Load Balancing Algorithms in Cloud Computing

Sameeksha Irkal[1], Vandana Sunag[2]

Assistant Professor, KLE BCA P C Jabin Science College, Vidyanagar, Hubballi, India[1]

Assistant Professor, KLE BCA P C Jabin Science College, Vidyanagar, Hubballi, India[2]

ABSTRACT: Cloud computing has become popular Internet based service that provide different services to users. By Virtualization on internet various applications and services can be offered to the end user. Cloud Computing has become on-demand information and technology service and product. The vast growth of cloud computing in recent year has led to a massive increase in both the amount of traffic and the number of service requests to cloud servers. This growth trend of load poses serious challenges to the cloud load balancer in efficient balancing of the load. Load balancing is the technique that tends to distribute work load over network, the major concern is to achieve optimal resource utilization, minimum response time, to maximize throughput, and avoid overload. The cloud load balancing is a highly researched field where numerous solutions to balance load have been proposed. This paper summarizes the cloud computing, load balancing and discusses load balancing algorithms.

KEYWORDS: Cloud Computing, Load Balancing, Load Balancing Algorithm.

I. INTRODUCTION

Today the amount of data storage on computer and exchange of data over internet is constantly increasing; this increase of data requires more equipment to meet the needs of an organization. To better capitalize their investment, the over equipped organizations open their infrastructure to others by exploiting the Internet and related technologies and other emerging technologies such as virtualization by creating a Cloud. Cloud Computing means instead of using all computer hardware and software on our computer or inside our company network, it provides us a service by using internet from other company. The location of where the hardware or software is located and how it works does not matter. The organization has to pay only for what resources it is using (i.e. Data storage, software packages, programming environments) by a third party provider through a private or public network, using virtualization.In cloud computing model the software developers can focus on building application instead of managing servers. Cloud computing removes the burden of software vendors by providing convenient method of deploying software. Today the software vendors develop their applications and deploy them on number of rented servers and deliver Software as Service(SaaS).The rented servers give them instant delivery for their software ,inefficient load distribution among server may result in poor performance of delivering software service. To improve the performance and efficiency the incoming traffic is distributed on multiple servers using Load balancer.

II. LOAD BALANCING IN CLOUD COMPUTING

Cloud computing is a federation of many resources interacting together by sharing and pooling resources efficiently. Day by day millions of data are created every second this leads in increase of web traffic in web world. Load balancing has become imperative and important need of balancing the load on this heavy traffic. Load balancing mechanism can be constructed by many strategies and techniques. Load balancing is done in such a way that the entire load is distributed efficiently among available server so that the server can handle it gracefully, consequently each system does almost equal amount of task at any point of time. There are dedicated servers allocated to handle this load balancing across the servers, if there is failure of any node in network it will redirect the load to the running instance or node. The Cloud vendors are based on automatic load balancing services which allows users to add or removed the number of servers based on users demand. Elastic Load Balancing also performs health checks on each instance. If an instance is

not responding, the load balancer will automatically redirect traffic to the healthy instances. Fig 1 depicts load balancing architecture used in cloud environment where load balancer load traffic using following steps
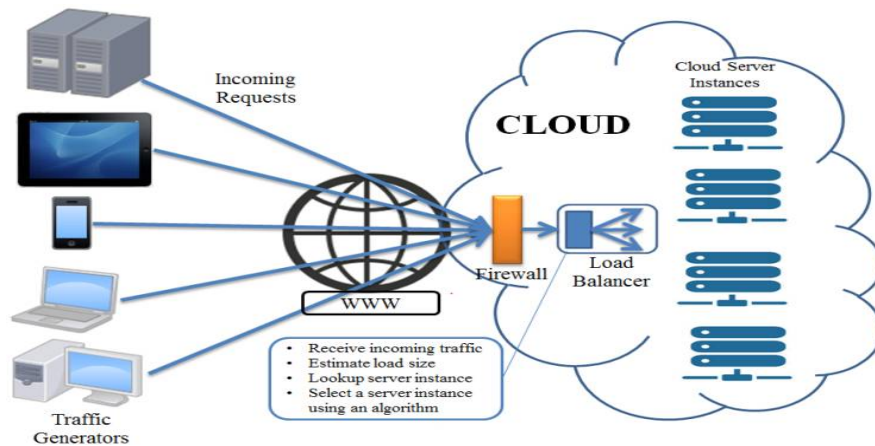

Fig 1: Load Balancong Architecture

1. Various incoming client request are received.
2. It checks the load size of incoming request and builds a request queue.
3. The current load of the server is checked in the server pool.
4. It uses load balancing algorithms to select appropriate server.

The various task performed by load balancing are as follows

1. Intercepts network-based traffic.
2. Split the incoming traffic into individual request and decided which serves receives the individual request.
3. Monitors the available of server ensuring that they are responding to the traffic, if not they are taken out.
4. Provided redundancy by employing more than one unit in fail-over case.

The major objective of cloud computing architecture is to provide on demand services. This means whatever the application is they have to scale at any point to meet the demand of user. There should be a device to distribute the request across multiple instance of application. This is achieved by Load balancer.

III. ADVANTAGES OF LOAD BALANCING

1. **Scalability:**If we have a website, we must be uploading content. And as the number of visitors on your site the amount of traffic on website is directly effect on performance of website. If there is sudden increase in traffic However, it is might become difficult for server to handle the excess traffic and the website may crash. By introducing load balancing, the traffic can be spread across multiple servers and the increase in the traffic can be handled in a much easier manner. Depending on how the site's traffic fluctuates, the server administrators can scale the web servers up or down depending upon your site's needs.
2. **Redundancy:** When we use load balancing for maintaining a website on more than one web server, the impact of hardware failure on a site's overall uptime can be limited significantly. By implementing load balancing you can achieve redundancy. This means that when the website traffic is sent to two or more web servers and one server fails, then the load balancer will automatically transfer the traffic to the other working servers. When you maintain multiple load balanced servers, you can be assured that a working server will always be online to handle site traffic even when the hardware fails.

3. **Flexibility:** By using multiple load balancer servers we can flexibility handle website traffic. It will perform several maintenance tasks on server without impacting site's uptime. This is achieved by directing all the traffic to one server and putting the load balancer in active/passive mode. We have the flexibility of having a staggered maintenance system, where at least one server is always available to pick up the workload while others are undergoing maintenance. This ensures that the site's users do not experience any outages at any time.

4. **Managing Failure:** Load balancing helps in detecting failures early and manages them efficiently, making sure that failure of any kind doesn't affect the servers or the workload. By using multiple data centers that are spread across a number of cloud providers, we can bypass the detected failures by re-distributing resources to other areas that are unaffected, thus causing minimal disruption.

IV. LOAD BALANCING ALGORITHMS

There are several algorithms that are used to balance the load of client request across server pools. The algorithm chosen will depend on the status of network and server at time of request or the type of service served. When the load is light then one of the simple techniques will be used .But when load is heavy complex methods are used to ensure even distribution of request. The methods outlined below will be used to determine the best server to service new request. Here are the most common load balancing algorithms

1. **Round Robin Algorithm**
   It is a simple mechanism in which the content access request is responded to by the load balance in a rotational basis, the first request grants access to the first available content server giving its IP address and the second to the second server IP address and so on. The moment a server IP address has been given its IP address is moved to the back of the list of available IP addresses and gradually it moves back to the top of the list and becomes available again. The frequency that it returns to the top depends on the number of available servers in the round robin server cluster being used.For example, a company has one domain name and three identical copies of the same web site residing on three servers with three different IP addresses. When one user accesses the home page it will be sent to the first IP address. The second user who accesses the home page will be sent to the next IP address, and the third user will be sent to the third IP address. In each case, once the IP address is given out, it goes to the end of the list. The fourth user, therefore, will be sent to the first IP address, and so forth.
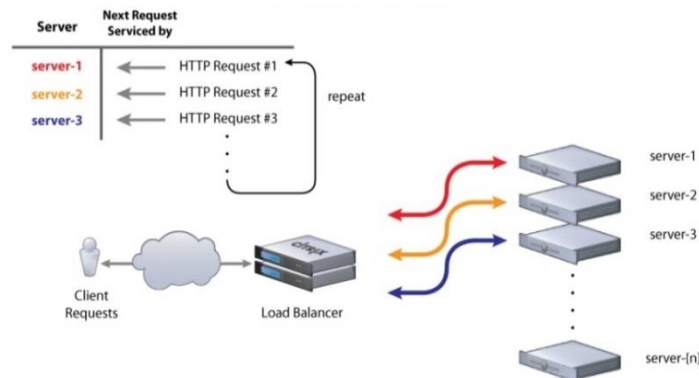
2. Weighted Round Robin Algorithm

This builds on the simple Round Robin load balancing method. In the weighted version each server in the pool is given static numerical weights. Servers with higher ratings get more requests sent to them. For example when a client's start coming in, the first 5 will be assigned to Server 1 and the 6th to Server 2. If more clients come in, the same sequence will be followed. That is, the 7th, 8th, 9th, 10th, and 11th will all go to Server1, and the 12th to Server 2, and so on.
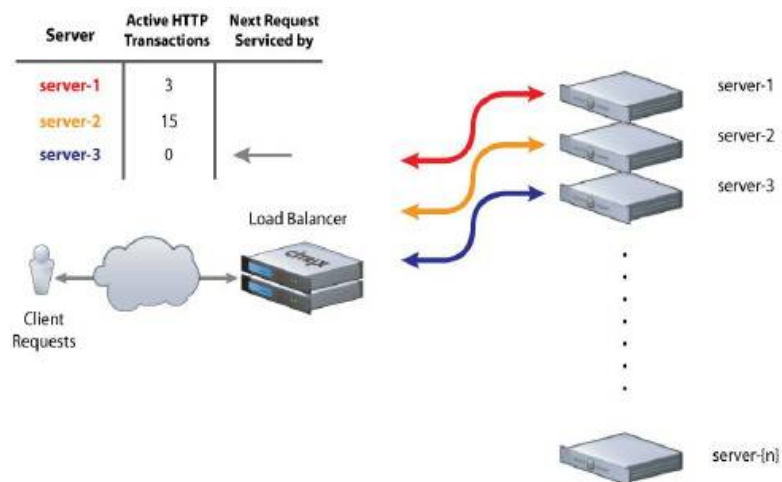
3. Least Connection Algorithm

The current server load is neither taken by Round Robin nor Weighted Round Robin when distributing requests. The Least Connection method does take current server load into consideration. The current request goes to the server that is servicing the least number of active sessions at the current time that is it checks the number of current connections each server has. For example say if client attempts to connect server 1 and server 2 but the server are already active in session, so the request is sent to Server 3 because it does not have any load.
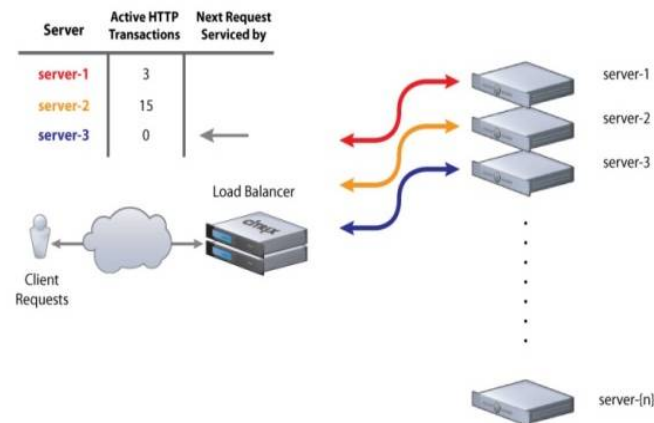
4. **Weighted Least Connection Algorithm**

It is builds on the Least Connection method. Like in the Weighted Round Robin method each server is given a numerical value. The load balancer uses this when allocating requests to servers. If two servers have the same number of active connections then the server with the higher weighting will be allocated the new request. Load balancer takes into consideration two things: the weights/capacities of each server AND the current number of clients currently connected to each server.
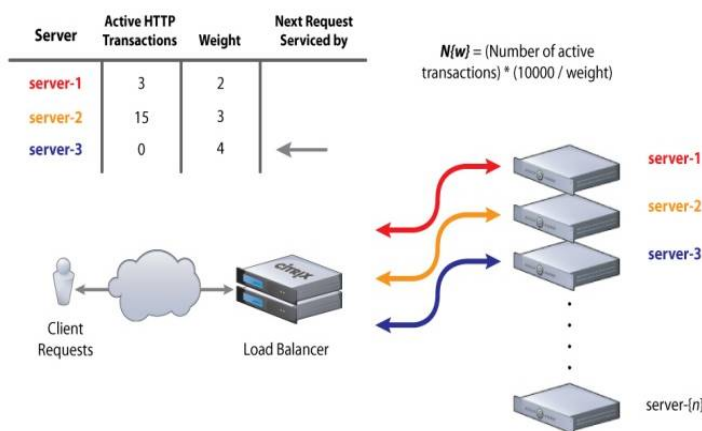
5. **Source IP Hash Algorithm**

Source IP Hash load balancing uses an algorithm that takes the source and destination IP address of the client and server and combines them to generate a unique hash key. This key is used to allocate the client to a particular server. As the key can be regenerated if the session is broken this method of load balancing can ensure that the client request is directed to the same server that it was using previously. This is useful if it's important that a client should connect to a session that is still active after a disconnection.
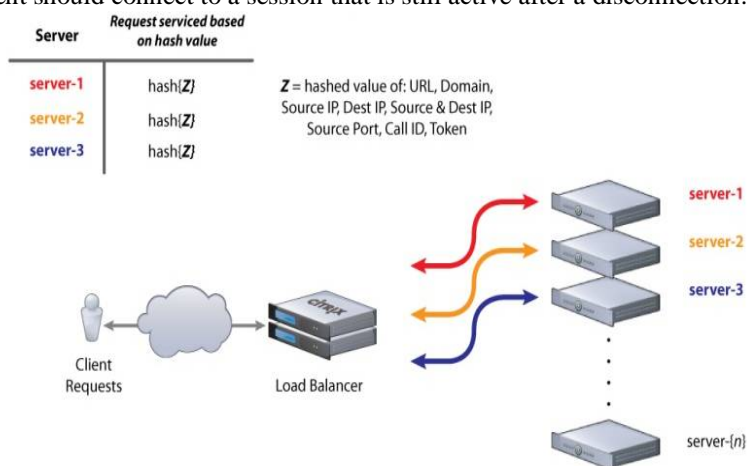
V. CONCLUSION

We have seen a rapid growth of cloud computing in recent years. But the technologies involved are still need to improve for the companies to believe and trust their business with cloud computing. There are many challenges; that

needs to be researched. In this paper we discussed about load balancing, advantages and algorithms. We conclude that this is a field that needs a lot of research and each cloud vendor will have to adopt the right technique to achieve optimal performance and business endurance.

REFERENCES

[1] Load Balancing. Available [Online]: http://aws.amazon.com/elasticloadbalancing/
[2]Amazon Elastic Compute Cloud (EC2)Available [Online] : http://www.amazon.com/gp/browse.html?
[3]Load Balancing – Least Connections Windows Azure Traffic Manager [Online]. Available: http://blogs.citrix.com/
[4] Advantage of load balancing [Online]Available -https://blog.wsol.com/the-benefit-of-load-balancing
[5]Weighted Least-Connection Scheduling [Online]. Available: http://kb.linuxvirtualserver.org/wiki/Weighted_Least-Connection_ Scheduling
[6]Least-Connection Scheduling [Online]. Available: http://kb.linuxvirtualserver.org/wiki/Least-Connection_Scheduling.
[7] Round-Robin Scheduling [Online]. Available: http://kb.linuxvirtualserver.org/wiki/Round-Robin_Scheduling.
[8] Round-robin Scheduling [Online]. Available: http://en.wikipedia.org/ wiki/Round-robin_scheduling.
[9]Weighted Round-robin Scheduling [Online]. Available: http://kb.linuxvirtualserver.org/wiki/Weighted_Round-Robin_Scheduling.
[10]Kumar, Pratik Sharma, Vishal Krishna, Chhavi Gupta, Kuwar Pratap Singh, N. Nitin, and Ravi Rastogi Nishant, "Load balancing of nodes in cloud using ant colony optimization," Computer Modelling and Simulation (UKSim), 2012 UKSim 14th International Conference, IEEE, pp. 3-8, 2012.
[11]Jasmin James and Bhupendra Verma, "EFFICIENT VM LOAD BALANCING ALGORITHM FOR A CLOUD COMPUTING ENVIRONMENT," International Journal on Computer Science & Engineering (IJCSE), vol. 4, no. 9, 2012.