


Realtime
publishers

"Leading the Conversation"

The Shortcut Guide[™] To



Optimized
WAN Application
Delivery

sponsored by

Blue  **Coat**[®]

Ed Tittel

Introduction to Realtimepublishers

by Don Jones, Series Editor

For several years, now, Realtime has produced dozens and dozens of high-quality books that just happen to be delivered in electronic format—at no cost to you, the reader. We’ve made this unique publishing model work through the generous support and cooperation of our sponsors, who agree to bear each book’s production expenses for the benefit of our readers.

Although we’ve always offered our publications to you for free, don’t think for a moment that quality is anything less than our top priority. My job is to make sure that our books are as good as—and in most cases better than—any printed book that would cost you \$40 or more. Our electronic publishing model offers several advantages over printed books: You receive chapters literally as fast as our authors produce them (hence the “realtime” aspect of our model), and we can update chapters to reflect the latest changes in technology.

I want to point out that our books are by no means paid advertisements or white papers. We’re an independent publishing company, and an important aspect of my job is to make sure that our authors are free to voice their expertise and opinions without reservation or restriction. We maintain complete editorial control of our publications, and I’m proud that we’ve produced so many quality books over the past years.

I want to extend an invitation to visit us at <http://nexus.realtimepublishers.com>, especially if you’ve received this publication from a friend or colleague. We have a wide variety of additional books on a range of topics, and you’re sure to find something that’s of interest to you—and it won’t cost you a thing. We hope you’ll continue to come to Realtime for your educational needs far into the future.

Until then, enjoy.

Don Jones

Introduction to Realtimerepublishers..... i

Chapter 1: Networking Issues and Complexity Increases as Technology Improves and Speeds Up1
In Terms of Network Performance2

The Intent of this Guide4

Original Networking Foundations4

 Over Time, Network Complexity Increases in Several Dimensions7

 WAN Technologies Emerge8

 Special Challenges for Application Measurement, Monitoring, and Optimization.....10

 Internal to Internal Access12

 Internal to External Access13

 External to Internal Access14

 Ways to Meet Challenges in Application Delivery15

 Measuring Complex Data and User Transactions17

 It’s Not a Challenge; It’s a Promotion!.....18

 Progressing Up the Value Chain.....20

Summary20

Copyright Statement

© 2008 Realtimepublishers.com, Inc. All rights reserved. This site contains materials that have been created, developed, or commissioned by, and published with the permission of, Realtimepublishers.com, Inc. (the "Materials") and this site and any such Materials are protected by international copyright and trademark laws.

THE MATERIALS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT. The Materials are subject to change without notice and do not represent a commitment on the part of Realtimepublishers.com, Inc or its web site sponsors. In no event shall Realtimepublishers.com, Inc. or its web site sponsors be held liable for technical or editorial errors or omissions contained in the Materials, including without limitation, for any direct, indirect, incidental, special, exemplary or consequential damages whatsoever resulting from the use of any information contained in the Materials.

The Materials (including but not limited to the text, images, audio, and/or video) may not be copied, reproduced, republished, uploaded, posted, transmitted, or distributed in any way, in whole or in part, except that one copy may be downloaded for your personal, non-commercial use on a single computer. In connection with such use, you may not modify or obscure any copyright or other proprietary notice.

The Materials may contain trademarks, services marks and logos that are the property of third parties. You are not permitted to use these trademarks, services marks or logos without prior written consent of such third parties.

Realtimepublishers.com and the Realtimepublishers logo are registered in the US Patent & Trademark Office. All other product or service names are the property of their respective owners.

If you have any questions about these terms, or if you would like information about licensing materials from Realtimepublishers.com, please contact us via e-mail at info@realtimepublishers.com.

[**Editor's Note:** This eBook was downloaded from Realtime Nexus—The Digital Library. All leading technology guides from Realtimepublishers can be found at <http://nexus.realtimepublishers.com>.]

Chapter 1: Networking Issues and Complexity Increases as Technology Improves and Speeds Up

Today's business environments have scaled to new heights and reached new plateaus. An increasingly globalized economy continues to transform modern enterprises to adopt more adaptive networking and business service management (BSM) models. At the same time, an increasingly mobile workforce demands remote connectivity for its people, processes, and resources. As organizations and employees become ever more distributed throughout branch offices and off-site locations, boosting productivity grows increasingly important in all kinds of interesting—and unobvious—ways. All users require efficient, secure, and unimpeded access to critical business applications wherever and whenever they create static or temporary workspaces: company headquarters, branch offices, and even off-site client locations.

Several distinct and high-impact business requirements and technology trends compel the need for organizations to further improve performance, while increasing security and managing risk in their business applications. To remain competitive within increasingly challenging markets, IT must streamline the business by ensuring superior application responsiveness and delivering an agile infrastructure without undue (or any) increases to operational costs or headcount. Then too, a burgeoning set of state and federal regulatory requirements and an increase in underlying threats and vulnerabilities continues to raise the bar when it comes to managing and accepting risk. At the same time, the intersection of consolidation, outsourcing, and mobility pushes applications and end users ever further apart. This creates extraordinary pressures on both network technologies and network traffic patterns.

Centralized applications and widely distributed users can impose huge performance penalties on modern business environments, where remotely connected users suffer most from network latency and bandwidth constraints. Although ongoing trends toward consolidating or outsourcing critical applications and servers to fewer locations may simplify administration tasks, consolidation and virtualization can also pose performance problems to a distributed user base. Such problems are often particularly, if not spectacularly, evident whenever business-critical applications and resources are scattered among a variety of remote global locations and across numerous servers. When combating these difficulties, many organizations turn to solutions to accelerate and secure the delivery of business applications for all users across a distributed enterprise—including those near Internet gateways, located within branch offices or data centers, and even at individual end-points.

In Terms of Network Performance

Network performance can be monitored and measured and is typically defined using any number of metrics and terminology such as link/line speeds, throughput and round-trip times, bandwidth utilization, and inherent delay. This last element captures network latency, or the minimum time required to move data between endpoints on a network—and is a crucial determining factor when it comes to calculating most performance penalties. Latency is an important component for network performance, and describes delays that are inherent in any connected computing environment. Whenever network latency exists (particularly when it's lumped into the common concept of user-visible latency to describe the total delay from end-user click to endpoint action), it can never be completely eliminated, only mitigated or reduced (and often, only for special, preferred classes of services and applications). Latency also represents an ugly truth and a difficult proposition, particularly for network designers as they struggle to meet service level requirements in a fast-moving, quickly expanding business environment that typically encompasses a mixture of unlike networking components, protocols, and technologies.

There are many types of latency—computer processing speed, distances traversed across global networks as signals propagate from senders to receivers, delays associated with round-trip times and communications frequency, and delays associated with device queues. Several types of latency discussed later may be addressed only by upgrading slower components for faster ones but can never be completely eliminated. Those that remain are an ongoing and constant challenge for network designers and end users alike.



We often refer to the speed of light—the absolute maximum transmission speed—as being a fundamental latency. For example, despite a “perfect” network link operating under ideal conditions between the United States and India, there is a 60ms latency just to account for the speed of light across the total distance and media traversed from “here” to “there.” This becomes glaringly obvious in satellite communications, where the round trip to a single satellite typically adds a half-second to latency, and where multiple satellite hops can add as much as 2 to 3 seconds to overall delays from sender to receiver (and back again).

Latency is usually understood in the following terms:

- Computational latency refers to the amount of time it takes to process a given workload, and depends upon the speed and capacity of the hardware in use at each step in the communications path between sender and receiver. Using “faster” hardware is generally the only way to reduce this type of latency.
- Signal propagation delay refers to the amount of time it takes for a signal to travel from one end of a connection to another, and depends on the length of the link and the number of bits that can be transmitted across the link at any one time. Although many types of network communication may be multiplexed, that is, may involve transmission of multiple channels of data at the same time, there is still a limit to how fast signals can move across any networking medium. There is little anyone can do to reduce this type of latency. Signal propagation delay is most noticeable when satellite communication links enter into consideration: given that the orbit of geosynchronous satellites is more than 20,000 miles above the earth, even a single up-and-down satellite link can introduce a half-second of delay between sender and receiver (this can double when satellite-to-satellite relays are also required).

- Serialization delay refers to the amount of time it takes to convert an n-bit wide signal into a corresponding series of individual bit values for transmission across a network medium. Thus, for example, if data shows up in 8-bit bytes at an interface, serialization involves stripping all bits from each byte in some specific order, then emitting each one in that order onto the network medium on the sending side (and works in reverse on the receiving end). Except for increasing the signal clock used to synchronize transmissions from sender to receiver (which again involves a hardware upgrade), serialization delay remains a constant source of latency in networked communications.
- Queue delay refers to how long a message element must “stand in line” to wait its turn for media access, and generally applies when transmissions must traverse a router or a switch of some kind. Here again, this is a case where latency depends on the type of hardware used to store and forward network transmissions, as well its queuing capacity. When link oversubscription occurs, in fact, sufficient congestion can occur to make users think that they have “run out of bandwidth.”




Generically, *latency* is a measure for the time that any one part of a system or network spends waiting for another portion to catch up or respond to communication activity. Latency describes any appreciable delay or the time that elapses between stimulus and response. Such delays occur virtually throughout all operational aspects of any given computing environment but not all forms of latency are perceptible in human terms. Once introduced into any computing environment—within a system or network—the cause itself must be removed, mitigated, or reduced to improve performance.

Latency is measured in a number of ways, include one-way transit time from sender to receiver as well as round-trip time (often the most useful measure of latency because a complete transaction from sender to receiver invariably involves transmission of a request of some kind from sender to receiver, followed by delivery of a response to the request back from the receiver to the sender). Round-trip latency also offers the advantage that it can be measured at any single point on a network. On a complex, far-flung network, in fact, variations in round-trip latency between minimum to maximum values may be more interesting from a quality control standpoint than average round-trip latency, because those users subject to maximum round-trip latency will be those whose experience of a system’s responsiveness and efficiency is worst.

But how do you accommodate latency incurred by the needs of users who are accessing resources cross-country or across the globe? Or to further compound the problem, how can you accommodate protocols that may require multiple round-trips to satisfy a single service request? What if a given protocol format specification doesn’t directly provide any means for protocol acceleration or traffic prioritization?

Consequently, network designers and implementers have had to consider performance and latency from a radically different perspective, as the networking landscape has shifted to include more services and applications, each with its own unique operational parameters and specific properties. Mobile workers, remote offices, and distant partnerships are an important aspect of this brave new networking world, and demands acceptable performance for a diverse set of applications, platforms, and users. And when the end-user experience suffers or degrades, network designers must shoulder the inevitable blame that follows in its wake. For remote users and remote applications, the Internet is your WAN, therefore Internet latency reduction should be a key factor when choosing any WAN optimization solution.

 We'll use the term *end-user experience* throughout this guide. The end user serves as our barometer for the overall status and well being of any business network, as they drive business operations and experience the most severe penalties whenever performance lags or falters. Thus, the “end-user experience” encompasses all aspects of their interactions with a company, its services, and its products. To deliver a truly superior user experience requires seamless integration among multi-disciplinary platforms and services, and a holistic, end-to-end view of networks that sets and manages user expectations and carefully monitors and manages network behavior and performance.

The Intent of this Guide

The intent of this guide is to discuss, detail, and decipher the issues, properties, and solutions associated with optimal, timely delivery of applications and services to end users across the globe. This guide also seeks to translate business application, process, and end-user needs into usable, intelligible techniques to help you improve performance goals and meet business objectives. Part of this process involves due diligence and demands that we dig into subject matters relevant to business-critical platforms and protocols that provide the basis for understanding—and ultimately improving upon—network performance.

This guide to optimized Wide Area Network (WAN) application delivery examines the scope of modern computing enterprises and their increasing needs to meet ever-expanding demand, while enhancing both efficiency and effectiveness. Sharing business assets and applications has fostered a renewed and invigorated quest for operational efficiency and applicability among the large numbers of people and processes that networks must serve across dispersed and often disjointed regions and territories.

Original Networking Foundations

Given that the original scope for Local Area Network (LAN) communications was purely local, starting with handfuls of machines that needed to communicate at a single location (often in a single room), the original vision for applications interacting with the LAN was understandably indifferent to latency. Short distances between data collection and processing end-points were the norm, where a narrowly limited scope of operations defined the then-current networking environment.


Early designers of networking technology neither considered nor accounted for long-haul linkages so common in modern enterprise network topologies, which can frequently span multiple regional territories and geographically dispersed locations. Such early designs were once well-suited to low-throughput traffic characteristic of bygone days. But today's rich networking environment must accommodate increasing levels of high-bandwidth, resource-intensive protocols and payloads. This has outstripped the tenets of basic network design, particularly for wide-area communications, and now burdens infrastructure designers with managing and prioritizing loads to restore balance and order. At the same time, these designers must ensure that security measures keep up with these performance enhancements and that they remain in compliance with security policy and applicable mandates.

Traditional network protocols favor short, bursty communications and chatty information exchanges along short, transitory paths—and include no formal concept of traffic shaping or prioritization. This includes well-known application support protocols such as NetBIOS as well as various distributed file services—most notably Microsoft’s Common Internet File Services (CIFS). Because of the erratic nature of so-called “best effort” delivery mechanisms and limited consideration for careful consumption of WAN resources, it becomes extremely difficult to predict the demands of the network and how much operational capacity may be necessary at any given moment. When sufficient resources are unavailable, only buffering can help offset demand—but this is only a stopgap measure and not a true resolution. Ultimately, enough bursty traffic at any given moment produces severe contention for network resources and introduces difficulties for modern globally distributed network designs, no matter how they might have been deliberately over-provisioned when initially specified and introduced.

Existing network services and applications operate primarily in terms of simple data exchanges and generally short message lengths and durations—such as HTTP. HTTP is notoriously chatty and exchanges numerous small bits of data (text files, graphics, style sheets, and so forth) to accommodate the client-server request-response cycle. Consequently, any good WAN optimization strategy seeks to address this issue, often by batching multiple requests into a single transmission, and doing likewise for all the responses produced to answer those requests. Lots of other client-server applications likewise employ protocols that utilize “chatty” response mechanisms and produce large amounts of traffic on a per-request basis. This works perfectly well on a local network in most cases, but remote Web-based applications and high-latency protocols typically suffer from performance degradation when employed across long-haul networks, particularly when large numbers of such traffic streams must share the same WAN links. Typically, these applications and services utilize parameters within the Transmission Control Protocol/Internet Protocol (TCP/IP) framework for session initiation, management, and tear-down.


Then, too, it is not uncommon for enterprises to recognize that as the level of WAN traffic increases, it becomes ever more necessary to regulate which protocols and applications may access the WAN and to what degree. Detailed inspection of protocol distributions for such traffic may, for example, reveal the presence of unauthorized and unwanted peer-to-peer (P2P) protocols such as BitTorrent, FreeNet, or KaZaA, which typically do not play an official role on enterprise networks and can be blocked at the gateway without affecting critical services or applications.

However, many of the protocols for important services and applications built atop TCP/IP lack native traffic prioritization schemes (or fail to exploit any such properties that TCP/IP may offer to developers) to alleviate some of the traffic burden so typical of streaming protocols and short-duration bursts of activity. This leaves the network medium exposed to saturation issues because both short- and long-term protocol sessions coexist in the same resource space with no real differentiation between importance and potential.

 TCP/IP is the protocol framework that defines the most prominent types of network interaction but is by no means the only format. Many Internet-based transactions occur via TCP/IP with some foreign or little-used protocols encapsulated as payloads. Our focus throughout this guide is primarily geared toward higher prioritization and enhanced performance in the existing TCP/IP framework. It's crucial to understand the operational parameters and properties of TCP/IP to properly design, implement, and utilize performance-enhancing programs, platforms, and procedures. One of the very best books to help guide you into this subject is by Geoff Hughes: *Internet Performance Survival Guide* (Wiley Computer Publishing, 2000, ISBN: 0471378089); despite its publication date, it offers the best in-depth analysis of WAN application and protocol behaviors we know of in print.

Efficiency in throughput hits a downward spiral as more applications, services, and end users share and increasingly occupy the same medium. Additional in-line network appliances and routing devices only increase the congestion burden because inherent performance issues are not directly addressed but compounded instead. And as the distance between end users and applications also increases, some network designers optimistically assume they can create a “one-size-fits-all” network solution for most scenarios, which is entirely incorrect when it comes to serious WAN optimization, where an understanding of the various factors that come into play is needed, and where different situations dictate different optimization approaches.

Research firm Gartner uses the terminology of application and vendor silos to explain that networking professionals are responsible for delivering more than just the bits and bytes involved in networked communications, and must also be sensitive to the quality of the end-user experience. Typically, an application or service is viewed as a standalone silo, which implies a challenging and severely limited reach or range of capability. The goal then becomes to find a common language, both in the real-world and world of technology, so that network architects and application architects can exchange information about overall performance and behavior. By way of explanation, an information silo is any management system incapable of reciprocal interaction with other related management systems. This, too, directly impacts and influences the end-user experience because it means that the systems used to monitor and assess that experience cannot provide a single, coherent, end-to-end view of that experience. This works against developing a broader understanding of network conditions and behavior, and often has to be forcibly offset by creating mechanisms to deliver measurements and analysis of the real end-user experience and of end-to-end activity and response times. In many cases, enterprises find themselves forced to add a set of probes or agents to deliberately create (or simulate) end-user activity, just so it can be measured and monitored. WAN optimization solutions can also offer such information because measuring and managing response time is such a key component in making such technology function properly.

 A system is considered a *silo* if it cannot exchange information with other related systems within its own organization, or with the management systems of its customers, vendors, or business partners. The term *silo* is a pejorative expression used to describe the absence of operational reciprocity.

The original networking model plays well for LANs where latency does not typically create a significant role in communication delays. Local transmissions often happen without much perceptual lag and usually have no substantial impact on throughput, where distances are kept at a minimum and technologies may operate at or near maximum speeds. Scale this into much larger, more modern networking environments and these seemingly insignificant factors soon impose significant hindrances on site-to-site communications.

This institutionalized chaos also puts forward a strong case for introducing quality or class of service mechanisms into enterprise networks. These mechanisms relegate protocols, applications, and services into specific well-defined categories and offer priority access to WAN bandwidth for mission-critical or time-sensitive applications according to such category assignments. However, a quality of service scheme can downgrade less time-sensitive data transfers so that they fit themselves in around such higher-priority traffic. It is important to recognize when using these methods that overall bandwidth and throughput for the entire network usually degrades slightly because of the processing overhead involved in classifying and organizing traffic by class or type of service. But indeed the end-user experience for important or time-sensitive services and applications should improve: why else would one abandon best-effort delivery in favor of priority mechanisms? At the same time, performance for less time-sensitive, lower-priority traffic will actually degrade, but if the right protocols, services, and applications are relegated to this category, the overall end-user experience should not change for them, or be too noticeable anyway.

Over Time, Network Complexity Increases in Several Dimensions

A complete perspective of the modern networking environment requires comprehensive knowledge of the elements and links and the behavioral properties involved in its operation. Dynamic business needs and an ever-increasing set of applications and services in turn ratchet up demand and put greater stress on existing networking capabilities, techniques, and technologies. Service agreements and a focus on the end-user experience also propel the growing need to enhance application performance, especially for remote and roaming employees, or for Internet-connected customers or partners.

Executive management almost universally applies pressure to accelerate those business processes that drive the bottom line. This includes product development, inter-regional team collaboration, supply chain management, virtual presence and services, and national or international sales. These business processes can never be too efficient nor their people too productive: their surrounding climate is one of constant, incremental improvement and change.

Consequently, the underlying network infrastructure must also accelerate to accommodate the pace of business in a transparent and performance-enhancing manner. Network design takes into account at least three primary disciplines: cost, performance and reliability. All three properties influence and are influenced by network scope, making individual roles and contributions entirely different for different networks and situations.


To put things into proper perspective, consider that the LAN is no longer the staple medium for communications within large scale enterprises. Large (and sometimes unwieldy) WANs and mobile networking become the most prominent means for linking sites or users together, and for forming partnerships among companies. Contention management for LAN and WAN, which occurs when multiple prospective consumers of such resources vie for their use at the same time, becomes geometrically complex with advances in throughput and processing capabilities within the underlying networking technologies. To further aggravate this situation, the ebb and flow of network traffic is best characterized in the form of asynchronous bursts of information that occur at irregular intervals. This causes randomness, unpredictability and confusion. Couple this with predictable, cyclic demand for backups, end-of-period reports and accounting, and other recurring network events, and demand and utilization can evolve from modest to major in a matter of seconds.

There is another, arguably more important, issue around iterative improvements to application delivery: going *up* the protocol stack. Any bit-pushing aspects of the network infrastructure have seen endless improvement at Layers 1 through 3, and even some at Layer 4 for UDP and TCP. Thus, though there is little left to optimize below Layer 4, there is ample opportunity to resolve issues at the application layer. The reason this hasn't already happened is because of the complex nature of managing application layer issues at line speeds. However, this is crucial in situations where performance remains a constant problem because there are no further iterative improvements available for optimization at Layers 4 and below.

The distance between points (A) and (B) now spans to include various regions, territories and continents. What was once an easily manageable network environment by a few on-site personnel has expanded to long-haul linkages between distant end-points. Many of these connections involve the use of satellite communications, making the problem of network management increasingly more difficult, thanks to delays that can be as long as several seconds as soon as one or more geosynchronous satellites enter into the latency equation.

WAN Technologies Emerge

WAN is quickly becoming a staple element of many modern business computing environments where it doesn't already exist. LAN distances are far too restrictive and limiting in scope, scale and capacity to support the burgeoning push of an increasingly globalized market space. Plus, they're inherently local in nature. The LAN environment has simply become the underlying infrastructure to meet local demand that must couple up to globalized resources to achieve a common and individually unattainable goals or business objectives.

 *Wide Area Network* is any computer-based communications network that covers a broad range, such as links across major metropolitan, regional, or national territorial boundaries. Informally, a WAN is a network that uses routers and public Internet links (or in some cases, private and expensive leased lines). WANs are the new-age bailing wire that ties and interconnects separate LANs together, so that users and computers in one location can communicate with those in another location, and so that all can share in the use of common resources, services, and applications.

Services and applications include complex transactions that occur among many multi-tiered applications, and employ multiple server and service hierarchies. "One size fits all" approaches and "end-all be-all/mother-of-all" methodologies simply cannot apply. Network designers are forced to re-examine their approaches to design and implementation best practices and reconsider what they know works best, as a shifting business landscape dictates new and different operational parameters and workplace requirements to meet changing or increasing business needs. "Add more bandwidth" is neither practical nor the panacea that it once was, because not all forms of delay are amenable to cure by increasing WAN throughput.

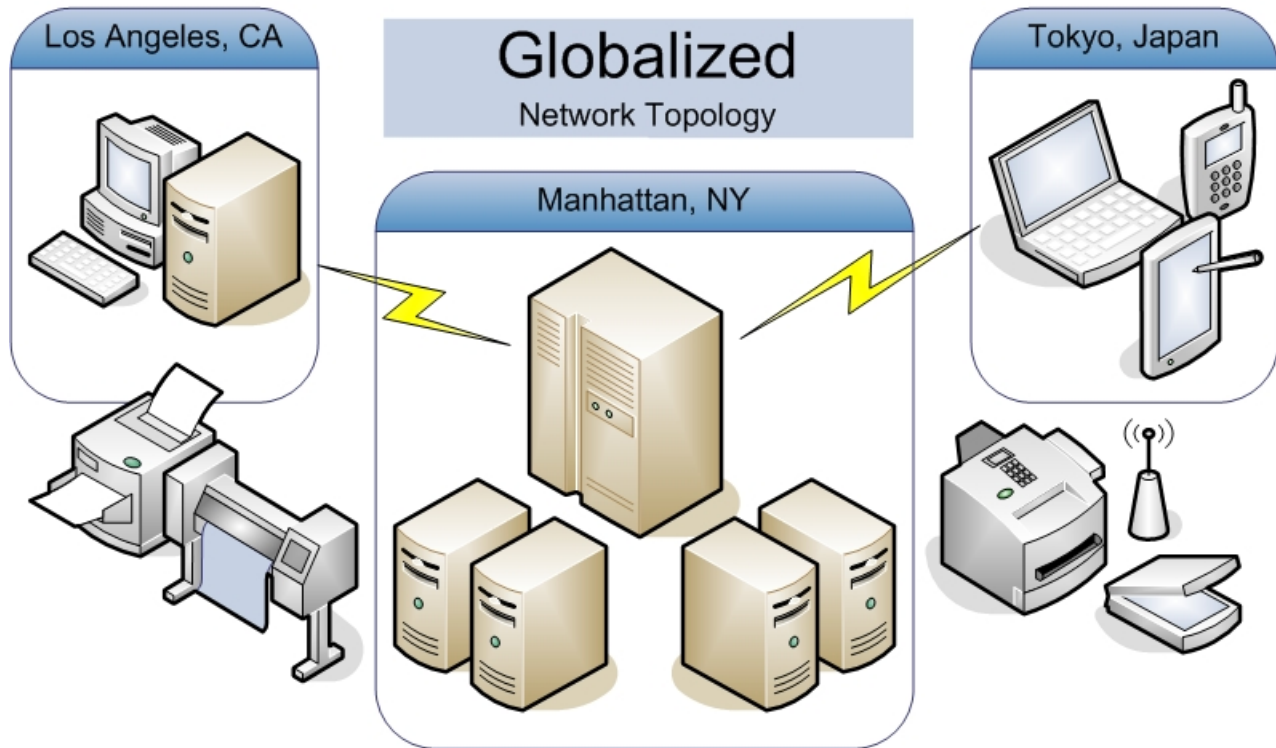


Figure 1.1: Typical enterprise architectures involve long-haul links between data and operations centers. Here, New York provides the servers and the computing facilities, while LA and Tokyo drive business activity.

As Figure 1.1 is meant to illustrate, many typical enterprise network architectures connect distant users (in Tokyo and Los Angeles, in this case) with data centers and other centralized resources (in NYC as depicted). Without careful attention to how the WAN links between the user centers and the datacenter get used, enterprises can easily find themselves woefully short on bandwidth, and dolefully long on response time. The end user experience can't help but suffer in such situations, without careful forethought, intelligent design, and judicious use of WAN optimization tools and technologies.

Multiple carriers and operators control various portions of the intervening infrastructure between sites (and also between computing resources and mobile users as well), which itself introduces trust and control issues. The diverse nature of hardware and software operating platforms and the introduction of dispersed and diverse employee bases—and all related permissions, parameters and properties—creates a complex management nightmare. It's all too easy to inherit problems associated with other platforms through newly-formed partnerships, or via mergers and acquisitions. As operational wants, needs and concerns are addressed for some given platform in a particular environment, they may still prove insufficient when applied to the much larger context in a business computing environment that encompasses many platforms, technologies, and computing strategies.

Anyone can control operational behavior in site-specific and localized contexts, but no single entity or organization can expect to completely control behavior from end-to-end. The management nightmare only worsens when public Internets become involved versus leased-line, site-to-site connections. As the old saying goes, “Jack of all trades is master of none,” and this holds truest in a global network context. Personnel may adapt and acclimate, and even become adept at handling site-to-site procedures for a specific set of operational conditions and criteria. Users and network staff may even get used to poor performance because that is the only kind they’ve ever experienced for certain applications. But when you introduce a global computing context for business networking, such expectations and specific adaptive behaviors developed in response to particular conditions will soon go by the wayside.

In essence this phenomenon explains why the one size/method fits all approach falls so drastically short of meeting business goals. Local and specific solutions targeting only packet delivery issues cannot and will never cover WAN application and management needs, and instead create only dysfunctional infrastructures that require remediation or redesign.

Special Challenges for Application Measurement, Monitoring, and Optimization



Traditional measurements of link latency and utilization are mismatched with perceived end-user service or application experiences and expectations. Perhaps more precisely, people often don’t realize that WAN latency is a primary determinant for the end-user performance experience. However, it remains only one component in latency end-to-end, and serves as a single metric building block for a much larger performance matrix.

The imposition of Service Oriented Architectures (SOAs) and related service delivery contracts increases requirements for managing response time, setting end-user expectations, and managing agreed-upon availability, throughput, and so forth. With SOA, systems are architected for life cycle management of business processes through specific IT infrastructure provisions and documented through end-user or provider expectations as outlined and defined in related service agreements or requirements. Basically, this contractual obligation ties infrastructure performance to end-user or provider expectations with regard to delivery of services and business application interactions. Many service agreements relate directly to time-sensitive, bandwidth-hungry applications such as Voice over IP (VoIP), streaming or on-demand video, and Citrix/Remote Desktop Protocol (RDP).

It’s also the case that SOAs, and the applications they deliver, can blur the “internal-external” distinction that profoundly affects WAN use and its optimization potential. That’s because even though an SOA application may reside inside organizational boundaries, at least some of the data it needs and uses may come from outside those boundaries, as in public or private Internet-hosted sources. This poses an interesting challenge for IT in that while the SOA application will undoubtedly be trusted, the data it presents to the user uses may include unwanted or unauthorized content simply because the data source is unwilling or unable to deliver clean, malware-free information. This puts the onus on the consuming organization to inspect and clean incoming data before passing it on to end users: a task for which WAN optimization devices are uniquely well-suited.

Interestingly, it's also sometimes the case that SOA-based applications load significant amounts of data that particular users may not need at any given moment. The nature of Web delivery is such that a page typically won't finish loading until all its data is available (transferred to the client), which can impose increasingly onerous delays when that data is unnecessary or irrelevant to the task at hand. This is another situation where the caching that WAN optimization devices provide can reduce delays, because as long as a local cached copy is current, it can be supplied at LAN speeds to users rather than forcing them to wait for transfer of that information across the WAN.

The notion of service contracts within an SOA context is similar to though distinctly different from the kinds of service level agreements, or SLAs, with which network professionals are already familiar—namely, packet delivery (latency, bandwidth, and loss). SOA service contracts involve application uptime, data coherency, and schema compliance in addition to response time. Network SLAs, in contrast, deal with link uptimes, packet losses, and average sustained throughput levels. An application that compiles data from multiple sources throughout a data center and across a distributed enterprise or global network, may fail to meet higher-level business service requirements. This occurs when a network imposes delays in getting access to the distributed data that SOA is supposed to compile and present in new and silo-busting ways. These kinds of applications must, however, be understood and prioritized within the overall application and services context, lest they rob precious bandwidth needed to process orders, synchronize databases, ferry payroll and payment information, and all the other critical business transactions that go straight to the bottom line (or not, as the case may be).

-  A Service Oriented Architecture (SOA) is a computer systems architectural style for creating and using business processes, packaged as services, throughout their life cycle. SOA also defines and provisions the IT infrastructure to allow different applications to exchange data and participate in business processes.
-  A Service Level Agreement (SLA) is part of a service contract where the level of service is formally defined, which is formally negotiated between two participating parties. This contract exists between customers and service provider, or between separate service providers, and documents common understanding about services, priorities, responsibilities and guarantees (collectively the *level of service*).

Related access issues can be logically divided into three separate elements: internal users accessing distant internal applications (internal to internal); internal users accessing distant external applications (internal to external); and external users accessing distant internal applications (external to internal). We omit coverage of the final case: external users accessing distant external applications, because those types of communications are not normally addressed with WAN optimization and are outside the scope of this guide. Here's a diagram for what we've just described, after which we take an individualized look into each of these different operational perspectives and see how they apply and affect business processes related to WAN optimization.



Figure 1.2: Different ways to involve the WAN between users and applications.

Internal to Internal Access

Internal users accessing internal or intranet resources stress demands on localized networking contexts—namely, the LAN. This is where the greatest performance and throughput is attained. In networking terms, latency is more easily controlled through implementing the fastest routers, utilizing the highest wire-rated cables, and organizing the most efficient network topology. When internal users in one place access internal applications in another place, they are also able to exploit the benefits of WAN optimization because a single company or organization definitely controls both ends of the WAN link and, when leased-lines are used, may even be said to “control” the link as well. This is the optimum situation for WAN optimization because it permits the owner to employ compatible WAN optimization equipment on both ends of such a connection (or all ends of such connections, where applicable) without involving any third parties.

Internal users may connect with internal applications and services such as an intranet Web site or Common Internet File System (CIFS), which provides shared access to files, printers and miscellaneous connection points between nodes on the network. Typically, there isn’t a substantial delay in satisfying local requests for local resources under this usage scenario. But CIFS is indeed a chatty protocol, and does not behave at all well if that file system is projected across multiple sites (and perforce also across multiple wide area links).

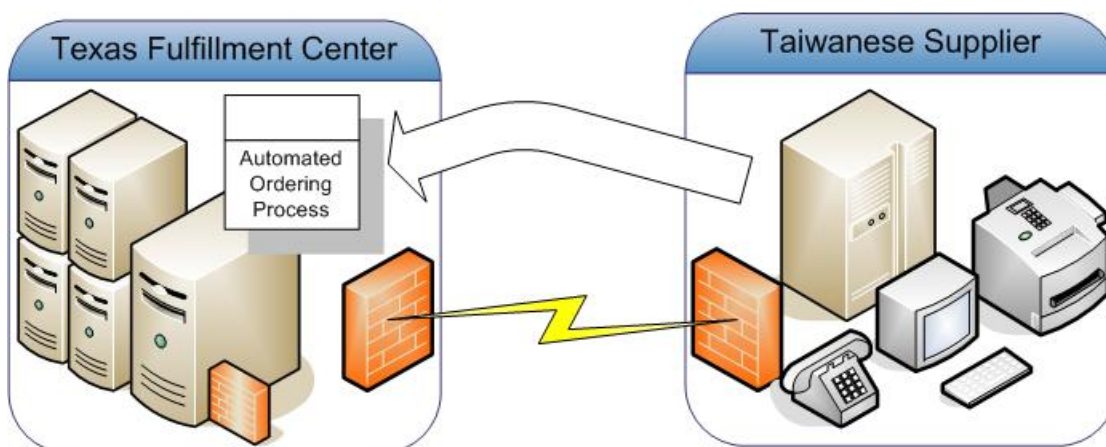


Figure 1.3: When an outside supplier has parts ready for delivery to the fulfillment center, it interacts with an automated ordering process to inform the center of pending delivery.

In general, scaling applications and services to make most effective use of WAN links means choosing protocols and services that behave more reasonably when using such links where possible. Alternatively, it means using technologies that act as local proxies for chatty protocols and services, while implementing more efficient, less chatty replacement protocols and services across wide area links in the background.

When making WAN-friendly protocol or service choices isn't possible, it becomes necessary to use local proxies to accommodate chatty, bursty services and applications. Then organizations can repackage and reformat WAN communications to behave more responsibly, to communicate less frequently, and to make best possible use of bandwidth when data must actually traverse a WAN link. This is also a case where shared cache data (identical information elements maintained in mirrored sets of storage at both ends of a WAN link) can speed communications significantly, because pairs of devices with such caches can exchange cache references (which may require only hundreds of bytes of data to be exchanged) rather than shuttling the actual data between sender and receiver (which may require exchanges at megabyte to gigabyte ranges).

Most organizations consider private WANs to be internal-to-internal application delivery issues, but it remains a subject of debate as to whether this view is strictly true or false for semi-public MPLS. It's definitely true for point-to-point links such as frame relay. Internal-to-internal acceleration is the basis for the WAN optimization market as it currently stands, whereas today's CIFS/MAPI issues will become the intranet/portal issue of tomorrow.

Internal to External Access

Internal users accessing external resources are where the rubber really meets the road. Utilizing high-speed local hardware to access external linkages and resources emphasizes the lowest common denominator. That is, your network is only as fast as its slowest link—in this case, the vast performance differential and bottlenecking that occurs when hitting a shared network medium (the Internet) rife with contention and collision (among other issues).

Latency between local and global networks is largely out of the hands and subsequent control of networking professionals. An inescapable fact is that it takes about 150ms or 1/8th of a second to transmit from New York city to Tokyo, Japan—on a good day with low utilization trends and high resource availability. That's nearly 100 times longer than the typical network latency on a LAN.

In practice, network latency turns out to be quite significant, because a badly designed application might require thousands or tens of thousands of seconds per interaction to process data across a WAN. Such an application might also make many round trips between sender and receiver using numerous small packets for each individual action or request/reply sequence. Each fraction of a second adds up and becomes an increasingly noticeable part of the end-user experience when one or more WAN links enters the picture. What's perfectly fine on local, low latency networks becomes impossible and impractical at WAN latencies. Although this effect is additive, with enough WAN activity involved in a long series of individual data exchanges, it soon takes on multiplicative if not exponential delay characteristics. That explains why LAN applications that also display these characteristics are never, or only very seldom, used in an Internet context.



These same observations hold true for private networks. There is about 120ms delay between internal-to-internal configurations where long distance transmission is involved.

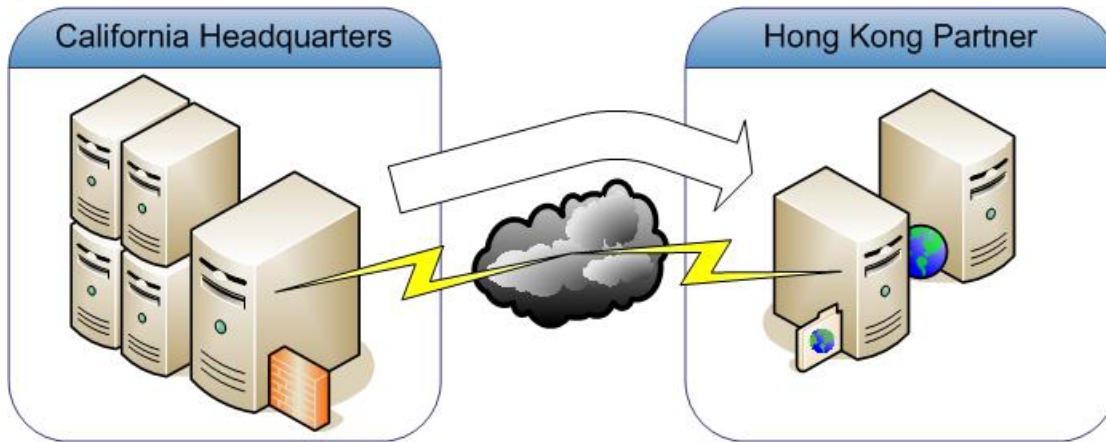


Figure 1.4: When the California HQ operation needs to exchange design information with a Hong Kong partner, it used a shared network link to ferry that data across the Pacific

In practice, this means that applications and services should be designed to minimize back-and-forth communications, and to stuff as much data as possible into messages whenever they must move between sender and receiver. Thus, as shown in Figure 1.4, when the HQ operation needs to share design plans with its Hong Kong partner, the mechanisms employed to manage and ensure their delivery must work as quickly and efficiently as possible, so that individual file transfers proceed rapidly, and so that transmission errors or failures can neither abort nor severely slow down or damage key data files and information. Here again, this requires judicious selection and use of protocols and services optimized for WAN situations.

External to Internal Access

At some point in their workday, a mobile end user will want to establish an inbound connection to the organization to make use of internal resources. Ironically, by taking the branch-side WAN links out of the equation, remote access often improves worker productivity, especially when those workers are furnished with special-purpose client software to help them maximize the performance of their remote access links and use applications designed to perform as well as possible in a remote access situation. This even presents a case where “one-sided WAN optimization” (on the remote server/application end) delivers useful and measurable performance improvements as well.

Imagine instead that the same remotely-connected client is making requests via CIFS, which is often limited on the client end (by Windows) to 4KB reads per request. Over the LAN, this is often imperceptible; introduce WAN link delays and the effects can quickly turn into a harrowing experience. A client requesting a 20MB file (at 4KB chunks) over a WAN with a 300ms delay requires 5,000 reads for this single request. That client will wait approximately 25 minutes for completion of the request. Thus, when a roaming client in Tokyo accesses the corporate Web site in NYC as shown in Figure 1.5, such access should be carefully crafted to let them navigate around the site, send messages, and handle financial transactions without incurring substantial delays during any step in that sequence (or within any particular activity, either). That’s where careful and thoughtful application design and good remote access tools really pay off.

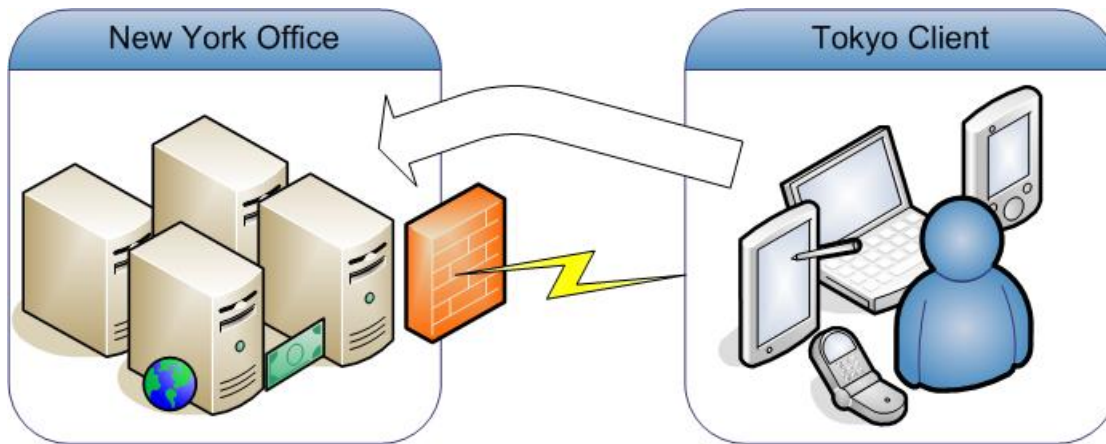


Figure 1.5: A client in Tokyo accesses the corporate Web site in the New York office to access messaging and financial services applications

Soon someone somewhere will notice that the existing state of affairs on this Internet-based WAN is insufficient and unsupportive in its operational limitations, capacity, and speed. The ability to provide a fast and safe connection to all users and applications efficiently and effectively, regardless of workstation and location will also prove problematic. This is particularly evident where no strategies are in place yet to integrate the Internet into the existing WAN topology beyond a basic VPN-based approach.

CIFS/Server Message Block (SMB), Real-Time Streaming Protocol (RTSP), VoIP, HTTPS, and various other protocols all present significant challenges for application monitoring, measurement and optimization. Additionally, Secure Socket Layer (SSL) and HTTPS acceleration is necessary to enhance speed and security, especially when traffic must traverse Internet and WAN links.

Ways to Meet Challenges in Application Delivery

A WAN optimization solution ultimately seeks to accelerate performance for distant branch, remote and roaming users, and to bring their end-user experiences into line with near-native LAN speeds. Policy-driven controls can be applied to better manage network resource utilization to keep users, applications, and data in line compliance with responsible usage policies and business priorities.

The benefits of WAN optimization include:

- Increased productivity levels driven by faster delivery of applications and data to remote users. In the simplest of terms, users who spend less time waiting for replies to their requests spend more time getting things done.
- Enable implementation of various IT mandates such as server consolidation, outsourcing, and SOA. WAN optimization lets data centers make more effective use of virtualization technologies, for example, so that a smaller number of (more centralized) servers can deliver a larger number of applications and services to end users, wherever they may be.
- Internet utilization versus costlier private WAN interconnects. WAN optimization delivers significant reductions in the bandwidth required to handle various levels of activity, so that public Internet links may be used (with access and equipment charges in the tens of thousands of dollars) rather than expensive leased lines (with access and equipment charges often in the hundreds of thousands). More importantly, the latency delta between private and VPN-over-public WAN (if any) is outweighed by strategies to mitigate risk associated with open Internet access.
- Cost reductions by delaying additional bandwidth purchases. Because WAN optimization makes more effective use of bandwidth already available, it can extend the usable life cycle for specific bandwidth allocations and keep costs the same. Though purchasing additional bandwidth can introduce economies of scale (unit costs) it seldom, if even, introduce outright savings (absolute costs).
- Traffic redundancy reduction to ensure greater availability. WAN optimization technology introduces shared caches that must be updated only as needed, and substitutes very short cache reference exchanges across the WAN thereafter for arbitrarily long data exchanges between senders and receivers. Though the cost and capacity of cache will limit how much efficiency will be introduced, it's not uncommon for traffic volumes to drop by 30 to 60% when common data elements can be abstracted from multiple or recurring traffic streams.
- Operational savings expenditures from reduced WAN utilization. When enterprises must pay as they go for metered WAN bandwidth, reduced WAN utilization or lower bandwidth consumption translates directly into cost savings—at least for satellite. The more pressing issue—bursting—involves exceeding the allotted bandwidth for MPLS, which can be limited through traffic shaping when corporate policy dictates that traffic is valued highly enough to justify the cost. These are offset to some extent by modest fixed costs for WAN optimization technology, but reduction of regular recurring costs for bandwidth quickly swamps one-time expenses and related support or maintenance costs.
- Internet usage instead of backhauling leased lines to headquarters or implementing a hub-and-spoke WAN topology. Use of WAN optimization permits use of local Tx/Ex lines to ISPs, rather than requiring expensive leased lines or dedicated circuits from branch locations to regional hubs, or from regional locations to corporate HQ operations. Here again, connection costs often drop from tens to hundreds of thousands of dollars per month for backhauled circuits to hundreds to thousands of dollar per month for high-bandwidth ISP connections.

- Controlled access to corporate network resources. Use of WAN optimization technology also enables consistent, controlled enforcement of access controls, and permits only authorized, useful services and applications to consume WAN bandwidth. Not only does this permit more efficient use of bandwidth, it also helps to avoid the kinds of headaches that unauthorized or unwanted protocols and services (and the data they convey) can cause.
- Corporate resource prioritization remains consistent with organizational policies and imperatives. Use of WAN optimization ensures coherent, consistent application of security policy, access controls, and traffic priorities. WAN optimization technology permits centralized control and management over all WAN links, and enables organizations to impose and enforce the controls and priorities congruent with their needs and priorities, and to make and control changes over time as needs dictate.

A well-designed WAN optimization solution may be delivered and deployed using appliances in remote and headquarters offices to improve performance in a variety of ways. Two early ways to achieve accelerated application delivery come from Quality of Service (QoS) classifications, and bandwidth management techniques to prioritize and groom traffic. Other performance-enhancing techniques involve caching algorithms, shared data dictionaries and cache entries, and protocol acceleration through proxy services. But how do these apply in a world where clients, servers, and data come in so many different types?

Measuring Complex Data and User Transactions

It suffices to say: that which cannot be measured cannot be monitored. Without a formal means of establishing benchmarks and performance metrics, there simply isn't any means for monitoring application delivery. The introduction of simulation/probe and passive measurement techniques becomes the basis through which performance benchmarks can be obtained and monitoring techniques applied and understood. Many connections are serially-oriented—operating as a progression of individual events—as opposed to creating several concurrent connections in parallel. This leaves much to be desired—and much more to be optimized.

Several significant considerations must be applied:

- Understand the protocols being used, then optimize them to improve performance and efficiency. WAN optimization technology uses proxies to manage local, chatty services and applications on the LAN, and to map this kind of frequent, recurring local network traffic into more efficient communications and data transfers across the WAN.
- Save and re-use recurring requests for data via objects and data string caching. Caches and symbol dictionaries may be created and maintained across multiple WAN optimization devices, and updated only when actual changes occur. In the meantime, references to arbitrarily large data objects or strings may be sent across the WAN instead of requiring transfer of the actual data involved. Because the size of such references seldom exceeds 10 KB, and each such reference can point to data that is significantly larger, data volume reductions up to 99% are possible. Consistent reductions in the 30-60% range are typical.

- Instant, predetermined compression and encryption of data before distribution across the WAN. WAN optimization devices employ sophisticated hardware compression and encryption devices to make sure that the communications that actually traverse WAN links are both compact and as indecipherable to unauthorized third parties as modern technology will allow.
- Data caching is challenging when considering that a majority of objects by count (and by size) are too small to fit in a byte cache as described here. Unless they happen to appear in exactly the same order, which is highly unlikely on a contended network, byte caching won't improve performance. The only improvement for large (that is, video) and small (that is, Web page) object performance is through an *object cache*. Byte caching is designed for CIFS and MAPI optimizations, where it continues to perform the best. Object caching, however, often delivers the most dramatic improvements in performance when WAN optimization techniques are properly employed.
- Establishment of data delivery priorities based on users, applications, and processes. WAN optimization technology lets enterprises determine what kinds of traffic gets to jump to the front of the queue and obtains the best quality of service or service level guarantees. This not only helps to make effective use of WAN links and bandwidth, it also helps to ensure that end-user experiences are as positive as their priority ranking and assigned importance will allow.

If you don't measure and model your network infrastructure through a well-constructed service commitment, SLA breaches may go undetected. Reasonable expectations cannot be stated or met in terms of the service commitment when informal or ad-hoc service architectures are in place. Enforcement of said commitments becomes an infeasible and impractical proposition.

If you don't monitor the end-user experience, end-user perception and end-to-end response time, unpleasant surprises lie in wait on the bumpy network path ahead. Expectations can neither be defined nor met without a formal understanding of these performance properties. Ultimately, it's the end user who suffers the most with an indirect but significant impact on business flow.

It's Not a Challenge; It's a Promotion!

Visualize and utilize the proposition of network performance enhancement as a promotion, not a challenge. Delivering applications, rather than just packets, requires an enhanced skill set for networking professionals that ultimately aligns them closer to the value-adding parts of the organization. Treating all network connections as a singular flow of packet traffic gives neither the flexibility nor the scalability to ensure that business-critical applications perform as desired and expected.

Protocol optimization requires in-depth protocol knowledge to accelerate end-user response time and enhance serially-oriented network requests. Optimization strategies can better anticipate user requests through by understanding the intricacies of how certain protocols function natively on the LAN, and how they can better function across the WAN. Applications that use serialized requests (e.g., HTTP, CIFS, etc.) and traditionally “chatty” applications (e.g., RPC, RTSP) or those designed for LAN environments (i.e., CIFS, MAPI) achieve considerable performance gains through by bundling or short-circuiting transactions, or using pre-fetch techniques to anticipate upcoming requests and data transfers. Essentially this translates into batching up groups of related requests on one side of the WAN link, then doing likewise for related responses on the other side of the WAN link. It also involves use of proxies to carry on conversations locally for chatty protocols, then switching to bulk transfer and communication mechanisms across the WAN to lower the amount of back-and-forth traffic required across such links.

Networking professionals ultimately inherit the responsibility of promoting service and performance levels because IT and Information Management System (IMS) are inherently problematic. Remote Windows branch office servers have proven unmanageable, and IT governance doesn’t mean the same thing to all people. Many organizations use spur-of-the-moment processes that are either too loosely or too rigidly adhered, often concentrating efforts on the wrong aspects and failing to focus on key operational factors that make the IT process work efficiently. Oftentimes, there’s no surefire direction or method of approach to ensure the right aspects of performance are maintained at reasonable levels. Sometimes this results in the end user pointing the accusative finger of blame directly to those hard-working network professionals.

Shortly thereafter follows all kinds of server proliferation as an interim solution that proves equally unmanageable. Many of these so-called solutions still require manual intervention to operate and maintain, which is neither a model of efficiency nor room for innovation to thrive. Router blades for Domain Name Services (DNS), Dynamic Host Control Protocol (DHCP), and Remote Authentication Dial-In User Service (RADIUS) largely rely on the data professional delivering these goods over time. Print, file and services delivery are also an integral component to this unmanageably complex nightmare.

Moreover, these services are not integrated into routers because it’s the optimal architectural place for them—the performance issues inherent in hosting high-level services in a store-and-forward appliance are obvious. Such services are integrated into routers because there is a profound organizational desire to have network administrators manage them, and for that purpose, there is no better obvious placement.

Get involved in the application and protocol format: deconstruct the entire application and analyze its format to perform protocol optimization and manipulation. It requires a keen programmer’s insight—well, almost—and fundamental understanding of protocol topics to design, implement and deliver the appropriate optimization solution.

Progressing Up the Value Chain

Network IT professionals are moving higher up the value chain as these issues raise significant performance problems on business operations. Once part of the transparent business process, network professionals are now a very visible and valuable aspect of the IT value chain. Even end-user awareness has increased to include these networking professionals, as their efforts directly affect the end-user experience. And we're not just talking about the accusative finger-wagging sort of end-user awareness.

Consolidating servers out of branch offices assumed that the network would consume any excess capacity resulting from their aggregation. In theory, this may have made for sound composition; in practice, reality paints an ugly picture. Consider the example of Messaging Application Programming Interface (MAPI). Using MAPI—a messaging architecture and a Component Object Model based API for MS Windows—allows client programs to become e-mail messaging-enabled, messaging-aware and messaging-based. MAPI subsystem routines interface with certain messaging services that are closely related to the proprietary protocol that MS Outlook uses for communications with MS Exchange. However, another ugly reality rears its head as the MAPI framework fails to scale meaningfully within much larger network contexts.

That's not to say that Microsoft was necessarily trying to lock customers into branch office servers. Protocols were written for contention management networks where small packets and short, bursty chattiness were acceptable, and desirable for reducing the impact of network collisions. Here again, impact of WAN level latency and round trip times played little or no role in protocol and service designs.

Originally, these conversational, busy protocols worked well when networks had relatively low latency and more demand for available bandwidth. Excessive round-trip times per request, particularly for large payloads as seen in video and audio-bearing communications channels, create inordinate amounts of wasteful line noise that invariably adds to such network latency. When mapped onto widely-dispersed long-haul connections, such latency often multiplies by one or more orders of magnitude (tenths of seconds to full seconds or longer), and delay quickly becomes a massive management nightmare.

Summary

This chapter lays the foundation and defines the concepts for WAN concepts and components, with an emphasis toward enhancing and optimizing its operation. By layering WAN topologies over LAN technologies, performance decreases in a dramatic and discernible way. There are methods of monitoring, measuring, and modifying operational aspects of WAN technologies to improve the end-user experience and alleviate strain on potentially overworked networking professionals. In the next chapter, we'll adopt a more focused perspective on the types of routing protocols, processes, and procedures used to address these performance issues.

Download Additional eBooks from Realtime Nexus!

Realtime Nexus—The Digital Library provides world-class expert resources that IT professionals depend on to learn about the newest technologies. If you found this eBook to be informative, we encourage you to download more of our industry-leading technology eBooks and video guides at Realtime Nexus. Please visit <http://nexus.realtimepublishers.com>.