Since the rise of the Internet, enterprises have done an excellent job of providing network-level infrastructure and security. However, an increasingly mobile workforce, globalization, and a growing dependency on Web-based applications are illuminating problems that a network-only focus cannot address. Application users need quick response times, high availability, and application-layer security for the mission-critical Web-based applications used today. Citrix NetScaler application delivery controllers can address these challenges and provide advantages to a broad range of applications from leading independent software vendors.

Identifying Web-based application challenges

Routers, switches, firewalls, and other networking infrastructure, along with advanced dynamic routing protocols, have created an efficient method of moving packets. Optimal paths are selected, networking protocols can retransmit dropped packets, and “connection-oriented” protocols help ensure packets are eventually delivered. Yet, for all the networking advances that have been made in Open System Interconnection (OSI) Layers 2, 3, and 4, some issues persist when this infrastructure is used for high-performance applications communicating across wide area networks (WANs). Moreover, many of today’s most pressing availability, performance, and security issues require Layer 7 visibility and processing.

TCP/IP, the protocol that Web-based applications are built on, originated with the premise of ensuring connectivity under any circumstances. HTTP was initially designed for academics to share research papers. Performance and security, especially for highly interactive, real-time, and security-sensitive applications, were not initial design points for the underlying infrastructure Web-based applications rely on.

Even using the fastest routers with wire-speed throughput and optical transport links, communicating to a global user base can add hundreds of milliseconds of latency to packet transport. The TCP/IP emphasis on availability magnifies the impact of this latency, in many cases increasing network congestion and frequently degrading application performance. Often, this delay causes network- or application-layer time-outs and forces traffic regeneration, which only further exacerbates the problem.

Only so much data (often less than 64 KB) can be transmitted before the receiver must send an acknowledgment to the transmitter. Although very few packets are typically dropped, these dropped packets force a retransmission of some packets previously sent successfully. Data rates from transmitters must frequently be throttled back because of wire congestion or the inability of the receiver (or routers along the way) to buffer the packets. When this throttling occurs, transmission rates only slowly return to normal.
levels. The net result is that a connection specified as 100 Mbps, for example, cannot carry more than a small fraction of this load. In many instances, a 45 Mbps T3 link with 90 milliseconds of round-trip time provides barely over T1-level throughput.

HTTP, the application protocol the Web is built on, is notoriously chatty. This inherent chattiness, when combined with the highly interactive nature of Web and Web 2.0 applications, exacerbates the inherent performance limitations of the TCP architecture, and has created inefficiencies when scaled globally. Applications originally meant to be used within an office, but now used globally, suffer from poor performance. Because the core problems are associated with essential infrastructure, rewriting applications, increasing bandwidth, and adding servers fail to resolve them.

Addressing networking and application challenges

Solving these problems requires addressing the core networking behaviors that do not align with the behavior of Web-based applications. Addressing these behaviors frequently requires operating not only at the networking level, but also at the application layer (see Figure 1). Because the packet delivery infrastructure is opaque to the applications themselves, a viable solution must have insight into, and be transparent to, both elements.

Citrix NetScaler meets this requirement by integrating traditional server and network load-balancing services with the application-aware functionality necessary to efficiently deliver highly interactive Web-based applications across the globe. Specifically, NetScaler incorporates extensive network and application-layer capabilities such as protocol optimizations, compression, caching, and application security to accelerate application performance, increase application availability, and enhance application security (see Figure 2).

Application acceleration

Citrix NetScaler incorporates networking and application techniques to accelerate application performance. To offset the negative consequences of the original TCP/IP architecture, NetScaler provides numerous networking optimizations that help minimize the impact of network latency, reduce network congestion, and ultimately increase application performance. Examples include the following:

- **TCP windows scaling**: Enables administrators to set the TCP window size at greater than 65 KB
- **Selective acknowledgment and retransmission**: Intelligently manages data acknowledgments and retransmits dropped packets, helping reduce wire congestion
- **TCP fast ramp**: Overcomes the standard TCP slow-start algorithm by initially setting the transmission rate equal to what the client can receive, helping increase the efficiency of bandwidth use and reduce response times
- **Client keep-alive**: Intelligently determines when a client connection can be kept open and when it must be closed; maintaining client connections helps reduce network congestion and response times

Using compression also helps accelerate data delivery. Compressing data helps reduce traffic, thereby minimizing the likelihood of packet loss and the number of data acknowledgments while increasing data transfer efficiency and application performance. By offloading this processing-intensive data compression task, NetScaler appliances enable servers to operate with increased efficiency. NetScaler is designed to compress data by three to five times, depending on content type, thus helping reduce bandwidth needs.

Content caching is designed to further reduce user response times. Multiple users often request the same information, which may

---

**Figure 1. Application and network traffic**

**Figure 2. Citrix NetScaler application delivery controller architecture**
be static or dynamically generated by the application. By caching data ahead of time on the appliance, NetScaler enables an immediate response, further enhancing user experience.

**Application availability**

Functionality such as load-balancing, content-switching, and offload techniques help ensure servers are reachable. When servers are load-balanced, each server gets its proper share of requests, helping maximize its efficiency and ability to respond in a timely manner. Load-balancing methods can be as simple as round-robin algorithms or as complex as load distribution based on real-time processor utilization rates. Health checks help ensure servers and databases are operational before requests are forwarded to them. If a session demands a persistent connection, it can be maintained by a variety of methods, including Secure Sockets Layer (SSL) IDs, cookies, or tokens created from an aspect of the initial request’s payload—for example, an application-level transaction ID within a Simple Object Access Protocol (SOAP) request.

With content switching based on application-layer content, load balancing is further enhanced as traffic is directed to servers optimized for specific content requests. For example, requests for dynamic content (such as URLs with a suffix of .asp, .dll, or .exe) can be directed to one server or farm, and requests for static content directed to another. Organizations can achieve excellent server farm efficiencies without duplicating content across servers.

The multiple offloading methods Citrix NetScaler provides help increase server availability and reduce application response times. SSL encryption, especially given the increased use of large key sizes and complex encryption algorithms, is incredibly processor intensive for application and Web servers. Offloading SSL handling onto purpose-built appliances frees servers to serve content, generally leading not only to reduced server processor utilization, but also to increased application availability and accelerated response times.

### Table: Test Results

<table>
<thead>
<tr>
<th>Workflow</th>
<th>Test</th>
<th>Response time (seconds)</th>
<th>Relative decrease in response time with NetScaler</th>
</tr>
</thead>
<tbody>
<tr>
<td>MySite</td>
<td>Baseline</td>
<td>6.65</td>
<td>47.7%</td>
</tr>
<tr>
<td></td>
<td>NetScaler</td>
<td>3.48</td>
<td></td>
</tr>
<tr>
<td>Search Query</td>
<td>Baseline</td>
<td>5.24</td>
<td>26.7%</td>
</tr>
<tr>
<td></td>
<td>NetScaler</td>
<td>3.84</td>
<td></td>
</tr>
<tr>
<td>Document Library</td>
<td>Baseline</td>
<td>6.49</td>
<td>48.2%</td>
</tr>
<tr>
<td></td>
<td>NetScaler</td>
<td>3.36</td>
<td></td>
</tr>
<tr>
<td>Team Site</td>
<td>Baseline</td>
<td>6.29</td>
<td>82.0%</td>
</tr>
<tr>
<td></td>
<td>NetScaler</td>
<td>1.13</td>
<td></td>
</tr>
</tbody>
</table>

*Note: Response time based on configuration with 512 Kbps bandwidth and 140 ms latency.*

**Figure 3. Test results when running Microsoft Office SharePoint with Citrix NetScaler**

The basic compute-intensive TCP/IP session setup and teardown process is also largely offloaded from the server. NetScaler terminates client TCP connections and, in turn, manages a set of long-lived sessions with the servers. This multiplexing is designed to reduce the number of TCP sessions a server must maintain by 100 times or more.

**Application security**

Networking-level security holes have largely been plugged. However, the applications themselves remain exposed, because the network must allow users to reach these applications. Application-level attacks take advantage of seemingly innocuous coding flaws.

Application-layer protection helps mitigate the application-layer attacks prevalent today. Network security cannot prevent attacks based on SQL injection, buffer overflows, form field tampering, or cross-site scripting. Moreover, these threats are also extremely difficult to mitigate using intrusion prevention systems.

To help deliver zero-day protection against these exploits and many more, Citrix NetScaler employs the positive security model. Rather than relying on signatures, this model combines knowledge of HTTP standards and protection against known application attack vectors with the ability to automatically learn an application’s correct behavior. Abnormal application behavior is treated as potentially dangerous and, depending on administrative preference, blocked and/or logged. The appliance automatically learns legitimate behavior and is designed to block known malicious usage without false positives.

**Testing performance enhancements with Citrix NetScaler**

Citrix NetScaler has demonstrated enhancements that not only address traditional server availability concerns, but also help increase application performance. Testing done in conjunction with Microsoft, among many other software vendors, has shown tangible advantages. Such independent testing occurred on-site in the vendor’s laboratories. For example, when running Microsoft® Office SharePoint® software on Dell™ PowerEdge® servers in conjunction with NetScaler, Microsoft showed up to an 82 percent reduction in latency for various workflows (see Figure 3).1

**Helping applications run as intended**

Applications once all but useless because of the inherent limitations of legacy protocols and WAN distances can excel when deployed with Citrix NetScaler. Using Citrix NetScaler can help organizations increase performance levels and server availability, maximize operational efficiency, and enhance security all the way to the applications themselves.

**John Gudmundson** is a senior product marketing manager in the Applications Networking Group at Citrix Systems.

---