SERVER CONNECTION OFFLOAD

MAXIMIZING SERVER PERFORMANCE USING LOAD BALANCERS TO OFFLOAD CONNECTION MANAGEMENT FROM SERVERS

Overview

Web-enabled IP applications are at the heart of economic activity, and their availability, performance and response time immensely affect the productivity and profits of various organizations around the globe. Traditionally, these mission-critical applications are implemented on servers running different types of operating systems like UNIX, Windows and Linux. Using standardized server and OS platforms to build and implement IP applications gives deployment flexibility, eases management and maintenance, and reduces the total cost of ownership. While the new server technology kept pace with Moore's law and offers increased performance year-over-year, OS implementation complexities still remain and limit servers performance, which is adversely impacted by the inefficient connection management between clients and application servers. Even with the benefits of Moore's law for newer servers, the "installed base" of server and application infrastructure continues to be dogged by poor performance. Studies show that 30 to 40% of the server resources are used for connection management and other overhead, which severely limits the application response time, scalability and availability. Additionally, placing the burden of connection management on the servers also makes them vulnerable to various types of malicious attacks, which take advantage of TCP/IP connection setup mechanisms to launch DoS (Denial of Service) attacks.

Offloading connection management from the application servers immediately increases their available capacity to handle application transactions more effectively. It also provides robust security to server and application infrastructure, which increases the overall application availability. Recovering and redeploying server resources for the real purpose of supporting application traffic will protect investment in server farms and increase the ROI (Return on Investment).

High-performance Layer 4-7 load balancing switches are widely used as the technology of choice to increase application availability, security and response time. The leading Layer 4-7 switches in the market, like the Foundry ServerIron, also support "Server Connection Offload" feature. By implementing this function on intelligent load-balancing devices, network and application managers can increase available capacity of the server infrastructure, and simultaneously take advantage of the application and server farm security and scalability. Load balancers, front ending the server farm, minimize connection management overhead on the servers by utilizing the new HTTP1.1 protocol to switch traffic from many client-side connections to a few server-side connections. By maintaining an optimal ratio of client-side to server-side connections, the load balancers ensure that the application performance and server utilization is maximized.

HTTP 1.0 and 1.1 Connection Management

Web applications use standards-based HTTP (Hyper Text Transfer Protocol) to exchange information between the client (browser) and the application servers. Web application transactions involve setting up a TCP connection between the client and the server, requesting and receiving application information, and closing the TCP connection.

There are two widely used versions of the HTTP protocol: HTTP 1.0 and HTTP 1.1. In the HTTP 1.0 version, only one request/response for application document is allowed per TCP connection. In other words, each application message exchange between the client and the servers requires a separate TCP connection.

Figure 1 below shows an example of two HTTP 1.0 connections between the clients and a server. The ServerIron Layer 4-7 switch sits in between the clients and the servers, and load balances client requests to multiple "Real Servers" (not shown in the figure). Each connection between the client and the server involves one application level message exchange, and is followed by the termination of the connection.

When there are thousands of clients, each requesting many application documents, the total number of TCP connections quickly multiplies, and becomes a significant burden to the real servers. The added load of

connection management can cause the servers to perform very poorly and respond slowly, which causes downtime and lost productivity. Using the ServerIron switches to load balance to multiple servers will help improve the response time and scalability. But, load balancing does eliminate TCP connection overhead on the servers. Additionally, because the TCP protocol uses Slow Start mechanisms that limit the throughput achieved initially on a new connection, using a new connection for each application exchange forces poor performance. Server Connection Offload feature on the load balancer is designed to maximize server utilization, and overall application performance and throughput.

HTTP 1.1 protocol version, when used by both the clients and the servers, helps improve connection efficiency by allowing multiple application requests and responses over a single TCP connection between the client and the server. This version has many more new features, but the focus in this document will be on the enhancement in HTTP1.1 that allows multiple exchanges over the same connection.

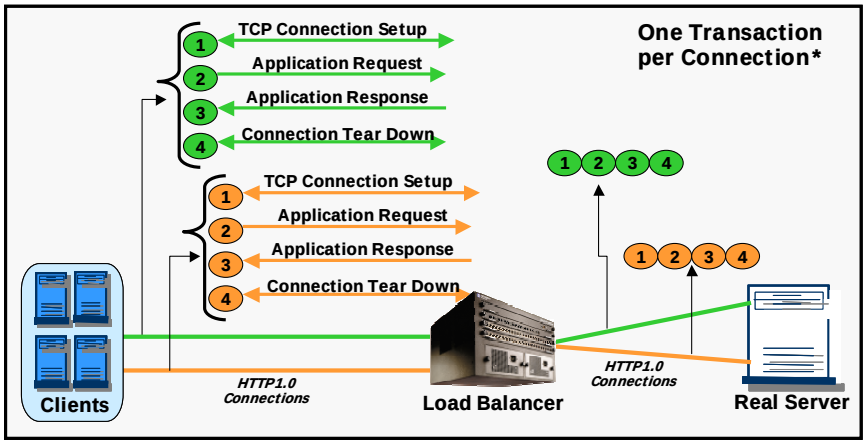

Figure 1: HTTP 1.0 Connections Between Client and Server

Figure 2 below shows a client connecting to a real server using an HTTP1.1 connection and exchanging multiple application messages over the same connection. Either the client or the server may gracefully terminate the connection at any time, or send keep-alive messages to continue using the same connection. Using HTTP 1.1 connections between clients and servers will reduce the connection management overhead to some extent, but it could still be very high when the servers and applications are supporting thousands to millions of clients, or when many clients are limited to using HTTP1.0 even though the servers support HTTP 1.1. Additionally, servers and applications are still vulnerable to security threats in the form of DoS, Virus and Worm attacks, and this threat cannot be mitigated by simply switching to HTTP1.1 connections.

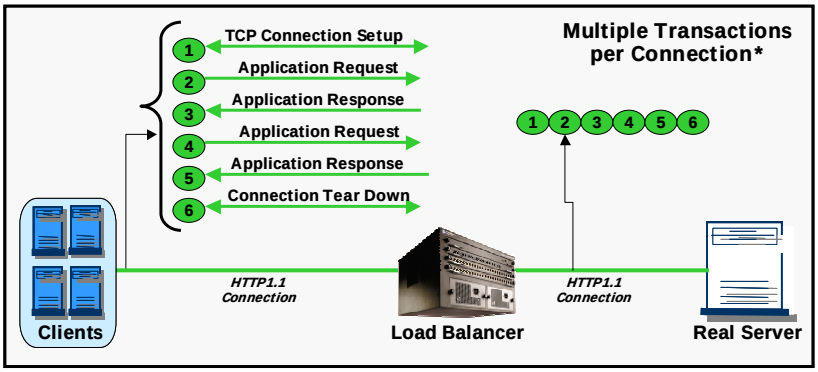

Figure 2: HTTP 1.1 Connections Between Client and Server

Server Connection Offload using Foundry ServerIron Layer 4-7 Switches

Foundry's industry leading ServerIron Layer 4-7 load balancing switches feature "Server Connection Offload" feature that eliminates the need to setup and teardown connections to the server for every client that needs to access IP applications. The ServerIron switches establish a few server-side HTTP1.1 connections, and switch traffic from many client-side connections to these semi-permanent server-side connections. The real servers are completely shielded from client connection requests and connection management, and need to only maintain few connections to the load balancer that are used for extended duration. All server resources are dedicated to process application transactions, and not for connection management, which maximizes application performance and response time. With the conserved resources, the servers can now support more clients than they otherwise could when burdened with connection management. Because clients do not directly establish connections to the servers, the servers are totally protected from DoS attacks. When malicious users launch attacks against the servers and applications, the ServerIron switch identifies and discards the attack messages without the servers ever receiving them.

Using the Server Connection Offload feature on the ServerIron switches, businesses can scale their server farms to support thousands to millions of clients with few tens and hundreds of connections to the servers. Tests and independent studies show that, on average, 20 to 40 percent of the server resources are conserved when using connection offloading. Combining Server Connection Offload a rich set of Layer 4-7 features and high security and connection performance, and ServerIron switches help maximize server utilization and deliver always-on applications.

Figure 3 below shows two client-side connections (either HTTP1.0 or HTTP1.1) being terminated by the load balancer, and switched over a single HTTP1.1 connection on the server-side. Once the server-side connection is established, only the application request response messages go to the server(s), irrespective of how many client-side connections are established. The load balancer treats client-side and server-side connections as independent, and supports the offload function for client-side connections that use either HTTP 1.0 or 1.1 versions. The load balancer always uses HTTP1.1 connections to the server, which allow multiple transactions over a single connection.

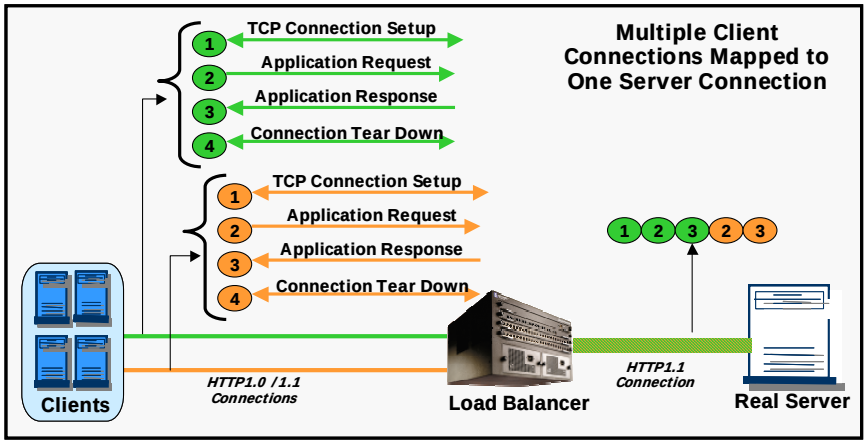

Figure 3: HTTP 1.1 Connections Between Load Balancer and Server with Multiple Client Connections

The load balancer terminates client connections fully, which allows it to examine clients' application messages and perform content switching like URL and HTTP Header switching on each message, and also maintain session persistence using the traditional methods like source IP, cookie and SSL Session ID.

Server Connection Offload function is completely transparent to the clients and the applications, and can be implemented in existing and new IP application environments without disruption to behavior or infrastructure changes as long as the application servers support HTTP1.1 protocol version.

ServerIron Key Benefits

Application Performance

ServerIrons switches, with their intelligent application-aware load balancing and content switching, significantly improve overall performance by optimally utilizing server resources. Using customizable load balancing methods and metrics, application performance can be tuned to achieve best response time and maximum throughput. By taking advantage of HTTP1.1 protocol mechanisms, the ServerIron switches support Server Connection Offload feature, which eliminates connection overhead from the servers and provides robust security. Server resources are truly dedicated to maximize application performance and user response time.

Application Availability

High-performance load balancing using ServerIron switches ensures always-on applications by intelligently distributing application traffic among all available servers, and dynamically monitoring the ability of serves and applications running on them to deliver optimal performance. Using customizable health checks at various levels of granularity like host, port, application and transaction, ServerIron switches instantaneously and transparently react to increases and decreases in server resources by re-directing client traffic as needed. To protect applications from catastrophic failures, the switches can be deployed in multiple high-availability modes with stateful session failover. Applications are completely transparent to switch failures, and continue to function uninterrupted.

Application and Server Farm Security

Security is a critical challenge for businesses, especially for the mission-critical applications where the stakes are very high. As reliance on the network to deliver the mission-critical applications increases, so does the threat posed by network-based attacks. ServerIron switches have many intelligent features and superior performance to reliably protect against many forms of DoS, Virus and Worm attacks. They protect application infrastructure and server farms against wire-speed Gigabit rate DoS attacks, which translates to 1.5 million attack messages in a one second duration. ServerIron product family features industry's most advanced security intelligence to provide high-performance IronShield™ security that meets the needs of even the most demanding networks and applications serving millions of clients.

Application and Server Farm Scalability

Scaling applications and server farms is one of the most fundamental requirements for continued business growth, and is easily and permanently met by the ServerIron load balancers. ServerIron switches provide unlimited scalability to any IP-based application, and allow businesses to leverage commodity servers to build highly sophisticated and secure application infrastructure. Massive scalability is achieved with complete transparency to existing clients and servers without downtime.

Higher Return on Investment (ROI)

Foundry Networks' ServerIron load balancers provide immediate ROI, and also improve the ROI of application and server infrastructure. By implementing the new "Server Connection Offload" feature in existing server farm and application deployments, customers can immediately improve the overall capacity by an average of 20 to 40%. The ServerIron switches support significantly higher application traffic and clients with existing resources by efficient utilization. Downtime associated with security breaches, and server and application maintenance is

eliminated, resulting in improved availability. Load balancers also simplify application and server farm management, which improves productivity and helps conserve valuable capital to address other critical problems in the network.

Summary

Servers are a necessary component of the mission-critical application infrastructure, but are not optimized for connection management. When the connection load surges, server performance suffers drastically, which affects application availability. Additionally, server farm vulnerability to security threats is the greatest when they are exposed to the clients directly for connection management. Malicious users can launch DoS attacks that take advantage of TCP connection setup mechanisms, and cripple the servers and the applications.

Layer 4-7 load balancing switches have evolved to be the technology of choice to scale IP applications, improve availability and security. By taking advantage of HTTP1.1 protocol mechanisms, the Layer 4-7 switches support Server Connection Offload feature, which eliminates connection overhead from the servers and provides robust security. Server resources are truly dedicated to maximize application performance and user response time. By conserving and re-covering server resources and allowing the resources to be (re-)deployed for valuable application support, the load balancers improve server farm and application ROI in existing and new deployments. By transparently deploying the Server Connection Offload function, IT managers can extend the life of exiting server farms and protect their investment.

ServerIron Platform Support

The Server Connection Offload feature is supported on the ServerIron 100/400/800 products in the TrafficWorks IronWare 9.1R software release.