimize response time, and avoid overload of resources.

One of the crucial issue related with Load balancing is to removing the condition in which some of the nodes are over utilized or some of the nodes are under utilized. Improper workload is distributed among the all computing resources and simultaneously removing a condition in which some of the nodes are over loaded while some others are under loaded. Other issues are like (a) resource utilization (b) scalability (c) energy efficiency (d) latency (e) throughput (f) performance (g) money (h) Achieving green computing. However with proper load balancing technique we can reduce resource consumption which is not only helps in reducing cost but making enterprises greener[3,10,11].

The rest of this paper is organized as follows. Section 2 presents the Background terminology with energy model. Section 3 presents the Related work or existing Load Balancing Techniques. Section 4 presents the comparison and discussion of load balancing techniques with the all performance metrics. Section 5 proposed method and Section 6 presents the Conclusion and future work related with this research.

2. BACKGROUND
In this Section we classify the Load balancing techniques in mainly two categories: static algorithms and dynamic algorithms [6, 12] that have been developed for cloud computing which is shown in fig.1.

Static load balancing algorithms assign the tasks to the nodes to process new requests. The process is based solely on prior knowledge of the node's properties and capabilities. Static algorithms do not change the attribute like node's processing power, memory and storage capacity at run time[2].

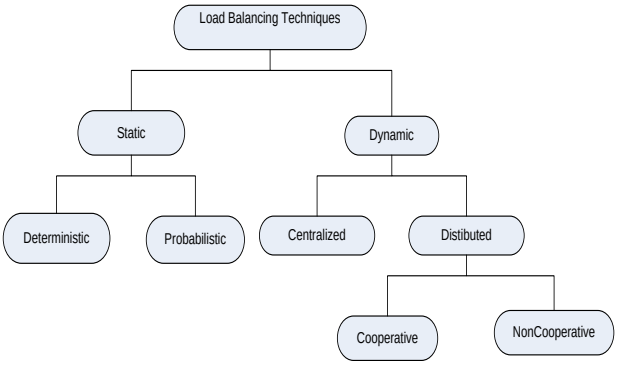

Fig.[1] Classification of Load balancing techniques.[13]

Static load balancing algorithms assign the tasks to the nodes to process new requests. The process is based solely on prior knowledge of the node's properties and capabilities. Static algorithms do not change the attribute like node's processing power, memory and storage capacity at run time[2]. There are some problem with static load balancing i.e. in the long run, static weight cannot be corrected and the node is bound

to deviate from the actual load condition, resulting in load imbalance so it can't handle long-connectivity applications well. Most of the Dynamic load balancing algorithms relies on a combination of knowledge and run-time properties. These algorithms assign the tasks to a node and may dynamically reassign them to another node based on the attributes gathered and calculated. Dynamic load balancing algorithms are more accurate and could result in more efficient load balancing. Such Algorithms require constant monitoring of the nodes and task[6,12].

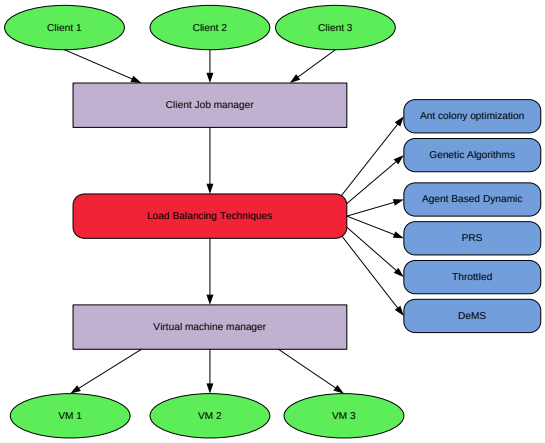

Fig.2 Load Balancing techniques in cloud computing [8]

Here fig.2 shows that the all Load balancing Techniques which is used for balancing the overall workload of all VMs into the system. Job manager having a several VMs, using this list of VM it assign the desire job to the appropriate VM. If not any VM is free at that time the job manager wait for the client request and place that job into queue for the fast processing[8].

A. Energy Model

Energy is the capacity to do the work. Energy consumption in data centers by computing nodes are mostly determined by the physical resources such as CPU, memory, disk storage, and network interfaces. This energy model is create on the basis of that processor utilization has a linear correlation with energy consumption. Measure the energy consumption for a particular task, to use the information as its processing time and processor utilization is sufficient[15]. For host Hi at any given time, the utilization Ui is Defined as,

$$U_i = \sum_{j=1}^{M_i} (u_{ij})$$

3. RELATED WORK

This section gives a brief review about the various existing Load Balancing Algorithms Which is used for Balancing the overall load of any Host in cloud computing environment.

Kum li et al.[4] suggested an algorithm called Load Balancing Ant Colony Optimization(LBACO) based on task scheduling policy. This algorithm used to finding out the optimal resource allocation in Dynamic cloud system in complex network. The LBACO algorithm chooses optimal resources to perform tasks according to resources status and the size of given task in the Cloud environment.

Kousik Dasgupta et al.[5] proposed an algorithm known as genetic algorithm based load balancing algorithm has been used as a soft computing approach, which uses the mechanism of natural selection strategy. The Advantages of this techniques is that it can handle a vast search space, applicable to complex objective function and can avoid being trapping into local optimal solution. It also guarantees the QOS requirement of customer job.

Jitender grover et al.[6] proposed an Agent based dynamic load balancing(ABDLB) scheme for cloud computing. After comparing with traditional algorithm the advantages of ABDLB algorithm is CPU time consumption is 1 unit whereas in traditional algorithm CPU time consumption is 10 unit.

Huangke Chen et al.[7] present a novel scheduling algorithm Proactive and Reactive Scheduling(PRS) which is dynamically exploits for scheduling real time, aperiodic, independent task. Benefits associate with this algorithm is, It can prohibit propagation of uncertainties throughout the schedule, This design allows each task waiting on LQ to start as soon as its preceding task has finished, so the possible execution delay for anew task is removed. This design enables overlapping of communications and computations overlapped to save time and improve scheduling performance[6,13], It also can reduce the overheads of task transfer among hosts when corresponding VMs need to migrate.

Vibhore Tyagi et al.[8] can propose the load balancing technique with throttled Algorithm to reduce the cost and response time across VM's in multi data center and optimize response time service broker policy. Throttled policy defines the work to finding the applicable virtual machine for assigning individual job.

Yu Liu et al.[9] propose a DeMS consist of hybrid scheme of task scheduling and load balancing technique having three algorithms,(1)On-Demand Scheduling, (2)Querying and Migrating task(QMT), (3)Staged task migration(STM).

4. COMPARISON AND DISCUSSION:

Load balancing in cloud computing is distributing the workload among the all node and transfer the load from heavily node to idle node. It helps to improve response time, migration time, throughput, Resource utilization, scalability and overall performance.

Table 1. Comparison Of Different Load Balancing Techniques With Various Metrics.

Metrics/tech.	Nature	Environment	Response Time	Migration time	Resource utilization	Performance
LBACO[4]	Dynamic	decentralized	-	-	high	Less
GA[5]	Dynam	Centralized	Less	-	-	High
ABDLB[6]		Centralized	Less	less	-	High(Less CPU time)
PRS[7]	Dynam	decentralized	Less	less	high	High

Throttled[8]		decentralized	Less	less	high	High
DeMS[9]		Centralized	Less	hig	-	Less

Response time: It is the time interval between sending request and receiving response. This time should be minimum for increasing the performance.

Migration time: It is time taken to migrating the task or transfer the task from one node to another. Minimum time having the maximum performance.

Throughput: This metrics is used to estimate the task, whose execution complete successfully. For increasing the performance, increase the value of this metrics.

Resource utilization: it is used to insure that the utilization of system resources. Better load balancing techniques give the better resource utilization.

Scalability: It determines the ability of the system to accomplish load balancing algorithm with a restricted number of nodes.

Performance: It represents the effectiveness of the system after performing the load balancing algorithm. When the above metrics satisfy optimally then the overall system performance will be increases.

5. OUR CONTRIBUTION

Grover et al [6] proposed a system Architecture which consists 'n' numbers of clients connected with cloud service providers via internet and service provider consists VM, managed unit and 'm' numbers of shared pool of resources which are considering as servers. At the shared pool of Hosts, agent complete one cycle in two walks:

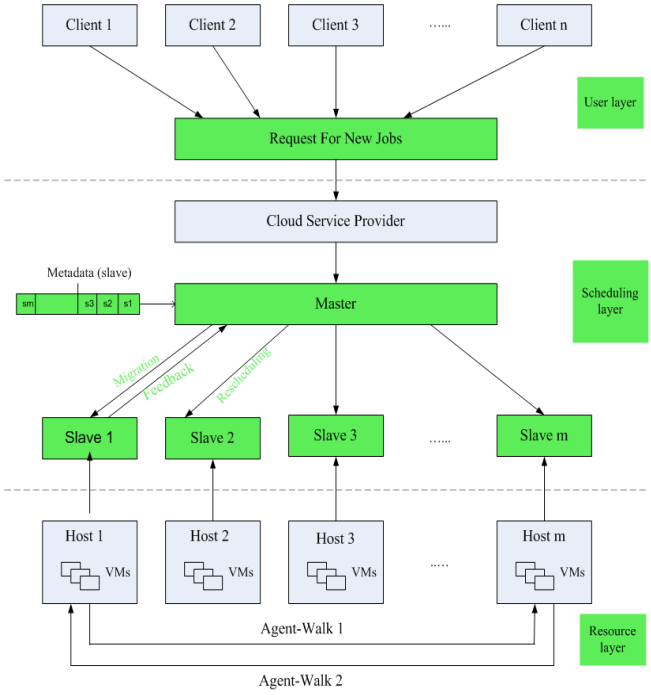

Fig. 3 System Architecture[6]

- In First walk it moves from first server to last server and gathers information from all servers, for making decision for Load Balancing and
- In Second walk it balances the host's load on the basis of Standard deviation method

Here, our contribution has been illustrated by green boxes. We divide this architecture in three layers, User layer, Scheduling layer and Resource layer. Also we use the master slave mechanism for migrating the jobs from overloaded server to underloaded server and for managing a new arrival jobs. Agent store the information of all host into that Slave and at the end all the Slave Metadata is stored into the Master.

Agent Walk1:
In the Agent walk1 Grover et al[6] describes the model in terms of flow chart. The working of the model can be explained in five steps.

Step1: Agent is activated at any random server and finds number of jobs in queue at that server.

Step2: Agent will repeat this process for all servers of shared pool.
Step3: After that it will calculate AVERAGE.
Step4: On the basis of AVERAGE, it will sense the server's status in terms of overloaded and underloaded.
Step5: Server's status will be decided as follows.

- AVERAGE, then transfer the server's status as overloaded.
- If the number of jobs at i^{th} server is less than the AVERAGE, then transfer the server's status as underloaded.

At the step1 we recommends to,

- Calculate, mean of utilization using

$$S = \sum_{j=1}^{M_i} (u_{ij})$$

- Calculate, mean of utilization of Hi

$$E = \frac{1}{P} \sum_{k=1}^{\mathrm{p}} (X_k)$$

Whereas, X_k is sum of utilization of VMs on host Hi in time frame k.

P is the total time frames.

- Calculate, variance of utilization for host Hi

$$V = \frac{1}{P} \sum_{k=1}^{\mathrm{p}} (X_k - \mathrm{E})^2$$

- The standard deviation equals to the square root of V.

At the Step3 we contribute, instead of calculating the AVERAGE, we calculate the predicted utilization using standard deviation method[15]. Because of calculating AVERAGE of all job into the queue it's not enough for finding the server is overloaded or underloaded we use this standard deviation method. Understanding of my best which is better than calculating only AVERAGE of queuing jobs.

PU = E+S*StdDev

At Step4 we use 'PU' for deciding the host is overloaded or underloaded. If the If 'PU' is greater than the current utilization of host than host status is overloaded otherwise unedrloaded and this information is stored into the slave.

At the last step5 Information of all Slave is stored into the Master.

Agent Walk2:

In the Agent walk2 Grover et al[6] describes the model in terms of flow chart. The working of this model can be explained below.

Agent will start backtracking from last server to first server for balancing load of servers. At each server it will check the condition. If the status of server is overloaded than transfer the jobs to underloaded server otherwise receive the jobs from overloaded server. Continue this process until the first server.

Here, we recommends to use master slave mechanism. Master have the all information about the slave. Agent is currently at the Master and ready for balancing the load and also assigning the new jobs. First, agent will check the state of slave in master having the id and state. For each host check the condition, if the host_slave_state = overloaded then master send a migration request to the slave. If slave send a positive response to the master then migrate jobs to underloaded host. If the host_slave_state = underloaded then Receive jobs from "overloaded" host or new job is directly assign on that host. Agent will perform this operation until it reaches at the first host with balancing all host's load including first server also.

6. CONCLUSION AND FUTURE WORK

Load balancing is the major issue in cloud environment. Cloud load balancing is the process of distributing workloads across multiple computing resources or data centers, network links, central processing units, disk drives, on the cloud. In this paper, we analyze various techniques for load balancing in cloud computing. We discussed the advantages and disadvantage of this algorithm. Using this technique, improper workload can distributed among the all nodes which are idle. It help to achieve the user satisfaction by improving the metrics like, Response time, migration time, throughput, resource utilization, Scalability, and overall performance of the system. Also using this load balancing techniques we can reduce the energy consumption and carbon emission to making environment greener. Also we propose a modified agent based technique which is used for Balancing a load of the all host and also manage the new arrival jobs.

As a future research direction, we implement this technique in real cloud environment and also we can design more efficient algorithm which will maintain a better trade-off between all the metrics of algorithm using,

a) A combination of two or more existing techniques

b) Improvement in one of the available techniques or

c) Completely a new approach.

7. REFERENCES

[1] Mell, P., & Grance, T. (2011). The NIST definition of cloud computing.

[2] Nuaimi, K. A., Mohamed, N., Nuaimi, M. A., & Al-Jaroodi, J. (2012, December). A survey of load balancing in cloud computing: challenges and algorithms. In Network Cloud Computing and Applications (NCCA), 2012 Second Symposium on (pp. 137-142). IEEE.

[3] Sreenivas, V., Prathap, M., & Kemal, M. (2014, February). Load balancing techniques: Major challenge in Cloud Computing-a systematic review. In Electronics and Communication Systems (ICECS), 2014 International Conference on (pp. 1-6). IEEE.

[4] Li, K., Xu, G., Zhao, G., Dong, Y., & Wang, D. (2011, August). Cloud task scheduling based on load balancing ant colony optimization. In Chinagrid Conference (ChinaGrid), 2011 Sixth Annual (pp. 3-9). IEEE.

[5] Dasgupta, K., Mandal, B., Dutta, P., Mandal, J. K., & Dam, S. (2013). A genetic algorithm (ga) based load balancing strategy for cloud computing. Procedia Technology, 10, 340-347.

[6] Grover, J., & Katiyar, S. (2013, August). Agent based dynamic load balancing in Cloud Computing. In Human Computer Interactions (ICHCI), 2013 International Conference on (pp. 1-6). IEEE.

[7] Chen, H., Zhu, X., Guo, H., Zhu, J., Qin, X., & Wu, J. (2015). Towards energy-efficient scheduling for real-time tasks under uncertain cloud computing environment. Journal of Systems and Software, 99, 20-35.

[8] Alakeel, A. M. (2010). A guide to dynamic load balancing in distributed computer systems. International Journal of Computer Science and Network Security (IJCSNS), 10(6), 153-160.

[9] Mata-Toledo, R., & Gupta, P. (2010). Green data center: how green can we perform. Journal of Technology Research, Academic and Business Research Institute, 2(1), 1-8.

[10] Lee, R., & Jeng, B. (2011, October). Load-balancing tactics in cloud. In Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2011 International Conference on (pp. 447-454). IEEE.

[11] Hu, M., & Veeravalli, B. (2013). Requirement-aware strategies for scheduling real-time divisible loads on clusters. Journal of Parallel and Distributed Computing, 73(8), 1083-1091.

[12] Sinha, P. K. (1998). Distributed operating systems: concepts and design. PHI Learning Pvt. Ltd..

[13] Cao, Z., & Dong, S. (2012, December). Dynamic VM consolidation for energy-aware and SLA violation reduction in cloud Computing. In Parallel and Distributed Computing, Applications and Technologies (PDCAT), 2012 13th International Conference on (pp. 363-369). IEEE.