A Survey on Existing Load Balancing Techniques

Trushar Gavit

Information Technology Department

Silver Oak Collage of Engineering and Technology

Ahmedabad, India

Asst. Prof. Brona Shah

Computer Engineering Department

Silver Oak Collage of Engineering and Technology

Ahmedabad, India

Abstract—**Cloud computing is emerging technology that uses virtualization of resources to fulfill user demands. Since the number of users are increasing day by day, the workload on cloud is hard to manage. There are some issues in cloud that should be taken to provide batter services on cloud such as load balancing, fault tolerance, security, migration etc. Load balancing is focused on distributing workload among different resources and utilizing resources to its maximum level for better performance. The paper is focused on the survey of different load balancing techniques that are currently used in cloud.**

Keywords— Cloud computing; Load balancing; Survey; Resource utilization;

I.   INTRODUCTION

Cloud is technology that is going to replace the traditional computing technology. It is the development of parallel computing, distributed computing grid computing, and is the combination and evolution of Virtualization, Utility computing, Software-as-a-Service (SaaS), Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS) [1]. Cloud can be described as a space on the web where data, applications, operating system, processing power, and storage exist and ready to be shared among users. Pay-per-use-On-Demand mode is used by the cloud computing which accesses shared resources through the Internet. Resources include server, network, application, storage, service and so on. Deploying these resources can be quick and easy and requires the least management. Cloud computing is more reliable than other computing technologies such as grid computing. The growth of cloud increases the availability of resources. Cloud creates benefit for users by avoiding the cost to buy new resource equipment that can be costly.
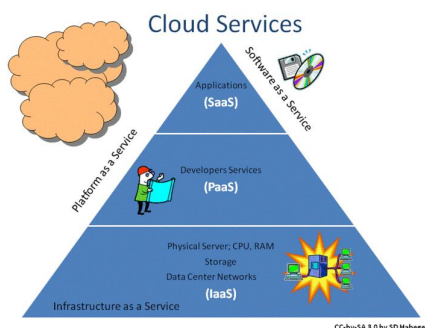
There are three types of services provided cloud as given in [1]:

- Software as a Service
- Platform as a Service
- Infrastructure as a Service

Software as a Service (SaaS): Cloud developers release their applications in a hosting environment that can be accessed via the Internet from client applications (Web browsers, PDA, etc.) by the end user. This application needs to be organized in a single environment to optimization in terms of security, availability, speed, disaster recovery, and maintenance. SaaS can be Office365, Google Mail, Google Sheets, etc.

Platform as a Service (PaaS): PaaS provides a platform for developers to build and test cloud applications and services. So the only difference between SaaS and PaaS is that the SaaS hosts cloud applications that are completed where PaaS provides a platform that hosts both in-progress and completed cloud application. To achieve these functionalities PaaS supports programming environment, application hosting environments, configuration management, tools and so forth. An example of PaaS is Google AppEngine.

Infrastructure as a Service (IaaS): End users directly use resources (processing, storage, networks and other fundamental computing resources) provided by the cloud service provider in the IaaS. To integrate/decompose infrastructure resources to meet increasing or decreasing demand virtualization used. The idea of virtualization is to isolate independent virtual machines (VM) from other VMs and the underlying hardware. Amazon's EC2 is an example of IaaS.

II.   LOAD BALANCING
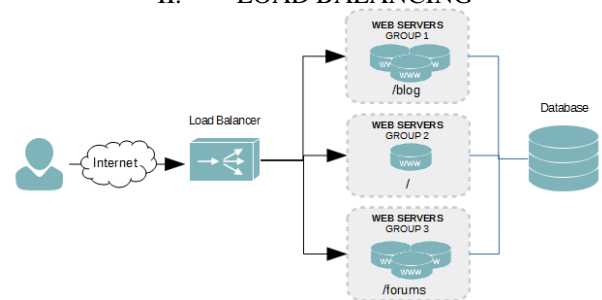
Load balancing is the process of reassigning the total load to the individual nodes of the system to achieve efficient resource utilization and to improve the response time of the jobs. Load balancing algorithms can be classified into two types (1) static and (2) dynamic. Dynamic load balancing algorithm considers the current state of the system to improve the response time. Static on another hand, works in predetermined instruction. Depending on who initiated the process, load balancing algorithms can be divided into three types as sender Initiated, receiver Initiated and symmetric.

Load balancing can be classified in two types [2]:

- Static load balancing
- Dynamic load balancing

Static load balancing: - Static load balancing uses predetermined instruction which cannot be changed if the system environment

changes. It does not depend on the current state of the system. It aims to minimize the execution time and reduce the communication overhead and delays. The drawback of static load balancing is that it cannot shift the load during execution to another underutilized virtual machine for balance the load.

Dynamic load balancing: - In dynamic load balancing, load balancing is done based on current state of the system, no prior knowledge is needed. These algorithms are more flexible and easily adapt changes and provide better performance [2]. It allows to move load from over utilized VM to underutilized VM dynamically for better execution.

Load balancing considers some parameters to measure effectiveness of load balancing algorithm in cloud.

Fault tolerance: Any kind of failure in system should not affect the entire system. Load balancing should be capable to handle failures and do not interrupt process in system.

Resource utilization: Resources should be utilized to handle in overloaded, under-loaded or idle condition. Allocation of resources should be consistent.

Response time: The algorithm should execute multiple task in minimum completion time.

Scalability: The algorithm must be scalable according to current condition in the system.

III.   LOAD BALANCING TECHNIQUES

In this paper we have considered few load balancing techniques and a brief detail about it.

Ant colony optimization: The author has considered the behavior of ant to find the food. Ants drops pheromone at some point on their path. This pheromone helps ants to come back to their colony. In colony optimization, mobile agents are used. Each agents maintain a pheromone table that contains path from source to destination. The table is updated at each new location. The distance value in the table allow other agents to choose shortest path among all paths. Author has used this concept to fulfill demands by finding resource's free capacity [3].

Honey bee foraging: Honey bee foraging algorithm is inspired from the behavior of honey bees. Honey bees perform a unique dance when they found food which is called waggle dance to tell the quality and quantity of the food to the beehive. Distance can be determined from the dance. In load balancing, the dance is advert board which is also used to advertise the overall colony profit. Servers here are grouped under the virtual server with their own virtual queue. Each server processing the demand from its queue first calculates the profit which is analogous to the quality that bees show in waggle dance. In load balancing this profit or the waggle dance is equal to the amount of time required to fulfill the request or the resources used to fulfill the request [4].

Min-min: The algorithm first evaluates the completion time of the unassigned tasks that are in queue. It's a static load balancing algorithm. Instruction to execute the tasks is already decided by the algorithm. The algorithm first assigns the task having minimum execution time. The tasks with maximum execution time have to wait for an unspecific period of time. The algorithm is efficient when a number of tasks having smaller completion time is more than the tasks having large completion time [5].

Max-min: It is developed to overcome the drawback of the min-min algorithm. The algorithm first assigns tasks with maximum execution time. It's also a static algorithm. Here tasks with minimum completion time have to wait for an unspecific period of time. The algorithm is efficient when a number of tasks having maximum completion time is more than tasks with minimum completion time [5].

Round robin: Static based load balancing algorithm which selects the first node randomly and allocates jobs to all other nodes in a round robin fashion. There is no priority, the tasks are assigned in circular order to the processors. Due to the non-uniform distribution of load, It is not suitable for cloud computing. The algorithm causes some nodes to be overloaded and some nodes to be under-loaded because the running time of any process is not known in advance. The weighted round robin overcomes this limitation. On basis of weight on the node, it will receive appropriate number of requests. This algorithm is not preferred because prior prediction of execution time is not possible [6].

Genetic Algorithm: Genetic algorithm is a search and optimization technique which can also be used for load balancing in cloud. The algorithm initializes with random population of chromosomes. Genetic operators such as selection, crossover and mutation is applied to generate new offspring. It uses fitness function as the base to find optimal solution. This algorithm tries to balance the load of the server and also try to minimize the time taken to fulfill the processing requests. Algorithm works better in comparison with the existing techniques like First come first server (FCFS) and Round Robin (RR) [7].

Join Idle Queue: The algorithm is used for distributed load balancing in large system. The Join Idle Queue algorithm is developed to overcome limitations of Power-of two algorithm to avoid communication overhead between the dispatcher and the processors at job arrival. The author analyzed that the JIG algorithm produces 30 fold reductions in the queuing overhead as compared to the Power-Of-Two algorithm at medium and high load [8].

Shortest Job Scheduling Algorithm: Shortest job scheduling algorithm selects first the task with shortest execution time. The algorithm reduces waiting time for the queued processes. In this paper author simulated various number of jobs with different execution time and analyzed the results for the model. Results are better and shows that the shortest job first algorithm decreases the average waiting time for the jobs [9].

Stochastic Hill Climbing: In comparison to hill climbing in which algorithm chooses the steepest uphill move, stochastic hill climbing chooses randomly form the uphill moves with effective probability. In this paper, author used the local optimization approach stochastic hill climbing for allocation of incoming jobs to the virtual servers or servers. Simulation is done in CloudAnalyst and the performance is analyzed both quantitatively and qualitatively and made a comparison with Round Robin (RR) and First Come First Serve (FCFS) [10].

The two phase scheduling load balancing algorithm: The algorithm is a combination of Opportunistic Load Balancing (OLB) and Load Balance Min-Min (LBMM) to maintain the load balancing system and for better execution of tasks. OLB scheduling algorithm focused on node to keep them working to achieve load balancing and LBMM scheduling algorithm is focused to minimize the execution of time of each task on the node. The algorithm utilizes the resources in more efficient manner [11].

Active clustering load balancing Algorithm: The algorithm creates group of similar nodes. Match maker node concept is used for the process of grouping. Match maker node forms connection between its neighbors which is like as the initial node. Then the matchmaker node disconnects the connection between itself and the initial node. The above set of processes is repeating again and again. The performance of the system is increases on the basis of high availability of resources, because of that, the throughput is also increasing. This increase in throughput is because of the efficient utilization of resources [12].

Particle swarm optimization: PSO is inspired from the behaviors of bird flocking or fish schooling. PSO is also an evolutionary algorithm which is based on population of solutions. A group of birds are randomly searching food in an area. There is only one piece of food in the area being searched. All the birds do not know where the food is. But they know how far the food is in each iteration. So the best strategy here id to follow the bird which is nearest to the food. It

considers two values pbest and gbest. The pbest value indicates the personal best fitness value of particle in population and gbest indicates the global best fitness value among all particles. In each iteration the particle velocity and position will be updated. The algorithm search for the gbest until the termination condition has been reached or optimal solution has been found [13].

IV. PERFORMANCE METRIC FOR LOAD BALANCING

There are some simulation tools which can be used for the performance evaluation.

Real-time platforms:

The load balancing algorithms can be executed in real-time platforms provided by IT companies.

Amazon EC2: It is commercially available web service platform. It is based on rental system to provide resources to the customer. It consist of processing, storage, web services that can be offered to the customers. EC2 provides a computing environment which helps customers to use web service interfaces to operate different operating system by launching instances.

Open stack: It is an open source cloud computing software that provides infrastructure as a service for public and private cloud. Several components of open stack are Nova, Swift, Keystone, Cinder, Quantum, and Glance.

- Keystone: The main purpose of the keystone is used for authentication and authorization of identity. It manages the roles like operator, admin, tenants etc.

- Horizon: It provides the web based interface to users to interact with open stack services like managing images, object storage and VM computation.

- Cinder: It provides the persistent storage in the form of volumes in the Virtual Machine. Cinder provides storage with high availability, fault tolerance.

- Glance: It is a catalogue for virtual disk images. To launch an instance images are utilized.

- Quantum: It deals with the network services. Quantum is responsible for the communication between the interface devices.

Eucalyptus: It is an open source platform which is used to manage the complexity and heterogeneity of large and distributed infrastructures. It consist of certain components such as cloud controller, cluster controller, node controller, and storage controller.

- Cloud Controller: It is an entry point into the cloud for administrators, developers, project managers and end users. Eucalyptus cloud only have one Cloud Controller.

- Cluster Controller: It basically runs on the host machine. Cluster controller gathers information about the node controller and schedules virtual machine on the node controller.

- Storage Controller: The main purpose of the Storage controller is to store the persistent data and past VM termination.

- Node Controller: Node controller executes on any machine that consist of VM instances. It control VM activities like execution, inspection and termination of instances of VM.

- Walrus: Walrus is used to create, list, delete buckets or to delete the objects.

V. COMPARISON OF ALGORITHM

	Throughput	Response time	Fault tolerance	Scalability	Performance	Resource utilization	Overhead
Ant colony optimization	No	No	No	No	Yes	Yes	No
Honey bee foraging	No	No	No	No	No	Yes	No
Min-Min	Yes	Yes	No	No	Yes	Yes	Yes
Min-Max	Yes	Yes	No	No	Yes	Yes	Yes
Round-robin	Yes	Yes	No	No	Yes	Yes	Yes
Genetic algorithm	No	No	No	No	Yes	Yes	No
Join Idle Queue	No	Yes	No	No	Yes	No	Yes
Shortest Job Scheduling Algorithm	No	No	No	No	No	Yes	Yes
Stochastic Hill Climbing	Yes	Yes	No	No	Yes	Yes	No
The two phase scheduling load balancing algorithm	Yes	Yes	No	No	Yes	Yes	Yes
Active clustering load balancing Algorithm	No	No	No	No	No	Yes	Yes
Particle swarm optimization	No	No	No	No	Yes	Yes	No

CONCLUSION

Cloud computing provide services over network to the user. User only needs to pay for the services that they have used. For better experience of these services load balancing is required in cloud to handle workload among resources. In this paper, different load balancing techniques have been discussed that are already exists. Also these algorithms are compared with different parameters. Due to every algorithms as some limitations, different load balancing algorithms are used for different cloud environment.

REFERENCES

[1] Prof. S.B. Bele, "An Empirical Study on 'CLOUD COMPUTING'", IJCSMC, Vol. 7, Issue. 2, February 2018, pg.33 – 41.

[2] Monika Kushwaha, Saurabh Gupta, "Various Schemes of Load Balancing in Distributed Systems- A Review", IJSRET, ISSN 2278 – 0882 Volume 4, Issue 7, July 2015.

[3] Ranjan Kumar, G Sahoo" Load balancing using ant colony in cloud computing" International Journal of Information Technology Convergence and Services (IJITCS).

[4] Ram Prasad Padhy, P Goutam Prasad Rao "Load balancing in cloud computing systems" Thesis from National Institute of Technology, Rourkela-769 008, Orissa, India May, 2011R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] T. Kokilavani, Dr. D. I. George Amalarethinam ―Load Balanced Min-Min Algorithm for Static Meta Task Scheduling in Grid computing‖ International Journal of Computer Applications Vol20 No.2, 2011.

[6] Pooja Samal, Pranati Mishra, ‖Analysis of variants in Round Robin Algorithms for load balancing in Cloud Computing‖ (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (3) , 2013, 416-419.

[7] Kousik Dasguptaa, Brototi Mandalb, Paramartha Duttac, Jyotsna Kumar Mondald, Santanu Dame "A genetic algorithm (ga) based load balancing strategy for cloud computing" International Conference on Computational Intelligence Modeling Techniques and Applications 2013.

[8] Yi Lua, Qiaomin Xiea, Gabriel Kliotb, Alan Gellerb, James R. Larusb, Albert Greenbergc" Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services"IEEE Conference on Foundations of Computer Science.

[9] Poonam Devi,Mr. Trilok Gaba"Implementation of cloud computing by using short job scheduling" International Journal of Advanced Research in Computer Science and Software Engineering.

[10] Brototi Mondala, Kousik Das Gupta, Paramartha Duttab"Load balancing in cloud computing using stochastic hill climbing-a soft computing approach"Science Direct.

[11] Karanpreet Kaur, Ashima Narang, Kuldeep Kaur, "Load Balancing Techniques of Cloud Computing", International Journal of Mathematics and Computer Research, April 2013.

[12] Ram Prasad Padhy ,P Goutam Prasad Rao, ―Load Balancing in Cloud Computing Systems‖, National Institute of Technology, Rourkela, India, 2011.

[13] Akash Dave, Prof Bhargesh Patel, Prof. Gopi Bhatt,   Yash Vora, "Load Balancing in cloud Computing Using Particle Swarm Optimization on Xen Server", IEEE, 2017.