A Various Load Balancing Techniques and Challenges in Cloud Computing – Survey

Avani Kansara[1] Ronak Patel[2]
[1,2]Department of Computer Engineering
[1,2]Ipcowala Institute of Engineering & Technology, Dharmaj, Anand, Gujarat, India- 388430

Abstract— Cloud computing is a virtualization of cloud program through internet connection, in which there is no need of installing application on everywhere. In cloud computing applications are provided and managed by the cloud provider. Cloud users may access the resources through computer notebook, smart phone or other devices, so they can store data on servers and can access data through Internet. It offers the facility of pay per use to their customers. In cloud computing, clients are demanding more services so for that load balancing of resourcing must be required. Load occurs when the number of job increases. The load can be a memory, CPU capacity, network, or delay load. Load balancing ensures that all the processors in the system do equal amount of work to make the best response time and also good utilization of the resources. Load balancing can be achieved by maximizing throughput or by minimizing response time. In this paper, it describes a survey on load balancing schemes in cloud environment. There are various load balancing techniques used in this paper and their corresponding performance metrics, advantages and disadvantages are studied. These scheduling algorithms focus on response time, make span, execution time and throughput as evaluation parameters.

Key words: Cloud computing, Load balancing, Load balancing algorithm

I. INTRODUCTION

A cloud refers to a distinct IT environment that is designed for the purpose of remotely provisioning scalable and measured IT resources [1]. In this type of computing, recourses are shared instead of owning local personal servers. This is used to handle applications on system. The word cloud in cloud computing is used as a symbol for internet, so we can define a cloud computing as the internet based computing. Cloud computing provides different services like storage, servers and application which are provided to organization computers and device using internet [2].

A. Cloud Computing Architecture:

In real time environment, cloud computing and the information about the cloud and services are growing rapidlly. In below figure 1, the three basic services layers are described. Those are Software as a Service, Platform as a Service and Infrastructure as a Service.

The cloud comprises three major components, that are Clients, Data centers and Destributed Servers. With the use of these components how they make cloud computing solution is showing in figure 2.

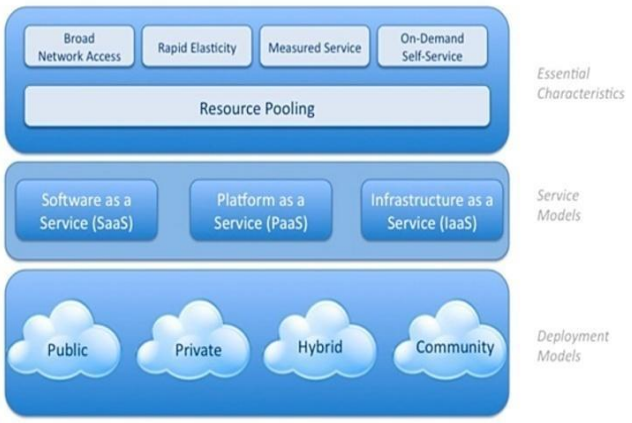

Fig. 1: cloud computing architecture[2]

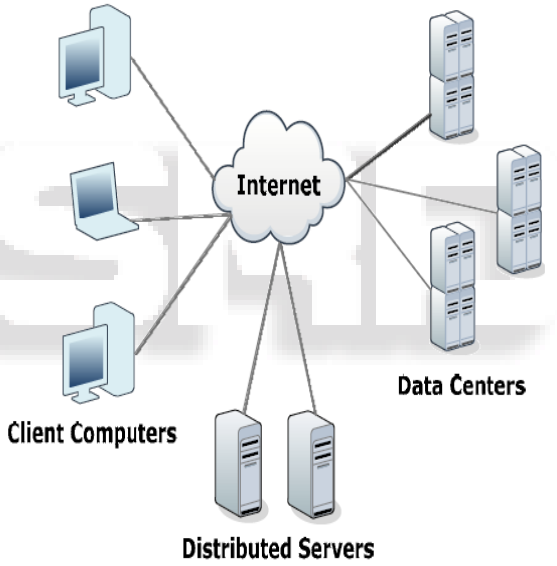

Fig. 2: Components make up of cloud computing solution[3]

B. Load Balancing in a Cloud Computing Environment:

Load balancing algorithms can be categorized into two types.

- Static
- Dynamic

Static algorithm requires the prior knowledge of system and it is independent of current system. Dynamic algorithm gives better performance than static algorithm and it is dependent of current system.

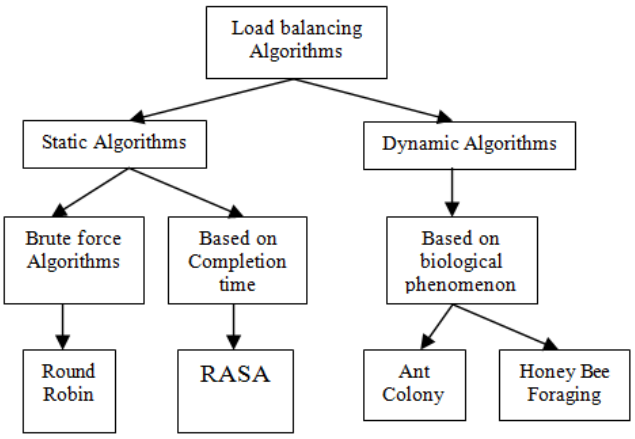

Fig. 3: Classification of load balancing algorithms.

This paper is organized as: section 2 describes load balancing techniques , section 3 describes comparison of load balancing algorithms, section 4 describes challenges for load balancing and section 5 decribes conclusion.

II. LOAD BALANCING TECHNIQUES

In cloud computing Load balancing is one of the main issue. There is a situation, where nodes are heavily loaded or under loaded in the network. So that load balancing ensures that every node in the system has equal amount of work. The Goals of Load Balancing are:

- Improving the performance
- Maintaining the system stability
- Building the fault tolerance system

Load balancing contains the following algorithms:

A. Round Robin scheduling:

This algorithm uses the concept of time slices. In this algorithm time is divided into multiple slices and each node and each slices is given a perticular time interval. In this time interval the node will perform its operations. This algorithm uses queue concept to store jobs. In this concept each job will be executed in its turn [4]. If in their turn, perticular job didn't complete their work then it will be stored back to the queue,waiting for the next turn. Main advantage of this algorithm is, here each job will be executed in turn and they don't have to be waited for the previous one to get completed. If the load is found to be heavy, this algorithm take long time to complete all jobs. This algorithm simply allocates the job in round robin manner, which does not consider the load on different machines [4].

B. Genetic Algorithm:

Genetic alorithm is mainly used because, it can handle a vast search sapce [5]. This algorithm can be applicable to complex objective function and can avoid being trapped into local optimal solution. This algorithm is composed of three operations:selection, genetic operation,and replacment. This algorithm can be described as follow:

- Selection: GA works on fixed bit string, so all the possible solutions in the solution space are encoded into binary strings. From this, intially random selection is done
- Crossover: In this step, the best fitted pair of individuals are selected ofr crossover. The fitness value of each individual cromosome is calculated using the fitness function as given in next step
- Mutation: Very small value, such as 0.05 is picked up as mutation probability. Depending upon this mutation value the bits of the chromosomes are toggled from 1 to 0 and 0 to 1. And output of this is a new mating pool ready for crossover [5].

C. Ant Colony optimization:

Ant colony optimizationis a random search algorithm which imitates the concept of behavior of ant colonies. It uses small-world and scale-free characteristics of a complex netwok to achieve better load balancing [6]. This technique overcomes heterogeneity,is excellent in fault toletance, is adaptive to dynamic environmets and has good scalability. Ants are trailing from their nest to food and connect eachother by pheromone. High denser pheromone attracts more ants, Because of this, performance of system is improved [6].

D. Honey Bee Behavior Algorithm:

Honey Bee behavior load balancing algorithm is decentralied, which is a nature-inspired algorithm. Through local server action, this algo achieves global load balancing [7].This algorithm is best suited for the conditions where the diverse population of service type is required. Performance of the system is enhanced with increased system diversity but throughput is not increased. This algorithm is derived from the behavior of honey bees that uses the method to find and gather food [7].

E. RASA Algorithm:

RASA algorithm is the combination of min-min and max-min algorithm. The algorithm builds a matrix C where C_{ij} represents the completion time of the task T_i on the resource R_i [8]. If the number of available resources is odd, the min-min algorithm is applied to assign the task first, otherwise the max-min algorithm is applied. With the use of one of this strategies, remaining tasks are assigned to their appropriate resources. RASA algorithm takes advantages of min-min algorithm and max-min algorithm, and ignore disadvantages of both. So waiting time of the small tasks in max-min algorithm and the waiting time of the large tasks in min-min algorithm are ignored [8].

F. Equal Spread Current Execution Algorithm:

In Spread Spectrum technique, load balancer makes effort to preserve equal load to all the virtual machines connected with the data centre. Load balancer maintains the index tableof virtual machines as well as number of requests currently assigned to the virtual machine(VM) [9]. In this, if the request comes from the data centre to allocate the new VM, it first scan the index table for finding least load VM. If there are more than one VM is found than first identified VM is selected for handling the request of consumer, and that VM id is also returned to the data centre controller. Than communication takes place. After completion of the request by VM, decreasing the allocation count. So Equally loads improves performance by transferring load from heavily loaded server, and in Spread Spectrum technique

load balancer makes effort to preserve equal load to all the VM connected with the data centre [9].

G. Stochastic Hill Climbing Algorithm:

For solving optimization problem, Stochastic Hill Climbing Algorithm is one of the incomplete approach. It is simply a loop that continuously moves in the direction of increasing value, which is uphill. It stops, when it reaches a peak, where no neighbour has a highr value [10]. This alternative chooses at random from among the uphill moves and the probability of selection can vary with the steepness of the uphill move. By making minor changes to the original assignment, it maps assignments to a set of assignment. To improve the evaluation score of the state, some criteria is designed to move closer to a valid assignment. This basic operation is repeated until a solution is found or a stopping criteria is reached [10].

III. Comparison of Load Balancing Algorithms

Table 1 and Table 2 shows the comparison of Load Balancing algorithms which were discussed above.

Load balancing methods	**Parameter**	**Merits**	**Demerits**
Genetic algorithm	Memory	1.Well suited for vast searching	1.Same Priority.
Ant colony	Response Time.	Guarantees the QoS requirement.	1.Fault tolerance issue does not consider
Honey bee	Makespan Task Migration Execution Time	1.Maximizing Throughput 2.Waiting Time of Task Minimum 3.Low Overhead	1.Low Priority 2.Load Become Stay Continuously on the Queue
RASA	Response time	1.Efficient resource allocation 2.Minimum execution time	1. No Ability to handle failover
ESCE	Response time Processing time	1.Improved response time and processing time	1. No priority.
SHC	Response Time.	1.Improve processing time	1.Problem with response time

Table 1: A comparison of LB algorithms

Algo	Static	Dynamic	Possible starvation	Optimal resource utilization	Ability to handle failover
A1	Y	Variations supports dynamic behavior	N	N	N
A2	N	Y	N	Y	Y
A3	N	Y	N	Y	Y
A4	N	Y	N	Y	N
A5	Y	N	Y	Y	N
A6	N	Y	N	Y	Y
A7	N	Y	Y	Y	N

Table 2: A comparison of LB algorithms

IV. Challenges for Load Balancing

There are some qualitative metrics that can be improved for better load balancing in cloud computing [11][12].

A. Throughput:

It is the total number of tasks that have completed execution for a given scale of time. It is required to have high through put for better performance of the system.

B. Associated Overhead:

It describes the amount of overhead during the implementation of the load balancing algorithm. It is a composition of movement of tasks, inter process communication and inter processor. For load balancing technique to work properly, minimum overhead should be there.

C. Fault Tolerant:

We can define it as the ability to perform load balancing by the appropriate algorithm without arbitrary link or node failure. Every load balancing algorithm should have good fault tolerance approach.

D. Migration Time:

It is the amount of time for a process to be transferred from one system node to another node for execution. For better performance of the system this time should be always less.

E. Response Time:

In Distributed system, it is the time taken by a particular load balancing technique to respond. This time should be minimized for better performance.

F. Resource Utilization:

It is the parameter which gives the information within which extant the resource is utilized. For efficient load balancing in system, optimum resource should be utilized.

G. Scalability:

It is the ability of load balancing algorithm for a system with any finite number of processor and machines. This parameter can be improved for better system performance.

H. Performance:

It is the overall efficiency of the system. If all the parameters are improved then the overall system performance can be improved.

V. Conclusion

Various load balancing tehniques for cloud omputing is surveyed in this paper. Main purpose of load balancing is to satisfy the requirements of customer, by distributing load among the nodes and make maximum resource utilization. Maximum resource utilization is achieved by reassigning the total load to individual node, by this it ensures that every

resource is distributed efficiently. So the performance of the system is increased.

REFERENCES

[1] Basic concept and terminology of cloud computing- http://whatiscloud.com
[2] L. Wang, J. Tao, M. Kunze,"Scientific Cloud Computing: Early Definition and Experience", the 10th IEEE International Conference Computing and Communications 2008.
[3] P. Mohamed Shameem, R.S Shaji. "A methodological survey on load balancing technique in cloud computing", in IJET, 2013, pp. 3801-3812
[4] Nusrat Pasha, Dr. Amit Agarwal, Dr. Ravi Rastogi, "Round Robin Approach for VM Load Balancing Algorithm in Cloud Computing Environment", in IJARCSSE, 2014, pp. 34-39.
[5] Kousik Dasgupta, Brototi Mandal, Paramartha Dutta, Jyotsna Kumar Mondal, Santanu Dam, "A Genetic Algorithm (GA) based Load Balancing Strategy for Cloud Computing", in ScienceDirect, 2013, pp. 340 – 347.
[6] Santanu Dam, Gopa Mandal, Kousik Dasgupta, and Paramartha Dutta, "An Ant Colony Based Load Balancing Strategy in Cloud Computing", in Springer, 2014 pp. 403-413.
[7] Dhinesh Babu L.D , P. Venkata Krishna, "Honey bee behavior inspired load balancing of tasks in cloud computing environments", in Elsevier, 2013, pp 2292–2303.
[8] S. Mohana Priya, B. Subramani, "A NEW APPROACH FOR LOAD BALANCING IN CLOUD COMPUTING", in IJECS, 2013, pp. 1636-1640.
[9] M. Aruna, D. Bhanu, R. Punithagowri, " A Survey on Load balancing Algorithms in Cloud Environment", IJCA, Volume 82, No. 16, November 2013.
[10] Brototi Mondala, Kousik Dasgupta, Paramartha Dutta, "Load Balancing in Cloud Computing using Stochastic Hill Climbing-A Soft Computing Approach", in Elsevier, 2012, pp. 783 – 789.
[11] Foster, I., Y. Zhao, I. Raicu and S. Lu, "Cloud Computing and Grid Computing 360-degree compared," in proc. Grid omputing Environments Workshop, pp: 99-106, 2008.
[12] Buyya R., R. Ranjan and RN. Calheiros, "InterCloud: Utilityoriented federation of cloud computing environments for scaling of application services," in proc. 10th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP), Busan, South Korea, 2010.

