The Three Components of Optimizing WAN Bandwidth

Ashton, Metzler & Associates

October 2002

P.O. Box 1640
Sanibel, Florida 33957

Introduction

There are two fundamental truths that characterize an enterprise Wide Area Network (WAN). The first of those fundamental truths is that an enterprise WAN brings indisputable value to businesses of all types. Because of this business value, the amount of traffic that transits an enterprise WAN has historically increased at around thirty percent per year. However, a number of business and technology factors has caused an acceleration in the use of wide area networking. In particular, the amount of traffic on the typical enterprise WAN is currently increasing at around forty to forty-five percent per year. At this rate of increase, the traffic on an enterprise WAN will double every two years.

One of the business factors that has been a key driver of this discontinuity in the use of enterprise WANs is the deployment of content rich applications that enable both B2B and B2C interaction. In addition, enterprises are continually adding more to their networks – more users, more applications, more sites, more business partners, and more customers. The applications are getting more demanding in their requirements (e.g. Web, streaming, VoIP, casual downloads), and can burst to consume significant amounts of bandwidth (e.g., e-mail, P2P, downloads), at the expense of other applications.

Enterprise IT organizations continue to use low speed Frame Relay (i.e., Frame Relay ports at T1 speeds or slower) to provide connectivity to branch offices, and to use high-speed services to provide access to the corporate data centers. One of the real challenges of this design is when information is being pushed from the data center over a 100 Mbps access link and is destined for a branch office that is supported by an access link that is running at less than one percent of that capacity. In those instances, it is a virtual certainty that the sheer volume of information trying to transit this slow speed access link will negatively impact the arrival of delay sensitive information.

When faced with performance issues, there has traditionally been the temptation to throw more bandwidth at the problem. But, as traffic demands have been steadily increasing, last mile bandwidth prices have not been falling as fast as many had predicted. In fact, in many markets they are stabilizing due to declining competition and lack of broadly available alternatives (e.g. Cable, DSL, fixed wireless). This makes throwing bandwidth at the problem at best an expensive solution. For example, consider a company that is supporting two hundred branches with a Frame Relay network that is comprised of T1 access links, 128 Kbps PVCs, and 256 Kbps Frame Relay ports. Further assume that in order to improve network performance the company decided to increase just the size of their Frame Relay ports to 512 Kbps. Based on current domestic US Frame Relay tariffs, that change would cost around $300 per month per office. For the entire branch office network, this equates to an increase of $60,000 per month, or $2,160,000 over a three-year life cycle.

However, even if the company is willing to spend in excess of two million dollars on upgrading its access network, it is highly unlikely that it will solve its performance issues. That follows because even with these upgrades, the access link into the branch office is still bursting at a speed that is roughly one half of one percent of the speed of the access link into the corporate data center. As such, there is still the very strong likelihood that information being pushed from the data center to the branch will congest the access link into the branch office, and cause delay sensitive, business critical applications to perform poorly.

The slow-speed Frame Relay access links that were described in the preceding example highlight the access bandwidth bottleneck that almost always exists when providing connectivity between the corporate LAN

and the core WAN. In particular, most enterprises have deployed LAN infrastructures with dedicated 10 Mbps connections to the desktop. This desktop connectivity is supported both within and between buildings by networks running at 100 Mbps or higher. Analogously, the low end of the WAN core runs at speeds of at least OC-3 (155 Mbps) or OC-12 (622 Mbps). It is extremely likely that connecting the corporate LAN and the core WAN with a much slower access link will result in severe congestion.

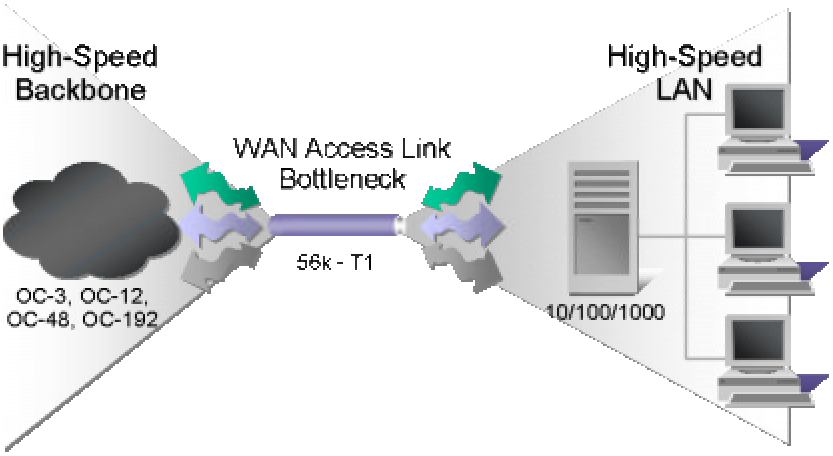

Figure 1: WAN access speeds are the bottleneck

The second of the fundamental truths that characterize an enterprise WAN is that throwing bandwidth at performance problems is seldom a viable approach for two reasons. The first reason is that supporting an enterprise WAN consumes considerable resources; both budget and people. However, given the need for profitability in these demanding economic times, most companies have put significant constraints on both the budget and the people that are allocated to support their WAN. Hence, throwing bandwidth at a performance issue is not likely to be an economically feasible option for most network organizations.

The second reason why throwing bandwidth at performance problems is not viable is because this approach is also imprecise. In particular, the added bandwidth is likely to be consumed by aggressive applications that are often the less business critical applications. Also, just providing more bandwidth does not address the added latency and jitter that these bursty, aggressive applications can introduce into the network.

In order to successfully support the aforementioned acceleration in the rate of increase of WAN traffic with highly constrained resources, enterprise network organizations need to increase the attention that they pay to better managing their network traffic and bandwidth. In this context, managing traffic and bandwidth refers to managing throughput, delay, and jitter.

There are three primary inter-dependent components to optimizing WAN bandwidth. They are:

1. Traffic Management/QoS

2. Caching
3. Compression

This document will discuss each of these techniques in detail, and will create a framework that the user can customize for their unique situation. This framework will discuss the types of situations in which the use of a given technique is likely to result in a significant benefit, as well as those situations in which the use of a given technique is not likely to result in a significant benefit. The framework will also highlight the dependencies among the management techniques.

1. Traffic Management/QoS

Traffic Management refers to the ability of the network to provide preferential treatment to certain classes of traffic. It is required in those situations in which bandwidth is scarce, and there are one or more delay sensitive, business-critical applications. Note that in this context, preferential treatment could mean limiting the bandwidth that certain applications (i.e., email) receive while simultaneously ensuring fairness for all the users of some business critical, delay sensitive application. One way to ensure fairness is to guarantee that all users of a specified application receive at least a given amount (e.g., 20 Kbps) of bandwidth.

In order to ensure that an application receives the required amount of bandwidth – or contrarily does not receive too much bandwidth, the traffic management solution must have application awareness. This often means detailed Layer 7 knowledge of the application, because many applications share the same port (i.e., Layer 4), or even hop between ports. One common example of a popular port is port 80, which often is used by web-based business applications, casual web browsing, Pointcast, streaming media and other applications. Many peer-to-peer applications such as Gnutella or iMesh constantly hop between ports making port-based bandwidth management rules ineffective. In particular, a solution that cannot distinguish between them cannot effectively manage bandwidth.

Another important factor in traffic management is the ability to effectively control inbound and outbound traffic. Queuing mechanisms control bandwidth leaving the network, but do not address traffic coming into the network, where the bottleneck usually occurs. Technologies like TCP Rate Control tell the remote servers how fast they can send content providing true bi-directional management.

One class of application that has become both pervasive and business-critical over the last few years is Enterprise Resource Planning (ERP) applications, such as SAP R/3. Several of the SAP R/3 modules are notably delay sensitive. An example of this is the Sales and Distribution (SD) module of SAP that is used for sales order entry. If the SD component is running slowly a company can compute the lost productivity of the company's sales organization as they waste time waiting for the SD module to respond. In addition, if the SD module times out, this can irritate the customer to the point where they hang up, taking their business elsewhere.

In order to quantify the bandwidth requirements of SAP R/3, Netigy [1] recently tested the SAP Create Sales Order Transaction. Among the observations that Netigy made were that the SAP R/3 Release 4.6C SAP

[1] Evaluation of Application Network Performance for SAP Sales and Distribution R/3 Release 4.6, November 2000, the Netigy Corporation

GUI for HTML produced more than 30 times the volume of traffic when compared to a Release 3.x client for the same Create Sales Order transaction. Netigy also noted that in a congested network, implementing QoS functionality reduced the SAP transaction time by approximately 75%.

Companies looking to deploy Traffic Management/QoS need to:

1. Instrument the network in such a way as they can monitor and classify the WAN traffic
2. Analyze the WAN traffic to understand the traffic patterns of the key applications
3. Implement an appropriate level of policies
4. Provide service-level monitoring and reporting to actively manage shifting network traffic patterns

Most organizations that are just getting started with deploying Traffic Management/QoS will get a major benefit by deploying just a few simple policies; i.e., control the bandwidth that email consumes, while assuring that SAP R/3 performs well. Over time, these companies may well want to deploy more sophisticated policies to gain additional benefits.

Traffic management is the most critical component of optimizing WAN bandwidth. Caching and compression provide additional benefits, but without traffic management, their benefits effectively act as a one-time bandwidth enhancement. The end result is that the same aggressive applications will consume the newly created bandwidth, just as they are consuming the existing bandwidth.

2. Caching

The current motivation for using caching is to both accelerate the delivery of content, as well as to optimize the use of WAN bandwidth. Viewed this way, it is possible to think of caching as a technique in which storage resources are deployed in such a way to cause the WAN to perform as if it had received a bandwidth upgrade. However, adding a relatively small amount of bandwidth will not solve the problem of poor application performance by itself. This follows because bandwidth will remain scarce and there is still a need to protect delay sensitive, business critical applications. The solution is to implement bandwidth enhancement techniques, such as caching, in conjunction with Traffic Management/QoS.

Caching is often present in several forms within a corporate environment. Most browsers contain a local cache for web pages they have already accessed. A proxy cache acts similar to a browser cache, but sits at the WAN edge and caches requests from many users, instead of just one. Finally, data center (or reverse proxy) caches offload servers and help speed content out of the data center. When referring to actively managing WAN bandwidth, proxy caches are the type of caches that are used.

The following illustrates one way in which proxy caches could be implemented.

- A user requests a Web page
- The network analyzes the request and decides to send the request to a local proxy cache

- If the proxy cache can fulfill the request, it does. If it can not, it will send a request to the original Web server
- The original Web server delivers the page to the proxy cache, which both stores the page, and delivers it to the requesting user

The third bullet in the preceding list highlights the primary factor that influences how much of an advantage proxy caching will provide to an organization. Sometimes referred to as data concurrency, that factor is the probability that when a user requests data, that the request can be fulfilled by a local proxy cache. The higher the probability that the request can be fulfilled by a local cache, the more advantage that proxy caching provides. Another limitation of proxy caches is that they only store web protocols (streaming, HTTP, FTP, NNTP), and do nothing for other applications, besides removing web protocols to free up extra bandwidth.

One situation in which proxy caching is likely to be highly advantageous involves employees at a remote office accessing a centralized suite of interactive web-based applications. In this case, since the majority of the information that gets displayed on the employees' screens is static, it can be cached locally. The dynamic information that populates the screens is still accessed from a centralized computer. Given the interactive nature of these applications, it is advised to also use Traffic Management/QoS on the dynamic content that goes back and forth between the remote office and the central data center, in order to ensure timely refreshing of the user's screens.

3. Compression

The basic function of compression is to reduce the size of a file of data to be transmitted over a WAN. Viewed this way, it is possible to look at compression in a very similar way to caching. In particular, compression can sometimes be used to cause the WAN to perform as if it had received a one-time bandwidth upgrade. However, as was the case with caching, deploying compression does not solve the problem of poor applications performance. As was previously stated, the solution to that problem is to implement bandwidth enhancement techniques, such as compression, in conjunction with Traffic Management/QoS.

One of the more common types of compression is referred to as dictionary compression. An example of dictionary compression is the Lempel-Ziv algorithm. This algorithm is based on a dynamically encoded dictionary that replaces a continuous stream of characters with codes. Note that variations of the Lempel-Ziv algorithm are used in many popular compression programs, such as Stac (LZS), ZIP and the UNIX compress utility.

A recent article from Cisco[2] pointed out issues to analyze when considering deploying compression. Those issues are:

- Number of remote sites
- Increase in latency as a by-product of compression

[2] Cisco IOS Data Compression, www.cisco.com/warp/public/cc/pd/iosw/tech/compr_wp.htm

- Throughput or congestion management
- Memory requirements
- Speed as a function of CPU cycles and compression algorithm instructions
- Compression ratio

The Cisco document explains each of the preceding issues in detail. This document will not repeat that discussion, but will focus on a couple of key concepts. One of these concepts is the compression ratio, which is a measure of the advantage that compression offers. The compression ratio is expressed as a ratio, x:1, where "x" is the number of input bytes divided by the number of output bytes. Clearly the larger the compression ratio, the more advantage that there is to deploying compression. The Cisco article discusses tests that were performed over a variety of different file types. Based on these tests, the article pointed out that in those cases in which compression adds value, that the reader should expect compression ratios in the range of 1.7:1 to 2:1.

While there certainly are times that compression can add value, there are many situations in which compression is either problematic or provides no value. For example, compression often does not scale well in terms of both memory and CPU cycles in a situation where a central site is communicating with a number of branch offices. The scaling issue with regards memory is that a dictionary-based compression algorithm requires that each end of a point-to-point connection must have dedicated memory. Clearly the greater the number of remote sites, the greater the memory that is required in the central site. The scaling issue with regards CPU cycles is that each additional connection to the central site increases the CPU utilization, which can also increase latency. New generation compression solutions are deployed on dedicated devices with large amounts of RAM and fast CPUs to address these issues, as well as improve on the compression ratios a Cisco solution may be able to obtain. One of the reasons these products are able to obtain higher compression ratios is that they utilize a larger dictionary across many packets, unlike a Cisco solution that only compresses within a single packet.

In addition, compression offers little value to certain classes of traffic such as VoIP traffic. That follows because VoIP traffic is already inherently compressed and cannot be further compressed, with the exception of the header information. In most cases, compression offers no value for any encrypted traffic. That follows because most encryption algorithms, such as 3DES, will produce few repeated sequences, and hence cannot be compressed by standard dictionary compression algorithms.

Finally, an effective application performance management solution needs to address both the throughput and latency concerns. Compression only addresses throughput and can, in fact, exacerbate latency. Traffic management provides latency management and should be combined with compression to control it.

The new generation of compression products have jitter buffers on each end to deal with the induced jitter caused by variable compressibility that exist in diverse traffic environments. Even though these buffers do a fairly good job of eliminating aggregate traffic jitter, they cannot effectively de-jitter any one traffic flow. Thus the overall traffic may flow smoothly but individual flows of traffic may incur more jitter as a result of compression. Traffic management, especially when using TCP Rate Control to manage TCP sessions, mitigates this problem by rate shaping each traffic flow so that bursty traffic cannot induce jitter into latency-sensitive traffic.

4. Single-Sided Compression

An alternative approach to deploying a box at every remote site is to deploy a new category of one-sided web acceleration products. These products, use a variety of technologies to optimize a web page for delivery, including standards based compression, image conversion and optimization techniques, delta encoding, and caching. The end result is compression ratios that range from 2:1 to 8:1 depending on the product and specific content.

These solutions have several advantages over dual-sided compression solutions, or proxy caching solutions. The most obvious is that they are considerably less expensive to deploy and manage than dual-sided solutions. This is because only a few boxes are required at the central site to optimize content to all of the users. The second benefit is that this technology can be used to optimize bandwidth to dial-in users as well as users at the other end of WAN links. Some of these solutions are able to recognize the connection speed and browser type that each user is coming in on and optimize content specifically for that user. A final advantage over proxy caches is that they optimize not only static content, but also dynamic content.

The primary drawback to this technology is that it is only used to accelerate web applications that the customer controls. Unlike dual-sided compression, which compresses nearly all traffic, and proxy caches, which cache all static web content (including public sites), single-sided solutions only optimize specified web sites.

The amount of business critical web-based traffic varies from one company to the next, but clearly is growing across all industries. Within the next several years, web traffic will likely dominate WAN traffic. In addition to the fact that web-enabled applications are growing as a percentage of the overall number of applications, the bandwidth consumed by web-enabled applications tends to be considerably larger than their client server predecessors. As previously noted, the SAP HTML GUI can consume over 30 times the traffic of the client-server version. These trends make some form of HTTP optimization (proxy caching, dual-sided, or single-sided), along with traffic management a requirement for effectively managing WAN applications.

5. Summary and Conclusions

One of the truths in our industry is that few organizations pay much attention to proactive network management of any type. However, the proactive optimization of WAN bandwidth will quickly become one of the primary ways that savvy network managers cope with the acceleration in their company's demand for WAN bandwidth, coupled with the severe resource constraints that they face.

As described in this paper, there are three primary inter-dependent components of optimizing WAN bandwidth. They are:

1. Traffic Management/QoS
2. Caching
3. Compression

The relative advantages and disadvantages of each of these components are summarized in Table 1.

	Traffic Management /QoS	Caching	Compression (dual-sided)	Compression (one-sided)
Protect mission-critical applications	Yes	No	No	Some (web applications)
Reduce bandwidth	No	Yes	Yes	Limited
Optimize bandwidth	Yes	No	No	No
Core/Edge deployment	Core or edge	Edge	Both core and edge required	Core
Cost to deploy	Variable	High	High	Low
Leading vendors	Packeteer, Allott	Network Appliance, Cisco	Expand, Peribit, (Cisco)	Redline, Packeteer, Fineground

Comparing Bandwidth Optimization Techniques
Table 1

When referring to Table 1, it is helpful to place performance management techniques into two distinct, but related classes. One class of performance management techniques, such as caching and compression, are concerned with bandwidth enhancement or WAN traffic reduction. In those situations in which these add value, they act like a one time increase in the WAN bandwidth.

The other class of performance management techniques is traffic management/QoS. Traffic management is crucial to the success of effectively managing WAN bandwidth. Without it, compression and caching will reduce the aggregate amount of traffic, but will not protect against aggressive applications, such as music downloads, bursting up and squeezing out mission-critical applications. Traffic management is the key technology that will allow enterprises to fully capitalize on the other bandwidth enhancement techniques.

To truly optimize WAN bandwidth, companies need to consider making their network “smarter”, not just “fatter”. When companies approaches the problem in this way, it will be able to get the most out of their network infrastructure.