SPIRENT Communications | Inspired Innovation

White Paper

Evaluating, Selecting and Deploying WAN Acceleration Solutions

Real-World Testing Sifts Through Conflicting Claims to Identify the True Performance of Competing Solutions

January 2008

Rev. A 01/08

Spirent Communications, Inc.
1325 Borregas Avenue
Sunnyvale, CA 94089 USA

Email: sales@spirent.com
Web: http://www.spirent.com

Americas
T: +1 800.SPIRENT
+818 676.2683

Europe, Middle East, Africa
T: +33 1 6137.2250

Asia Pacific
T: +852 2511.3822

Evaluating, Selecting and Deploying WAN Acceleration Solutions

Real-World Testing Sifts Through Conflicting Claims to Identify the True Performance of Competing Solutions

Contents

Overview

As corporate networks continue to expand worldwide, the efficiency of the global network becomes critical. The question arises, "How do you expand the corporate network to a national or global scale without sacrificing usability and productivity?"

The most obvious solution for optimizing the wide area network (WAN) is the WAN accelerator. But the wide range of technologies and solutions are often difficult to differentiate and evaluate. And once a solution is selected, IT organizations face the challenge of determining the optimum configuration.

Real-world testing allows IT organizations to cut through hype and reveal the true performance of a technology, implementation or brand. During deployment, real-world testing isolates factors affecting performance in a production network. Insight into these factors enables proper configuration of a solution for optimum performance.

This white paper highlights WAN performance issues, explains technologies developed to address those issues and demonstrates the importance of real-world testing when evaluating, selecting and deploying WAN acceleration tools or technologies.

Alleviating WAN Constraints

The distributed workplace is a given in the 21st century. Beyond the corporate office we find regional, branch or remote offices, off-site data centers (both primary and backup), mobile users, telecommuters, off-shore offices and international offices. At each location, a local area network (LAN) connects local users. Between locations, a WAN connects the LANs.

Organizations rely on the WAN to support a highly diverse set of applications, from best-effort delivery of email and Web pages to real-time traffic like Voice over IP (VoIP) and video conferencing. Many other business-critical functions lie between these two extremes including data synchronization, off-site backup, distributed applications, collaboration applications, file sharing and others.

The bandwidth and latency constraints posed by the WAN often create performance and responsiveness problems affecting top and bottom lines. Slow response from online retail systems can result in abandoned shopping carts and reduced revenues. Slow response from back-office systems, extended time for backups and mirroring, and poor audio and video conferencing quality can translate into higher operating costs and reduced profits.

While some of these issues can be addressed simply by increasing bandwidth, others require Quality of Service (QoS) classes to prioritize traffic. Others still are resolved only by increasing the efficiency of the current network. WAN acceleration is one method commonly used to increase network efficiency.

WAN Acceleration and Response Time

WAN acceleration reduces costs by increasing the effective throughput of the existing network. Most WAN acceleration solutions use an array of techniques to get a 10X to 40X improvement in response time or transaction processing rate. Some specialized techniques targeted to specific protocols, such as the Messaging Application Programming Interface (MAPI) and the Common Internet File System (CIFS), can achieve up to a 100X improvement in performance.

Consider the experience of one North American organization that partnered with a team in Asia. They discovered that it took 160 times longer to send a document over the WAN, as transfer time increased from 30 seconds to 80 minutes. TCP optimization and compression reduced the overseas transfer time to 2 minutes. With dramatic improvements like this, it is not surprising that WAN accelerators are attractive to organizations facing WAN issues.

Technology in the WAN

The primary effects the WAN introduces are bit errors, packet loss, and latency also known as delay. To a lesser degree, packets may be reordered or duplicated while traversing the WAN. These impairments have varying effects on different types of traffic. For example, as latency increases beyond 80 ms, throughput on a TCP session begins to degrade regardless of available bandwidth. VoIP, on the other hand, handles twice that latency without problems as long as the delay variation is low. As a result, the technique used to increase performance and throughput varies based on the type of impairment and the type of traffic.

Traffic Shaping through QoS

Traffic shaping improves throughput for preferred traffic at the expense of lower-priority traffic. Delay-sensitive traffic such as video or voice is prioritized and guaranteed delivery. Other data traffic, such as email or Web page requests, are delivered if available bandwidth permits or are dropped if higher priority traffic has consumed available bandwidth. Some WAN accelerators offer QoS features to prioritize traffic.

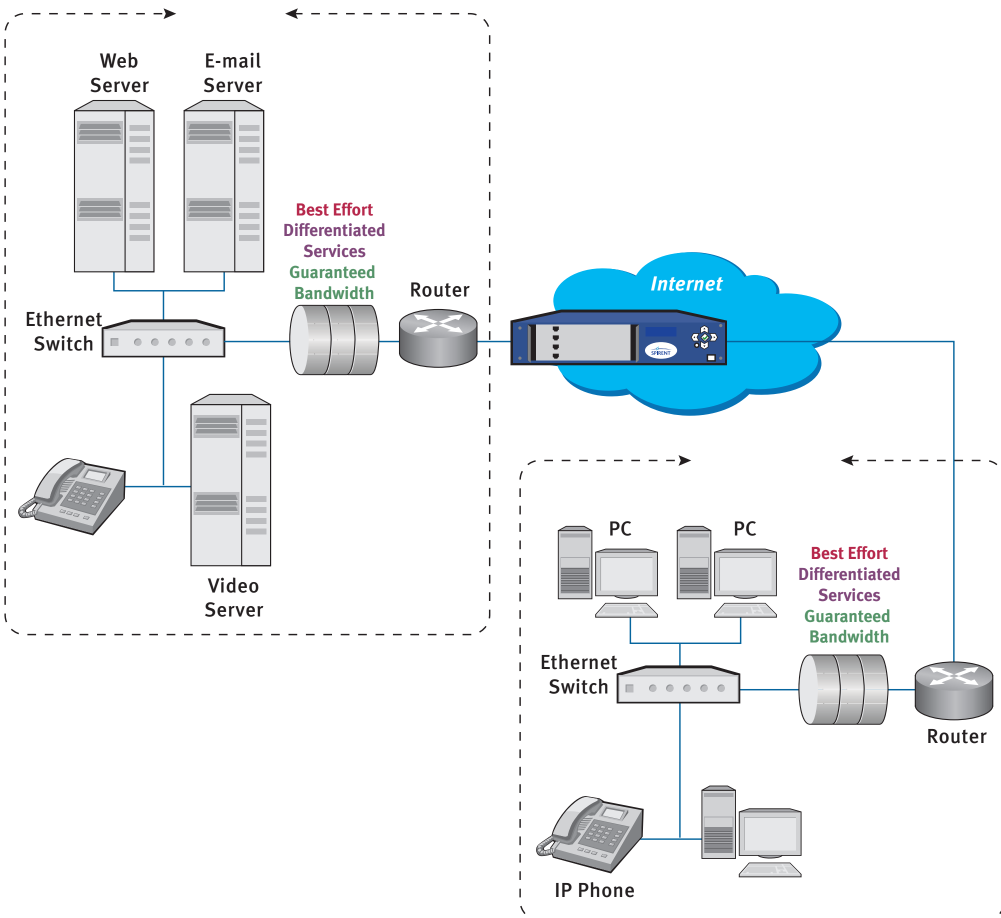


Figure 1. Network QoS

TCP Optimization

Internet Protocol (IP) sends packets to the destination without checking to see if they arrived. The Transport Control Protocol (TCP) maintains the end-to-end connection by requiring periodic acknowledgments from the destination. TCP uses a sliding window to avoid congestion. The window size denotes the number of packets that can be sent before an acknowledgement from the receiver is required. A larger window size means more data can be transmitted before waiting, which results in greater throughput.

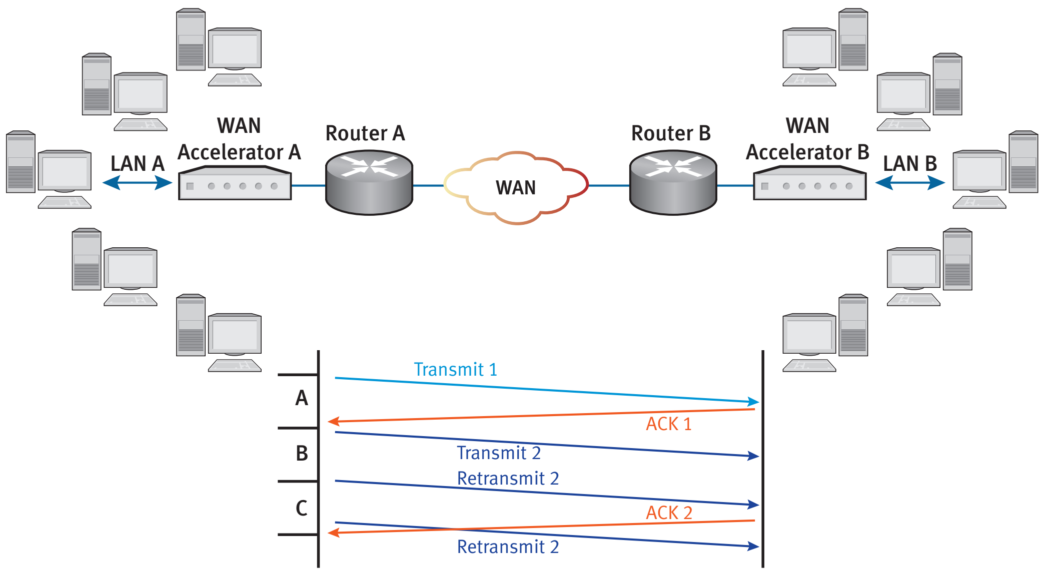


Figure 2

When the sender receives an acknowledgement (ACK) before it times out, it increases the window size for the next transmit. (Scenario A in Figure 2) If the transmit or ACK is lost, the sender times out and re-transmits. (Scenario B in Figure 2) If the ACK is delayed due to network latency and arrives after the time-out value, the sender assumes the packet was lost and re-transmits. (Scenario C in Figure 2) Re-transmissions cause TCP to reduce the window size and/or increase the time-out value, which can result in inefficient use of bandwidth.

TCP uses a slow-start method to gradually ramp up the window size, which can cause sluggish performance. Enabling RFC 1323 TCP Extensions for High Performance can improve throughput, but for high-latency connections the round-trip time can cause long waits for acknowledgement even if bandwidth is available and there is no congestion.

WAN acceleration solutions use connection optimization and aggressive windowing methods to improve performance. Rapid accelerate-and-fallback windowing can achieve 700 Mbps throughput even on intercontinental links. Connection optimization reduces the wait for acknowledgements by terminating the TCP session at the local accelerator rather than at the destination, resulting in near-instantaneous responses rather than long waits for acknowledgments (Figure 3).

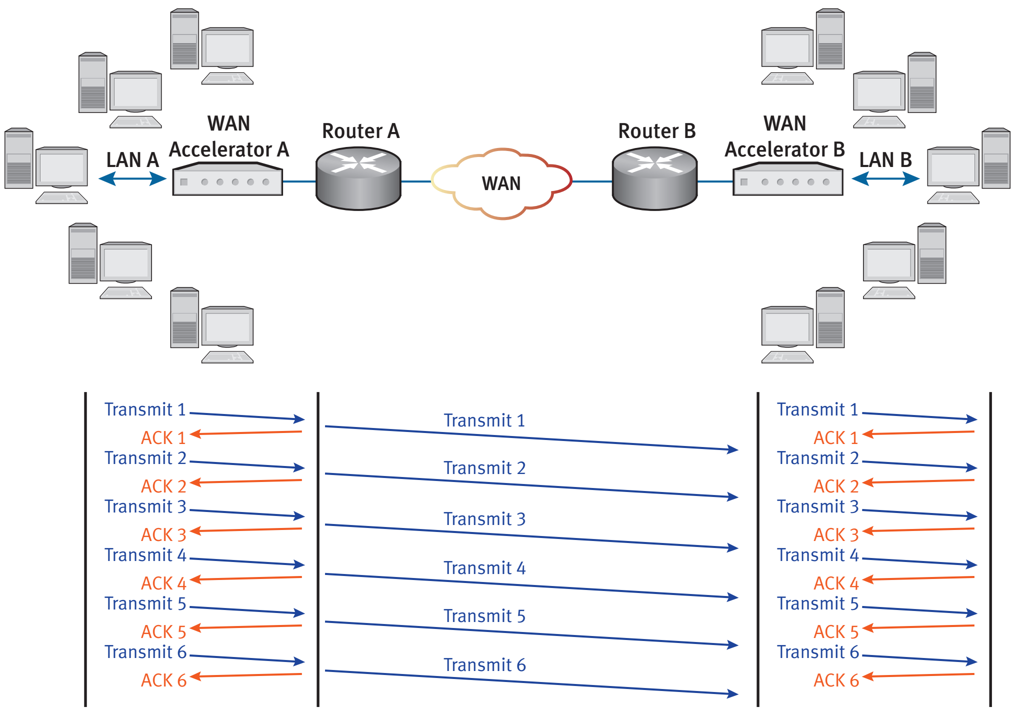


Figure 3

In addition, connection pooling re-uses WAN-side connections for multiple requests, saving setup and teardown time and CPU cycles. Unlike many of the other acceleration techniques, TCP optimization doesn't necessarily require a pair of devices, one on each end of the connection. It can be implemented on a server-side appliance, improving throughput for anyone who accesses the server – be it an entire remote office or just a single remote user.

Compression

Typical compression techniques replace repeating bit patterns with a short label before sending the data across the WAN. On the receiving side the label is removed and the original pattern is inserted back into the bit stream (Figure 4). Depending on the compression scheme and the data type, the number of bits transmitted can be reduced by 20% to 70%. Some proprietary techniques that use parsing and hierarchical trees claim to reduce data size by up to 98%. In addition, by offloading compression to the WAN accelerator, servers can serve more requests thereby improving throughput and response time.

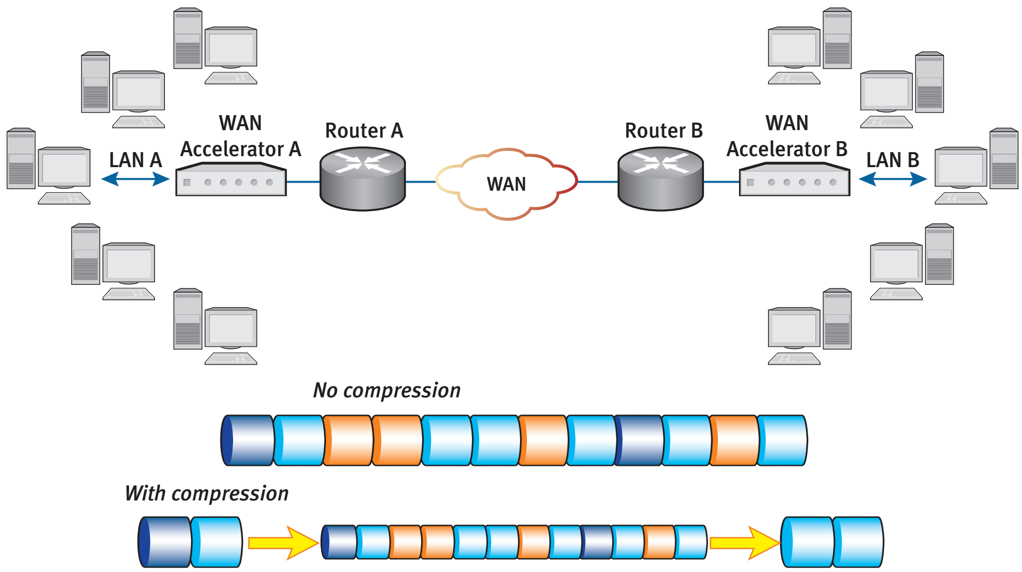
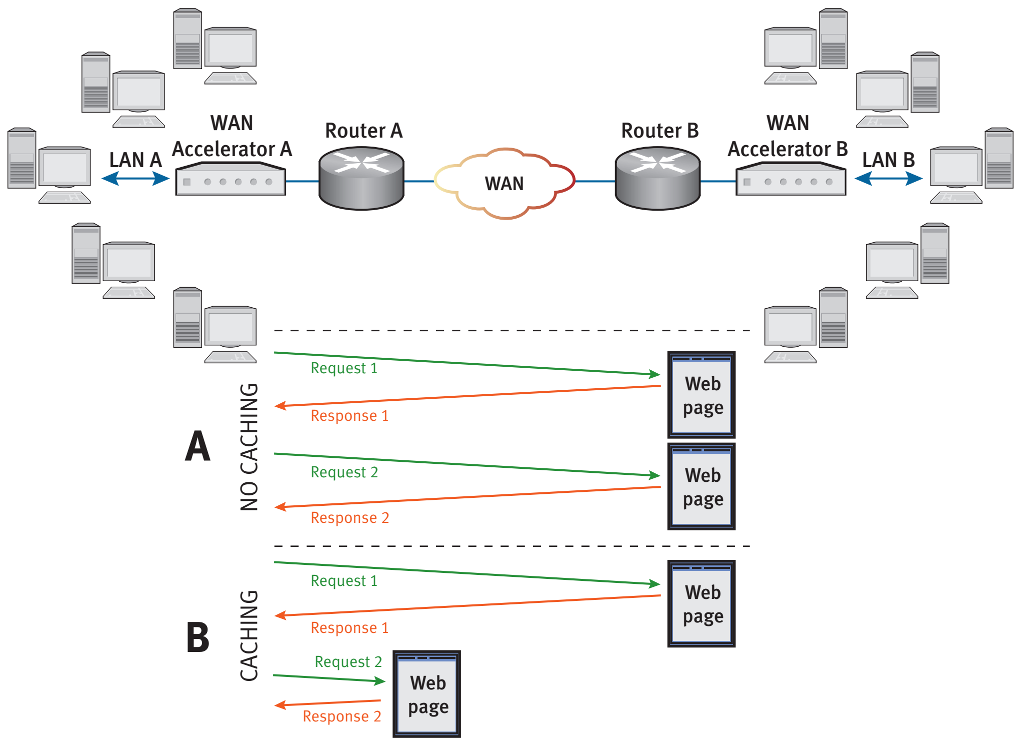


Figure 4

Caching and Mirroring

Caching is used at many levels in computing and networking, from onboard caches in system processors to print job spooling. Client-side WAN accelerators store Web pages and other data as it passes through so subsequent requests can be answered locally by the accelerator rather than requiring retransmission from the remote server. Database applications can provide read-ahead caching to transmit additional pages for a query before they are requested. Caching improves response time and reduces demand on the servers (Figure 5).

High-availability mirror systems provide synchronized copies of entire databases at multiple locations. When a change is made locally, it is replicated at all the other sites in the background so that requests from users in those locations can provide the most recent data without having to query a remote server.



A. Caching
B. No caching

Figure 5

Off-loading CPU-Intensive Tasks

Multi-processor systems are typically licensed per processor, adding hardware and software costs as more processors are required. As the number of transactions increase, so increases the demands on processors that may already be pushed to their limits with ancillary tasks such as compression, encryption and XML processing. By offloading CPU-intensive tasks to an appliance, organizations improve WAN performance and handle increased traffic without the expense of licensing additional processors (Figure 6).w

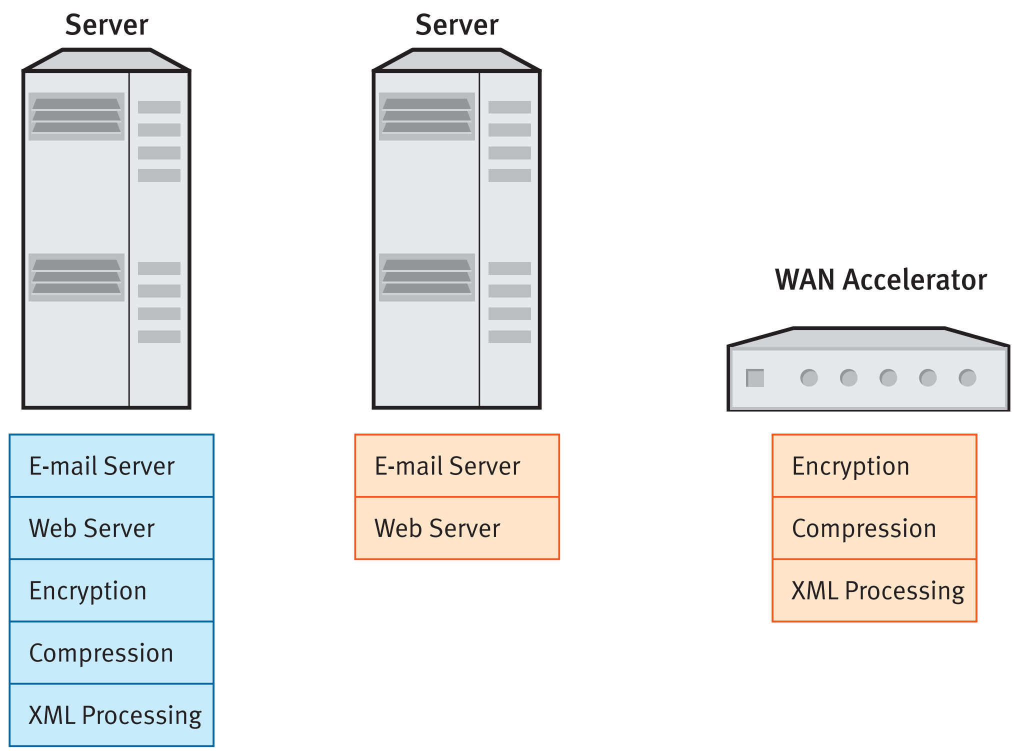


Figure 6

Forward Error Correction (FEC)

Some WAN acceleration solutions use FEC to reduce the need for retransmissions on lossy networks. FEC sends recovery information in-band with data for indexing to allow for reconstruction of lost packets.

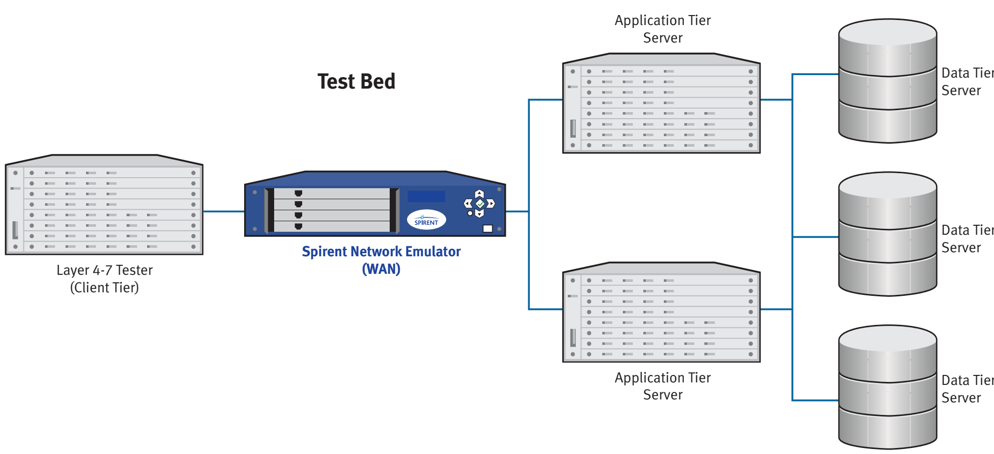
Evaluating WAN Acceleration Solutions

During the early stages of acquiring a WAN acceleration solution, efforts are focused on identifying products with the desired features and capabilities, comparing costs and other off-line methods of reducing the wide array of options to a manageable short list. Between the short list and the purchase is a critical step – performance evaluation in a test lab.

All the various WAN acceleration techniques have one thing in common: They improve the performance of systems running across a WAN. Consequently, testing the solutions under real-world WAN conditions is the only true method of validating the claims of the various vendors. While WAN acceleration solutions usually promise a significant ROI, they also pose a significant CapEx investment. An objective evaluation reduces dependence on a vendor's claim for ROI and increases confidence in meeting ROI goals.

Unfortunately, live networks (or production networks) even if available are not suitable for testing. Conditions at any given time cannot be controlled or, in some cases, cannot even be known. This reality poses significant problems in evaluating or benchmarking one WAN Acceleration solution vs. another when the same scenario cannot be repeated between tests.

However, designing a test bed to replicate the "real-world" is possible. It provides an environment that produces meaningful and repeatable tests and results. The test bed includes the system under test, a traffic generator/analyzer and a network emulator. The test-bed emulates the behavior of the production network – including delay, jitter, packet loss, corruption, etc. – through user, traffic and network profiles.



A test methodology is a definitive procedure to identify, measure, and evaluate a product, system, or service that produces a test result. Test solution vendors can provide a set of methodologies targeted at testing specific capabilities. The configuration of the user, traffic and network profiles will vary based on the capabilities under test. For example, TCP optimization is evaluated using a network profile that emulates a range of realistic settings for latency and packet loss. FEC is evaluated using a network profile that emulates a range of realistic bit error rates.

Following the methodology provides a test result that includes a set of observed metrics. The results are compared to target metrics. The specific metrics of interest vary based on the capabilities under test. For example, when testing TCP optimization, data throughput rates are the primary metric of interest. If the observed throughput doesn't meet or exceed the target throughput, other metrics such as the number of concurrent sessions or the number of timeouts correlated to latency and packet loss are reviewed during troubleshooting.

When the IT department has hard data in the form of test results obtained under known and repeatable conditions, a purchasing decision is based on recorded metrics rather than vendor promises. The performance of competing solutions can be directly evaluated in an apples-to-apples comparison. Consequently, the risk of unpleasant surprises after purchase and deployment is reduced, and the confidence that ROI targets will be met significantly increases.

Configuring Real-World Network Emulation Profiles

There are two ways to configure realistic network emulation profiles.

Profiler/Playback

To replay actual network conditions on a production network, use Profiler/Playback to capture the end-to-end packet loss, delay and jitter for each profile. Then play back the conditions dynamically during testing so the emulator replicates actual production network conditions on a packet-per-packet basis.

Standards-based Models

IP networks don't present deterministic, periodic disruptions to traffic. Instead, impairments vary over time, presenting problems in bursts resulting from such issues as route flaps, queue discards and buffer overruns. The Network Model for Evaluating Multimedia Transmission Performance over Internet Protocol, adopted by TIA as TIA-921 and by ITU-T as Recommendation G.1050, is a time-varying model that emulates the dynamic nature of impairments in an IP network.

This model has been adopted by several standards organizations for testing real-time applications and protocols. It is a statistical model based on actual network information obtained from anonymous service providers. It uses an impulse-driven time series model to emulate impairments introduced by each leg of an end-to-end network. The dynamic nature of the emulated conditions reflects the time-varying conditions found on actual production networks.

Deploying WAN Acceleration Solutions

The value of real-world testing extends beyond evaluation and purchase. WAN accelerators are integrated solutions that implement multiple configurable technologies to improve WAN throughput and performance. To achieve performance goals, network engineers configure the parameters to address issues specific to a production network. Refining the settings is often an iterative process. The devices must be configured under production network conditions. However, due to the risk of adverse effects from configuration errors, establishing the optimum configuration should not be done on the production network.

The test lab's controlled environment provides the ability to precisely and accurately emulate the environment where the solution will be deployed, but without risk to productivity on the production network. In addition, the test lab provides repeatable emulation of network conditions, making it possible to experiment with multiple configurations to identify the optimum settings for a specific network. The result is an optimum configuration that increases the likelihood of meeting ROI goals.

Network Emulation Enables WAN Acceleration

WAN acceleration devices address the issue of maintaining usability and productivity while expanding the corporate network into the global workplace. IT departments looking to purchase and deploy WAN acceleration solutions face a complex and confusing array of technologies, vendors and performance claims. Once a solution is selected, the challenge of determining the optimum configuration can be equally daunting.

By creating a test bed of traffic generator/analyzers, network emulators, and production applications, testing under "real-world" testing can take place – providing the benefits of testing on a production network without the inherent risks. These components, combined with a test methodology that incorporates realistic user behavior, traffic mixes and network conditions produce a controlled, precise and repeatable environment that makes meaningful test results possible. This enables proper selection of WAN Acceleration solutions to meet your needs.

The use of Network Emulators in your test-bed provides a cost-effective and efficient method of identifying the true performance of a solution. In addition, real-world testing assists in determining the proper configuration of a solution for optimum performance. Real-world testing increases confidence in meeting performance and ROI targets. It also reduces the likelihood of buying an ineffective solution while minimizing risk from performance and productivity issues due to configuration errors.

Inspired Innovation