

# TechBrief

## Extreme Networks

# Server Load Balancing

## 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 web site. 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 solutions to support all applications at all times.

However, the availability of these applications is often threatened by network overloads as well as server and application failures. Resource utilization is often out of balance, resulting in the low-performance resources being overloaded with requests while the high-performance resources remain idle. Server load balancing is a widely adopted solution to performance and availability problems.

## Server Load Balancing and its Benefits

Server load balancing is the process of distributing service requests across a group of servers. The following diagram shows load balancing within a server farm.

Server load balancing addresses several requirements that are becoming increasingly important in networks:

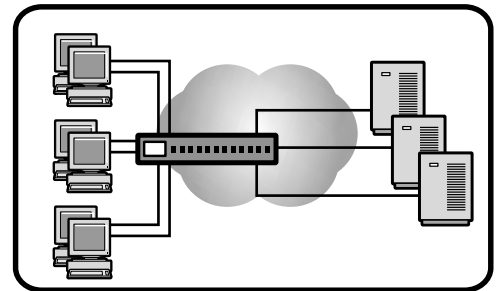
- Increased scalability
- High performance
- High availability and disaster recovery

Many content-intensive applications have scaled beyond the point where a single server can provide adequate processing power. Both enterprises and service providers need the flexibility to deploy additional servers quickly and transparently to end-users. 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 power 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.

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

Distributed server load-balancing products can also provide disaster recovery services by redirecting service requests to a backup location when a catastrophic failure disables the primary site.



*End-user requests are sent to a load-balancing device that determines which server is most capable of processing the request. It then forwards the request to that server. Server load balancing can also distribute workloads to firewalls and redirect requests to proxy servers and caching servers.*

# 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 cost of network ownership.

Layered as a service on top of these powerful functions, Extreme Networks has integrated the F5 Networks industry-leading server load balancing source code into its own award-winning wire-speed switching solutions for Internet service provider, web content provider and enterprise networks. To help companies migrate existing networks to meet today's requirements, Extreme Networks offers its family of Summit™ stackable switches, the BlackDiamond® chassis switch and the ExtremeWare™ software suite to scale speed, bandwidth, network size and Policy-Based Quality of Service.

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

- Hardware integration for wire-speed server-to-client performance
- Web cache redirection techniques for full, wire-speed traffic redirection capabilities across single or multiple web caches or other types of caches
- Coordination of high-availability server load balancing features with Layer 3 and Layer 2 resiliency techniques for simple and effective redundancy
- Sophisticated high-availability capabilities, such as exchanging session information between active and standby server load balancing services and “active/active” configurations
- Flexible “persistence” options to preserve session integrity with servers and optimize hits on web cache servers
- Layer 1-7 server health checking, including the ability to leverage external devices that perform health checking on customized applications
- Wire-speed access control lists for increased security
- Policy-Based QoS bandwidth management and DiffServ capabilities to control and prioritize server applications or access by specific customer classes
- Several load-balancing algorithm options
- Global load-balancing and site recovery functions through integration with the F5 3DNS solution
- Management visibility by integration with the F5 SeeIT management application

## Load Balancing Algorithms

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

- **Round robin** – A simple algorithm that distributes each new connection/session to the next available server
- **Weighted round robin with response-time as weight** – An enhancement of the round robin method where response times for each server within the virtual service are constantly measured to determine which server will take the next connection/session
- **Fewest connections with limits** – determines which server gets the next connection by keeping a record of how many connections 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 times or number of active connections as the selection criteria can optimize user response times.

# The Benefits of Integrated Server Load Balancing

---

The ExtremeWare software suite leverages the capabilities of Extreme Networks' new “i” series switch hardware by enabling wire-speed server load balancing and web cache redirection. They are overlaid on the Extreme Networks infrastructure as “just another service” along with Wire-Speed IP Routing at Layer 3, wire-speed Layer 2 switching, Layer 1-4 access control lists and Policy-Based QoS with bandwidth management.

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

- 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
- Wire-speed performance for server load balancing and transparent web cache redirection applications
- True integration provides a simpler and more resilient solution for link, switch, router and load-balancing capabilities
- Coordinated capabilities for Policy-Based QoS, access policies and system security
- Fewer devices to manage and less training required
- Lower cost of network ownership

Extreme Networks is further leveraging its non-blocking, high-throughput architecture, while continuing to lead the market in delivering Wire-Speed IP Routing at Layer 3, wire-speed Layer 2 switching and Policy-Based QoS. Customers reap the benefits of a cost-effective switching solution that delivers wire-speed server load-balancing capabilities.

The following “before” and “after” diagrams demonstrate the benefits of combining server load balancing with Layer 3 and Layer 2 switching in a highly resilient configuration.

This “before” scenario shows a configuration based on point products, sometimes referred to as the “six-pack” approach. Layer 2 switches provide connectivity between server load balancing devices and the server farm. An additional set of Layer 2 switches is necessary to connect the server load balancing devices to the redundant routers that provide Internet access.

This approach is complex and expensive because each product is an “island of functionality” that must be replicated to eliminate single points of failure. Also, the sheer number of devices in this example increases administration and management complexity.

Each device has its own distinct redundancy mechanism such as the Virtual Router Redundancy Protocol (VRRP), Spanning Tree and the server load balancing redundancy protocol. None of these layered protocols interact with each other, which leaves no resiliency in the event of a failure. If a server load balancing device fails, the standby server load-balancing device may take over. But unless an actual link fails, the upstream and downstream Layer 2 switches will continue to forward server session traffic to the failed server load balancing device.

Another drawback to this approach is that server load-balancing devices can cause performance problems because, unlike LAN switches, they generally do not run at wire speed. Overall, this network can only run at the rate of the slowest device.

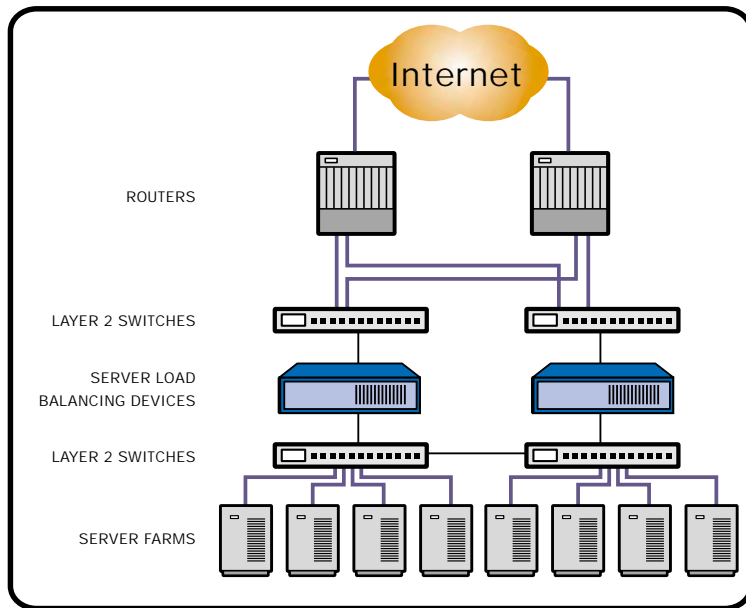
A more effective solution is shown in the “after” scenario, which is based on Extreme Networks “i” series switches with integrated server load balancing.

In this example, only two Summit7i switches are required to build a highly redundant Layer 2 and Layer 3 infrastructure that delivers best-in-class server load balancing at wire speed. Coordinated redundancy between switching, routing and server load balancing within a consistent and easy-to-manage platform offers a much better solution.

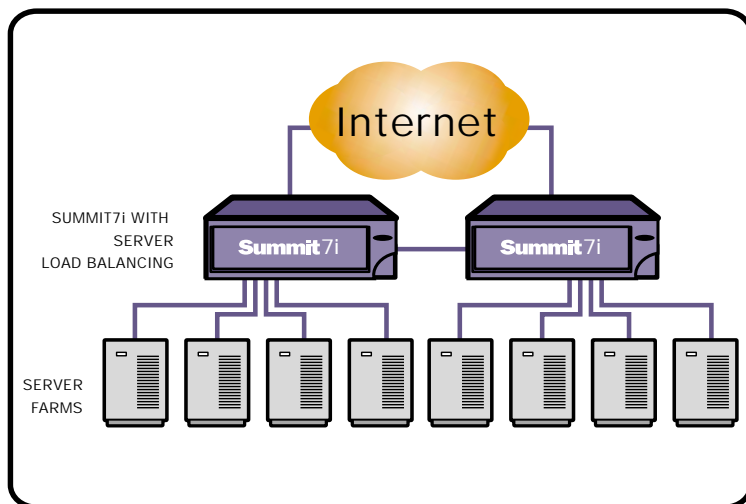
High availability can be ensured when multiple switches take on primary/secondary responsibilities that converge in seconds to support multiple virtual services. This fail-over mechanism is coordinated using Layer 2 and Layer 3 redundancy, which makes sure the entire network can route traffic correctly in the event of a failure.

The Extreme Networks switches perform wire-speed Layer 3 switching with associated routing protocols to the Internet. They also implement default router redundancy using the Extreme Standby Router Protocol™ (ESRP™) to attached devices. ESRP enables host devices to continue communicating even if there is a physical router failure or upstream failure in the routed path. This is vital in today’s service provider, web content provider and enterprise networks, where physical redundancy is essential to ensure uptime, fault tolerance and high availability.

By fully integrating its resilient, wire-speed Layer 3 switching solutions with high availability server load-balancing, Extreme Networks delivers the network and server availability required by mission-critical and revenue-generating applications.



*Before: With point products*



*After: Server load balancing as a layer-independent switch service*

## Web Cache Redirection

---

In addition to server load balancing, Extreme Networks integrates web cache redirection with a twist – it is done transparently and at wire speed. Traffic is redirected at wire speed using Layer 4 criteria, such as HTTP port 80, to one or more load-shared ports across one or more web cache servers. All this occurs transparently, which means users do not need to reconfigure browser applications.

Extreme Networks' wire-speed web cache redirection capability is the first to be integrated with Wire-Speed IP Routing at Layer 3 and wire-speed Layer 2 switching on a single hardware platform. This integration allows for effective web caching integration without requiring radical changes to the network design. Again, the ExtremeWare software suite enables Layer 4 web cache redirection as just another overlaid service on your existing network infrastructure and does not require any network redesign.

For e-businesses and web content providers, transparent caching significantly reduces repetitive hits on servers and allows content-rich pages to be served more quickly to customers. For service providers, this lowers the cost of WAN bandwidth consumption outside the point of presence (POP). And for enterprise networks, frequently accessed web content stays local, thus conserving WAN bandwidth and reducing its associated costs.

## The Benefits of Integrated Server Load Balancing

---

Some applications require persistent sessions between clients and servers. Persistence is the ability to ensure that a user's session with a server will continue to be connected to that particular server. The reasons to preserve a specific session to a particular server can vary from optimizing the cache performance of the server to ensuring a session is not broken. A broken session can result in a shopping cart losing its contents on an e-commerce site.

Persistence based on IP destination address enables service providers and web content providers to optimize repetitive web hits to specific content. Persistence based on source IP address ensures that a client remains connected to a specific server for the duration of a statefull transaction.

Simple persistence based on source IP address works for nearly all Internet applications, except for those clients that might be located behind web proxy farms. It is possible in this scenario for a user's source IP address to change during a single session. This can be overcome using persistence based on the source IP address with a mask. As a result, any sessions from a given set of web proxies will be aggregated to single server.

Some applications may have special needs, such as URL/cookie persistence or SSL session ID persistence for more secure transactions. The most common motivation for considering URL/cookie persistence is to improve database cache hits on servers and preserve session integrity for clients situated behind proxy server farms. This persistence requires all client sessions to terminate on the server load-balancing device and then be reconnected from the server load-balancing device to all servers.

The performance impact of this approach is significant. So much so that the negative aggregate performance on the server load balancing device may be greater than any "cache hit" benefits to the server. Session integrity is often more easily handled by utilizing the previously discussed IP source address persistence with an appropriate mask. In this way, each group of proxy server farms maintains session integrity with individual servers.

The same motivations for doing URL/cookie persistence apply to secure sessions, except that secure sessions encrypt URLs and cookies. Persistence can be based on SSL session IDs. However, session integrity is again easier to handle by utilizing IP source address persistence with an appropriate mask.

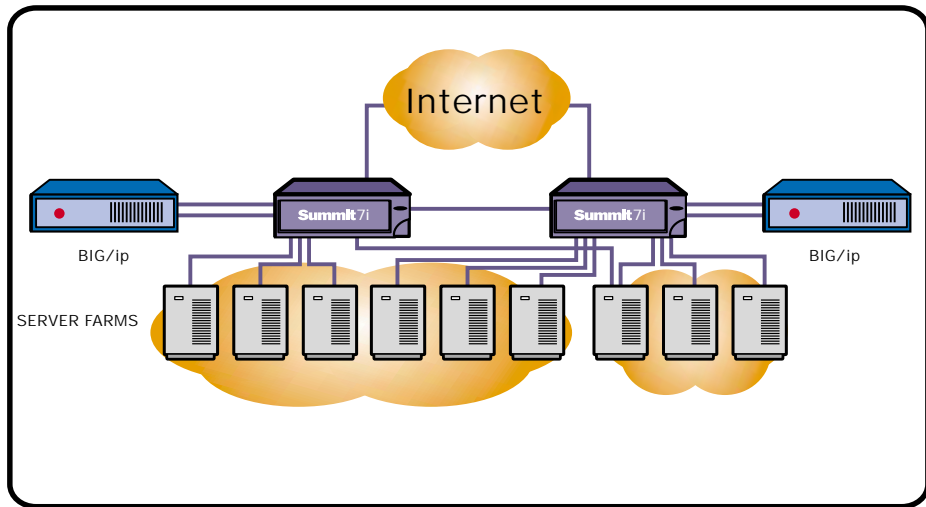
## Server Load Balancing in Specialized Environments

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
- The greater performance of a wire-speed switching solution
- Lower overall system cost

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.

In this example, the Summit7i handles the mainstream server load-balancing applications with the highest load requirements, along with Layer 3 routing functions and associated routing protocols to the Internet for all virtual services. The Summit7i also provides default router redundancy using ESRP to attached devices. The specialized application-level functions, if required, are directed to the F5 BIG/ip network appliance. With this solution, F5's SEE/IT management application and 3DNS distributed load balancing product work with all the depicted load balancing functions, both integrated within the Extreme switch and on the Big/IP external appliance.



*An example of a combined high availability solution consisting of an F5 network appliance solution and a layer-independent switch*

## Distributed Load-Balancing and Disaster Recovery

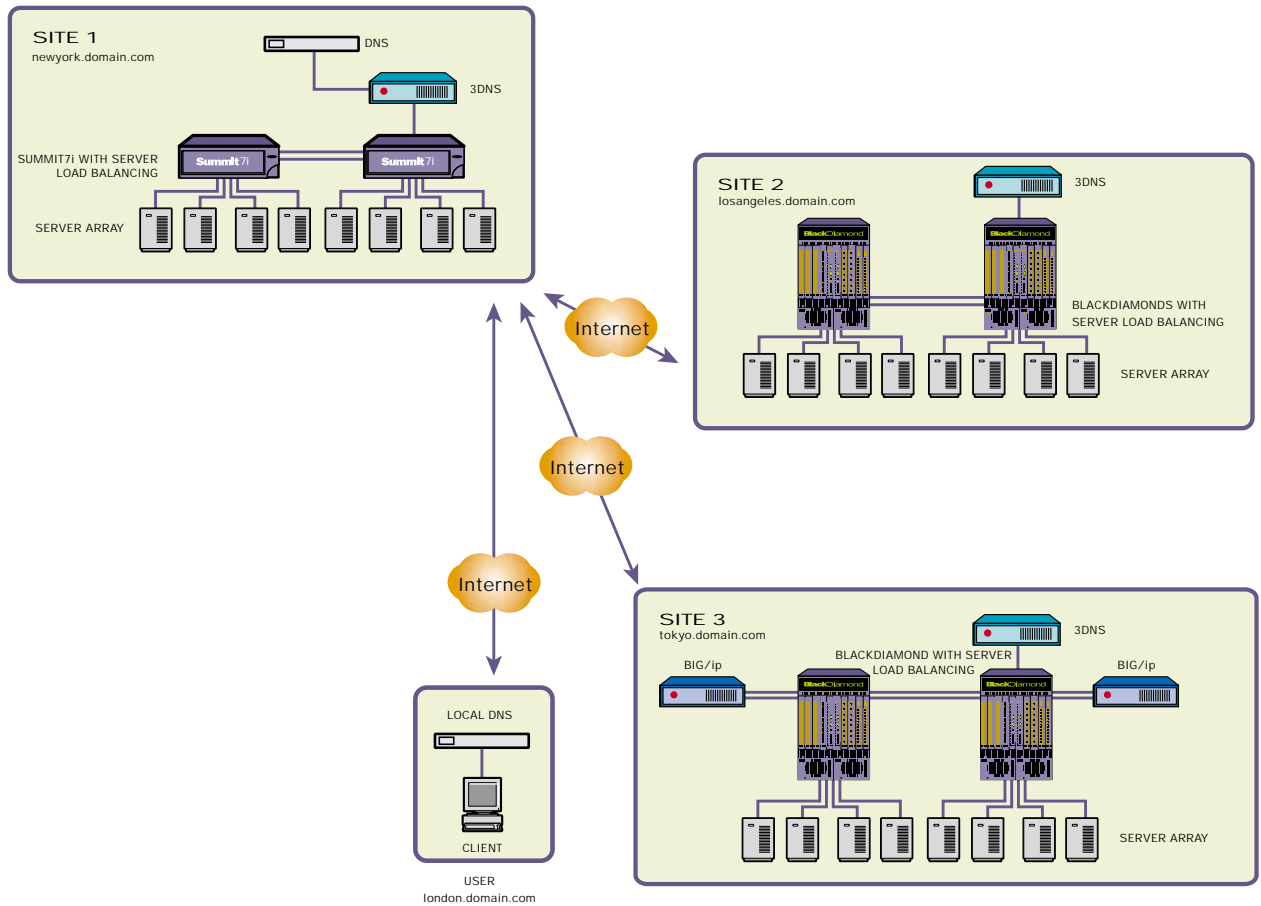
Extreme Networks' partnership with F5 also enables another special capability called distributed server load balancing. Distributed load-balancing gives users access to servers that are geographically distributed across a WAN.

There are several reasons for distributing servers geographically. First, servers can be located closer to end users to minimize WAN delays. Distributed server load balancing can also provide disaster recovery services, also known as wide area failover. Mission-critical applications can be replicated at a disaster recovery site and if the primary site becomes unavailable, the workload is automatically redirected to the backup site.

Extreme Networks switches with integrated server load balancing provide the real-time performance and availability of information necessary for the F5 3DNS controller to intelligently balance traffic on a global scale and provide site redundancy. 3DNS is a wide area traffic manager that extends the capabilities of the Internet's Domain Name Service (DNS). 3DNS maps domain names into the IP address of the appropriate server or server farm based on criteria such as the proximity of the servers to end users, or the availability of applications at each site.

This creates a single "virtual site environment" that centralizes the management of geographically distributed Internet sites and data centers, giving end users a single URL with transparent access to multiple servers in multiple geographic locations.

The diagram below shows F5 appliances and Extreme Networks switches with integrated server load balancing in a distributed network.



## Summary

Server load balancing is a powerful technique for improving application availability and performance in service provider, web content provider and enterprise networks, but piecemeal implementation can also increase network cost and complexity. Server load balancing devices that don't operate at wire speed can also create their own performance bottlenecks.

Extreme Networks 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, Extreme Networks eliminates the need for the extra "islands of functionality" that increase cost and complexity. Even with all server load-balancing functions enabled, Extreme Networks switches continue to operate at wire speed on every port and will not become a bottleneck.

# How Extreme Networks Solves Real-World Server Availability Problems

**Problem:** Customers cannot access servers due to latency in the router interface.

**Solution:** Using the integrated solution, traffic is switched across the Gigabit Ethernet infrastructure at wire-speed so users don't experience delays due to network congestion.

**Problem:** Low-performance servers receive many requests, while high-performance servers are underutilized.

**Solution:** The integrated solution uses advanced server load-balancing techniques at wire-speed to properly direct traffic to the server that is best able to handle it. The result is increased efficiency and reduced workload on servers, which protects the capital investment made in high-performance server arrays.

**Problem:** An upstream router failure causes connectivity problems for servers using their default gateway.

**Solution:** Extreme's ESRP provides servers with redundant default gateway services, including protection from upstream router failures.

**Problem:** A complex mix of point products makes the deployment of a high-availability solution difficult to manage.

**Solution:** The integrated Extreme solution provides a greatly simplified network design with all the necessary power, link, switch, route and server load-balancing resiliency on one stable embedded platform.

**Problem:** There's a server failure. A server becomes unavailable due to a hardware failure or an operating system failure.

**Solution:** Traffic is automatically routed around any server that fails or becomes unavailable, while keeping failures transparent to users. Once a server responds properly, it's automatically added back to the server farm to ease administration.

**Problem:** There's a software failure. Individual applications stop responding, even though other applications are healthy.

**Solution:** The integrated Extreme solution provides proactive monitoring at the protocol port level and detects the failure. The request is then sent to another server where application services are running properly.

**Problem:** There's a content problem. Servers and applications are working properly, but are responding to requests with "404 Object Not Found."

**Solution:** Using content verification capabilities, Extreme Networks switch will actively query individual servers at the application level. If an application does not return the right content, it redirects requests to applications that are responding properly. Users never receive a "404 Object Not Found" message.

**Problem:** There's way too much traffic. As traffic increases, servers reach a critical point where they are unable to respond to requests promptly.

**Solution:** Using the integrated or network appliance solution enables you to set thresholds for acceptable performance. Requests are automatically redirected if a server, service or application is not responding within an acceptable threshold. A maximum number of connections can also be set for each server to eliminate overload. Users experience acceptable response times and a desirable level of QoS is ensured. Bandwidth may also be reserved for specific customers or for specific server applications.