Advanced load balancing: 8 must-have features for today's network demands

Application availability and scalability are no longer enough. Today's enterprises require an integrated solution that also delivers the highest levels of security, performance and adaptability for their business critical Web applications.

Overview

Early-generation server load balancing technology has proven to be an invaluable asset, especially for organizations hosting widely utilized Web applications. By operating as a virtual entry point to such applications, load balancing provides an opportunity to execute a variety of algorithms for splitting the processing load among back-end servers. In addition, periodic polling to establish the status of participating nodes can be used not only to fine tune the load distribution but also to avoid directing traffic to servers that are actually offline. In other words, server load balancers (SLBs) are a simple yet highly effective means to scale an application environment while simultaneously ensuring its availability.

Time marches on, however. Business requirements evolve, as do the processes and technologies used to fulfill them. In fact, the following are just a handful of the key changes and trends that have taken hold since SLBs were first introduced:

- Organizations have become heavily reliant on e-commerce and e-business and the use of the Internet, in general, as a legitimate business tool.
- Traffic volumes have risen dramatically, often creating contention for constrained resources (e.g., network bandwidth, system capacity).
- Applications have become more complex. Support for real-time interaction and multimedia content has placed even greater demands on computing infrastructure at the same time that sensitivity to latency has become the status quo.
- Computing resources have become increasingly centralized (e.g., due to datacenter consolidation) at the same time that users have become increasingly decentralized (e.g., due to mobility, globalization and offshoring).
- The proliferation of regulatory requirements has significantly elevated the business importance of ensuring data privacy and having a comprehensive information security program.
- A shift in hacker motivation has led to a significantly more dangerous threat landscape characterized by a growing percentage of highly elusive application layer attacks.

What these changes and trends expose is the need for enterprises to step up from a simple load balancing solution to a more comprehensive application delivery solution—a solution that addresses not just scalability and availability of the application environment, but application performance, security and adaptability as well. Accordingly, this paper is intended to serve as a guide for organizations looking to replace their early generation SLBs. Details on the top eight criteria to use during an evaluation process are provided, along with numerous examples of how Citrix® NetScaler® meets and often exceeds the associated requirements.

Citrix NetScaler in a nutshell

Citrix NetScaler is an enterprise-class solution for server and global server load balancing. However, it is much more than that. Because NetScaler also incorporates comprehensive application performance and security functionality, it is appropriately classified as a full-featured Application Delivery Controller. A market-proven solution, NetScaler is used by eight out of the 10 largest Web sites, with an estimated 75 percent of Internet users hitting a NetScaler daily.

8 must have features for today’s network demands	
1	Layer 4 load balancing
2	Layer 7 load balancing
3	Global server load balancing
4	Application acceleration
5	Comprehensive application security
6	A purpose-built platform – The key to superior scalability
7	An integrated, modular design – The key to superior agility
8	Unified, simplified management – The key to superior usability

Core load balancing capabilities still an essential starting point

These days, placing greater emphasis on enhancing application performance, security and adaptability is indeed appropriate. By no means, however, does this obviate the need to address fundamental requirements pertaining to application availability and scalability. To ensure these baseline objectives are met, it is recommended that organizations begin their evaluation of an SLB replacement by considering the presence and strength of the feature sets for layer 4 (L4) load balancing, layer 7 (L7) content switching and other L7 traffic management functionality, and global server load balancing.

1. Layer 4 load balancing

The ability to direct traffic based on L2-L4 information (e.g., MAC/IP address and TCP port) should be considered a prerequisite for all load balancing solutions. Related functionality that should also be present is concerned with health monitoring, session persistence and network integration.

- Health monitoring entails using various mechanisms (e.g., ping, SNMP, scripts) to continuously establish the availability and relative health—from a performance perspective—of virtually every part of the application infrastructure: intermediate network links and devices, server hardware, operating system services and even individual modules of the application itself. The gathered information can then be used to help distribute sessions in a manner that avoids bottlenecks or downed components.
- Session persistence is necessary for designs where back-end state information is not being shared and, therefore, any given user's session needs to be handled by the same server from start to finish. In this case, various options (e.g., source IP address, cookies or hashing of various attributes) should be available to ensure follow-on requests continue to be directed to the server node chosen to process the initial request.
- Network integration and compatibility are easy to overlook, but equally important. The load balancing platform should simply fit in to the existing environment without the need for modifications. As a result, it should support a wide range of routing protocols (e.g., OSPF, RIP, BGP) and common networking techniques (e.g., 802.3ad link aggregation, 802.1q VLAN tagging).

A leading solution such as NetScaler can be identified by its superior breadth of coverage, measured in terms of the protocols that are supported (e.g., TCP, UDP, FTP, HTTP, HTTPS and SIP), the load balancing options and algorithms that are available to choose from (e.g., round robin, least packets, least bandwidth, least connections, response time, SNMP monitoring of back-end resources) and the scope of health attributes that can be monitored. Having the option to deploy the solution using purpose-built appliances, virtual appliances or a combination of both is also an important characteristic when it comes to *fitting in*.

2. Layer 7 load balancing

Also referred to as content switching, L7 load balancing is essentially an extension of the traffic distribution, health monitoring and session persistence capabilities discussed above. The difference is that routing decisions can also be based on application layer data and attributes, such as HTTP header, uniform resource identifier, SSL session ID and HTML form data. This difference enables more-efficient utilization of resources because all of the services and components of an application no longer need to be implemented on all of the server nodes. As a result, each physical system can now be tailored to the functions it will be supporting.

When evaluating solutions against this criterion, emphasis should be placed on the breadth and depth of L7 load balancing and content-switching policies that can be established, as well as the ease with which they can be constructed or configured. Organizations should also consider the value of a variety of advanced L7 content features not strictly associated with distributing traffic. For example, NetScaler enables content to be rewritten (e.g., to mask sensitive data) and includes a responder module for configuring custom responses (e.g., redirects, error messages) to specified types of inbound requests.

3. Global server load balancing

The general concept of global server load balancing is to extend the core L4 and L7 capabilities so that they are applicable across geographically distributed server farms. The primary objective is to provide an additional degree of availability by accounting for site level disruptions and outages. Secondary benefits include: (a) being able to further enhance performance for remote users by routing their sessions to the closest or best-performing datacenter; and (b) being able to balance and optimize resource utilization on an enterprise wide basis.

Unlike many other solutions on the market, NetScaler incorporates global server load balancing as an optional feature. A separate, standalone device is not required. NetScaler's other distinct advantage, once again, is that it offers an extensive array of options when it comes to the site level health attributes that can be monitored, as well as the mechanisms and algorithms that can be used to distribute sessions among an organization's different datacenters.

Stepping up to application delivery

The point has already been made that simple, early generation load balancers are not sufficient. Overall, they leave organizations in the undesirable position of having to acquire and implement an additional set of products to achieve adequate levels of application performance and security. The deficiencies in these early load balancers also explain why leading industry analysts strongly encourage organizations to embrace advanced Application Delivery Controllers (ADCs) when replacing their server load balancers. The intent with ADCs in general, and NetScaler in particular, is to have a single device that incorporates not just a core set of load balancing capabilities but a comprehensive set of application performance and security services as well. The next two sections elaborate on what this means in terms of specific functionality.

4. Application acceleration

Compensating for obvious deficiencies and otherwise enhancing application performance can be a tricky proposition. Sub-optimal application performance can be the result of resource constraints at virtually any point in the path that a user's session traverses. A few of the more likely bottlenecks are inadequate client hardware, insufficient bandwidth at either the client or server end of the connection and overloaded server infrastructure. Alternately, there can be problems with the application itself. This is frequently the case when the underlying protocols or application logic have not been optimized for operation over a wide area network. The resulting condition, referred to as chattiness, is a highly inefficient behavior whereby it takes numerous back-and-forth exchanges between client and server to complete a single, user level action.

The diversity of potential issues is why an ideal solution should incorporate an overlapping set of features that enhance application performance. These include caching, compression, TCP communications management and SSL offload.

- Caching techniques enable frequently requested content to be served from the load balancer platform. This technology accelerates delivery to the user while relieving some of the processing demand placed on back-end servers. These gains are maximized with NetScaler, based on the fact that its Citrix® AppCache™ functionality provides in-memory caching not just for static data, but for dynamically generated HTTP application content as well.
- Compression is all about reducing the amount of data that must traverse the connection in the first place—even for encrypted sessions. The next generation of Web 2.0 applications frequently includes large numbers of cascading style sheets and JavaScript, making compression even more important. Compression helps alleviate network congestion and can accelerate transactions by three to five times.
- TCP communications management covers two major items. At the front end (i.e., between the client and ADC), TCP optimization techniques (e.g., forward-error correction, window scaling and buffering) help make more efficient use of available bandwidth and reduce the amount of chattiness. At the back end (i.e., between the ADC and server nodes), TCP multiplexing enables the aggregation of a large number of HTTP requests over a much smaller number of long-lived TCP connections. The impact on server load and response time can be quite dramatic, as this significantly reduces the processing demand associated with connection setup and teardown.
- SSL offload similarly relieves back-end servers by performing compute-intensive encryption and decryption processes on their behalf—ideally, by taking advantage of hardware that is specialized to the task.

Of course, having a comprehensive set of application acceleration features is really just table stakes. With NetScaler, organizations also benefit from having highly granular control over the configuration of these capabilities. This control is particularly important for caching and compression mechanisms since there are often scenarios where: (a) it is preferable to not cache certain content or (b) the use of compression incurs a greater penalty than the benefit it provides (e.g., for low-latency, high-bandwidth connections).

Pulling double duty

All of the application acceleration capabilities discussed above contribute to a significant, secondary benefit. Specifically, by offloading network and server infrastructure, these capabilities often enable organizations to make do with fewer resources, delaying the need for further investments in network bandwidth, routing and switching platforms, and server hardware.

5. Comprehensive application security

As an intermediary between users and back-end resources, the SLB/ADC is also an ideal place to implement much needed security measures. Recalling the trends highlighted earlier—especially those pertaining to the evolution of threats, user mobility and inter-connectivity—it should be clear that SSL VPNs and application firewalls are two countermeasures, in particular, that deserve attention.

Aside from facilitating remote access, the benefit of having SSL VPN technology as an integral component of an ADC is that it provides fine-grained control over which users have access to which functions in which applications, and under which conditions (e.g., based on type and configuration status of the client device). When properly utilized, this capability can substantially reduce the risk of providing application access to a vast population of remote, mobile and third party users.

The shortcomings of network firewalls, which concern themselves primarily with network addresses and port-level information, are well documented. In general, they do not understand the inner workings of protocols and languages such as HTML and XML; they do not understand HTTP sessions; they cannot validate user inputs to an HTML application; they cannot filter or obfuscate sensitive data included in server responses; they cannot detect maliciously modified parameters in a URL request; and they are incapable of inspecting SSL-encrypted traffic. In contrast, it is specifically this depth of visibility and control that enables an application firewall to protect Web applications against a wide range of both known and unknown attacks.

Of course, having robust, application layer controls does not obviate the need to provide protection at other layers of the stack. This is another area where NetScaler outshines the competition. For example, NetScaler features a customized TCP/IP stack that: (a) enforces a positive security model, dropping all traffic that deviates from common guidelines for packet formation and content; and (b) prevents leakage of low level information by zeroing the unused portions of reused packets. In addition, NetScaler provides robust connection handling routines to automatically thwart many types of DDoS/flood attacks.

Meeting and exceeding expectations

The final three criteria are what set superior application delivery solutions such as NetScaler apart. Although many solutions may, in fact, incorporate all of the aforementioned functional capabilities, those that fail to thoroughly address the need for a choice of platform, an integrated, modular design and unified management will not be nearly as effective and efficient as those that do.

6. A purpose built platform – The key to superior scalability

Application delivery is substantially more compute intensive than ordinary load balancing. Not only is the scope of functionality greater, but so is the depth of processing that needs to be conducted to provide the requisite level of application visibility and control. Less clear, though, is how to account for this difference, especially in ensuring the solution is able to scale appropriately.

One key is having a platform where the hardware—and more importantly, the system level software—has been constructed and optimized explicitly for the higher-level services that define an ADC. Some of the more significant features of such a platform are:

- A customized operating system – General purpose operating systems are interrupt driven and designed to provide equitable treatment for the widest possible set of applications. However, because it has complete control over functions such as process timing, memory management and network access, the customized system in NetScaler is able to optimize resource allocation for the tasks at hand. The result is a far more deterministic processing model with lower latency and greater overall scalability.
- A customized TCP/IP stack – A logical extension of the previous item, this one ensures even greater processing efficiency, and also provides an opportunity to implement the aforementioned stack-level security mechanisms.
- An intelligent HTTP parsing engine – Ideally, packet processing tasks should not need to be repeated for each individual function (e.g., caching, compression).
- Appropriate appliance design – This does not imply that custom silicon (i.e., ASICs) should be used for everything, or even most things. Indeed, when it comes to L7 operations, general-purpose hardware (e.g., the Intel x86 platform) has proven to be more efficient, adaptable and therefore economical. However, where massive scalability is required for lower-layer processes that are highly deterministic and repetitive (e.g., cryptographic functions or flow control), using an ASIC to accelerate this lower layer processing is appropriate.

Equally important to being able to scale, however, is being able to do so in a manner that is both affordable and agile. This is where the flexibility of the NetScaler system architecture has a distinct advantage, since it makes feature-complete NetScaler virtual appliances possible. With the addition of Citrix® NetScaler® VPX to the NetScaler product family, IT organizations have a choice. They can implement purpose-built NetScaler appliances to achieve maximum scalability; implement NetScaler VPX virtual appliances to reduce their total cost of ownership, and increase the flexibility and responsiveness of their application delivery infrastructure; or implement a combination of both platforms to achieve an optimum balance between both sets of objectives.

With NetScaler VPX—a full-featured virtual appliance version of NetScaler that can be deployed on any hardware platform running the Citrix® XenServer™ server virtualization system—there is no physical appliance to deal with. As a result, IT departments can deploy application availability, security and accelerations services on-demand, anywhere within private, hosted or cloud-based networks and datacenters. Not only is a more thorough implementation of critical application delivery services possible, but it can be done in a way that takes full advantage of virtualized servers and off-the-shelf hardware that already in place—all while facilitating the longer-term objective of having a fully dynamic datacenter.

7. An integrated, modular design – The key to superior agility

For most organizations, having options is a firm requirement. So is having a solution that is adaptable and, therefore, future proof. Consequently, a top consideration for an SLB replacement is that it feature a modular design. This way individual capabilities (e.g., application firewall, SSL VPN) can be added as needed when the organization is ready to take the next step in the evolution of its application delivery infrastructure. Furthermore, new modules that account for ever changing conditions can be developed and implemented over time without having to resort to deploying a fleet of additional, standalone devices.

Equally important is that the modules be truly integrated components of the overall system. For instance:

- Each module should take full advantage of the embedded scalability, performance and security features of the underlying platform.
- The presence of any given module should not prevent other functional modules from taking advantage of a given system's features (e.g., support for multi-core processing).
- Modules should be intelligent and selective—for example, if the application firewall requires full, deep-packet inspection of specific traffic flows, then it should not automatically force all other flows to be handled this way.
- Individual modules should not require their own, separate management consoles.

NetScaler fully meets these requirements. Its design is highly modular, yet the individual functional capabilities are tightly integrated and completely compatible. Furthermore, all features are available on all models all of the time.

8. Unified, simplified management – The key to superior usability

Ultimately, the ability to unleash the full power of an ADC depends heavily on the strength and usability of the associated management capabilities. Three elements of the NetScaler solution are particularly helpful in identifying the specific features to look for when considering management capabilities.

- The intuitive AppExpert Visual Policy Builder enables application delivery policies to be created without having to code complex programs or scripts. In addition, the unification and consolidation of multiple capabilities in a single solution keep administrators from having to jump between different consoles and policy models.

- Citrix® EdgeSight™ transparently instruments HTML pages, providing granular visibility into how Web applications are behaving from the user's perspective. Detailed results can then be used to fine tune individual policies and take further advantage of the system's acceleration capabilities to ensure a superior application experience.
- NetScaler Command Center enables efficient, centralized administration of system configuration, event management and performance management for organizations that elect to operate multiple NetScaler appliances.

Summary

Early generation server load balancers are tried and true solutions for improving the availability and scalability of an organization's application infrastructure. Nonetheless, enterprises that persist in using such products run the risk of exposing themselves and their customers to increasingly poor application performance and a seemingly endless stream of application layer security threats.

One option to overcome these shortcomings would be to implement multiple standalone devices that address each of the underlying issues. However, a much more efficient and effective approach is to replace old server load balancers with new Application Delivery Controllers. These tightly integrated physical and virtual appliances not only provide core load-balancing capabilities, but also deliver the highest levels of security and performance for today's business critical Web applications. Furthermore, the eight criteria detailed in this paper can be used to help ensure that enterprises select a solution that is truly best of breed.

Worldwide Headquarters
Citrix Systems, Inc.
851 West Cypress Creek Road
Fort Lauderdale, FL 33309, USA
T +1 800 393 1888
T +1 954 267 3000

www.citrix.com

Americas
Citrix Silicon Valley
4988 Great America Parkway
Santa Clara, CA 95054, USA
T +1 408 790 8000

Europe
Citrix Systems International GmbH
Rheinweg 9
8200 Schaffhausen, Switzerland
T +41 52 635 7700

Asia Pacific
Citrix Systems Hong Kong Ltd.
Suite 6301-10, 63rd Floor
One Island East
18 Westland Road
Island East, Hong Kong, China
T +852 2100 5000

Citrix Online Division
6500 Hollister Avenue
Goleta, CA 93117, USA
T +1 805 690 6400

About Citrix
Citrix Systems, Inc. (NASDAQ:CTXS) is the leading provider of virtualization, networking and software as a service technologies for more than 230,000 organizations worldwide. Its Citrix Delivery Center, Citrix Cloud Center (C3) and Citrix Online Services product families radically simplify computing for millions of users, delivering applications as an on-demand service to any user, in any location on any device. Citrix customers include the world's largest Internet companies, 99 percent of Fortune Global 500 enterprises, and hundreds of thousands of small businesses and prosumers worldwide. Citrix partners with over 10,000 companies worldwide in more than 100 countries. Founded in 1989, annual revenue in 2008 was $1.6 billion.