LOAD BALANCING IN CLOUD COMPUTING: THE ONLINE TRAFFIC MANAGEMENT

Navleen Kaur*, Dr. Jaspreet Singh1, Shanky Goyal2, Bharti Duhan4

1*Research Scholar, Chandigarh University, Gharuan, Mohali District, Punjab, India,
2Assistant Professor in IT department, Chandigarh Engineering College, Landran, Mohali, Punjab.
3Research Scholar, IKGPTU, Kapurthala, Asst. Prof. IT dept., Chandigarh Engineering College, Landran, Mohali, Punjab
4Assistant Professor, Chandigarh University, Gharuan, Mohali (Punjab)

ABSTRACT

Cloud computing is a kind of Internet-based computing. Clouds are highly configured infrastructure, which allow the users to make the payment only for the services procured. Cloud computing is basically delivery of different type of services to an organization’s computers and devices over the Internet. These resources can include tools and applications like data storage, servers, databases, networking, and software. As opposed to keeping records on a selective hard drive or other nearby storage devices, cloud-based storage makes it conceivable to spare them to a remote database or server farms in a protected and versatile manner. The electronic device just needs to access the web, using which it gets to access the data and the software programs to run it. Cloud computing is now days becoming a popular option for people and businesses for a number of reasons including cost savings, increased productivity, speed, efficiency, performance and security. For the most part the cloud is placed on data centers which are amazing to deal with countless clients. Load balancing mechanism comes into role to manage the workloads, which further defines the efficiency of the cloud. The basic aim behind balancing the load over the machines is to distribute workloads and then computing resources across one or more servers by reducing the energy consumption and providing maximum resource utilization. This paper gives concise information about load balancing, necessity to maintain the load and basic algorithm used to maintain the load over the cloud. In addition to this Static and dynamic load balancing algorithms have also been compared.

Keywords: Cloud computing, Load Balancing, LoadBalancing Algorithms; Load Balancers.

I. INTRODUCTION

Cloud computing is computing as-a-service which means that rather than owning own computing infrastructure or data centers, one needs to pay to use someone else's equipment with the amount of capacity needed, it’s provided as a service by another company and accessed over the Internet, usually in a completely seamless way. Cloud computing offers pay-as-you-go method in which there is no wastage of resources, users only needs to pay for services used. Cloud Computing has become one of the most popular technologies now days which is being adopted by both the scholastic and industry, as it provides an efficient and adjustable method for storing and accessing the files.

Cloud computing is an Internet based computing of the delivery of on-demand computing services from applications to storage and processing power in which the resources, software, information are shared to computers and to the other devices [1]. The main objectives of cloud are to reduce the cost, enhance response time, provide better performance; hence cloud is also called as a pool of Services [2]. Cloud allows the users to access the information at any time and from any place. The user need not be present at the same location as the hardware on which the data is stored. The user just needs the internet connectivity to access the services of the cloud. It delivers all the services needed by the user dynamically through the internet.
The architecture of cloud computing consists of two components: the front end and the back end which are connected through a virtual network or the internet. Front end comprise customer part of cloud computing system. It involves interfaces and applications that are required to get to the cloud computing platform. While back end alludes to the cloud itself, it contains the resources that are required for cloud computing services. It consists of virtual machines, servers, data storage, security mechanism etc. It is under provider’s control. The cloud computing scenario is shown in figure 1. It shows how the different Cloud services can be accessed through different browsers.

![Cloud Computing Architecture](image1)

Figure 1: Cloud computing Architecture

Computing is an emerging new technology that is quickly strengthening itself as the next big step in the development and deployment of an increasing number of distributed applications [4]. The cloud has created a new look to place together IT and the business visions. The different type of cloud services are shown in figure 2.

**The three major Cloud Computing services are:**

A. **Software as a Service (SaaS):** Software distribution model in which applications are encouraged by a seller or service provider and are made accessible to clients over a network (internet). SaaS applications are known as Web-based software, on-demand software and hosted software.

B. **Platform as a Service (PaaS):** It gives a stage and condition to permit designers to construct applications and services. This service is facilitated in the cloud and got to be used by the clients by means of internet. PaaS services are continually refreshed and new highlights are added.

C. **Infrastructure as a Service (IaaS):** IaaS, is a cloud model, which provides computing infrastructure like virtual server space, network connections, transfer speed, load balancers and IP addresses. The pool of equipment asset is removed from various servers and systems normally disseminated over various data centers.

![Cloud Computing Services](image2)

Figure 2: Cloud computing Services

Cloud deployment model indicates demonstrates how the cloud services are made accessible to clients. The four deployment models related with cloud computing are as per the following as appeared in figure 3.

A. Public Cloud: A public cloud supports all different type of users who want to access the computing resource such as hardware (OS, CPU, memory, storage) or software (application server,
database) on a subscription basis. The services provided by the service providers may be free or the amount can be paid for the number of hours the cloud is used and can exit whenever the work is completed. The most common use of the public clouds is file-sharing and e-mail services.

B. Private Cloud: Private Cloud: Private cloud as the name proposes is the infrastructure which is utilized by a single organization. Such infrastructure might be overseen by the organization itself to help different user groups, or it could be overseen by a service provider that deals with it either on location or off-site.

C. Community Cloud: A community cloud supports multiple organizations which share computing resources that are part of a community. The community cloud environment is restricted to the members of the community only.

D. Hybrid Cloud: Hybrid cloud is combination of both private and public cloud infrastructure. Using hybrid cloud both the non-critical and critical tasks can be done. Non-critical tasks are done using public cloud whereas critical tasks are done using the private cloud.

In this paper, we provide an overview of load balancing in cloud computing. Though, there are various challenges being faced by cloud computing like task scheduling, security, resource migration, and energy efficiency. The basic idea of Cloud is that it manages the online traffic by distributing the workload over different servers and resources.

II. REQUIREMENT OF LOAD BALANCING IN CLOUD COMPUTING

Different users may try to access a website or any web-application at the same time due to which it becomes very difficult for the website to manage the request of all the users at the same point of time. Many times it leads to system crash. In such situation load balancing is required and the major role is played by the load balancer.

The work which an individual computer has to do is divided among two or more other computers, resulting in fast execution and efficient utilization of the resources. By this way the work is divided among different computers due to which a single computer is not over burdened while other computers remain free or have less work to do. Load balancing ensures that the workload is somewhat equally distributed over all the computers connected through network. Thus Load balancing is the prenecessities for boosting the cloud Performances and using the resources.
Load balancing is basically a process of distributing the tasks and computing resources within the environment of the cloud [7]. The Workload of an organization or any enterprise is managed among multiple systems or servers by allocating the resources to them. Load balancing in the cloud also manages the workload traffic that inhabit over the internet. Load balancing removes the tasks from the machines which are overloaded due to work and assigns them to the machines which are idle or don’t have much work to do.

A. Necessity of Balancing Load in Cloud

The basic need of Load balancing is to divide the tasks to be performed among different servers. There are multiple features in balancing the load which includes: - evenly distribution of the workload among different nodes, satisfying the user, boosting up the performance of the system, decrease the response time and to deliver the services to achieve full resource utilization [5]. Figure 4 shows the load balancing process in cloud computing. For example, if an application is made on cloud and multiple users are assumed to use it at same time, hence the response time for different users would be very moderate and the servers would become engaged very quickly. This would result in slow response time and the users will be left unsatisfied. Here the load balancing comes in action, if we use it on our application, then work will be divided among all the other nodes resulting in high performance and better response [6]. A load balancer stands in between client and server, accepts multiple tasks from different clients and then distributes the tasks among various servers.

Load balancing in cloud computing main aims at distributing the workload dynamically to the computer nodes. It is used to utilize the resources efficiently and to make the user satisfied with the service provided. To improve the system reliability over the cloud environment it is important to reduce the power consumption which also reduces the operating cost [8]. Using load balancing the work gets divided among different computers due to which multiple users can do the work at the same time and at a faster speed. The workload is evenly distributed among multiple nodes [9]. It is the task of load balancer to decide which web server needs to serve the request. Different types of scheduling algorithms are used to determine which server would serve or pass the request on to other server [10].

A well specified load balancing algorithm aims at using the available resources efficiently, by ensuring that all the nodes have the load equally distributed and that no node is over burdened or under burdened. Targets of load balancing include [11]:

1. Most appropriate resource usage
2. Improved throughput
3. Quick response time
4. Maintain Website traffic
5. Handle sudden traffic burst

B. Load balancing Measuring Parameters

Load balancing is evaluated over different parameters. Such as [12] [13]:

a. Resource Utilization: This parameter is used to look over the utilization of resources. A load balancing algorithm is considered efficient only if the resources have favorable resource utilization.

b. Performance: This metric is utilized to check the productiveness of the system. It must be high.
c. Scalability: The capability of an algorithm to uniformly distribute the tasks of the system with limited number of nodes. This metric must be enhanced.

d. Throughput: It illustrates the count of jobs whose execution is finished. Throughput must be high for finer performance.

e. Response time: It is the measure of time a specific load balancing algorithm takes to reply in a distributed system.

f. Expense: It discovers the running cost and inter-process communication involved while carrying out the load balancing algorithm.

III. BASIC ALGORITHMS IN CLOUD COMPUTING FOR LOAD BALANCING

Different types of algorithms are used to implement the load balancing [13]. Many researchers have been done to balance the load in order to enhance the performance and to use the resources efficiently. Load balancing algorithms in view of the present condition of the system are divided in two main categories, namely static and dynamic [14], [15], [16], [17].

Figure 5: Categorization of load balancing algorithms.

Figure 5 shows the classification of load balancing algorithms. The algorithms are grouped into two category dependent on the system's current situation and the other who started the procedure. Based on who starts the task, load balancing algorithms are classified into three types [11]:-

a. Sender Initiated: In this the sender finds out those nodes which are over burdened, following which it begins the implementation of load balancing algorithm.

b. Receiver Initiated: The server initiates the execution of the load balancing algorithm if over the cloud any variation is discovered by the collector/server.

c. Symmetric: It is a blend of both sender initiated and receiver initiated algorithms. In view of the current condition of the system load balancing are arranged into static and dynamic algorithms

In this section, we provide a detailed analysis on the present load balancing algorithms for cloud computing. The load balancing algorithms are categorized as static algorithms and dynamic algorithms. We initially examine the static load balancing algorithm, later we will talk about the dynamic load balancing algorithm.

1. Static Load Balancing:

In this methodology load balancing is accomplished by giving information about the system memory, processing power and execution well ahead of time. The performance of the node is
resolved at the time of compilation. Individual nodes ascertain their allocated work and then submit the outcome to remote node. Without thinking about the present burden on a specific node, the work load is distributed in starting only by just considering the performance of the node. Static load balancing techniques are non-preemptive which means that when the load is assigned to a node it cannot be shifted to some other node.

This strategy requires less correspondence henceforth diminishes the execution time. Such kind of load balancing is utilized in those systems where there is less variation of the data. The principle detriment of this methodology is that it doesn’t take current condition of the system while settling on allotment choices. This has the significant effect on the overall performance of the system due to load variation in distributed system.

The four distinct kinds of Static load balancing methods are Round Robin algorithm, Central Manager Algorithm, Threshold algorithm and randomized algorithm. [18].

2. Dynamic Load Balancing:
The changes in the work load of the system are noticed carefully and then the work is redistributed in like manner.

<table>
<thead>
<tr>
<th>Static Load Balancing</th>
<th>Dynamic Load Balancing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information about the state of the system is provided well in advance.</td>
<td>Information about the state of the system is not required in advance.</td>
</tr>
<tr>
<td>Performance of the node is determined at compile time.</td>
<td>Performance of the node is determined at run time.</td>
</tr>
<tr>
<td>Current state of the system is not considered.</td>
<td>Current state of the system is considered.</td>
</tr>
<tr>
<td>Non–Preemptive Algorithms</td>
<td>Preemptive Algorithms</td>
</tr>
<tr>
<td>Less Communication</td>
<td>More communication</td>
</tr>
<tr>
<td>Less resource utilization</td>
<td>Greater resource utilization</td>
</tr>
<tr>
<td>Execution time is reduced</td>
<td>Execution speed is increased.</td>
</tr>
<tr>
<td>Free from processor thrashing</td>
<td>Considerable process thrashing occurs</td>
</tr>
<tr>
<td>No fluctuation</td>
<td>Significant Fluctuation occurs</td>
</tr>
<tr>
<td>Changes in the system are not monitored, hence no load re-distribution.</td>
<td>Changes in the system is monitored, accordingly the load is re-distributed.</td>
</tr>
</tbody>
</table>

It does not require any prior information about the system’s state. Thus it just relies on the current state of the system. To increase the speed of execution, the processes are shifted from a heavily busy machine to an idle machine, resulting in improvement of the overall working of the system [19], [20]. Dynamic load balancing methods are preemptive unlike the static load balancing which are non-preemptive.
This algorithm is normally made out of three techniques: transfer strategy, location strategy and information strategy [11]. Transfer strategy decides on which tasks are eligible for transfer to other nodes for processing. Location strategy nominates a remote node to execute a transferred task. Load balancing algorithm uses Information strategy as the information center. It is responsible for providing location and transfer strategies to each node.

Dynamic algorithms can take three different controlling forms: centralized, distributed, or semi-distributed. In centralized load distribution, a single (central node) node in the network is nominated to be responsible for all load distribution in the network. In distributed the responsibility is divided among all nodes equally. In a semi-distributed,The network is fragmented into groups where each group is brought together. Load balancing of entire system is accomplished through the collaboration of central nodes of all the groups. There are three sorts of dynamic algorithms: focal line, nearby line and least connection [21].

IV. CONCLUSION
Load balancing is an important issue in cloud computing. Load balancing is needed to distribute the load uniformly across various nodes over the network. If the nodes get overloaded this would result in decreasing the efficiency of the nodes. A productive, well planned load balancing algorithm would be required for efficiently utilizing the resources. In this paper we have studied about cloud computing, the services it provides and about the necessity of load balancing in the cloud. The basic static and dynamic load balancing algorithms have been analyzed showing their dissimilarity. Future work can be done by inspecting different types of algorithms to balance the load in the cloud computing to attain better performance by securing the system.

REFERENCES


