HTRC

GROUP, LLC

Global Load Balancing Solutions

The HTRC Group
P.O. Box 2087
San Andreas, CA 95249
www.htrcgroup.com

About The HTRC Group, LLC

The High-tech Resource Consulting Group, LLC focuses on service provider networking, providing consulting, custom market research, and market research studies to service providers and product manufacturers.

Summary

The Internet is growing at an incredible speed, facilitating rapid innovation and the development of new revenue models. New technologies such as cable modems and Digital Subscriber Line (DSL) are continually increasing access speeds to accommodate the growing demand for more sophisticated, bandwidth-intensive, interactive content. As first-time users flock to get connected, the growing numbers can place considerable demand on a company's Web servers. At the same time, Web site performance and reliability have become two of the most important determinants of online user satisfaction and customer retention, making an increasingly significant impact on the bottom line.

Today, roughly half of the large companies in the U.S. have more than one data center housing a complicated array of redundant Web servers and network performance enhancement devices. A key component of this ever-growing Web site network architecture is global load balancing technology, which directs users of a specific Web site to the optimal global data center and to the optimal server within that center.

This paper explores the requirements and benefits of global load balancing solutions.

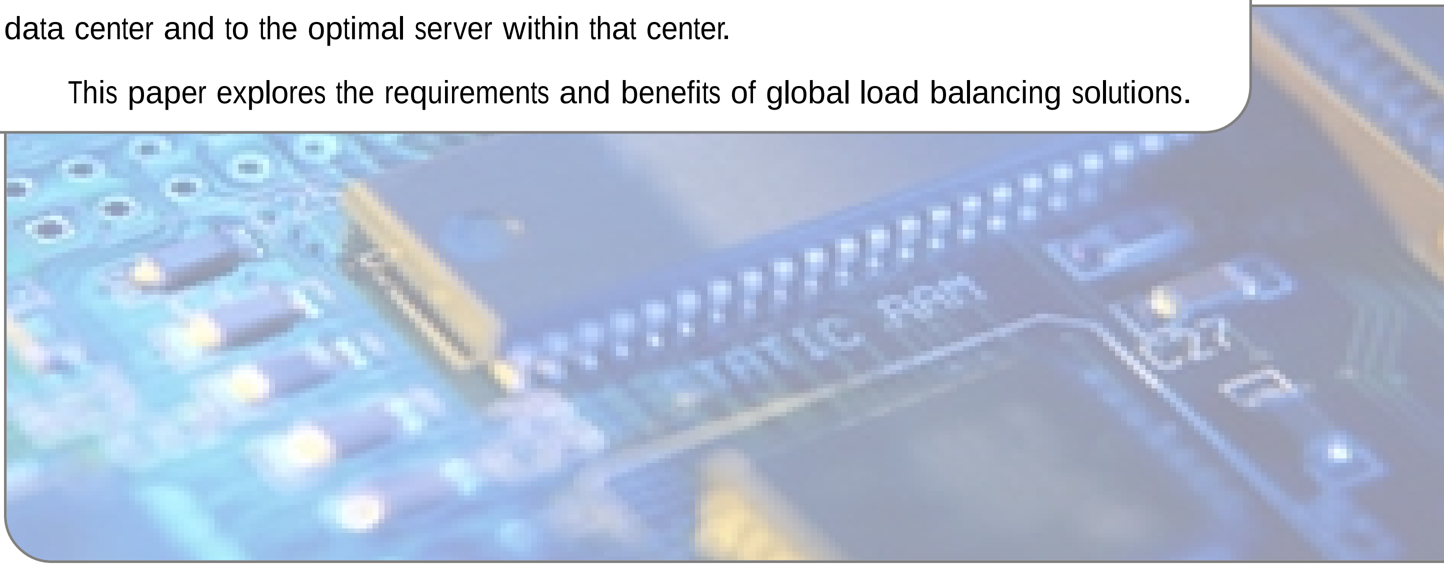


Table of Contents

The Demand for Web Site Performance and Reliability

Each day, new Internet users are appearing in droves, accessing sites from all over the world, across all time zones. There are also daily deployments of new broadband services to business and residential online customers. All of them demand that content and services be delivered immediately, around the clock. Companies that can't satisfy requests for information and complete transactions promptly and consistently will quickly lose a significant competitive advantage. Even as the demand for Web site performance and reliability increases, so do the sources of degradation and outages. Our interviews with large companies (those with more than 500 employees), published in the March issue of The HTRC Group's Rapid Business Intelligence (RBI) Research Service, revealed many serious ongoing causes:

- The chief cause, named by 63% of the respondents, was problems with basic connectivity to the service provider.
- Hardware failure, primarily servers and some network equipment, was cited by 54% of the respondents.
- All devices connected to the network need power, and 22% of the respondents identified power outages as a problem.
- Bandwidth capacity restrictions was named by 27%. Sites that do not outsource Web site hosting or collocate Web servers with a service provider are dependent on the local exchange carrier and ISP in order to maintain their own Web site Internet connection. According to The 1999 Content Delivery Service Study, Web site bandwidth increases an average of 8.4% per month. Web site owners provisioning additional capacity through data connections, rather than collocation facilities, are at the mercy of local exchange carriers (LECs) and Internet Service Providers (ISPs) for increased site bandwidth.

These are the most prevalent causes of Web site outages and service degradation. Many more exist, including software failure, hacker attacks, and bad firewall and security configurations.

The Costs of Downtime

Web users are very intolerant of slow-loading content and are less likely to purchase while browsing a lower performing e-commerce site. Web site disruptions or slowdowns can therefore have a direct, and often dramatic, effect on the bottom line. In fact, according to The 1999 Content Delivery Service Study, e-commerce Web sites lose an average of $16,201 in sales for every hour the site is down. In addition, respondents in The March RBI Research Service were asked to indicate on a scale of 1 to 7 how concerned their company was regarding lawsuits when Web site outages occur, where 1 was "not concerned" and 7 was "extremely concerned." Thirty-one percent gave a rating of 5, 6, or 7, indicating that there is significant liability concern regarding Web site performance.

New Technology Solutions

Faced with loss of revenue and liability concerns, content site owners can easily justify the cost for monitoring site performance 24x7, striving for 100% availability every day of the year. According to the March issue of The HTRC Group's RBI Research Service, 51% of large companies have more than one data center and use global load balancing products to increase the resiliency of their Web site environment. Global load balancing involves using network metrics—including latency, persistence, static mapping, controlled failover, and origin server load balancing—to guarantee Web site performance and serve content closer to users, thus enhancing the overall Web site experience.

e-commerce Web sites lose an average of $16,200 in sales for every hour the site is down

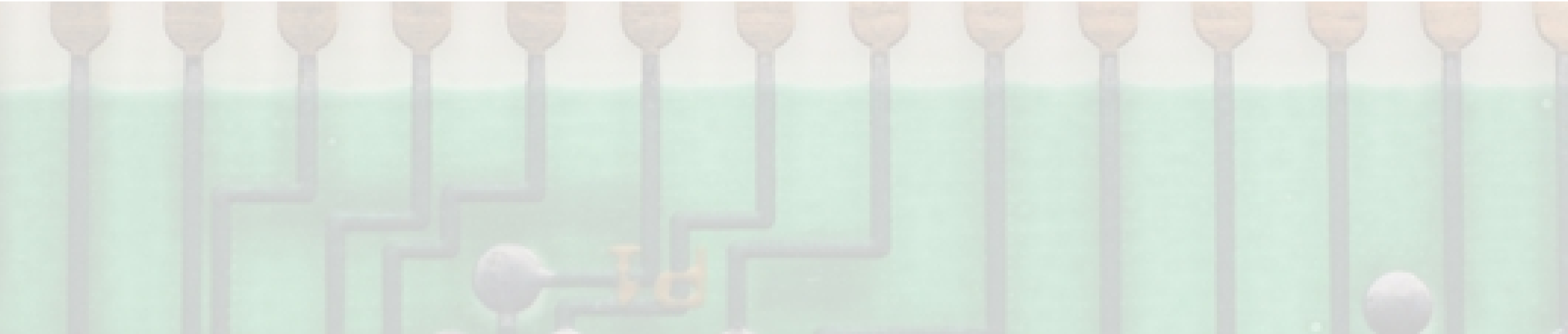
In addition, 95% of large companies responding to the survey said they plan to invest in content delivery technology products or services, which are complementary to global load balancing. They provide increased performance by intelligently placing cacheable content closer to users in the Internet.

Global Load Balancing Solutions

When users type in a Web site address or Universal Resource Locator (URL), they rely on Domain Name System (DNS) servers to direct them through the Internet's haphazard maze of interconnected networks and connect them to a Web server. (DNS is the service that maps IP addresses to their host names distributed throughout the Internet.) Global load balancing solutions optimize the navigation process by using specific network and server metrics to direct users to the best performing data center and Web server for a particular Web site.

Users can be directed to servers in their own geographic region or away from a congested local network, effectively reducing response times and increasing the quality of the end user's experience. In addition, selected global load balancing solutions check the various elements that create dynamic content to ensure the total data center site is operational before Web users are directed to any site.

Content site owners have many global load balancing solutions to choose from, each having distinct strengths and weaknesses. The two primary categories are product-based solutions—which are based on software applications, appliances, or switches—and service-based solutions, which are outsourced to a provider such as Speedera Networks.



Product-Based Solutions

Product-based global load balancing solutions are owned by a company and operated by its Web site professionals. There are three major categories:

- Software solutions are global load balancing applications designed to run on general purpose operating systems (OS) and include software from product manufacturers such as Resonate.
- Appliances are network devices built specifically to perform a single function, global load balancing, and include products such as 3DNS from f5 Networks.
- Switches, which are network elements that "switch" traffic to destinations based on layer 4 (Transport layer) information, are continually evolving to encompass the scale of network layers and provide more intelligent switching decisions. Switch-based solutions include products such as WebOS that reside on the Alteon 180, and AceDirector switches from Alteon WebSystems.

Figure 1 shows the network topology for a dual data center redundant global load balancing solution. Web users connect to the Internet through their ISP, and first connect to the local DNS to determine the location of the Web site that resides in the data centers. The Internet is made up of many interconnected service provider networks, and presents numerous routes to connect to the data center Web site. In this example, the product-based global load balancing solution can send the Web user to the San Jose or Boston data centers based on the network information available.

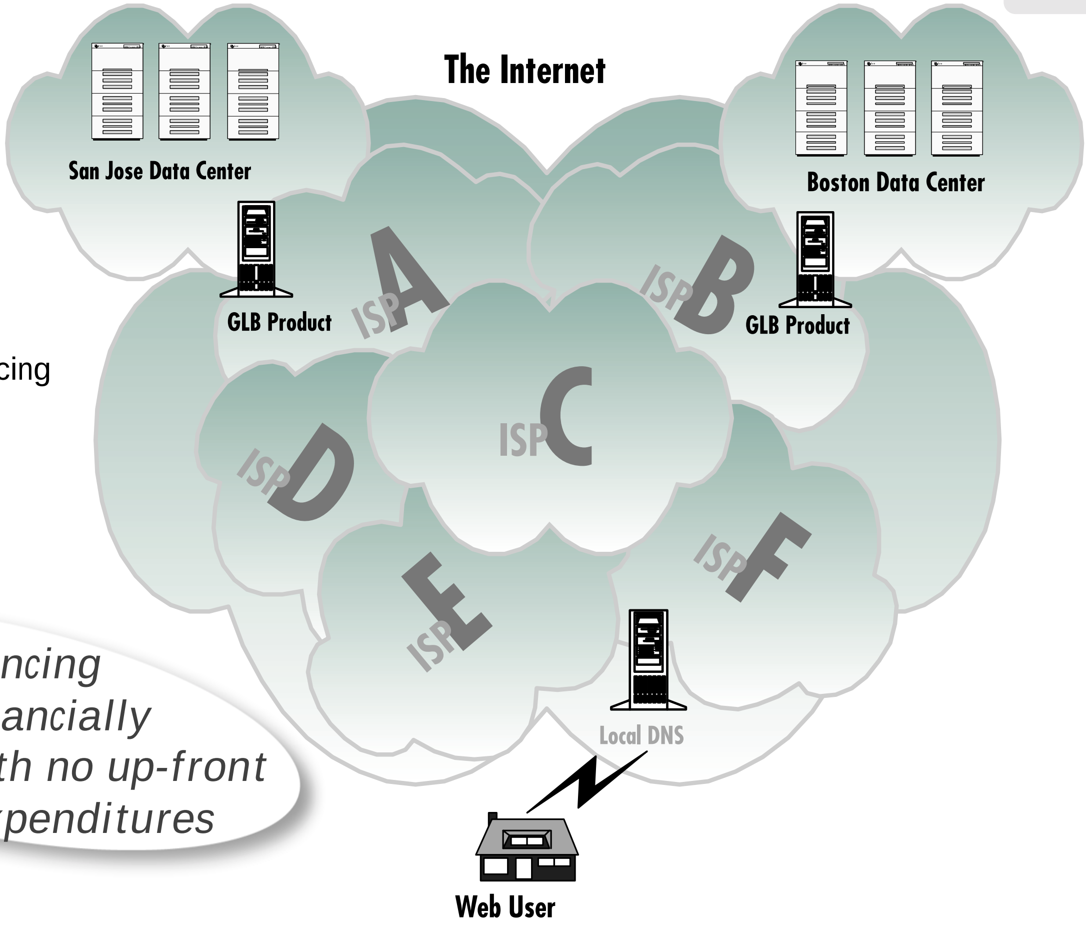


Figure 1: Product based implementations typically have two global load balancing products for redundancy.

Global load balancing services are financially attractive, with no up-front capital expenditures

Service-Based Solutions

Global load balancing services outsource all of the functions of a product-based solution, thus avoiding the heavy capital investment costs. These services include Speedera Network's Universal Delivery Network, which includes Global Traffic Management Service. As with other outsourced services, Web site owners receive a number of benefits:

- Fewer high-level experts are required to maintain the solution, which is an advantage as expertise remains scarce due to the Internet's rapid growth.
- Content professionals can focus on core expertise rather than the continual monitoring and management of global load balancing products.
- Services can be quickly provisioned and easily managed. Providers such as Speedera Networks offer "expertise on tap" so that any problems can be solved quickly without the need to retain many high-level IT professionals to manage the content site.
- The services are financially attractive, with no up-front capital expenditures and with reasonable recurring costs.
- Service level agreements can be negotiated to include compensation for violated agreements.

Subscribers, however, are fully dependent on outsourced global load balancing services, and must rely on the provider to develop and implement new service features.

Figure 2 below shows a global load balancing service solution with dual data centers and many redundant redirection elements. Web users connect to the Internet through their ISP, and first connect to the local DNS to determine the location of the Web site that resides in the data centers. As in the previous example, the Internet is made up of many interconnected service provider networks, and presents numerous routes to connect to the data center Web site. The service-based global load balancing solution can send the Web user to the San Jose or Boston data centers based on network information from DNS redirection elements strategically placed in many networks throughout the Internet.

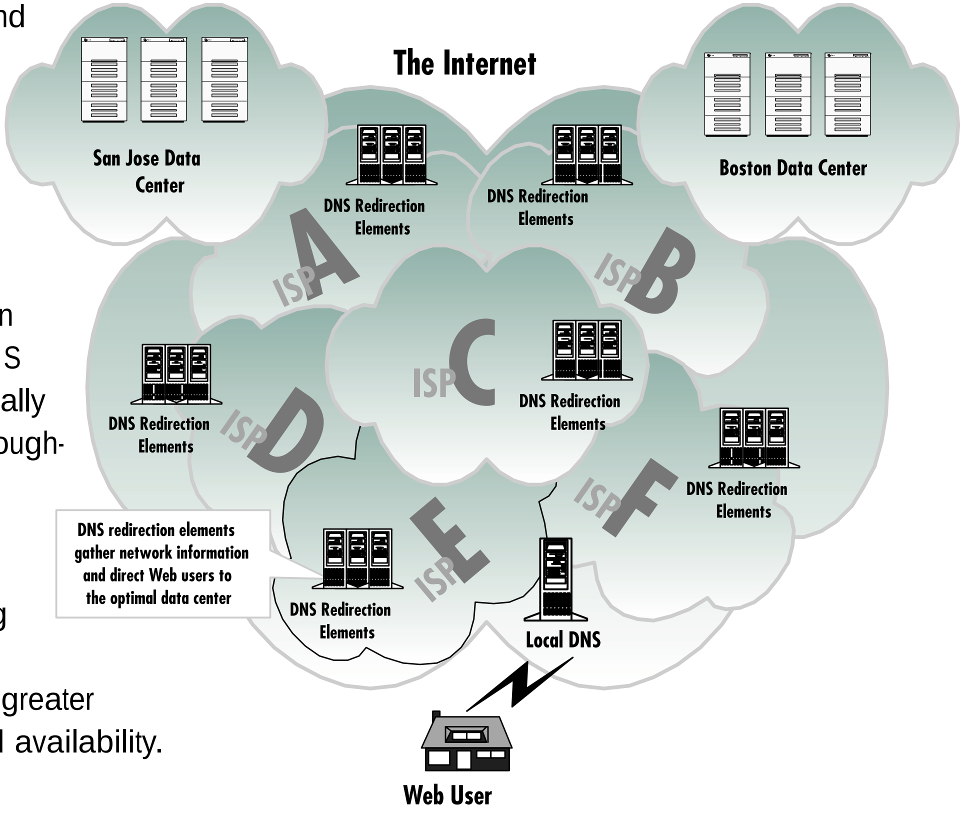


Figure 2: A service-based global load balancing solution has many DNS redirection elements offering greater scalability, performance, and availability.

Evaluating Global Load Balancing Requirements

Both product and service based categories offer advantages and disadvantages, depending on the prospective buyer. With so much at stake, owners must carefully evaluate solutions in the following key areas: manageability, support, performance, network latency metrics, server metrics, advanced routing, extensibility, network operations center (NOC), security, and availability.

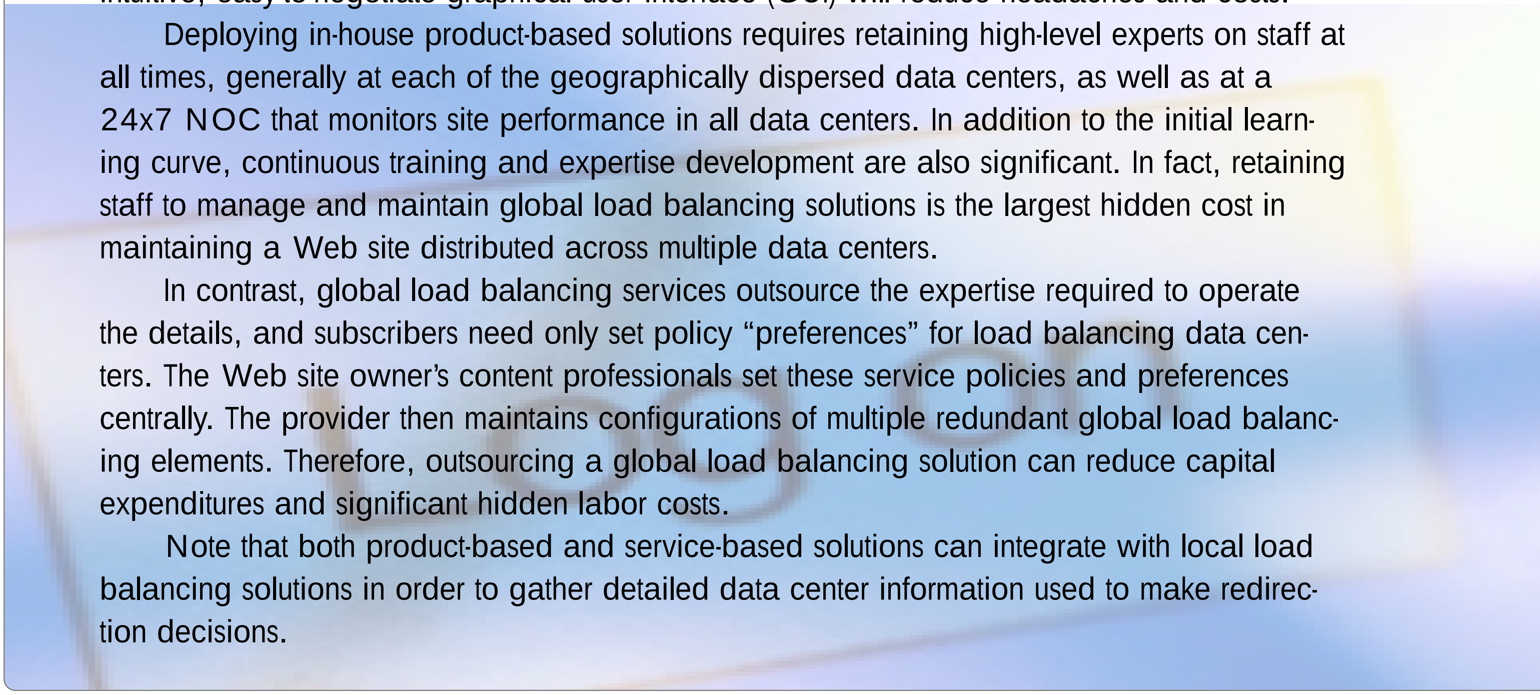
Manageability

Solutions must have the capacity to monitor the health of the network between data centers

In a product-based global load balancing solution, Web site owners are responsible for day-to-day operations, maintenance, and monitoring, as well as for all configurations of DNS, load balancing metrics, routing algorithm, hot fail-over, and weighting based on network performance. In addition, many product-based solutions are individually managed, with one box configured at a time.

In order to maintain global site performance, management solutions must have the capacity to monitor the health of the network between data centers, as well as the servers in the data centers. Most organizations today employ existing management platforms in addition to off-the-shelf and custom applications. SNMP is widely used to send information from load balancing solutions to management platforms and applications. Solutions should support SNMP, enabling communications with existing management platforms and applications.

It should be noted that global load balancing solutions involve configuring many settings in a complicated environment. Managing these configurations can be a challenge, so an intuitive, easy-to-negotiate graphical user interface (GUI) will reduce headaches and costs.



Deploying in-house product-based solutions requires retaining high-level experts on staff at all times, generally at each of the geographically dispersed data centers, as well as at a 24x7 NOC that monitors site performance in all data centers. In addition to the initial learning curve, continuous training and expertise development are also significant. In fact, retaining staff to manage and maintain global load balancing solutions is the largest hidden cost in maintaining a Web site distributed across multiple data centers.

In contrast, global load balancing services outsource the expertise required to operate the details, and subscribers need only set policy “preferences” for load balancing data centers. The Web site owner’s content professionals set these service policies and preferences centrally. The provider then maintains configurations of multiple redundant global load balancing elements. Therefore, outsourcing a global load balancing solution can reduce capital expenditures and significant hidden labor costs.

Note that both product-based and service-based solutions can integrate with local load balancing solutions in order to gather detailed data center information used to make redirection decisions.

Support

Redirecting users to the optimal site location is a complicated process involving global load balancing products, the hosted network, and the network by which the end user connects to the Internet. Product manufacturers offer a range of support options, from free, 90-day call-in support to costly incident-based pricing models. Eighty-five percent of the Webmasters and content managers interviewed for the 1999 Content Delivery Service Study rated service and support a critical feature when choosing a provider. Without the product manufacturer's intervention, problems may not be resolved quickly, if at all. Problems that occur in loosely-defined areas of responsibility are difficult to resolve; finger-pointing about support responsibility can delay the solution for days.

Packet loss is as important as latency when determining the optimal performing data center

With service-based solutions, the responsibility of support falls entirely on the provider.

Performance

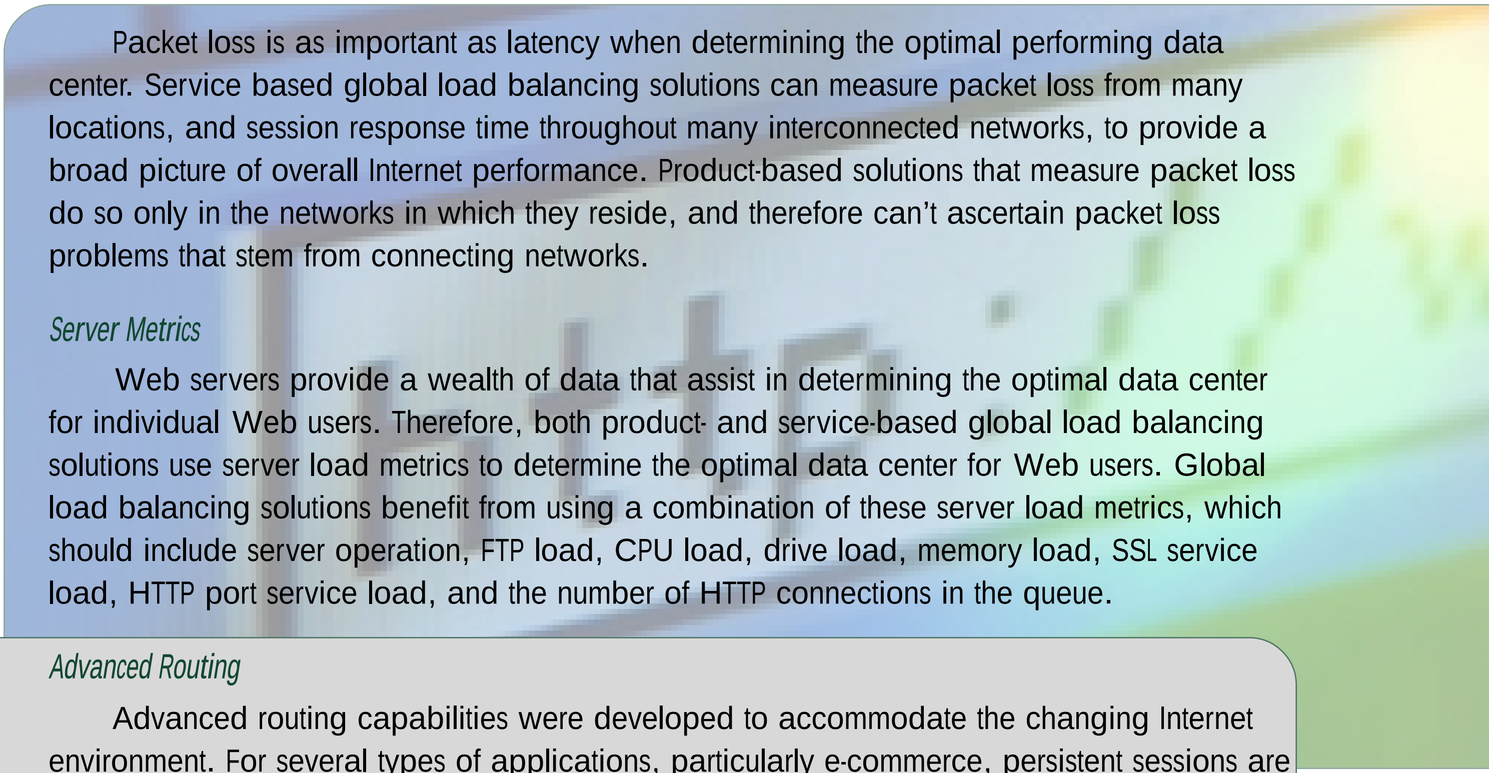
Web traffic demands are difficult to predict and flash crowds can dramatically affect the health of a Web site. Many Web sites implement redundant redirection products in each data center to provide critical redundancy to address any problems that these sudden onslaughts may cause. Virtual IP (VIP) addressing enables multiple layers of VIP addressed network elements. Assigning VIPs enables additional manageability features, multi-homed sites, and increased redundancy.

In addition, as technology innovation continues, more high-performance applications, such as e-commerce, and richer multi-media content will be integrated into Web sites. Both content site performance and application performance will increasingly weigh heavily in determining the best global Web site data center that users are directed to. Some applications send additional connection instructions using TCP in order to create additional connections to interact with the application. Solutions that understand application port load information can assist in determining the best data center for both application and content usage.

To monitor performance, service providers test user sessions at a configurable interval from their own network of DNS redirection servers. For example, the solution from Speedera Networks test HTTP requests and performs an actual HTTP response to verify content and response time for the overall request. For an e-commerce site, Speedera can also check a transaction and examines a response in order to detect backend failures and route accordingly.

Latency and Packet Loss

Network latency, one of the most commonly used metrics for evaluating Web sites, can be determined by calculating round trip time or the number of BGP hops. Both product- and service-based global load balancing solutions use one or more combinations of latency measurements to assist in ascertaining the optimally performing data center for individual Web users.



Packet loss is as important as latency when determining the optimal performing data center. Service based global load balancing solutions can measure packet loss from many locations, and session response time throughout many interconnected networks, to provide a broad picture of overall Internet performance. Product-based solutions that measure packet loss do so only in the networks in which they reside, and therefore can't ascertain packet loss problems that stem from connecting networks.

Server Metrics

Web servers provide a wealth of data that assist in determining the optimal data center for individual Web users. Therefore, both product- and service-based global load balancing solutions use server load metrics to determine the optimal data center for Web users. Global load balancing solutions benefit from using a combination of these server load metrics, which should include server operation, FTP load, CPU load, drive load, memory load, SSL service load, HTTP port service load, and the number of HTTP connections in the queue.

Advanced Routing

Advanced routing capabilities were developed to accommodate the changing Internet environment. For several types of applications, particularly e-commerce, persistent sessions are critical to providing Web users with a consistent experience. Persistent sessions are HTTP sessions that require continual Web user connectivity with the same server. Without this capability, user information such as shopping cart data may be lost because it is not immediately updated on all servers. In a multi-server e-commerce site, the local load balancing solution's traffic cop sends users to the same server by tagging the users or utilizing cookie technology.

Global load balancing solutions must also support persistence in order to direct a Web user to the correct data center and the server on which that user's information resides. As a service from Speedera Networks, Web site owners can turn on persistence for a given domain. Therefore, a client DNS for a given hostname for that domain will always return the same IP address, unless the service of that IP address fails, in which case Speedera will send the user to the next best-performing server.

Proximity to the data center can also be a factor in determining the nearest, best performing location for a Web user. Static mapping of the Web user's geographic location is also used to direct that user to the best-performing network in which the data center resides.

Web site owners should require the ability to weight static routing based on network cost or performance. For example, users can be routed to a data center with the lowest bandwidth cost. Web site owners can set performance threshold policies for least-cost bandwidth networks and direct Web users to higher-cost networks when congestion occurs. In addition, unattended disaster recovery policies can be set, whereby data centers are monitored for service degradations and outages and users are directed to a better performing data center.

Extensibility

The effective global load balancing technology architecture must include an extensible framework that supports the integration of current and future load balancing features, products, and services. The development of standardized “interface-compatible” components enables quick integration of future features, as well as product and service compatibility.

Network Operations Center

For most businesses, keeping the Web site up all the time requires continual monitoring. Most large businesses today have a NOC for enterprise network operations and support and many now staff their NOC 24x7 with content professionals that have the expertise to remedy major site problems. Companies that don’t maintain a 24x7 NOC can leverage the site monitoring capabilities of a service based global load balancing solution.

Security

Security has become a major issue, especially with on-going, high profile denial of service (DOS) attacks against major Web sites. Global load balancing solutions must include mechanisms to defend against DOS attacks by methods such as monitoring idle connections and monitoring suspicious activity (e.g., unauthorized access attempts). Other security mechanisms include source route tracking to reduce IP spoofing and packet filtering to allow authorized access based on source and destination of the packet.

Managing global load balancing solutions conveys sensitive information. Global load balancing solutions require a secure managed interface; a handful of skilled hackers can feed droves of “script kitties,” proving that lax security practices can result in a crippled Web site.

Availability

Data centers will vary in performance and, under similar loads, do not offer the same performance experience to the end user. To provide network redundancy and thus maximize availability, data centers are generally connected to several service provider networks. In the event one network experiences service degradations or connectivity problems, there is always back-up network connectivity. True redirection redundancy should include at least one redundant redirection device for each data center and reside in multiple networks.

The Advantages of Global Load Balancing Services

Global load balancing services can improve performance for Web site visitors while offering unique advantages. Service based solutions leverage an existing network of strategically placed global load balancing elements and content delivery servers to generate a near real time performance map of the Internet. The use of this map to direct Web site visitors is more comprehensive than single products deployed at each data center, and provides better data center location decisions. Customers can access reports detailing Web site performance information to gain a better understanding site usage. Reporting enables a Web site to analyze the content usage of the entire Web site to assist with future Web site content and performance decisions.

Maintaining qualified network engineers can prove difficult, and at times, near impossible. Another unique advantage service based solutions have is “expertise on tap.” As bandwidth and Web site demands increase over time, providers of global load balancing services maintain the expertise required to scale with the largest sites in the Internet. Aside from the service, a global load balancing service provider’s core competency is building the engineering organization that maintains an entire network of global load balancing elements. Web site professionals can focus on the content, rather than managing global load balancing products.

Operating at the speed of the Internet includes the risk of falling victim to success. Time to market factors may pressure hasty data center deployments in response to competitive threats and differentiation on the Internet. Global load balancing services can quickly be provisioned and easily managed to meet rapid data center expansion plans.

SLAs for global load balancing services offer a unique advantage over product-based solutions: compensation. When global load balancing performance does not meet expectations, subscribers are compensated. SLAs to customers may include performance measured and validated by third parties. Web sites have varied performance requirements, therefore, SLAs are expected to be negotiable. Fundamentally, SLAs should guarantee faster, more reliable services.

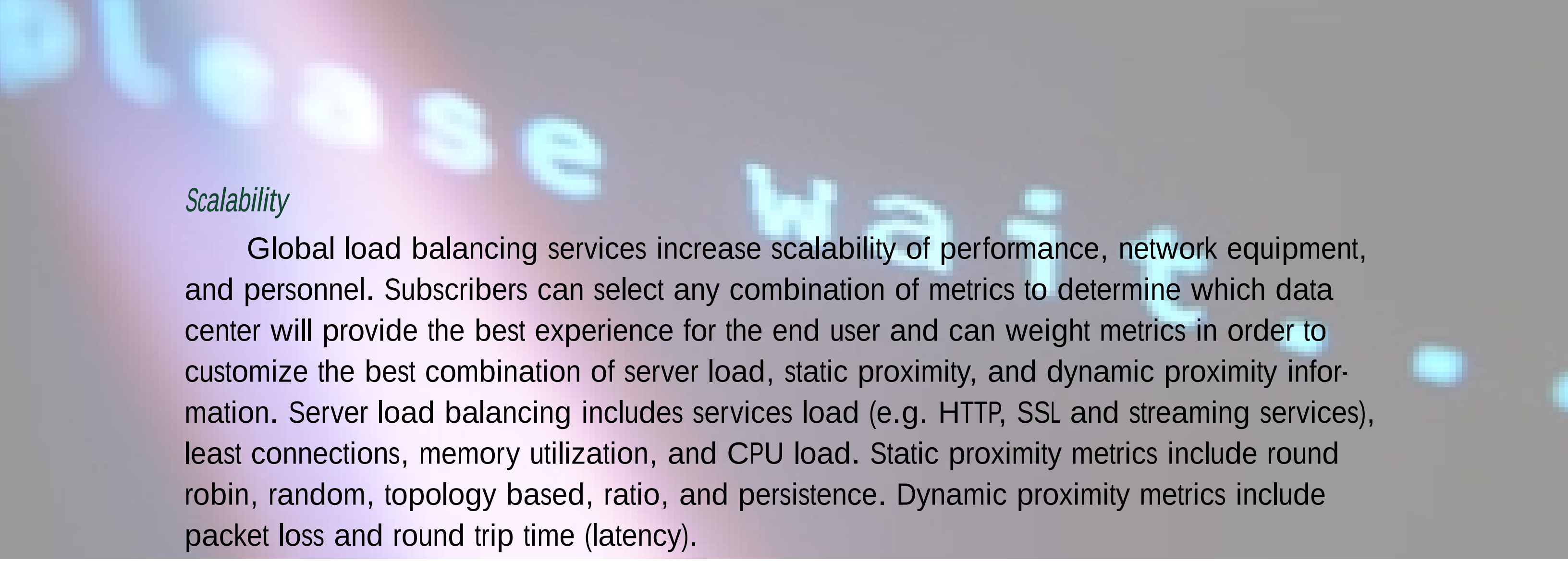
Global load balancing services can improve the experience of visitors to a content site by assisting in three basic areas: speed, reliability, and scalability. In addition, these services require few resources from the content site. By eliminating the burden of administration and maintenance of the global load balancing solution, subscribing companies can focus on their core competencies, including content development.

Speed

Broadband technologies such as wireless, cable modems, and DSL are significant factors in the increase of bandwidth demand from both consumers and businesses. The growing number of broadband service subscribers directly affects the load placed on a Web site data center – when end users with faster connections browse a content site, they consume more of that content site’s bandwidth and server resources. Using global load balancing services, content sites can scale to handle enormous traffic demands by directing users to the best-performing data center, reducing response times and increasing overall performance.

Reliability

Increasing the number of geographically-dispersed data centers increases the overall reliability and consistency to the Web site. Global load balancing providers offer a fault-tolerant, resilient network architecture that is designed to scale to accommodate the kinds of flash crowds experienced by the most popular Internet Web sites. Providers of global load balancing services have deployed networks of hundreds of servers that continuously monitor network congestion and server load and availability.



Scalability

Global load balancing services increase scalability of performance, network equipment, and personnel. Subscribers can select any combination of metrics to determine which data center will provide the best experience for the end user and can weight metrics in order to customize the best combination of server load, static proximity, and dynamic proximity information. Server load balancing includes services load (e.g. HTTP, SSL and streaming services), least connections, memory utilization, and CPU load. Static proximity metrics include round robin, random, topology based, ratio, and persistence. Dynamic proximity metrics include packet loss and round trip time (latency).

End user ecommerce transactions equate to increasing bottom line revenue. End users may not always be sent to the same data center and persistence maintains end user transactions among dispersed data centers in conjunction with local load balancing products. The need is greater than ever for global load balancing solutions that increase Web site speed, reliability, and scalability.

Conclusion

In today's increasingly demanding and competitive Internet business environment, Web site performance and availability have a strong effect on the revenue and reputation of a content site provider. Internet users are likely to shun a slow-performing content site in favor of a faster competitor. Service degradation and outages—including the inability to meet demands of flash crowds—mean not only the loss of current and future revenues, but possible liability for the business losses of customers. Subscribing to global load balancing services holds significant advantages in maintaining performance and uptime for growing Web sites.

The providers of global load balancing services gain a real time map of Internet performance through the many network probes and global load balancing redirection servers deployed within their worldwide networks. Customers of the service solution utilize the performance metrics gathered throughout the global Internet rather than only those points where a global load balancing product resides. As a result, service providers are able to build a comprehensive network map that reflects the current overall condition of the Internet. In contrast, Web site owners that deploy a product-based solution cannot benefit from the detailed performance map.

Global load balancing services utilize the service provider's extensive network expertise, guaranteeing correct configurations of complex Web site environments. Operating at the speed of the Internet is synonymous with rapid change. Constantly changing Web site configurations are monitored around the clock for performance. Direct support is a phone call away, any time of day.