WHITE PAPER

SERVER LOAD BALANCING IN TODAY'S WEB-ENABLED ENTERPRISES

FOUNDRY NETWORKS

FOUNDRY NETWORKS WHITE PAPER

Introduction

As more and more businesses have turned to using Web-based applications for everything from generating customer sales and communications to internal HR and sales applications, the spotlight has increased on IP networks to provide reliable and timely communication links to support mission-critical applications. With the resultant boom in TCP/IP traffic generated by these applications, new network traffic management devices have emerged, supplying smart content switching capabilities that monitor systems and distribute incoming traffic for optimal response.

Web, content, or Layer 4-7 switches, as these devices have come to be known, provide capability to intelligently route Internet traffic to application servers using load-balancing techniques. The more advanced Web switches can provide Layer 7 traffic routing by examining IP packets in more detail and forwarding based on HTTP header, URL, and cookies. For global companies, Web switches can distribute traffic to servers located anywhere in the world, providing users with the best possible response times and unmatched overall reliability.

In this brief we examine the role Server Load Balancing (SLB) and Global Server Load Balancing (GSLB) serves in today's enterprises running Web-based applications and how Web switches can provide an effective solution to manage and improve server performance related issues.

Contents

SERVER LOAD BALANCING IN TODAY'S WEB-ENABLED ENTERPRISES

Issues with Web-enabled applications

The Internet's communication protocol, HTTP, has allowed distributed client/server computing at unprecedented levels. Powerful applications for the exchange of information and transaction processing have emerged and been successfully implemented by manufacturing and services companies alike.

In this world of Internet computing, Web browsers provide the onramp to access applications, distributed across multiple sites and servers. The ubiquity of Web browsers on enterprise desktops combined with industry-wide standardization around the HTTP communications protocol has encouraged enterprises to invest in developing and/or porting applications that utilize data stored in OLTP systems and Supply Chain Management (SCM), and Customer Relationship Management (CRM) databases onto the Web using a wide range of Web languages such as HTML, DHTML, Java, CGI and Active Server Pages. Discussion of Web application development is beyond the scope of this brief but it is important to understand issues related to operating applications developed to run over the Web. Running on top of TCP, HTTP is a connectionless protocol allowing it to thrive over the Internet; but there are unique characteristics to be aware of in deploying these applications:

Scalability and Management

In addition to new application development, organizations have invested in upgrading their servers, installing additional servers and implementing redundant systems, hoping to provide their customers with faster response times, leading to better overall customer satisfaction. But introducing multiple servers brings forth additional requirements such as quickly propagating address changes when bringing servers offline for maintenance and doing so transparently to users. Effective server management requires control of IP addressing where multiple virtual IP (VIP) addresses are being used for server farms. VIP addresses are translated to registered IP addresses before traffic is routed to the public network, thereby providing virtually unlimited availability to non-registered addresses for adding to the server farm. The ability to implement VIPs also provides quicker network updates by eliminating the need to re-address a network that requires public network interaction.

For multinational organizations that have data centers in multiple locations, managing traffic load between sites takes on additional complexity. Directing users requests to sites based solely on DNS tables limits the number of sites available to users. It also reduces full and efficient utilization of all computing resources. Worldwide data center traffic management requires intelligent information about site load, server health, and even requesting client location.

Availability

As enterprises rely on Web-enabled applications to communicate and transact with their customers, these applications by definition become mission critical. Obstacles to accessing applications due to server latency, downtime or errors are therefore unacceptable. A widely cited statistic concerning e-commerce transactions shows why: response times greater than 7 seconds result in loss of 30% of prospective customers. For other applications such as online banking and manufacturing systems, slow response times leads to significant business disruption that is not acceptable. The ability for Web switches to monitor server health and performance

and respond by redirecting requests to better performing servers is a crucial component to Web-enabled applications.

Persistence

Web browser clients are effective in retrieving distributed information and in posting client side data. This dual role is integral to client/server applications and most familiar in e-commerce applications where a client will have multiple interactions usually with more than one server during the process of browsing, selecting product for purchase, and checkout. Developers usually separate these applications in separate servers and add additional servers as demand increases.

These applications are known as "stateful" since the user is returned to the same server, or group of servers when specialized applications are needed, to resume the state (e.g., shopping cart contents) and finish the transaction.

During the interaction, multiple HTTP sessions will be established and torn down by the client and server. Higher, content level information is therefore needed to return the client to the same server, differentiating it from a multitude of machines in a typical Web farm implementation that can fulfill the request. This requirement for stateful applications is know as persistence and is key to successfully processing today's Web-enabled applications.

Security

Features that have made Web servers easy to deploy and ubiquitous in the growth of the Internet have also introduced security vulnerabilities. Public access to Web services has raised the specter of Denial of Service (DoS) attacks intended to overwhelm, compromise and paralyze computing infrastructure. Being able to stop and thwart DoS attacks is the difference between successful online implementations with high QoS and low customer satisfaction due to slow response times caused by timed out requests.

In addition, unauthorized users can hijack confidential data off of IP addresses made publicly available through broadcasts. Web site administrators need to be able to hide IP addresses to protect themselves from unauthorized access.

Differentiated Services

Not all customers are alike and therefore not all requests are alike. The ability to direct users to different servers based on access - such as members vs. non-members - and service requested, such as database queries or image requests, is an important feature of Web applications.

The ability to offer differentiated services comes from information that is embedded in HTTP requests made from the client. In addition to the URL, the HTTP header contains cookie information set by the server and returned by the user's client to differentiate it from other clients making requests. The ability to read and act upon this information allows Web farms to distribute workload across application-optimized machines and efficiently manage traffic.

Organizations are now faced with showing return on investment, as well as protecting their investment in these fault tolerant, redundant server farms they have created to service Web applications.

This is where server load balancing (SLB) can help. Serially extending installed servers adds more computing power, allowing more applications to be hosted but it does not address issues specific to Web traffic such as persistence and stateful fail-over. SLB works together with enterprise server farms to maximize and protect IT investments and deliver high availability for mission critical applications.

SLB in Action with Web-enabled Applications

Server farms today contain a variety of machines with different performance characteristics and applications. Maximizing utilization of each server requires information about both the server and the incoming client request for services. SLB is the general term assigned to techniques intended to provide:

- maximum server utilization,
- high overall availability to applications,
- transparent load distribution across networked servers to make them seem as one to user, and
- manageability for servers that need to be removed from service and returned to operations following maintenance.

Let's review some examples of SLBs in operation to further examine their purpose in today's enterprises.

SLB During Server Failure

Server farms distribute applications across multiple servers for greater performance, availability and ease of maintenance. Distributing traffic between them is the work of load balancers. In the event that one machine fails, 1/n, with n being the number of servers, processing capacity is lost. With software fail-over capability, backup servers can assume the transaction for the failed server but this creates an uneven load distribution. SLB functionality redirects traffic to all of the remaining servers avoiding the situation where some backup servers become over utilized.

Further, SLBs also serve a role in returning the response to the client by translating the responding servers' IP address, which the enterprise may not want to make public, to the publicly registered address. This is known as Network Address Translation or NAT. This capability allows administrators tremendous flexibility in installing and managing application-dedicated servers, and demarcating public and private networks for protection against unauthorized access.

SLB During Peak Traffic Load

Cyclical or event driven traffic bursts can play havoc on enterprise server farms. Front-end congestion can prevent users from accessing applications. SLBs using header information from HTTP client requests can direct users around congested servers and to their intended destination server, which may not be as much in demand. In addition, SLBs can prioritize packets using a combination of destination address, destination port number, and HTTP header information, ensuring that traffic destined to critical applications get priority over other traffic.

Conversely, SLB can ensure that processing-intensive traffic, such as requests that are SSL encrypted, are directed to high performance servers thereby not impacting response times for static page requests.

SLB Techniques

Server load balancers use predictor techniques and advanced configurable application groupings to achieve desired traffic directing outcomes.

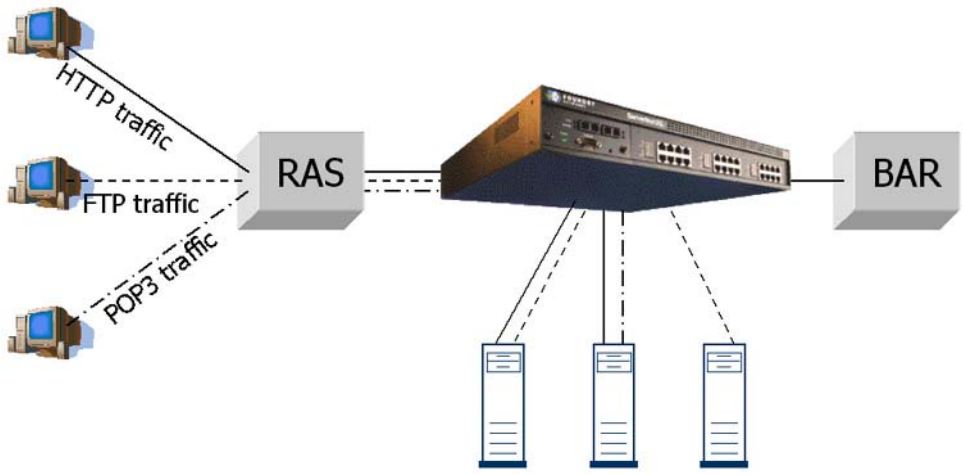
Predictor techniques are lower layer techniques that use statistical assignments to divide traffic flow to servers. Predictors are not content-aware and therefore have limitations when used by themselves for Web-enabled applications. Examples include: Round-robin where connections are assigned sequentially among servers in a logical community; Least Connections per server; Weighted Distribution, where user specified percentages are used for weight servers; fastest Response Time only; and a combination technique that uses least connections and response time.

As discussed earlier, Web applications often require clients to continue accessing the same real server for subsequent requests. A number of techniques are available to assist in load balancing stateful applications. Ports can be configured to return TCP/UDP requests to the same port related to the same real server. These are known as "sticky" ports.

Further enhancements can be made to augment SLB decision making about forwarding traffic based on higher layer HTTP protocol information. Known as Web switching or Layer 4-7 switching, today's server load balancers can receive requests from remote access servers (RAS) make decisions using URLs, cookies, header hashing, and SSL Session IDs.

- URL switching directs HTTP requests to a server group using information in the text of a URL string;
- Cookie switching directs HTTP requests to a server group based on information embedded in a cookie in the HTTP header;
- HTTP header hashing utilizes mathematical values to compare and map HTTP header information to a server and to direct all requests to that server; and
- SSL Session ID switching which connects a client to the same server to which it had previously established an SSL or Secure Sockets Layers connection.

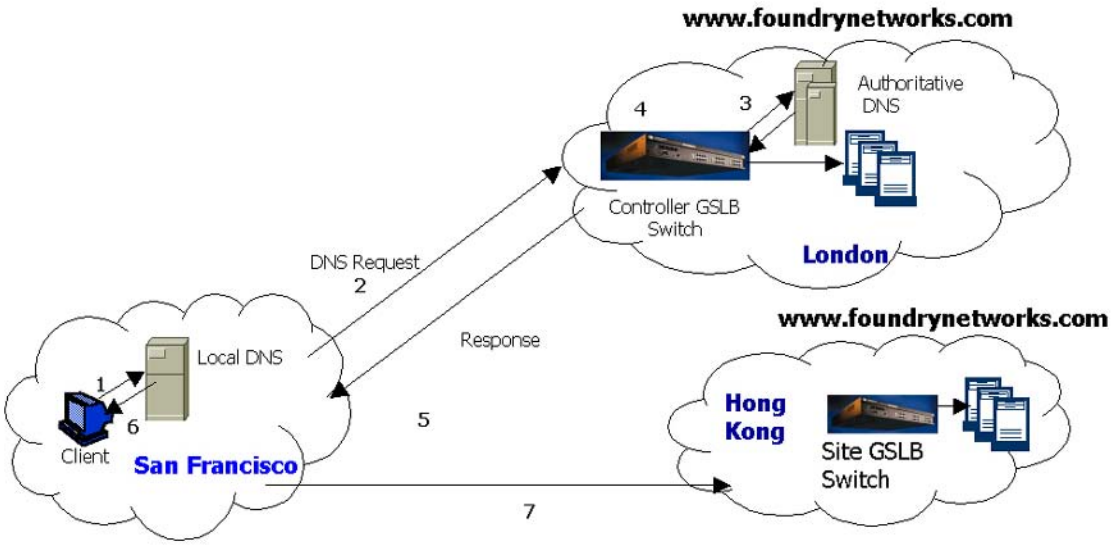
Clients send requests to the Web switch that distributes requests to real servers bases on load and requests



Global Server Load Balancing (GSLB)

Global Server Load Balancing or GSLB is a more powerful implementation of SLB. Whereas SLB works within data centers GSLB works on a global basis. With GSLB, Web response time can be reduced and server failures made invisible to customers around the world.

The basic premises of GSLB is to improve the process used in the Internet to match client requests with appropriate servers. This happens through a process known as DNS lookup. DNS stands for Domain Name Server, and is a repository of host names such as www.foundrynetworks.com and IP addresses maintained by individual companies and organizations. The DNS lookup and resolution process is shown in the figure below.



DNS Lookup Process

1. A client in San Francisco makes a request for www.foundrynetworks.com
2. The information is not found on the local DNS, so the client DNS makes the request to a higher authority or the Authoritative DNS in London.
3. The request from the Local DNS in San Francisco arrives and is intercepted by the Controller GSLB switch (CGS). The CGS sends this packet to the Authoritative DNS server. The DNS server processes this request and sends the IP address information for www.foundrynetworks.com back. (Note: in non-GSLB implementations, the request would go directly to the DNS server.)
4. The CGS intercepts this packet and based on a set of criteria (detailed later), selects the best IP address to send to the client in San Francisco. (Note: in non-GSLB implementations, this reordering does not happen. Addresses are simply returned without any qualification.)
5. The CGS then sends this response back to the client DNS.
6. The Local DNS sends the information back to the client. The client will select the first address.
7. A connection attempt is made to the www.foundrynet.com site in Hong Kong for which the Authoritative DNS provided the IP Address.

GSLB improves the DNS lookup process by rank ordering IP addresses stored in a DNS. The ordering is done on the basis of several metrics that measure the health of a particular site that corresponds to an IP address. In the figure above, this ordering happens in step 4. What is known as the Controller GSLB switch (CGS) receives responses from the Authoritative DNS (the master DNS) with IP addresses for a particular domain such as www.foundrynetworks.com. The CGS selects the best IP address based on several GSLB Metrics™:

- Health conditions of each site, server, and application that has an IP address residing on the Authoritative DNS server.
- The capacity and availability of each Site GSLB switch (SGS).
- The round-trip time between the SGS and client networks.
- The geographic locations of the server and client networks.
- The FlashBack™ speed measuring how quickly the SGS responds to health checks.
- The Least Response metric measuring the site address that has been selected the least.
- The administrative priority that can be configured on the SGS.

The addition of these metrics provides a powerful solution to improving the DNS system. It has the added benefit of not replacing existing DNS implementations that continue to function normally. DNS is at the heart of the Internet and safeguards must be taken by every company to ensure that their DNS has fault tolerance and backup functionality built in. With GSLB, companies can rest assured that their DNS systems are not tampered with but rather enhanced to provide better response times.

SLB Products

Now that we've seen what SLBs do for today's Web applications, we examine different SLB solutions and the issues with how they work.

SLB products have been implemented in software and hardware and can be categorized as: 1) switch-based, 2) PC-based, and 3) server-side software/OS level load-balancing agents.

PC-based product offerings are generally Intel-based devices with software specialized to load balance and a general-purpose operating system, such as Unix. PC-based products are intended for smaller, less computing intensive implementations and can become performance bottlenecks in concurrent connection capacity and throughput – the reason organizations deploy multiple servers for an application.

Software based solutions provided by OS vendors run as network drivers on each server beneath higher-level application protocols. Unaware about HTTP and FTP traffic, these solutions cannot intelligently route based on content. Further, these solutions generally cannot be used to directly load-balance client requests across stateful servers. Additional proprietary logic is needed and even then can interoperate only with applications supporting it. Software load balancers also do not directly monitor server applications, such as Web servers, for continuous operation. Instead, they provide the mechanisms needed by application monitors to control operations of clustered or servers that have been configured to share resources to manage server as a single system, detect downtime and provide recovery. Monitoring services are widely available for most client/server applications but add additional complexity to load balancing.

Switch-based load balancers are purpose built and are offered in two varieties. Multi-port, stackable switches come with a specialized operating system, and specialized chips for Layer 2 wire speed switching. These devices include ASICs for fast packet forwarding and provide

increased speed and resiliency. Load balancing switches offer a low-cost approach to managing server farms, and operate on industry standards without any special software installed on the servers.

Chassis-based web switches provide Layer 2 and 4-7 switching and Layer 3 routing capabilities though specialized chips put on every port for packet forwarding and session processing. These are multifunction devices that go beyond load balancing only and offer other traffic management capabilities such as bandwidth management, and routing.

Benefits of Web Switches

Web switches give IT managers the ability to choose the load balancing approach of their preference, thereby shaping network traffic without surpassing the capacity of any server within the enterprise server architecture.

Switch-based server load balancing implementations deliver high levels of performance both in terms of processing speed, and granularity of control in managing traffic. SLB switches are built with application specific computer processors (ASICs) in what is known as a distributed architecture fashion that allows very high-speed intake, processing, and forwarding.

In addition to processing and traffic forwarding speed, processors are also used for Layer 4-7 load balancing. A significant step up from server capacity-only load balancing, Layer 4-7 capabilities are integral to today's enterprises deploying Web, HTTP-based applications. Using HTTP header and SSL Session ID information, Web switches optimize server farms by directing traffic to application-specific machines, designed to handle specific requests.

A further benefit of Layer 4-7 intelligence is the ability to perform stateful switching and failover that ensures transactions requiring multiple connections with the same server are completed. Without this capability, clients would have to continually reestablish themselves and reenter the process or reenter information from the beginning.

Key Benefits of Web Switches

- Distributed processing architecture for optimum processing and forwarding performance
- Granular settings for directing HTTP traffic
- Stateful failover to ensure integrity of session information between client and server
- Cost effective scalability for growing computing architecture
- Familiar CLI/GUI based setup and SNMP management for compatibility with network management tools
- Transparent server maintenance
- No additional software to be loaded onto servers

Web switches providing high throughput capacity can front-end server farms. Server health monitoring allows statistics to be gathered, compared to set values for configured addresses, and prescribed action taken to avert forwarding traffic to overloaded or failed servers. By the same process, servers can be shutdown gracefully and brought online transparently to users. In all, Web switches deliver advanced server load balancing and intelligent traffic management in a stackable or chassis solution using familiar CLI syntax or GUI, speeding implementation and facilitating maintenance.

Foundry Networks Solutions

Foundry Networks award-winning ServerIron family of Web switches provide high performance, Layer 2 through 7 switching, enabling IT managers to control and manage today's exploding Web transaction, Web applications, and e-commerce traffic flows.

SERVER LOAD BALANCING IN TODAY'S WEB-ENABLED ENTERPRISES

Internet IronWare, Foundry's software suite of Internet traffic management capabilities powers ServerIron Web switches to direct requests to the right server and application based on the information that resides beyond the traditional Layer 2 and 3 packer headers. IronWare provides full support for TCP-based protocols including HTTP, Proxy, FTP, POP3, SSL, and SMTP.
ServerIron eases escalating Internet traffic overload, dramatically increases service availability, reduces the burden of server farm management, and allows the entire Web facility to scale to its full potential.

Foundry Networks' Web Switches				
	Servers Supported	**Connections per Second**	**10/100 Ports | Gigabit Ports**	
ServerIron XL	1024	80,000	24	8
ServerIron 400	2048	600,000	72	24
ServerIron 800	2048	600,000	168	56

High Availability for Server and Application

ServerIron switches ensure service availability by offering switch, server, link, and session level redundancy. In the event of a server or application outage, ServerIron switches provide detection and sub-second fail-over to the next server that supports a like service. ServerIron switches provide active-standby or active-active redundancy capability that allows administrators to establish primary and secondary load balancing switches to support identical configurations parameters and provide 100 percent availability.

Maximum Scalability

TrafficWorks™ IronWare™ running on ServerIron simplifies network design by enabling IT managers to create a server farm, represented by a single IP address known as a virtual IP address (VIP). ServerIron acts as a single server for incoming Web traffic, controlling, monitoring and directing client requests to the most appropriate real server in a server farm.

ServerIron's firewall load balancing (FWLB) eliminates firewall bottlenecks by distributing load across multiple firewalls. FWLB is Check Point certified, and can load balance up to 32 firewalls for a scalable and highly available Web-site deployment.

Ease of Management and Powerful Reporting

Foundry Network Management Solutions (NMS) provide a full suite of configuration, management, and monitoring applications. Using NMS – Monitoring, extensive accounting and statistics allows network managers to easily collect and display detailed information about network traffic destined to server farms. In addition, ServerIron supports Remote Monitoring (RMON) and SNMP device management and integrates into HP OpenView, providing capabilities such as topology discovery, network capacity planning, and network alarms.

Additional Information

More information is available about the features and services offered by the Foundry Networks family of ServerIron web switches. In addition to industry-leading performance for server load balancing and traffic management, the ServerIron family offers Global Server Load Balancing (GSLB), Firewall Load Balancing (FWLB), and Transparent Cache Switching (TCS). Please visit Foundry Network's ServerIron Web site for more information:

http://www.foundrynet.com/products/webswitches/serveriron/index.html

Foundry Networks, Inc.
Headquarters
2100 Gold Street
P.O. Box 649100
San Jose, CA 95164-9100

U.S. and Canada Toll-free: (888) TURBOLAN
Direct telephone: +1 408.586.1700
Fax: 1-408-586-1900
Email: info@foundrynet.com
Web: http://www.foundrynet.com