IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 15-19

Load Balancing in a Network

Uday Mane
Senior Network Engineer, FIS Global Solutions. Pune , India

Abstract: *This paper introduces mechanisms for load balancing in a network. Load balancing is a networking method for distributing load across multiple servers. Load Balancing is usually provided by dedicated software or hardware, such as a multilayer switch or Domain Name System Server Process. There are various algorithms to perform load balancing. In this paper we will discuss how to perform load balancing using load balancers and the advantage and disadvantages of the same.*

Index Terms**:** *APM, ASM, GTM, LTM, Load balancers, Nodes, VIP.*

I. Introduction

Today's drastically increasing internet traffic includes Multimedia as well as Critical traffic which does not afford any discrepancy in service. To provide flawless services irrespective of heavy traffic, companies need to assemble multiple servers in their data centers. These data centers are called as server farms now a days. To achieve redundancy and load sharing between servers using traditional devices is insufficient and limited. Load balancers are the solution for the same. Load balancers are designed to boost the capacity and provide redundancy of the network while allowing traffic connections.

Load balancing has been studied extensively in the literature [2]. From the time when Internet was being established and traffic load was less compared to present, traditional methods were used to load balance the traffic. One website cannot rely on single data server for service consumers. Network Address Translation (NAT) was implemented at the routers to route the traffic to intended servers. Types of NAT i.e. Static, Dynamic, Port Address Translation (PAT) were efficiently being used to distribute traffic between server according to need [3]. But Router's main function is routing traffic between core area which lead to have limitations to serve load balancing. .Load Balancers were manufactured by companies like F5 and Citrix to fulfill these requirements with advanced secure mannerThis paper presents techniques for Load Balancing in Network domain, absence of which is an insidious factor that can reduce the performance of a simultaneous applications significantly. For some applications, load is easy to predict and does not vary dynamically. However, for a significant class of applications, load representations by pieces of computations vary over time, and may be harder to predict.

This is becoming increasing prevalent with emergence of new-sophisticated applications. It enables Network Engineers to achieve greater levels of fault tolerance and improved performance seamlessly providing the required amount of load balancing capacity needed to distribute network traffic.

Rest of the paper is organized as follows: Section II gives an overview, background and real-life applications of load balancing using different scenarios. Advantages and disadvantages are discussed in Section.

In any data center environment, multiple Data servers are hosted to provide fluent and robust service. If we study any simplest data center which provides load balancing between multiple servers. Figure 1 shows the simple data center network topology to understand load balancing.

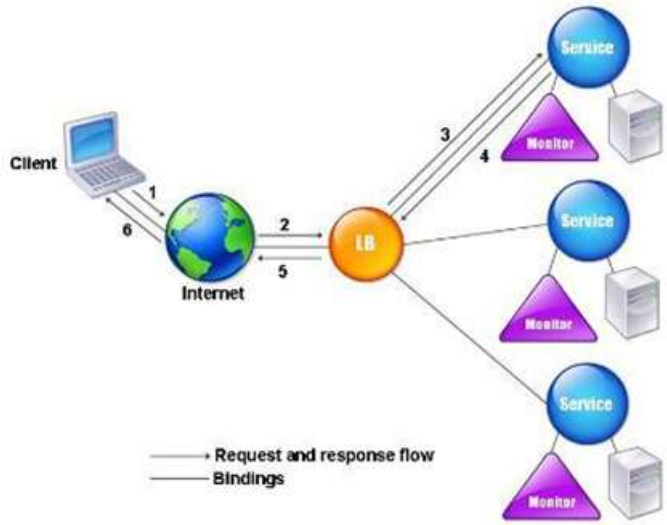

Fig 1 Basic Data Center Network Topology

Network topology in Figure 1 includes Network devices called Firewalls, Load Balancer (Here example is F5). We can have multiple data servers in environment where every request from User/ Client would decide where to forward traffic after Load Balancer. Here multiple types of request can be sent from client machines. HTTPS, FTP, Proxy, SSH are few popular examples of the request running now a days. Load balancer has One Virtual IP (VIP) address which is Public IP address generally reachable from internet. Any client can hit that after DNS resolution. Once Load Balancer (LB) receives request from user it redirects it e.g. If User wants to connect to http://sabc.com LB will redirect traffic according to configuration into internal server pool. This internal pool will include Private IP addressed servers. Multiple Load Balancing algorithms are developed [2] to distribute traffic between internal servers. This helps Network developers to publish only one public IP address which is configured as VIP on LB to Internet; while actual internal Private IP addressed servers are hidden for internet users. Also, multiple servers boost site capacity of users/load and avoids disconnections. Few majorly used load balancing algorithms are described later in this document. Load Balancer also acts as Firewall being Access Policy Manager (APM) to allow secured connections only.

II. Load Balancing Algorithms

A. Round Robin:

Round Robin passes each new connection request to the next server in line, eventually distributing connections evenly across the array of machines being load balanced. Figure 2 shows Round Robin algorithm;

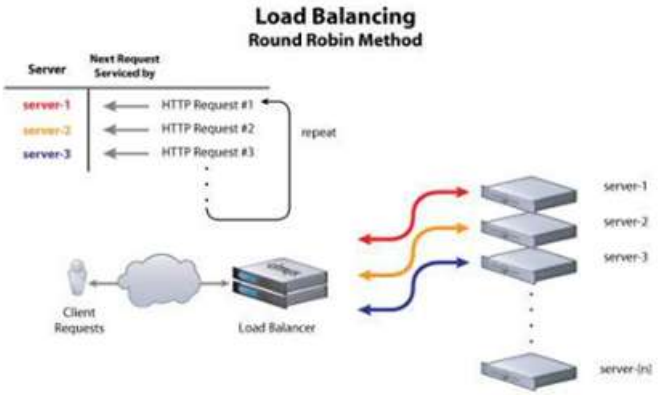

Fig 2 Round Robin algorithm for load balancing

B. Weighted Round Robin:

With this method, the number of connections that each machine receives over time is proportionate to a ratio weight defined for each machine on LB. This is an improvement over Round Robin because any internal

machine can have better performance compared to other members. Figure 3 describes the process of Ratio / Weighted round robin method effectively.

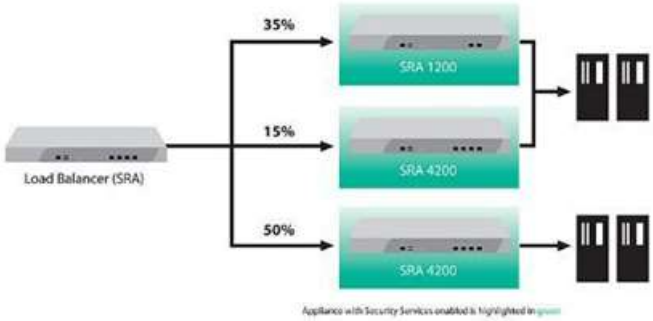

Fig 3 Weighted round robin / Ratio method for load balancing

C. Fastest:

The Fastest method passes a new connection based on the fastest response time of all servers. This method may be particularly useful in environments where servers are distributed across different logical networks. On the BIG-IP, only servers that are active will be selected

D. Least Connections:

With this method, the system passes a new connection to the server that has the least number of current connections. Least Connections methods work best in environments where the servers or other equipment that are load balancing have similar capabilities. Figure 4 shows the least connection method for load balancing. This is a dynamic load balancing method, distributing connections based on various aspects of real-time server performance analysis, such as the current number of connections per node or the fastest node response time.

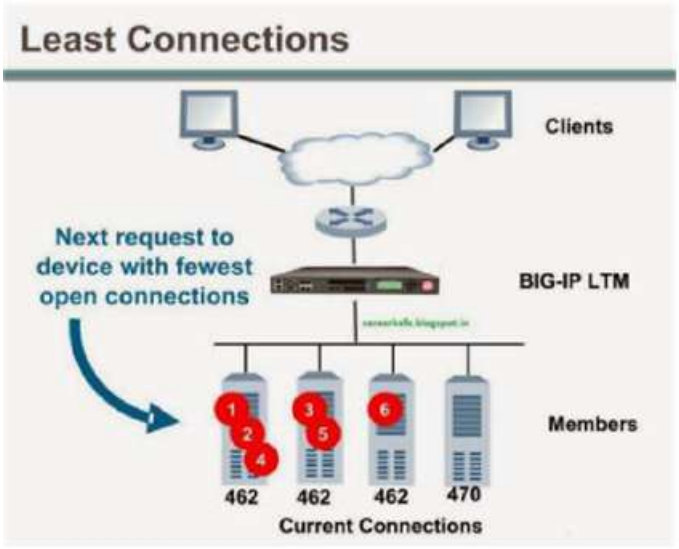

Fig 4 Least connections method for load balancing

E. Observed:

The Observed method uses a combination of the logic used in the Least Connections and Fastest algorithms to load balance connections to servers being load-balanced. With this method, servers are ranked based on a combination of the number of current connections and the response time. Servers that have a better balance of fewest connections and fastest response time receive a greater proportion of the connections.

F. Predictive:

The Predictive method uses the ranking method used by the Observed method, however, with the Predictive method, the system analyzes the trend of the ranking over time, determining whether a server's performance is currently improving or declining. The servers in the specified pool with better performance

rankings that are currently improving, rather than declining, receive a higher proportion of the connections. The Predictive methods work well in any environment.

III. Features Used In Load Balancers

A. Access Policy Manager (APM):

Today, local and remote employees, partners, and customers often access applications without context or security. A central policy control point delivers access based on context and is critical to managing a scalable, secure, and dynamic environment. It protects public-facing applications by providing policy-based, context-aware access to users while consolidating the access infrastructure. It also provides secure remote access to corporate resources from all networks and devices. Conclusively, it can act as a Firewall to allow secure connection.

B. Local Traffic Manager (LTM):

LB turns network into an agile infrastructure for application delivery. It's a full proxy between users and application servers, creating a layer of abstraction to secure, optimize, and load balance application traffic. This gives the flexibility and control to add applications and servers easily, eliminate downtime, improve application performance, and meet security requirements.

C. Take control over application delivery

The F5 TMOS platform gives complete control of the connection, packets, and payload for applications. Using event driven configuration statements can customize intercept, inspect, transform, and direct inbound and outbound application traffic.

D. Reduce servers, bandwidth, and management costs

Advanced TCP connection management, TCP optimization, and server offloading enables to optimize the utilization of existing infrastructure—tripling server capacity and reducing bandwidth costs by up to 80 percent. BIG-IP LTM helps to simplify system management. By using fewer servers, less bandwidth, less power, and less cooling, while reducing the time spent managing the infrastructure, it reduces operational costs.

E. Application Security Manager (ASM):

It is a flexible web application firewall that secures web applications in traditional, virtual, and private cloud environments. BIG-IP ASM provides unmatched web application and website protection, helps secure deployed applications against unknown vulnerabilities, and enables compliance for key regulatory mandates with data center firewall capabilities, and network and application access control.

F. Global Traffic Manager (GTM):

Global Traffic Manager (GTM) distributes DNS and user application requests based on business policies, data center and network conditions, user location, and application performance. High-performance DNS Services with visibility, reporting, and analysis; scales and secures DNS responses geographically to survive DDoS attacks. It delivers a complete, real-time DNSSEC solution; and ensures global application high availability which is the redundancy while accessing services.

G. Application Acceleration Manager (AAM):

Its Purpose is to overcome WAN latency, maximizes server capacity, and speeds application response times. AAM decreases the need for additional bandwidth and hardware so users get fast access to applications.

DISADVANTAGES OF LOAD BALANCER'S USAGE

Management – Difficulty in direct access to hardware, for example, specific cards or USB devices.

Great RAM consumption: Great use of disk space, since it takes all the files for each operating system installed on LB which increases cost of the devices.

IV. Conclusion

Cloud computing allows wide range of users to access distributed, scalable, virtualized, hardware and software resources over the Internet. Load balancing is one of the most important issue of cloud computing. It is a mechanism which distributes workload evenly across all the nodes in the whole cloud. Through efficient load balancing, we can achieve a high user satisfaction and resource utilization. Hence, this will improve the overall performance and resource utility of the system. With proper load balancing, resource consumption can be kept

to a minimum which will further reduce energy consumption and carbon emission rate. Through hierarchical structure of system, performance of the system will be increased.

Acknowledgment

Author would like to thank for Mrs. Shilpa Kharche and colleagues for constant guidance and unconditional support.

References

[1]. CCNA Routing and Switching Complete by Todd Lammle, "The Official guide for CISCO Networks" (2014, 2nd Edition, Kindle Edition) http://www.m5zn.com/newuploads/2014/06/30/pdf/m5zn_082644c04a97cdd.pdf

[2]. Phillip Jonsson, Steven Inveson, "F5 Networks Application Delivery Fundamentals", Kindle Edition, 2017. https://support.f5.com/csp/article/K29900360

[3]. "Mastering Netscaler VPX", Rick Roetenberg, Marius Sandbu

[4]. Ccna Security, 1st Edition (Sybex Publications , Tim Boyles),staffweb.itsligo.ie/staff/pflynn/Telecoms%203/CCNP%202%20Secure%20WAN's/Secure%20Converged%20Networks/CCNA%20Security.pdf