

Best Practices in Optimizing WAN Performance

by **Dr. Jim Metzler**

Ashton, Metzler & Associates

exinda
networks

Introduction

As recently as a few years ago, managing application performance was not an important topic for most IT organizations. The results of a recent survey¹, however, highlight the fact that that situation is no longer the case. The survey asked IT professionals to indicate how the importance of managing application performance was changing within their organization. Their responses are contained in Figure 1.

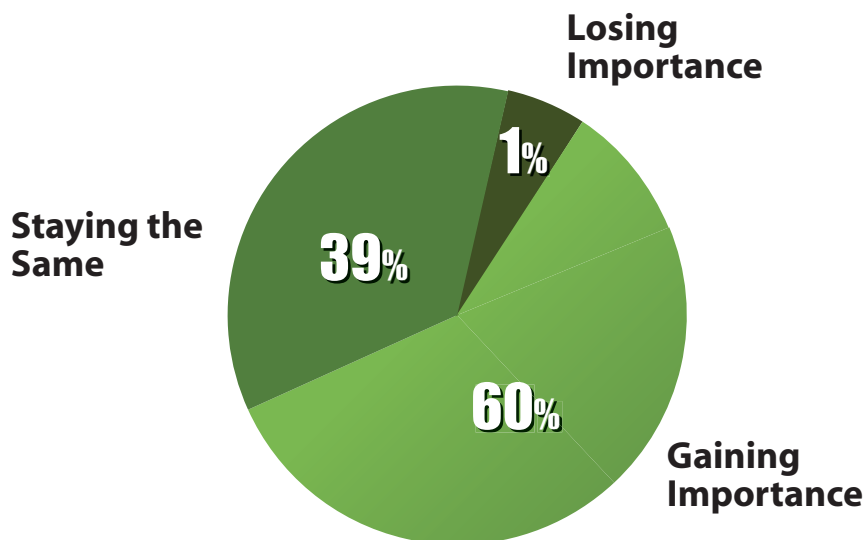


Figure 1: The Importance of Managing Application Performance

One of the key reasons why managing application performance is becoming so important is that companies today have a more distributed workforce than they had in the past. For example, currently there are over 6 million branch offices and 30 million branch office workers in the U.S.². In most cases, one or more of the applications that these branch office workers need to access are hosted in a central facility such as a data center. Because they are remote, branch office employees must access these applications over a wide area network (WAN) that has significantly lower bandwidth and notably more latency than does the local area network (LAN) that the employees would use to access these applications if they were located in a central facility.

One goal of this white paper is to outline a framework for application delivery. The second goal of this white paper is to use this framework to describe some best practices in WAN optimization. The third goal of this white paper is to explain how this framework and the best practices apply in specific industries.

A Framework for Successful Application Delivery

Because managing application performance is so complex, IT organizations will not be successful with an approach that is narrow in scope. Rather, IT organizations must approach application delivery in a broad, holistic fashion that is comprised of the four key activities: planning, optimization, management and control.

¹The Application Delivery Handbook, January 2007, www.kubernan.com

²Addressing Operational Efficiencies in Branch Offices, IDC, May, 2006

Planning

IT organizations will not be successful with application delivery if they wait until WAN links are fully saturated before they increase the capacity of those links. To avoid this situation, IT organizations need to implement effective capacity planning processes.

In addition, application performance is too important to the business and to the IT organization to proceed with an approach that is predicated on making changes and hoping that everything is acceptable. What is needed is an approach that allows the IT organization to model the change before it occurs. This approach enables IT organizations to understand in advance the impact of a change such as deploying a new application or modifying the network infrastructure.

Optimization

One of the primary goals of optimization is to improve the performance of applications that are delivered to branch offices over a WAN. As demonstrated in Table 1, there are multiple techniques that can be used to mitigate the negative impact that the indicated WAN characteristics have on application performance. The WAN optimization techniques listed in Table 1 will be explained in the next section of this white paper.

Table 1: WAN Optimization Techniques

WAN Characteristic	WAN Optimization Techniques
Insufficient Bandwidth	Data Reduction: Data Compression Differencing Caching
High Latency	Protocol Acceleration: Enhanced TCP Window Size Packet Aggregation Reduced Ch chattiness

Management

One of the primary roadblocks to successful application delivery is the fact that in the vast majority of instances in which the performance of an application is degrading, the end user notices the degradation before the IT organization does. To reverse this situation, IT organizations need granular visibility into application behavior so that performance related issues can be identified before they impact the users.

Application visibility at Layer 4 is helpful, but it does not provide IT organizations with the detailed level of information that they need. For example, data gathered at Layer 4 will allow IT organizations to determine how much traffic transited port 80, which is the well-known port for Web traffic. However, looking only at Layer 4, IT organizations cannot distinguish one Web-based application from another. This deficiency is becoming increasingly important as a growing number of peer-to-peer (P2P) applications such as Skype and AOL's Instant Messaging begin to use port 80.

Control

IT organizations exert control for multiple reasons. For example, an IT organization might exercise control as part of their security strategy or they might implement control as part of ensuring acceptable application performance. Throughout this white paper, control will refer to the functionality required to ensure acceptable application performance.

One of the factors that complicates the task of ensuring acceptable application performance is the deployment of applications that are both business critical and delay sensitive. One such application is Voice over IP (VoIP). The majority of IT organizations have deployed VoIP³. In order to be of acceptable quality, VoIP requires very low levels of delay, jitter and packet loss. Given the dynamic nature of IP networks, these stringent network parameters cannot be guaranteed without implementing Quality of Service (QoS). QoS refers to the ability of the network to provide preferential treatment to selected traffic classes. For example, an IT organization that has deployed VoIP would typically give VoIP traffic priority over other traffic types such as file transfer or email.

Best Practices

Based on the framework for successful application delivery that was outlined in the previous section, this section will develop a set of best practices for WAN optimization. Best practices are a set of functions that IT organizations need to be able to implement holistically in order to ensure acceptable application performance.

Application Visibility

As noted, in the vast majority of instances in which the performance of an application is beginning to degrade, the degradation is noticed first by the end user and not by the IT organization. In order to be regarded as being successful, IT organizations must reverse this situation. The first step in reversing this situation is to have the ability to identify the applications that are using the network and to be able to track usage and network utilization by application, host or conversation.

In order to do effective troubleshooting, it is important to be able to view application flows in real-time and to be able to drill down to specific branch offices, users or applications. Historical reporting is also important as it facilitates IT organizations being able to demonstrate the value that they provide to the business units. It also enables IT organizations to perform key planning functions such as capacity planning.

Application Performance Measurement

Being able to identify the applications that are using the network is a critical component of successful application delivery. However, just identifying the applications that are using the network is not sufficient. In order to be able to identify application degradation before the end user does, IT organizations also need the ability to measure the application response time as seen by the end user.

In order to minimize the amount of time that it takes to resolve performance problems, IT organizations also need to quickly determine how much of the application delay is due to the network and how much of the delay is due to the servers.

³ 2005/2006 VoIP State of the Market Report, Steven Taylor, www.webtorials.com

WAN Optimization

As shown in Table 1, there are multiple techniques that can be used to mitigate the negative impact that the WAN has on application performance. One class of techniques falls under the category of data reduction. These techniques result in less data having to transit the WAN. The primary data reduction techniques are:

Compression

The most common data compression algorithms look for redundancy in a data stream and use encoding techniques to remove the redundancy, creating a smaller file. A number of familiar compression tools are based on Lempel-Ziv (LZ) compression. This includes zip, PKZIP and gzip algorithms. Gzip is the primary compression algorithm used by HTTP V1.1. LZ based compression algorithms develop a codebook or dictionary as it processes the data stream and builds short codes corresponding to sequences of data. Repeated occurrences of the sequences of data are then replaced with the codes.

Differencing

In many cases, an organization repeatedly transmits large files with only minor differences between subsequent versions of the file. Differencing algorithms are used to update files by sending only the changes that need to be made to convert an older version of the file to the current version. Differencing algorithms partition a file into two classes of variable length byte strings: those strings that appear in both the new and old versions and those that are unique to the new version being encoded. The latter strings comprise a delta file, which is the minimum set of changes that the receiver needs in order to build the updated version of the file.

Caching

The motivation for caching is to store frequently used data locally and avoid having to fetch the data over a WAN each time it is referenced. In a typical situation, a user's Web browser in a branch office requests some information from a remote server. The request is first sent to a local cache. If the information is contained in the local cache and is current, the information is retrieved from the cache over a LAN. If the information is not available in the cache or is not current, the information is fetched from the remote server over a WAN and then both provided to the user and stored in the cache.

Another class of techniques that can be used to eliminate the negative impact that the WAN has on application performance falls under the category of protocol acceleration. By that is meant that these techniques mitigate the impact that data networking protocols have on application performance. Some of the primary protocol acceleration techniques are:

Enhanced TCP Window Size

One of the features of TCP (Transmission Control Protocol) that can cause poor application performance is the TCP window size. This refers to the maximum amount of information that TCP will allow to be transmitted without TCP receiving an acknowledgement that the data has been successfully received. In some cases, the TCP window size causes the WAN link to be idle even though there is data to be transmitted. The negative impact of the TCP window size can be reduced, and possibly eliminated, by implementing techniques that increase the window size.

Packet Aggregation

Another feature of many protocols that impacts application performance is the significant amount of overhead that is associated with each packet. As a result, if there are many small packets being transmitted, an unacceptable amount of the WAN bandwidth is consumed by the overhead in each packet. An effective way to reduce this overhead is to combine several small packets into one larger packet.

Reduced Ch chattiness

Another common feature of many protocols such as HTTP (Hypertext Transfer Protocol) and CIFS (Common Internet File System) is that they tend to be chatty. Chatty protocols typically require tens or even hundreds of application turns to complete a single transaction. For example, assume that a given protocol requires 100 application turns to complete a transaction and further assume that the round trip WAN delay is 150 ms. This results in an application delay of 15 seconds just due to the chattiness of the protocol. The actual application delay as seen by the end user would be higher due to other sources of delay (i.e., bandwidth induced delay and server delay). The most effective way to reduce the impact of chatty protocols is to implement techniques that reduce the number of round trips that the protocol requires.

Management

The usefulness of any WAN optimization solution is significantly impacted by the IT organization's ability to manage the solution. A weak management system limits the usefulness of the solution while a powerful management system enhances the usefulness. The primary characteristics that determine the impact of a management system are the system's functionality, ease of use and cost.

Some of the most important functionality of a management system for a WAN optimization solution was already discussed. In summary, that includes the ability to identify the applications that are using the network as well as the ability to track usage and network utilization by application, host or conversation. It also includes the ability to measure the application response time as seen by the end user. Other important management functionality includes the ability to produce customized reports as well as the ability to produce standardized reports on factors such as utilization, top applications, top URLs, and top hosts. As previously noted, IT organizations need reports that are both real-time and historical in orientation.

Given that WAN optimization devices are deployed in branch offices that typically have no IT personnel, a major component for ease of use of a management system is the ability of the system to manage the WAN optimization devices from a centralized location. Another factor that determines the ease of use of a management system is the ability of the WAN optimization devices to auto-discover each other.

Cost is a major criterion when choosing any IT product or service. One way to reduce cost is to have alternative ways to deploy the product or service. For example, an IT organization that deploys a WAN optimization solution may choose to take on the cost and responsibility of implementing a centralized management system for the solution. Alternatively, the IT organization may be able to save cost, time and reduce risk if the management system is available using the software as a service business model.

Control

In the context of managing application performance, one of the goals of control is to ensure that business-critical, delay-sensitive applications such as VoIP are not negatively impacted by other applications. Another goal is to either eliminate or severely limit the usage of the network by applications that do not have any business value such as Internet radio or gaming.

The most common way that IT organizations exert control over their network is by implementing QoS typically by manipulating the TOS (Type of Service) byte in the IP (Internet Protocol) header. Over the last several years the TOS byte has had various purposes. The current definition of the TOS byte (see RFC 3168) is a six-bit Differentiated Services Code Point (DSCP) and a two-bit Explicit Congestion Notification field. The DSCP value indicates the preferred QoS as the packet traverses the network.

Implementing Best Practices

This section of the white paper will discuss the use of WAN optimization in two key industries: retail and education.

The Retail Industry

A lot of market research⁴ describes the business pressures facing the retail industry. One of these pressures is that the vast majority of retailers report that they are concerned about customer complaints concerning their in-store experience. Another pressure is that retailers have not been able to add employees to their stores. In particular, 82 percent of retailers indicated that their payroll as a percentage of sales remains either constant or decreasing⁵. Historically, companies have deployed various forms of IT in order to provide better service without adding headcount. An example of that phenomenon in the retail industry is the deployment of self-check out systems. These systems are typically comprised of scanners that read a bar code on every product; and for every bar code have the price of the item. The deployment of these systems allows retailers to reduce how long customers have to wait in checkout lines without having to add staff.

While the retail industry has a track record of deploying technology to increase customer satisfaction and overall efficiency, the industry also has a track record of spending a smaller percentage of its revenues on IT than do most industries. The primary goal of WAN optimization is to maximize the use of the WAN, which has a high recurring monthly cost. This makes WAN optimization an important technology for the retail industry.

To put this in context, consider BigHoller. BigHoller empowers restaurants to offer on-line food ordering for customers. They do this by providing an ordering system for restaurants and in some instances they also provide a Web site for the restaurant. In the typical scenario, a customer goes to the restaurant's Web site and clicks on a link that pops up the restaurant's menu on a server hosted by BigHoller. The customer indicates the food and beverage that they want and BigHoller sends the information to the restaurant.

According to G.R. Homa, managing partner of BigHoller, the on-line ordering industry is going through a period of rapid growth. He stated that whereas only about five percent of restaurants currently offer on-line ordering, he expects this to grow to 80% in the next few years. He added that up until this year, BigHoller's data center was serviced by a 2 Mbps WAN link. Based on having a WAN optimization tool that allowed them visibility into application usage, they were able to plan for when they would need to add capacity to that link. Unfortunately, the deployment of the WAN upgrade was delayed three months.

As G.R. stated, "The worst thing for me is a slowdown in the system that impacts customers." Fortunately, their WAN optimization tool was able to identify a number of applications that were not time critical and that were also consuming WAN capacity. To avoid having the lack of WAN bandwidth impact their customers, BigHoller also used their WAN optimization tool to implement QoS and a number of data reduction techniques.

According to G.R., "What is important to me is that when something is slowing down the system, I can go to a tool, see what the problem is and take action."

⁴ Retail Trends and Forecasts: Winning Customers in the Age of Choice, http://cisco.com/web/strategy/retail/downloads/Retail_Trends_2007_031607a.pdf

⁵ RGAG Research: "Workforce Optimization: Boosting Store-Level Productivity & Top-Line Performance", August 2006

Education

The cost of education tends to increase rapidly due in large part to the labor-intensive nature of teaching. In an effort to stay affordable, educational institutions are continually looking for ways to control costs. In addition to being cost conscious in general, educational institutions are also very conservative relative to how much they spend on IT. As a result, educational institutions try to keep IT operating costs low, and hence are receptive to deploying technologies such as WAN optimization that can help them postpone expensive WAN upgrades.

Jason O'Rourke is the Managing Director of Pavilion Connexions, a company that provides Internet services to students who live in privately managed residences at seven sites at two universities in the United Kingdom. O'Rourke began to look at WAN optimization because of a problem that many educational institutions experience – severe network problems due to P2P file sharing. According to O'Rourke, "The students were abusing the network by downloading huge files using P2P applications such as LimeWire and BitTorrent to get their free music and videos. It brought the academic use of the network to a screeching halt and we needed some way to curb this activity."

Part of the challenge facing O'Rourke is that the policy of the universities deemed it acceptable to use the network for social purposes. As a result, O'Rourke could not just block all music and video downloads. Instead, in a manner similar to what G.R. of BigHoller did, O'Rourke implemented a solution that allowed him to prioritize interactive traffic (i.e., Web browsing) and email over P2P downloads.

As previously noted, G.R. said, "What is important to me is that when something is slowing down the system, I can go to a tool, see what the problem is and take action." In a very similar fashion, O'Rourke stated, "When we first installed the application optimization solution we got a fair idea of what was happening on the network. This information helped us to design the best network plan for shaping traffic. Without the knowledge of how the network was being used, the project would not have been successful."

Summary

Over the last few years, managing application performance has gained significantly in importance. That situation is not likely to change any time soon. In fact, it is highly likely that managing application performance will continue to increase in importance for the foreseeable future. This follows in part because organizations increasingly run their key operations using a growing set of applications. If these applications are not performing well, the organization's operations are negatively impacted.

Today most IT organizations deploy WAN optimization in a tactical fashion. They deploy the fewest appliances that they can to solve a very well defined problem. Given the growing importance of WAN optimization as well as the rapid pace of change of most networks, IT organizations need to plan for how the WAN optimization solutions that they implement today can support future requirements. As part of the tactical planning, IT organizations need to recognize that in most cases the total cost of ownership of a WAN optimization solution is far greater than the initial cost. As such, IT organizations that are evaluating these solutions need to understand not only the initial cost of a WAN optimization solution, but also the total cost of ownership over the planned lifecycle of the solution.

Too often IT organizations associate the phrase WAN optimization just with functionality such as data reduction and protocol acceleration. While this functionality is clearly necessary, it is not sufficient. For example, neither data reduction nor protocol acceleration will change the fact that in the vast majority of instances, the end user notices application degradation before the IT organization does.

To ensure acceptable application performance, IT organizations must deploy a WAN optimization solution that integrates planning, optimization, management and control. Both BigHoller and Pavilion Connexions illustrate the value of this approach. The WAN optimization solution that BigHoller implemented allowed them to identify when their WAN link would be saturated in time to upgrade the link. When BigHoller's upgrade was delayed for three months, their WAN optimization solution had the ability to optimize the use of the current WAN link. The solution also provided BigHoller with the application visibility that enabled them to identify applications that could be given a lower priority so that the performance of the customer-facing applications was not impacted. Similarly, the WAN optimization solution that Pavilion Connexions implemented gave them the visibility and control to implement the organization's policy of allowing students to use the network for social purposes while ensuring that the network could still support the intended academic uses of the network as well.