

LOAD BALANCING IN WEB SERVER

Renu Tyagi¹, Shaily Chaudhary², Sweta Payala³

*UG,^{1,2,3} Department of Information & Technology ,
Raj Kumar Goel Institute of Technology for Women,
Gautam Buddh Technical University, Lucknow, (India)*

ABSTRACT

The wide growth of internet caused a huge increment on the number of users requesting services through the network, as well as the number of servers and the amount of the services they offer. We have to minimize this problem through use of various load balancing techniques. In this paper we represent the load balancing result of the web server. There are basic terminologies, open flow technology integrated server load balancing at wire speed which influence the load balancing of web server. The performance of the server is greatly improved through round robin algorithms. We will analyze the efficiency of the various approaches and their trade off.

Key Words: Load Balancing, Over Flow, Integrated System.

I INTRODUCTION

The IT infrastructure is playing an increasingly important role in the success of a business. Market share, customer satisfaction and company image are all intertwined with the consistent availability of a company's website. Network servers are now frequently used to host ERP, e-commerce and a myriad of other applications. The foundation of these sites – the e-business infrastructure – is expected to provide high performance, high availability, and secure and scalable solution to support all applications at all times. However, the availability of these applications is often threatened by network overloads as well as server and applications failures. Resource utilization is often out of balance, resulting in the low - performance resource being overloaded with requests while the high performance resources remain idle. Server load balancing is the distributing service requests across a group of servers. Server load balancing addresses several requirements that are becoming increasingly important in network problem.

II INTERNET SIZING FOR GROWTH REASON

The load on an internet site is unlikely to remain constant. The number of access on a Web server or FTP server can increase for several reasons.

- i. Most companies add their Web site's address to television, radio and print advertising and to product catalogues and brochures. As these "Web-aware" publications circulate and replace the previous "URL-free" versions, awareness of the Web site grows.
- ii. As time passes, the Website gains better coverage in the on- line search engines such as Yahoo or Alta-Vista.
- iii. Assuming the site is providing useful information or a useful service to customers, repeat visitors should increase.
- iv. Most Web sites begin simply, with fairly modest content – mostly text, with some images. As the site designers grow in confidence, more resources are allocated, and as Web users in general increase their modem speeds, most sites move towards richer content. Thus, not only do hit rates increase, but the average data transfer per hit also rises.
- v. Most sites begin as presence sites – providing corporate visibility on the internet and making information about the company available to potential customers. Most present sites use predominantly static HTML pages. Static pages are generated in advance and stored on disk. The server simply reads the page from the disk and sends it to the browser. However, many companies are now moving towards integration applications that allow users of the Web site to directly access information from the company's existing applications. This could include checking the availability of products, querying bank account balances or searching problem databases. These applications require actual processing on the server system to dynamically generate the Web page. This dramatically increase the processing power required in the server.

III DEALING WITH THE GROWTH

There are several ways to deal with the growth of your internet site:

- i. Purchase an initial system that is much too large.
- ii. Replace the server with the larger system.
- iii. Purchase an upgradeable SMP system.
- iv. Perform load balancing between multiple servers.

IV BASIC LOAD BALANCING TERM

Most load balancers have the concept of a node, host, member, or server; some have all four, but they mean different things. There are two basic concepts that they all try to express. One concept- usually called a node or a server- is the idea of the physical server itself that will receive traffic from the load balancer, would be the IP address of the physical server and, in the absence of a load balancer, would be the IP address that the server name(for example, www.example.com) would resolve to. For the remainder of this paper, we will refer to this concept as the Host.

The second concept is a member (sometimes, unfortunately, also called a node by some manufactures). A member is usually a little more defined than a server/node in that it includes the TCP port of the actual application that will be receiving traffic. For instance, a server named `www.example.com` may resolve to an address of `172.16.1.10`, which represents the server/node, and may have an application (a web server) running on TCP port 80, making the member address `172.16.1.10:80`. Simply put, the member includes the definition of the application port as well as the IP address of the physical server. For the remainder of this paper, we will refer to this as the service.

Why all the complication? Because the distinction between a physical server and the application services running on it allows the load balancer to individually interact with the applications rather than the underlying hardware. A host (`172.16.1.10`) may have more than one service available (HTTP, FTP, DNS, and so on). By defining each application uniquely (`172.16.1.10:80`, `172.16.1.10:21`, and `172.16.1.10:53`), the load balancer can apply unique load balancing and health monitoring (discussed later) based on the services instead of the host. However, there are still times when being able to interact with the host (like low-level health monitoring).

V SERVER LOAD BALANCING AND ITS BENEFITS

Server load balancing is the process of distributing service requests across a group of servers. Server load balancing addresses several requirements that are becoming increasingly important in networks.

- i. Increased scalability
- ii. High performance
- iii. High availability and disaster recovery

Server load balancing makes multiple servers appear as a single server- a single virtual service- by transparently distributing user requests among the servers.

The highest performance is achieved when the processing powers of servers is used intelligently. Advanced server load balancing products can direct end user service requests to the servers that are least busy and therefore capable of providing the fastest response times. Necessarily, the load balancing device should be capable of handling the aggregate traffic of multiple servers. If a server load balancing device becomes a bottleneck it is no longer a solution, it is just an additional problem.

Another benefit of server load balancing is its ability to improve application availability. If an application or server fails, load balancing can automatically redistributes end users service requests to other servers within a server farm or to server in another location. Server load balancing also prevents planned out ages for software or hardware maintenance from disrupting service to end users.

VI SERVER LOAD BALANCING IN SPECIALIZED ENVIRONMENTS

For applications that still require processing intensive features such as URL/cookie and SSL ID persistence, it makes sense to use an external server load-balancing appliance to supplement Extreme Networks integrated wire-speed server load balancing capability. This approach provides the best of both solutions. The specialized functionality of an appliance-

- i. Greater performance of wire-speed switching solutions
- ii. Lower overall system costs

An ideal solution consists of combining Wire-speed IP routing at layer 3 and wire-speed layer 2 switching with specialized devices working side-by-side for a best-of-breed solution.

VII INTEGRATED SERVER LOAD BALANCING AT WIRE SPEED

Leveraging gigabit speed, Extreme Networks scales Ethernet by allowing managers to build larger, fault-tolerant networks while controlling bandwidth based on the relative importance of each application. Extreme networks delivers Wire-Speed IP Routing at Layer 3 and Wire-Speed Layer 2 switching as well as end-to-end Policy-Based Quality of service (QOS) and wire speed access policies at Layer 4 with resiliency options designed to reduce the network ownership.

7.1 A summary of the advanced server load balancing capabilities Extreme Networks offers includes:

- i. Hardware integration for wire-speed server-to-client performance
- ii. Web cache redirection techniques for full, wire-speed traffic redirection
- iii. Capability's across single or multiple web caches or other types of caches.
- iv. Coordination of high-availability server load balancing features with Layer 3 and Layer 2 resiliency techniques for simple and effective redundancy.
- v. Sophisticated high-availability capabilities, such as exchanging session information between active and standby server load balancing services and "active/active" configuration.
- vi. Flexible "persistence" options to preserve session integrity with servers and optimize hits on web cache servers.

7.2 This approach provides significant benefits when compared to point products or special purpose appliances:

- i. Server load balancing is delivered as an overlaid service on the existing network infrastructure. There is no need to redesign the network to accommodate server load balancing.
- ii. Wire-speed performance for server load balancing and transparent web cache redirection applications.

- iii. True integration provides simpler and more resilient solutions for link, switch, router and load balancing capabilities.
- iv. Coordinated capabilities for Policy-based Quos, access policies and system security.
- v. Fewer devices to manage and less training required.
- vi. Lower cost of network ownership.

VIII OPEN FLOW

The open flow technology allows the testing of experimental protocols in real networks concomitantly with production traffic. It is an abstraction and virtualization technology for networking, allowing control over network traffic through data flow.

This technology enables the management of programmable switches, allowing the isolation of different types of traffic by means of an external controller. This is possible due to the flow tables present in switches and routers, which are used to implement NAT (Network Address Translation), QoS (Quality of Service) and other features.

Despite the variation of these tables among different manufactures there is a standard set of functions present on all devices.

An open flow based network is composed of three main components: Open Flow switches, the Open Flow protocol and the open flow controller. The relationship among network components can be viewed in more detail. An open flow switch is composed by three main components:

8.1 Flow tables: these contain the information used by the switch to process frames of a given flow. For flows under the control of Open Flow, possible actions are defined.

8.2 Group tables: it allows a flow to point to a group, increasing the forwarding options of frames;

8.3 Secure communication channel: It is used for the communication channel between a switch and a remote controller, enabling control commands to be sent to the device responsible for managing network traffic. The protocol defines the standard for communication between a switch and a controller, allowing the addition, removal and update of entries in flow tables. The third component, the controller, is responsible for managing the switch through the open flow protocol.

The main function of the controller is to add and remove flows from the switch flow tables. Upon receiving a frame, an open-flow enabled switch searches its flow tables for any actions defined for the flow of the received frame. If an entry is found, the specified actions are performed and the frame is forwarded. Otherwise, the frame is sent to the controller, which will determine the action to be performed for the received frame, possibly adding an entry into the

flow table for the next frames of the same flow. This flexibility of the controller allows the specification of a given flow and its actions with great level of detail.

IX BALANCING ALGORITHMS

A key feature of server load balancing is its ability to intelligently direct service requests to the most appropriately server. Extreme Networks Switches offer the following integrated server load-balancing algorithms to successfully accomplish this:

9.1 Round robin: A simple algorithm that distributes each new connection/session to the next available server

9.2 Weighted round robin with response- time as weight: An enhancement of the round robin method where response time for each server within the virtual service are constantly measured to determine which server will take the next connection/session.

9.3 Fewest connections with limits: Determine which server gets the next connection by keeping a record of how many connection each server is currently providing. The server with fewer connections gets the next request.

The round robin algorithm can be effective for distributing the workload among servers with equal processing capacity. When servers differ in their processing capacity, using response time or number of active as the selection criteria can optimize user response time.

X CONCLUSION

Load balancers have become integral to application delivery. In fact, they have become so essential in modern networks that the technology has evolved into more powerful solution commonly dubbed application delivery controllers, or ADCs. In addition to providing very rich load balancing capabilities, ADCs include advanced functionality such as SSL offload, HTTP compression, content caching, application firewall security, TCP connection management, URL rewriting and application performance monitoring. Extreme network provides the key benefits of server load balancing while eliminating the potential cost, complexity, and performance issues. By fully integrating server load balancing into its wire-speed multilayer switches. Open Flow enables the creation of new techniques for load balancing, with

Greater flexibility and control over each server and also over each data flow, providing the ability to control the load of the servers at any time. Hence, we proposed three flow based load balancing policies. Through our measurements, weighted balancing achieved the best results over the other policies.

REFERENCES

- [1] AWEYA, J.; OUELLETTE, M.; MONTUNO, D.Y.; DORAY, B.; Felske, k.. *“An adaptive load balancing scheme for web servers”*. 2002
- [2] CHIN, M.L.; TAN, C.E.; M.. *“Efficient load balancing for bursty demand in web based applications services with domain named services”*. 2010
- [3] M.F. Arlitt, C.L Williamson, *“web server work load characterization. The search for invariants”*, IEEE /ACM Trans. On networking, Oct 1997
- [4] P.Barford, M.Crovella, *“Generating representative Web workloads for network and server performance evaluation”*, June 1998