Analysis of Load Balancing Mechanisms in Public Cloud Service

R.Ramya, B.Seetha, S. Puvini Vigneshwari, N.Angayarkanni

Abstract— Cloud computing is defined as the resources that can be shared or accessed by the local host from the remote server via the internet. Cloud computing mainly focus on sharing of resources to achieve consistency and economies of scale. The evolution of cloud computing has led to the evolution of modern environment due to the advancement and abundance in the computing infrastructure. Due to more and more user, the server in the cloud center can be heavily loaded. Without load balancing, there may be some effects to the server with respect to the workload distribution. So load balancing plays a vital role in cloud computing. Load balancing provides high availability of resource with less performance overhead to the server. Our objective is to review the existing load balancing approach proposed till now. The general characteristics of the approaches are summarized in this survey. And with this survey, the related studies in this area are well understood based on the characteristics of the system.

Index Terms— Cloud Computing, Load Balancing Approach.

1. INTRODUCTION

Firstly, we go for cloud computing in order to reduce paperwork manually. And because of the following reason, it is very difficult to own a proprietary infrastructure. One needs to upgrade the software and appropriate hardware every time. Then there is a need to upgrade the operating system which best suits for the above hardware and software up gradation. Thus cloud computing is the practice of accessing, storing, managing and processing the data from the remote server hosted on the internet, instead of using a local server or a personal computer. Many organizations spend more on a resource in computer infrastructure and maintenance and so cloud computing aims are to focus on the organization core business.

Due to the difference in the computing capabilities and uneven request arrival pattern, the server may vary greatly [2]. This will leads to high response time and low service availability. So in computing, load balancing is used to distribute the workload across multiple computing resources. Cloud load balancing high performance in the lower cost than the traditional load balancing technology. And also load balancing achieves high scalability and high availability of resources. The workload is divided into many servers or computing resources to enable better cloud resources utilization in which no single node is overloaded.

Cloud Computing plays a vital role in computing filed due to the virtual machine. A virtual machine is an operating system or an application environment which is implemented on the physical machine using a hypervisor or virtualization and also has its own processor, I/O devices, and the hard disk which is shared by the physical hardware. Virtual machine technology allows multiple of virtual machine to run on a single physical machine.

Virtual Machine Monitor (VMM) is a software program that can create and manage the multiple virtual machines which are running in the top of the host machine. VMM is also known as the hypervisor. A computer in which a hypervisor run is called as a host machine, and the virtual machine runs on the physical machine is called as the guest operating system Each guest OS has its virtual operating platform and the virtual machine monitor manages the execution of a virtual machine.

A load balancer is a device that distributes the traffic to multiple servers or computer. This helps the server to be equally loaded. By using the load balancer, it will increase the cloud users and also the reliability of applications. And also improves the performance of the application by reducing the overload of the instances by managing and maintaining the application and network sessions, and also performing the application-specific tasks[8].

One of the methodologies used in cloud computing is auto scaling which is the idea of load balancing, where it measures the number of active servers in the server farm and it automatically scales according to the load on the farm. In order to distribute traffic across servers, load balancing is used in data center networks. Using load balancing, the cloud users can efficiently use the network bandwidth and also reduce the provision costs [5]. By maintaining and managing the cloud application and network session, the load balancer decreases the burden of the server and thus it improves the performance of the application.

Goals of Load Balancing

1. Availability of services needed to be increased
2. A response time of the services need to be reduced
3. The total used and freed capacity of data centers to need to be optimal

Revised Version Manuscript Received on March 08, 2019.

R. Ramya, PG Scholar, Department of Computer Science & Engineering, PSNA College of Engineering & Technology, Dindigul, India.

B. Seetha, PG Scholar, Department of Computer Science & Engineering, Shanmuganathan Engineering College, Pudukottai, India.

S. Puvini Vigneshwari, Assistant Professor, Department of Computer Science & Engineering, Shanmuganathan Engineering College, Pudukottai, India.

N. Angayarkanni, PG Scholar, Department of Computer Science & Engineering, Shanmuganathan Engineering College, Pudukottai, India.

Requirements of Load Balancing

1. Distribute client request across multiple servers.
2. Ensure high availability and reliability by sending request only to the server that is online. Provides flexibility to add and subtract the instances as per the demands.

Advantages of Load Balancing

There are few advantages of load balancing and some of them are

1. Increased scalability
2. Redundancy
3. Reduced downtime and increased performance
4. Efficiently manages failures
5. Increased flexibility

Disadvantage of Load Balancing

Some of the cons of load balancing are

1. Require additional configurations to maintain the connection between the client and the server.
2. Hardware-based load balancer costs more.

II. LITERATURE SURVEY

To achieve better load balancing, the researcher developed many new loads balancing approaches to solve the complexity of cloud computing and there is a number of load balancing approaches. Thus, this section describes the related work of load balancing algorithms and it is obtained in some of the following ways. Game theoretic approach, task scheduling approach and task allocation approach. Figure 1 represents the type of load balancing approach.

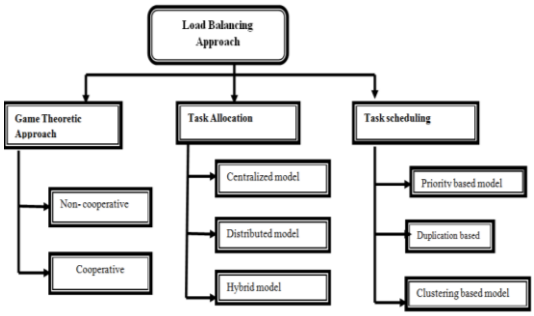

Figure 1- Types of Load Balancing

III. GAME THEORY BASED ON LOAD BALANCING IN CLOUD COMPUTING

Game theory is most widely used as the tool in the study of economics. It generally deals with how and why peoples make decisions. Game theory[7] is defined as the tool or techniques to take a decision even in the unpredictable events between the two or more opponents. An opponent is referred to as the players in the game. The strategy behind the game theory is that the any of the player's option that is used in the setting where the outcome of the action not only depend on that player but also on the action of the other players. The player strategy will determine the stage of the game. By using game theory in load balancing, it provides the fairness to the users and the user jobs ie it reduces the response time of the server. Some of the game-theoretic approach used in load balancing are given below and table I depicts the comparison of the approaches

1. Non-cooperative game theory
2. Cooperative game theory

1. Non- Cooperative Game Theory

A non-cooperative game theory[1] is a game where the competition is held between the individual server not among the group the server so that each server may come to know even the incomplete information of others. It finds the solution using Nash equilibrium due to the absences of Cooperative game theory. The Non-cooperative game theory provides the low-level procedural details of the game. a non-cooperative game theoretic approach can be used to balance the load in the heterogeneous distributed environment. Hence, all the static load balancing algorithm can be solved using non-cooperative game theory for a heterogeneous environment. Thus the main advantage of using non-cooperative game theory is that it provides low complexity and optimal allocation of loads to the cloud user. This type of game theoretic approach is used in the distributed system where it consists of the heterogeneous resource with multiple consumer and the system is worked in an independent manner. Thus, the objective of this approach is to minimize the mean time failure of the system.

2. Cooperative Game Theory

A cooperative game theory[3] is a game where the competition is held between a group of a server but not among the individual servers. It always provides high-level approaches like structure, strategies, and payoffs of the players. Due to the external enforcement of cooperation, this game theoretic approach produces the optimal solution. A cooperative game theory approach can be used to solve the problem having a large number of the server through the simplified approach about bargaining power. Cooperative game theory provides the Pareto optimal allocation of load to the user using Nash Bargaining Solution (NBS). This approach provides the cooperative method to all nodes to maximize the overall performance of the applications[4]. All nodes in the distributed system work together in this kind of approach. Thus, the cooperative game theory provides an efficient way to self organize the server. To solve bandwidth allocation between application, cooperative game theory is mainly used in networking. It mainly focused to derive the fair and Pareto-optimal allocation scheme.

Comparison Between Non-Cooperative and Cooperative Game Theory:

Table I – Comparison between Non-Cooperative and Cooperative Game Theory

Sr. no	Non-Cooperative Game Theory	Cooperative Game Theory
1	It predicts the player's individual strategies and payoffs and also finds the Nash equilibrium	It predicts which group will form together, the joint action the group takes and find the payoffs

2	It provides the low-level approach of the game	It provides the high-level approach of the game
3	Provide accurate result	Sometimes provides inaccurate results
4	Players are not supposed to take a binding agreement	A binding agreement is possible using an agent
5	Non-cooperative game theory is called as procedural	Cooperative game theory is called coalitional.

IV. TASK SCHEDULING BASED ON LOAD BALANCING

Task scheduling [10] is defined as the ability to schedule the load to the user at a pre-defined time or at a specified interval. Without efficient task scheduling, the performance of the system may be degraded. It mainly focuses on the communication cost, execution time and task resources allocation. The main aim of the task scheduling is to increase the resource utilization and to minimize the task completion task. Therefore, the task scheduler is the tool with some predefined action that is automatically executed whenever some set of the condition is met. The scheduled task can be distributed and managed across the different network through an administrative backend. To perform task scheduling, two modes are important. One is space sharing modes where the resources are not pre-empted and another one is time sharing modes where the resources are pre-empted until the completion of the task. Some of the task scheduling approaches are given below and table II depicts the comparison of the approaches[11]

a. Priority based scheduling
b. Duplication based scheduling
c. Cluster-based scheduling

a) Priority Based Scheduling

The scheduling process is done based on the priority. In this type of approach, the priority assignment is assigned to each process. The highest priority process is served as First Come First Serve(FCFS) and the lowest priority process is served as Last Come Last Serve(LCLS). That is, the task which comes first will be executed first to the Virtual Machine. As a whole, the highest priority is assigned to the user who pays more and they always enjoy better service. The priority of each queue is determined. There are two types of the priority-based scheduling algorithm. One is pre-emptive priority scheduling and another one is a non-preemptive scheduling algorithm. The preemptive scheduling algorithm is commonly used in the much real-time system. It always executes the task in the time slice mechanisms. Therefore, no task runs longer than the time slice. With non-preemptive priority scheduling, the priority number is assigned to each and every process. So with the use of priority number, processes are scheduled. Until the completion of the task, the process will run. Thus, the given task cannot be taken from the CPU until it gets completed. This approach is the best fit for the application which requires time and resources. With the increase in time, the priority of the process increases.

b) Duplication Based Scheduling

The main purpose of using duplication (replication) based scheduling is to achieve a DAG scheduling with minimized make-span time of the task and high efficiency of the task in the cloud service. For each new processor, the earliest finish time of the task is minimized so that the approach can achieve a less make-span time of the task. This approach always tries to duplicate the task so that the approach minimizes the make-span time in the cloud services. Now, the processor has some task to be scheduled and thus called as particle schedule. Then eliminate and merge the particle schedule task so that the number of processor in the system is reduced. Thus the efficiency of the task is improved. There exist some of the algorithms in duplication based scheduling for both homogeneous and heterogeneous environment. Communication cost is minimized by placing the tasks on the same processors.

c) Clustering Based Scheduling

It is defined as grouping a set of a task which needs to communicate among themselves and form clusters. This type of approach can be useful in the heterogeneous environment. A cluster of a task is created and it is assigned to the fixed number of available processor. To improve process sing power, clusters are interconnected with the symmetric multiprocessing. Generally, K-means clustering techniques are used for virtual machine clustering method. It divides the large number of VM into a small number of VM and groups them based on similar characteristics. In load balancing, clustering scheduling is used in order to improve performance. In addition, this approach can provide high availability of services to the cloud user. This type of approach works by the group of servers in the backend by sharing the resources by getting single or more cluster load.

Comparison between Approaches of Task Scheduling:

Table II – Comparison between Approaches to Task Scheduling.

Sr. No	Approach of Task Scheduling	Advantages	Disadvantages
1	Priority based scheduling	-As time increases, increase in the priority of the process -Suitable for application with varying time and requirement	-Starvation -low priority process get lost, when a system crashes
2	Duplication based scheduling	-Minimize make-span time of the task -High efficiency	-Scalability bottlenecks -Not suitable for some dynamic problem

3	Clustering based scheduling	-Better resource utilization and efficiency of a task.	-Scheduling must be scalable.

V. TASK ALLOCATION IN LOAD BALANCING & RESULTS

One of the important aspects of cloud computing is task allocation. The on-demand basis of cloud user, the hardware and software resources are allocated to the cloud environment. In a distributed system, all the applications are divided into some number of task and assigned to different nodes and the application purely dependent on the allocation of a task in the node. This is referred to as task allocation[6] problem in a distributed system. According to the definition of load balancing, if too many tasks are crowded on a single node, then the heavily loaded tasks are subdivided and assigned to different nodes to make the node as the lightly loaded task. By balancing the node, it will minimize the waiting time and response time of the task. Some of the task allocation approaches are given below and table III depicts the comparison of the approaches[9],

1. Centralized control model
2. Distributed control model
3. Hybrid control model

1. Centralized control model

This type of model is used only for a small system to calculate the allocation of tasks. Like in a real-time application, a central controller is used to know the status of the entire system and allocate the task based on the information that is sent from the other node. It is simple and efficient and achieves good performance level in the task allocation.

2. Distributed control model

As the name suggests, the distributed system can be used in the large and dynamic distributed environment. The workloads are distributed to all nodes independently without the need of a central controller. In here, the coordination and negotiations among tasks are crucial. Therefore some techniques are used to coordinate and negotiate the tasks well. One of the approaches is Market based approach where the aims are to produce the best distribution of goods by making an exchange within the system values and time tradeoffs in the economics. The node trade task and resource price are determined by the auction protocol after that all nodes are assigned to some tasks. Using this technique will minimize the response time of the task and some of the examples are a game theory, auction protocol. Thus this technique reduce the execution time of the tasks and some of the example are group mechanisms, strategy diffusion etc.

3. Hybrid control model

Hybrid control approaches are used to solve the centralized and decentralized problems. Here some nodes may act as a controller and some node can act as a the autonomy role to allocate the tasks. This approach retrieves the good performance and robust level.

Comparison of task allocation approaches:

Table III – Comparison of Task Allocation Approaches

Sr.no	Type of task allocation approach	merits	demerits
1	Centralized control model	-Simple -achieves global optimal results	-infeasible in reality
2	Distributed control model	-Self-adaptable -work well in a complex system	-performance bottleneck -high computational cost
3	Hybrid control model	-Simple and efficient	-not suitable for a large environment

VI. CONCLUSION

Load balancing is the major challenge in the cloud computing environment. The load balancer distributes the workload evenly among different computing resources to balance the load. So this will avoid the situation like overloaded or underloaded. Hence the performance and resource utilization of the system will be increased. Thus every computing resources will be distributed efficiently. Along with the research challenge in load balancing, the concept of cloud computing is also reviewed. Finally, this study reviewed the major thrust of the load balancing approaches, followed by the comparative study of the above-mentioned approaches in cloud computing with respect to throughput, performance overhead, communication delays, response and waiting time of the task in cloud services.

REFERENCES

1. Daniel Grosu, Anthony T," Noncooperative load balancing in distributed systems", Elsevier, vol.65, 2009
2. Deepak B S Shashikala S V Radhika K R NIE, "Load Balancing Techniques in Cloud Computing: A Study", International Conference on Information and Communication Technologies, ISSN: 0975-8887,2014
3. Grouse .D et.al, "Load balancing in distributed systems: an approach using cooperative games", IEEE, ISSN:0-7695-1573-8, 2002
4. jeyaprakash rajasekaran, visa kolivenen," Cooperative game-theoretic approach to load balancing in smart grids with community energy storage", 23rd European Signal Processing conference,ISS:2076-1465, 2015
5. Karanpreet Kaur, Ashima Narang , Kuldeep Kaur, "Load Balancing Techniques of Cloud Computing", INTERNATIONAL JOURNAL OF MATHEMATICS AND COMPUTER RESEARCH, vol.1, ISSN:2320-7167,2013
6. Larry Rudoploh et.al, " A Simple load Balancing Algorithms for Task Allocation in Parallel Machines", ACM on Parallel and Distributed Systems,2014
7. Riky subtra et.al , " Game Theoretic Approach for Load Balancing in Computational Grids", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS", vol.19, ISSN:1045-9219, 2012

8. Sheetanshu Rajoriya, "Load Balancing Techniques in Cloud Computing: An Overview", International Journal of Science and Research (IJSR), ISSN:2319-7064, 2013
9. Yichuan Jiang, "A Survey of Task Allocation and Load Balancing in Distributed Systems", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS", vol.20, 2016
10. yiqiu fang, fei wang, " A Task Scheduling Algorithm Based on Load Balancing in Cloud Computing", Springer, pp:271-221, 2010.
11. Zhang Qian, Ge Yufei, Liang Hong, Shi Jin," A Load Balancing Task Scheduling Algorithm based on Feedback Mechanism for Cloud Computing", International Journal of Grid and Distributed Computing, vol.8, 2016