

The Handbook of Application Delivery

Everything you wanted to know, but didn't know you needed to ask

SPONSORED IN PART BY

riverbed™

The Handbook of Application Delivery

Contents

1.0	Executive Summary	3
2.0	Introduction and Purpose	5
3.0	Focus, Scope and Methodology	6
4.0	The Applications Environment	7
5.0	Planning	11
6.0	Network and Application Optimization	20
7.0	Management	30
8.0	Control	40
9.0	Conclusion	45
10.0	Bibliography	47
11.0	Interviewees	54
12.0	Riverbed	55

IT Innovation Report

Published By
Kubernan
www.Kubernan.com

Cofounders
Jim Metzler
jim@ashtonmetzler.com

Steven Taylor
taylor@webtutorials.com

Design/Layout Artist
Debi Vozikis

Copyright © 2007
Kubernan

**For Editorial and
Sponsorship Information**
Contact Jim Metzler
or Steven Taylor

Kubernan is an analyst
and consulting joint
venture of Steven Taylor
and Jim Metzler.

Professional Opinions Disclaimer
All information presented and opinions
expressed in this IT Innovation Report
represent the current opinions of the
author(s) based on professional judg-
ment and best available information
at the time of the presentation.
Consequently, the information is
subject to change, and no liability for
advice presented is assumed. Ultimate
responsibility for choice of appropriate
solutions remains with the reader.

1.0 Executive Summary

IT organizations have two primary functions: application development and application delivery. Within most IT organizations, the application development function is highly formalized. In contrast, within most IT organizations there is typically nascent recognition of the existence of an integrated application delivery function. One key symptom of the lack of a formalized effective application delivery function is that in the vast majority of instances that a key business application is degrading, that degradation is noticed first by the end user and not by the IT organization. Another key symptom is that when application degradation does occur, most IT organizations are unsure how to best resolve the issue.

This report's goal is to help IT organizations develop the ability to minimize the occurrence of application performance issues and to both identify and quickly resolve those issues when they do occur. To achieve this goal, Kubernan synthesized its own knowledge with that of roughly a dozen of the industry's leading vendors and a similar number of IT organizations. Kubernan also surveyed hundreds of IT organizations. Given the breadth and extent of the input from both IT organizations and leading edge vendors this report represents a broad consensus on an application delivery framework that IT organizations can modify for use within their organization. To make the framework even more actionable, this report contains roughly 40 conclusions that IT organizations can use to shape how they modify the framework.

Below is a listing of some of the factors that complicate the application delivery function. The impact of these factors is not dissipating any time soon. If anything, the impact of each of these factors will increase.

- In the majority of cases, there is at most a moderate emphasis during the design and development of an application on how well that application will run over a WAN.
- There is a requirement to identify and classify an organization's applications based on its network requirements and business criticality.
- The performance of the user's desktop tends to degrade over time.
- Shifting traffic patterns make it more difficult to both manage and optimize traffic flows.
- The deployment of increasingly distributed applications increases the number of sources of application degradation.
- The Webification of applications tends to greatly increase the amount of traffic that the IT infrastructure must support.
- The movement to consolidate servers out of branch offices and into fewer data centers can result in significant performance issues.
- Both the movement to reduce the number of data centers and the movement to host a given application in a single data center increase the amount of WAN delay associated with accessing an application.
- The vast majority of people who access an application do not reside in a headquarters location. This increases the difficulty in managing the performance of the users' desktops and insures that the latency, jitter and packet loss of the WAN will impact the application's performance.
- The typical IT environment is highly dynamic. For example, new users, sites and applications are added on a regular basis.
- Because there are changing paths through an IP network, IT organizations need visibility into the operational architecture and dynamic behavior of the network.

- In many cases an application resides on multiple servers, which necessitates balancing the load across these servers.
- In many cases, server performance degrades due to the requirement to perform tasks, such as processing SSL traffic.
- The typical IT organization is highly fractured. For example, an IT organization rarely plans, funds and manages the components of IT in an integrated fashion.

Today, when most IT organizations think about application delivery, they think somewhat narrowly about just network and application optimization. While network and application optimization is very important, it is just one component of an effective application delivery function. The other components of effective application delivery are planning, management and control.

This report does not set out to cover all aspects of application delivery. Rather, it will analyze each of the four components of application delivery (planning, network and application optimization, management and control), discuss some of the key elements of each component, draw a number of key conclusions, and suggest criteria that IT organizations can use to choose an appropriate solution.

The planning components this report discusses are the ability to:

- Profile an application prior to deploying it.
- Baseline the performance of the network.
- Perform a pre-deployment assessment.

Relative to planning, one of the conclusions this report draws is that a primary way to balance the requirements and capabilities of the application development and the application delivery functions is to create an effective architecture that integrates these two functions.

The components of network and application optimization that are discussed in this report are:

- Branch Office Optimization solutions.
- Data Center Solutions.

Relative to network and application optimization, one conclusion this report draws is that in order to understand the performance gains of any network and application optimization solution, organizations must test solutions in an environment that closely reflects the environment in which it will be deployed.

The components of management this report discusses are:

- The organizational barriers.
- Discovery.
- End-to-end visibility.
- Network and application alarming.
- Measuring application performance.
- Route analytics.

Relative to management, one of the conclusions this report draws is that organizational discord and ineffective processes are at least as much of an impediment to the successful management of application performance as are technology and tools.

The components of control this report discusses are:

- Desktop control.
- Traffic management and QoS.
- Route control.

Relative to control, one conclusion this report draws is that one of the primary characteristics of a next generation desktop support system is automation.

2.0 Introduction and Purpose

IT organizations have two primary functions: applications development and applications delivery. Applications development involves a variety of tasks including developing new software, acquiring software from a third party, integrating software modules and maintaining software. Applications delivery involves any task required to ensure a secure, cost effective and acceptable level of application performance.

If you work in IT, you either develop applications or you deliver applications.

One of the IT professionals interviewed for this report is the COO of an electronic records management company¹. He put the value of application delivery in context when he stated “The days in which an IT organization can show business value merely by implementing and managing the infrastructure are over. Today, IT organizations absolutely must ensure the effective performance of applications.”

Over the last year, Kubernan has conducted extensive market research into the challenges associated with application delivery. One of the most significant results uncovered by that market research is the dramatic lack of success IT organizations have relative to managing application performance. In particular, Kubernan recently asked 345 IT professionals the following question. “If the performance of one of your company’s key applications is beginning to degrade, who is the most likely to notice it first – the IT organization or the end user?” Seventy three percent of the survey respondents indicated that it was the end user.

In the vast majority of instances when a key business application is degrading, the end user, not the IT organization, first notices the degradation.

The fact that end users notice application degradation prior to it being noticed by the IT organization is an issue of significant importance to virtually all senior IT managers.

In addition to performing market research, Kubernan also provides consulting services. Jim Metzler was recently hired by an IT organization that was hosting an application on the east coast of the United States that users from all over the world accessed. Users of this application that were located in the Pac Rim were complaining about unacceptable application performance. The IT organization wanted Jim to identify what steps it could take to improve the performance of the application. Given that the IT organization had little information about the semantics of the application, the task of determining what it would take to improve the performance of the application was lengthy and served to further frustrate the users of the application. This report is being written with that IT organization and others like them in mind.

The goal of this report is to help IT organizations develop the ability to minimize the occurrence of application performance issues and to both identify and quickly resolve issues when they do occur.

To achieve that goal, this report will develop a framework for application delivery. It is important to note that most times when the industry uses the phrase *application delivery*, this refers to just network and application optimization. Network and application optimization is important. However, achieving the goal stated above requires a broader perspective on the factors that impact the ability of the IT organization to assure acceptable application performance.

Application delivery is more complex than just network and application acceleration.

Application delivery needs to have top-down approach, with a focus on application performance.

¹ A summary of the IT professionals who were interviewed for this report can be found in section 11.

With these factors in mind, the framework this report describes is comprised of four primary components: Planning, Network and Application Optimization, Management, and Control. Some overlap exists in the model as a number of common IT processes are part of multiple components. This includes processes such as discovery (what applications are running on the network and how are they being used), baselining, visibility and reporting.

Conceptually, application delivery is not that difficult. To be successful with application delivery, all that an IT organization must do is:

- Have a deep understanding of the volatile environment that is the organization's applications, servers, users and networks.
- Implement the appropriate techniques to optimize the performance of the network, the servers, and the application.
- Measure everything and automate as much as possible.
- Develop control over the desktop as well as control over what traffic can enter the network and how that traffic is routed through the network.

As this report will show, however, the devil is definitely in the details.

3.0 Focus, Scope and Methodology

As sections 4 and 7 will discuss, a number of factors impact application delivery. These factors include the:

- Availability and performance of the desktop.
- Deployment of chatty applications.
- Widely varying requirements and business criticality of applications.
- Shifting traffic patterns.

- Deployment of distributed applications.
- Movement of employees out of headquarters sites.
- Webification of applications.
- Consolidation of servers and of data centers.
- Fractured nature of the typical IT organization.
- Deployment of protocols that are chatty (i.e., CIFS), dense (i.e., XML) or computationally intense; i.e., SSL.
- Changing paths through an IP network.
- Volatility of the IT environment.
- Impact of specific components of the IT infrastructure such as the WAN or servers.

This is a lengthy report. It does not, however, require linear, cover-to-cover reading. A reader may start reading the report in the middle and use the references embedded in the text as forward and backwards pointers to related information.

To keep the report a reasonable length, it does not contain a detailed analysis of any technology. To compensate for this, the report includes an extensive bibliography. In addition, the body of this report does not discuss any vendor or any products or services. The Appendix to this report, however, contains material supplied by Riverbed one of the leading application delivery vendors.

The report does not discuss how the design of components of the infrastructure such as the LAN, WAN, SAN or how the choice of servers, storage, or operating systems impacts application delivery. In addition, discussion of the growing impact that security has and will continue to have on application delivery is limited. This report does not discuss topics such as the role of managed service providers, the value of service level management or the role of ITIL in improving management processes. Future reports will cover those topics.

To allow IT organizations to compare their situation to those of other IT organizations, this report incorporates survey data that Kubernan has gathered over the last year. This report also contains input gathered from interviewing roughly a dozen IT professionals. As a general rule, IT professionals cannot be quoted by name or company in a report like this without their company heavily filtering their input. To compensate for that limitation, section 11 contains a brief listing of the people who were interviewed, along with the phrase the report uses to refer to them. The report sponsors provided input into the areas of this report that are related to their company's products and services. Both the sponsors and the IT professionals also provided input into the relationship between and among the various components of the application delivery framework.

Given the breadth and extent of the input from both IT organizations and leading edge vendors this report represents a broad consensus on a framework that IT organizations can use to improve application delivery.

4.0 The Applications Environment

This section of the handbook will discuss some of the primary dynamics of the applications environment that impact application delivery. It is unlikely any IT organization will exhibit all of the dynamics described. It is also unlikely that an IT organization will not exhibit at least some of these dynamics.

No product or service in the marketplace provides a best in class solution for each component of the application delivery framework. As a result, companies have to carefully match their requirements to the functionality the alternative solutions provide.

Companies that want to be successful with application delivery must understand their current and emerging application environment.

The preceding statement sounds simple. However, less than one-quarter of IT organizations claim they have that understanding.

The Application Development Process

In the typical application development environment, the focus is on delivering the promised software functionality on time and with relatively few bugs or security vulnerabilities.

In the majority of cases, there is at most a moderate emphasis during the design and development of an application on how well that application will run over a WAN.

This lack of emphasis on how well an application will run over the WAN often results in the deployment of chatty applications. *Chatty applications* are applications in which a given transaction requires tens or possibly hundreds of round trips, a.k.a., application turns. To exemplify this, assume that the round-trip WAN delay between a user and the application the user is trying to access is 120 ms. Also assume that a transaction requires 100 application turns. As a result, for a given transaction the user will experience 12 seconds of delay on average due to the application turns. The user will experience additional delay due to factors such as the transmission delay as well as the delay in the application and database servers. The peak delay will be even higher.

Taxonomy of Applications

The typical enterprise has tens and often hundreds of applications that transit the WAN. One way that these applications can be categorized is:

1. Business Critical

A company typically runs the bulk of its key business functions utilizing a handful of applications. A company can develop these applications internally, buy them from a vendor such as Oracle or SAP,

or acquire them from a software-as-a-service provider such as Salesforce.com.

2. Communicative and Collaborative

This includes delay sensitive applications such as Voice over IP and conferencing, as well as applications that are less delay sensitive such as email.

3. Other

This category contains the bulk of a company's data applications. While these applications do not merit the same attention as the enterprise's business critical applications, they are important to the successful operation of the enterprise.

4. IT Infrastructure-Related Applications

This category contains applications such as DNS and DHCP that are not visible to the end user, but which are critical to the operation of the IT infrastructure.

5. Recreational

This category includes a growing variety of applications such as Internet radio, YouTube, streaming news and multimedia, as well as music downloading.

6. Malicious

This includes any application intended to harm the enterprise by introducing worms, viruses, spyware or other security vulnerabilities.

Since they make different demands on the network, another way to classify applications is whether the application is real time, transactional or data transfer in orientation. For maximum benefit, this information must be combined with the business criticality of the application. For example, live Internet radio is real time but in virtually all cases it is not critical to the organization's success. It is also important to realize an application such as Citrix or

SAP is comprised of multiple modules with varying characteristics. Thus, it is not terribly meaningful to say that Citrix is real time, transactional or data transfer in orientation. What is important is the ability to recognize application traffic flows for what they are, for example a Citrix printing flow vs. editing a Word document.

Successful application delivery requires that IT organizations are able to identify the applications running on the network and are also able to ensure the acceptable performance of the applications relevant to the business while controlling or eliminating applications that are not relevant.

The Desktop

As section 8.2 will discuss, once a desktop or laptop computer is placed into service it begins to accumulate unnecessary applications and files. These unnecessary applications and files will be referred to in this report as *detrimental additions*.

The detrimental additions that accumulate on computers come from numerous sources. In some cases, the user is not directly involved in adding the detrimental additions that accumulate on their PC. Well-known examples of this phenomenon include worms, viruses and spyware that attach themselves to the user's desktop once a user unwittingly opens an email from a malicious source. Some less well-known examples of this phenomenon, such as registry creep and bad DLL files, however, can also have a significant long-term impact on desktop performance.

In most cases, these detrimental additions will affect the availability and performance of the desktop, and hence directly affect application delivery.

Traffic Flow Considerations

In many situations, the traffic flow on the data network naturally follows a simple hub-and-spoke design. An example of this is a bank's ATM network where the traffic

flows from an ATM to a data center and back again. This type of network is sometimes referred to as a one-to-many network.

A number of factors, however, cause the traffic flow in a network to follow more of a mesh pattern. One factor is the wide spread deployment of Voice over IP (VoIP)². VoIP is an example of an application where traffic can flow between any two sites in the network. This type of network is often referred as an any-to-any network. An important relationship exists between VoIP deployment and MPLS deployment. MPLS is an any-to-any network. As a result, companies that want to broadly deploy VoIP are likely to move away from a Frame Relay or an ATM network and to adopt an MPLS network. Analogously, companies that have already adopted MPLS will find it easier to justify deploying VoIP.

Another factor affecting traffic flow is that many organizations require that a remote office have access to multiple data centers. This type of requirement could exist to enable effective disaster recovery or because the remote office needs to access applications that disparate data centers host. This type of network is often referred as a some-to-many network

Every component of an application delivery solution has to be able to support the company's traffic patterns, whether they are one-to-many, many-to-many, or some-to-many.

Application Complexity

Companies began deploying mainframe computers in the late 1960s and mainframes became the dominant style of computing in the 1970s. The applications that were written for the mainframe computers of that era were monolithic in nature. *Monolithic* means that the application performed all of the necessary functions, such as

providing the user interface, the application logic, as well as access to data.

Most companies have moved away from deploying monolithic applications and towards a form of distributed computing that is often referred to as *n-tier applications*. Since these tiers are implemented on separate systems, WAN performance impacts n-tier applications more than monolithic applications. For example, the typical 3-tier application is comprised of a Web browser, an application server(s) and a database server(s). The information flow in a 3-tier application is from the Web browser to the application server(s) and to the database, and then back again over the Internet using standard protocols such as HTTP or HTTPS.

The movement to a Service-Oriented Architecture (SOA) based on the use of Web services-based applications represents the next step in the development of distributed computing.

Just as WAN performance impacts n-tier applications more than monolithic applications, WAN performance impacts Web services-based applications more than n-tier applications.

To understand why that is the case, consider the 3-tier application architecture that was previously discussed. In a 3-tier application the application server(s) and the database server(s) typically reside in the same data center. As a result, the impact of the WAN is constrained to a single traffic flow, that being the flow between the user's Web browser and the application server.

In a Web services-based application, the Web services that comprise the application typically run on servers that are housed within multiple data centers. As a result of housing the Web services in multiple data centers, the WAN impacts multiple traffic flows and hence has a

² 2005/2006 VoIP State of the Market Report, Steven Taylor, www.webtorials.com

greater overall impact on the performance of a Web services-based application than it does on the performance of an n-tier application.

Webification of Applications

The phrase *Webification of Applications* refers to the growing movement to implement Web-based user interfaces and to utilize chatty Web-specific protocols such as HTTP, HTML, XML and SOAP. Similar to the definition of a chatty application, a protocol is referred to as being chatty if it requires tens if not hundreds of turns for a single transaction. The next subsection of this report will describe the impact of a chatty protocol.

In addition, XML is a dense protocol. That means communications based on XML consume more IT resources than communications that are not based on XML.

The webification of applications introduces chatty protocols into the network. In addition, some of these protocols (e.g., XML) tend to greatly increase the amount of data that transits the network and is processed by the servers.

Server Consolidation

Many companies either already have, or are in the process of consolidating servers out of branch offices and into centralized data centers. This consolidation typically reduces cost and enables IT organizations to have better control over the company's data.

While server consolidation produces many benefits, it can also produce some significant performance issues.

Server consolidation typically results in chatty protocols such as CIFS (Common Internet File System), Exchange or NFS (Network File System), which were designed to run over the LAN, running over the WAN. The way that CIFS works is that it decomposes all files into smaller blocks prior to transmitting them. Assume that a client

was attempting to open up a 20 megabyte file on a remote server. CIFS would decompose that file into hundreds, or possibly thousands of small data blocks. The server sends each of these data blocks to the client where it is verified and an acknowledgement is sent back to the server. The server must wait for an acknowledgement prior to sending the next data block. As a result, it can take several seconds for the user to be able to open up the file.

Data Center Consolidation and Single Hosting

In addition to consolidating servers out of branch offices and into centralized data centers, many companies are also reducing the number of data centers they support worldwide. HP, for example, recently announced it was reducing the number of data centers it supports from 85 down to six³. This increases the distance between remote users and the applications they need to access. Many companies are also adopting a *single-hosting* model whereby users from all over the globe transit the WAN to access an application that the company hosts in just one of its data centers.

One of the effects of data center consolidation and single hosting is that it results in additional WAN latency for remote users.

Changing Application Delivery Model

The 80/20 rule in place until a few years ago stated that 80% of a company's employees were in a headquarters facility and accessed an application over a high-speed, low latency LAN. The new 80/20 rule states that 80% of a company's employees access applications over a relatively low-speed, high latency WAN.

In the vast majority of situations, when people access an application they are accessing it over the WAN.

³ Hewlett-Packard picks Austin for two data centers <http://www.statesman.com/business/content/business/stories/other/05/18hp.html>

Dynamic IT Environments

The environment in which application delivery solutions are implemented is highly dynamic. For example, companies are continually changing their business processes and IT organizations are continually changing the network infrastructure. In addition, companies regularly deploy new applications and updates to existing applications.

To be successful, application delivery solutions must function in a highly dynamic environment. This drives the need for both the dynamic setting of parameters and automation.

Fractured IT Organizations

The application delivery function consists of myriad sub-specialties such as desktop, LAN, SAN, WAN, storage, servers, security, operating systems, etc. The planning and operations of these sub-specialties are typically not well coordinated within the application delivery function. In addition, typically little coordination exists between the application delivery function and the application development function.

Only 14% of IT organizations claim to have aligned the application delivery function with the application development function. Eight percent (8%) of IT organizations state they plan and holistically fund IT initiatives across all of the IT disciplines. Twelve percent (12%) of IT organizations state that troubleshooting an IT operational issues occurs cooperatively across all IT disciplines.

The Industrial CIO described the current fractured, often defensive approach to application delivery. He has five IT disciplines that report directly to him. He stated that he is tired of having each of them explain to him that their component of IT is fine and yet the company struggles to provide customers an acceptable level of access to their Web site, book business and ship product. He also said

that he and his peers do not care about the pieces that comprise IT, they care about the business results.

The CYA approach to application delivery focuses on showing that it is not your fault that the application is performing badly. The goal of the CIO approach is to identify and then fix the problem.

5.0 Planning

Introduction

The classic novel *Alice in Wonderland* by the English mathematician Lewis Carroll first explained part of the need for the planning component of the application delivery framework. In that novel Alice asked the Cheshire cat, "Which way should I go?" The cat replied, "Where do you want to get to?" Alice responded, "I don't know," to which the cat said, "Then it doesn't much matter which way you go."

Relative to application performance, most IT organizations are somewhat vague on where they want to go. In particular, only 38% of IT organizations have established well-understood performance objectives for their company's business-critical applications.

It is extremely difficult to make effective network and application design decisions if the IT organization does not have targets for application performance that are well understood and adhered to.

The Manufacturing Manager indicated that his IT organization does offer an application SLA, but that it is primarily focused on the availability of the application, not the application's performance. The Consulting Architect stated his organization does not currently have well-understood performance objectives for their business-critical applications, and that this was part of what they were trying to accomplish. The CIO of that company highlighted the need for application performance objectives when he said,

“What we have now is garbage. We do not have the right metrics.”

One primary factor driving the planning component of application delivery is the need for risk mitigation. One manifestation of this factor is the situation in which a company’s application development function has spent millions of dollars to either develop or acquire a highly visible, business critical application. The application delivery function must take the proactive steps this section will describe. These steps protect both the company’s investment in the application as well as the political capital of the application delivery function.

Hope is not a strategy. Successful application delivery requires careful planning, coupled with extensive measurements and effective proactive and reactive processes.

Planning Functionality

Many planning functions are critical to the success of application delivery. They include the ability to:

- Profile an application prior to deploying it, including running it over a WAN simulator to replicate the performance experienced in branch offices.
- Baseline the performance of the network.
- Perform a pre-deployment assessment of the IT infrastructure.
- Establish goals for the performance of the network and for at least some of the key applications that transit the network.
- Model the impact of deploying a new application.
- Identify the impact of a change to the network, the servers, or to an application.
- Create a network design that maximizes availability and minimizes latency.

- Create a data center architecture that maximizes the performance of all of the resources in the data center.
- Choose appropriate network technologies and services.
- Determine what functionality to perform internally and what functionality to acquire from a third party.

The Sourcing Decision

This section of the handbook will detail three of the primary planning functions. Those functions being:

- Profiling an application prior to deploying it.
- Baselining the performance of the network and key applications.
- Assessing the infrastructure prior to deploying a key new application such as VoIP.

Kubernan asked 200 IT professionals if they prefer to perform these functions themselves or use a third party. Table 5.1 contains their responses.

Function	Perform it Themselves	Performed by 3rd Party	No Preference
Profiling an application prior to deploying it	79.5%	11.4%	9.1%
Baselining the performance of the network	79.6%	15.5%	5.0%
Baselining the performance of key applications	75.6%	13.9%	10.6%
Assessing the infrastructure prior to deploying a key new application such as VoIP	76.0%	16.8%	7.3%

Once conclusion that can be drawn from Table 5.1 is that between 75 and 80 percent of IT organizations prefer to perform the indicated planning functionality themselves. Conversely, another conclusion that can be drawn is that between 20 and 25 percent of IT organizations either prefer to have the indicated functionality performed by a third party or are receptive to that concept.

Application Profiling

Some factors that can impact application performance include WAN latency, jitter and packet loss. The effect of these factors on application performance varies widely, depending on the particular application. For example, email is the most common application the vast majority of companies use. Email is extremely tolerant to reasonably high levels of latency, jitter and packet loss.

One purpose of application profiling is to quantify how an application will perform when faced with WAN latency, jitter or packet loss. Table 5.2, for example, depicts the results of a lab test that was done to quantify the affect that WAN latency has on an inquiry-response application that has a target response time of 5 seconds. Similar tests can be run to quantify the affect that jitter and packet loss have on an application.

application server or the database server. As network latency is increased up to 450 ms., it has little impact on the application's response time. If network latency is increased above 450 ms, the response time of the application increases rapidly and is quickly well above the target response time.

In addition to identifying the impact of WAN latency, jitter and packet loss, another purpose of application profiling is to both characterize the information flow within the application and to quantify the processing time spent within each component of an application. As part of characterizing the information flow, IT organizations should quantify the *chattiness factor* (how many application turns for a given transaction) of the application.

To put application profiling in context, refer to the IT organization Section 2 of this report referenced. That IT organization had deployed an application on the east coast of the United States that users around the world accessed. Users in the Pac Rim were complaining about unacceptable application performance. In response to the complaints from the end users, the IT organization performed tests on the application to quantify the impact of WAN latency, similar to the tests reflected in Table 5.2. As a result of these tests, the IT organization:

- Validated that the application was performing poorly.
- Was able to estimate the improvement that they could make to the performance of the application by changing the WAN design.

However, the IT organization did not have a complete profile of the application. In particular, the company did not test the application in such a way as to characterize the information flow within the application or the processing time spent within each component of the application. As such, the company could not estimate whether or not implementing any of the network and application optimization techniques discussed in section 6 would be beneficial. The IT organi-

Table 5.2: Impact of Latency on Application Performance

Network Latency	Measured Response Time
0 ms	2 seconds
100 ms	2 seconds
150 ms	2 seconds
250 ms	2 seconds
350 ms	4 seconds
450 ms	4 seconds
500 ms	12 seconds

As Table 5.2 shows, if there is no WAN latency the application has a two-second response time. This two-second response time is well within the target response time and most likely represents the time spent in the

zation was also not in a position to estimate the affect of hosting the application somewhere in the Pac Rim.

The ability to understand how to optimally improve the performance of an application requires a complete profile of that application.

To put the actions of this IT organization into a larger context, Kubernan recently surveyed IT professionals relative to whether they quantify the impact of network parameters (i.e., loss, delay, jitter) on the performance of an application. We found:

- 42% of IT organizations currently perform this task.
- 65% of the companies that perform this task claim that they do it well.

The application profiling this IT organization performed is reactive. That means the organization profiled the application only after users complained about its performance. Some IT organizations profile applications more proactively. Some IT organizations, for example, profile an application shortly before they deploy it. The advantages of this approach are that it helps the IT organization:

- Identify minor changes that can be made to the application that will improve its performance.
- Determine if some form of optimization technology will improve the performance of the application.
- Identify the sensitivity of the application to parameters such as WAN latency and use this information to set effective thresholds.
- Gather information on the performance of the application that can be used to set the expectations of the users.
- Learn about the factors that influence how well an application will run over a WAN.

Since companies perform these tests just before they put the application into production, this is usually too late

to make any major change. The Automotive Network Engineer provided insight into the limitations of testing an application just prior to deployment. He stated that relative to testing applications prior to putting them into production, “We are required to go through a lot of hoops.” He went on to say that sometimes the testing was helpful, but that if the application development organization was under a lot of management pressure to get the application into production, that the application development organization often took the approach of deploying the application and then dealing with the performance problems later.

While the value of testing an application just prior to placing it into production has some limitations, it does position the application delivery function to be able to set expectations about how well the application will perform and to set the type of performance alarms that will be described in section 7.6. When done over a long enough period of time, application profiling enables both the application development and the application delivery functions to learn what type of factors cause an application to run poorly over a WAN.

An even more proactive approach is to involve the application delivery function at each stage of the application development lifecycle. Based both on its knowledge of what causes an application to perform badly over a WAN and on its ability to profile development versions of the application, the application delivery function can identify issues with the application early enough in the development cycle that major changes can be made.

The application delivery function needs to be involved early in the applications development cycle.

The Engineering CIO said that if his organization is acquiring an application from a third party, it talks to other users of the application, in part to get an understanding of how well the application performs. He went on to say that for many of the applications that it develops on its own, it

tests the application prior to deployment by putting it onto the production network and identifying how well it performs. He pointed out that his organization was an early adopter of some of the network and application optimization techniques section 6.2 of this report describes. As a result, it often tests the performance of application two ways – once with these techniques turned on and once with them turned off. This approach enables his organization to better understand the role that these techniques play relative to application delivery.

The Team Leader stated that one role of his group is to offer help with the design of key applications. He stated that his organization has typically profiled an application reactively. By that he meant that it profiled an application after it was deployed and failed to perform to expectations. His organization is becoming more proactive, in part by profiling applications before they are put into production.

The Team Leader also pointed out that the decision to profile an application prior to deployment depends on the size and the impact of the application. He said that overall, the number of times that they profile an application is increasing. As part of its standard approach to application profiling, they try to determine the type of sites in which the application will perform well and the type of sites in which it will not perform well. This approach often results in their identifying the need to make a change to the infrastructure or to the configuration of an application; i.e., to turn on caching. A second approach that his organization takes is to test possible ways to improve the performance of an application. As the Team Leader pointed out, however, this can be costly. He gave an example of an application in which it took two members of his organization three months to test the various techniques the organization could use to improve the application's performance.

The Consulting Architect pointed out that his organization is creating an architecture function. A large part of the motivation for the creation of this function is to remove the finger pointing that goes on between the network and

the application-development organizations. One goal of the architecture function is to strike a balance between application development and application delivery. For example, there might be good business and technical factors drive the application development function to develop an application using chatty protocols. One role of the architecture group is to identify the effect of that decision on the application-delivery function and to suggest solutions. For example, does the decision to use chatty protocols mean that additional optimization solutions would have to be deployed in the infrastructure? If so, how well will the application run if an organization deploys these optimization solutions? What additional management and security issues do these solutions introduce?

A primary way to balance the requirements and capabilities of the application development and the application-delivery functions is to create an effective architecture that integrates those two functions.

Baselining

Introduction

Baselining provides a reference from which service quality and application delivery effectiveness can be measured. It does this by quantifying the key characteristics (e.g., response time, utilization, delay) of applications and various IT resources including servers, WAN links and routers. Baselining allows an IT organization to understand the normal behavior of those applications and IT resources.

Baselining is an example of a task that one can regard as a building block of management functionality. That means baselining is a component of several key processes, such as performing a pre-assessment of the network prior to deploying an application (section 5.6) or performing proactive alarming (section 7.6).

Baselining also forms the basis of capacity planning. With that in mind, Kubernan recently surveyed IT profes-

sionals relative to both baselining network performance and performing capacity planning. The results of that market research indicated:

- Less than half (43%) of IT organization baseline their network.
- Just less than two-thirds of the companies that do baseline their network (63%) claim that they do it well.
- Half of IT organization performs capacity planning.
- Just less than two-thirds of the companies that perform capacity planning (64%) claim that they do it well.

The Team Leader stated that his organization does not baseline the company's entire global network. They have, however, widely deployed two tools that assist with baselining. One of these tools establishes trends relative to their traffic. The other tool baselines the end-to-end responsiveness of applications. The Team Leader has asked the vendor to link the two tools together so that he will know how much capacity he has left before the performance of a given application becomes unacceptable.

The Engineering CIO stated that his company uses a WAN service from a carrier that offers a service-level agreement relative to delay. As a result, this carrier provides reports that the IT organization uses to understand the delay characteristics of their WAN. In addition, his organization has just deployed a network and application-optimization solution that provides what the CIO referred to as rich reporting. By that he meant that these solutions, while designed primarily to optimize application performance, could in some cases provide input into both planning and ongoing management.

The Key Steps

Four primary steps comprise baselining. They are:

I. Identify the Key Resources

Most IT organizations do not have the ability to baseline all of their resources. These organization must determine which are the most important resources and baseline them. One way to determine which resources are the most important is to identify the company's key business applications and to identify the IT resources that support these applications. By definition, these are the most important IT resources.

II. Quantify the Utilization of the Assets over a Sufficient Period of Time

Organizations must compute the baseline over a normal business cycle. For example, the activity and responses times for a CRM application might be different at 8:00 a.m. on a Monday than at 8:00 p.m. on a Friday. In addition, the activity and response times for that CRM application are likely to differ greatly during a week in the middle of the quarter as compared with times during the last week of the quarter.

In most cases, baselining focuses on measuring the utilization of resources, such as WAN links. However, application performance is not directly tied to the utilization of WAN links. Application performance is tied directly to factors such as WAN delay. Since it is often easier to measure utilization than delay, many IT organizations set a limit on the maximum utilization of their WAN links hoping that this will result in acceptable WAN latency.

IT organizations need to modify their baselining activities to focus directly on delay.

III. Determine how the Organization Uses Assets

This step involves determining how the assets are being consumed by answering questions such as: Which applications are the most heavily used? Who is using those applications? How has the usage of those applications changed? In addition to being a key component of baselining, this step also positions the application-delivery function to provide the company's business and functional managers insight into how their organizations are changing based on how their use of key applications is changing.

IV. Utilize the Information

The information gained from baselining has many uses. This includes capacity planning, budget planning and chargeback. Another use for this information is to measure the performance of an application before and after a major change, such as a server upgrade, a network redesign or the implementation of a patch. For example, assume that a company is going to upgrade all of its Web servers. To ensure they get all of the benefits they expect from that upgrade, that company should measure key parameters both before and after the upgrade. Those parameters include WAN and server delay as well as the end-to-end application response time as experienced by the users.

Kubernan recently surveyed IT professionals relative to measuring the performance of an application before and after a major change. The market research indicated:

- 41% of IT organizations perform this task.
- 57% of the companies that perform this task claim that they do it well,

An IT organization can approach baselining in multiple ways. Sampling and synthetic approaches to baselining can leave a number of gaps in the data and have the poten-

tial to miss important behavior that is both infrequent and anomalous.

Organizations should baseline by measuring 100% of the actual traffic from the real users.

Selection Criteria

The following is a set of criteria that IT organizations can use to choose a baselining solution. For simplicity, the criteria are focused on baselining applications and not other IT resources.

Application Monitoring

To what degree (complete, partial, none) can the solution identify:

- Well-known applications; e.g., e-mail, VoIP, Oracle, PeopleSoft.
- Custom applications.
- Complex applications; e.g., Microsoft Exchange, SAP R/3, Citrix.
- Web-based applications, including URL-by-URL tracking.
- Peer-to-peer applications.
- Unknown applications.

Application Profiling and Response Time Analysis

Can the solution:

- Provide response time metrics based on synthetic traffic generation?
- Provide response time metrics based on monitoring actual traffic?
- Relate application response time to network activity?
- Provide application baselines and trending?

Pre-Deployment Assessment

The goal of performing a pre-deployment assessment of the current environment is to identify any potential problems that might affect an IT organization's ability to deploy an application. One of the two key questions that an organization must answer during pre-deployment assessment is: Can the network provide appropriate levels of security to protect against attacks? As part of a security assessment, it is important review the network and the attached devices and to document the existing security functionality such as IDS (Intrusion Detection System), IPS (Intrusion Prevention System) and NAC (Network Access Control). The next step is to analyze the configuration of the network elements to determine if any of them pose a security risk. It is also necessary to test the network to see how it responds to potential security threats.

The second key question that an organization must answer during pre-deployment assessment is: Can the network provide the necessary levels of availability and performance? As section 5.1 pointed out, it is extremely difficult to answer questions like this if the IT organization does not have targets for application performance that are well understood and adhered to. It is also difficult to answer this question, because as section 4 described, the typical application environment is both complex and dynamic.

Organizations should not look at the process of performing a pre-deployment network assessment in isolation. Rather, they should consider it part of an application-lifecycle management process that includes a comprehensive assessment and analysis of the existing network; the development of a thorough rollout plan including: the profiling of the application; the identification of the impact of implementing the application; and the establishment of effective processes for ongoing fact-based data management.

The Team Leader stated his organization determines whether to perform a network assessment prior to deploy-

ing a new application on a case-by-case basis. In particular, he pointed out that it tends to perform an assessment if it is a large deployment or if it has some concerns about whether the infrastructure can support the application. To assist with this function, his organization has recently acquired tools that can help it with tasks such as assessing the ability of the infrastructure to support VoIP deployment as well as evaluating the design of their MPLS network.

The Engineering CIO said that the organization is deploying VoIP. As part of that deployment, it did an assessment of the ability of the infrastructure to support VoIP. The assessment was comprised of an analysis using an excel spreadsheet. The organization identified the network capacity at each office, the current utilization of that capacity and the added load that would come from deploying VoIP. Based on this set of information, it determined where it needed to add capacity.

The key components of a pre-deployment network assessment are:

Create an inventory of the applications running on the network

This includes discovering the applications that are running on the network. Section 7.4 will discuss this task in greater detail.

In addition to identifying the applications that are running on the network, it is also important to categorize those applications using an approach similar to what section 4.2 described. Part of the value of this activity is to identify recreational use of the network; i.e., on-line gaming and streaming radio or video. Blocking this recreational use can free up additional WAN bandwidth. Section 7.4 quantifies the extent to which corporate networks are carrying recreational traffic.

Another part of the value of this activity is to identify business activities, such as downloads of server patches or security patches to desktops that are being performed during peak times. Moving these activities to an off-peak time adds additional bandwidth.

Evaluate bandwidth to ensure available capacity for new applications

This activity involves baselining the network as described in section 5.5. The goal is to use the information about how the utilization of the relevant network resources has been trending to identify if any parts of the network need to be upgraded to support the new application.

As section 5.5 described, baselining often refers to measuring the utilization of key IT resources. That section also contained the recommendation that companies modify how they think about baselining to focus not on utilization, but on delay. In some instances, however, IT organizations need to measure more than just delay. If a company is about to deploy VoIP, for example, then the pre-assessment baseline must also measure the current levels of jitter and packet loss, as VoIP quality is highly sensitive to those parameters.

Create response time baselines for key essential applications

This activity involves measuring the average and peak application response times for key applications both before and after the new application is deployed. This data will allow IT organizations to determine if deploying the new application causes an unacceptable impact on the company's other key applications.

As part of performing a pre-deployment network assessment, IT organizations can typically rely on having access

to management data from SNMP MIBs (Simple Network Management Protocol Management Information Bases) on network devices, such as switches and routers. This data source provides data link layer visibility across the entire enterprise network and captures parameters, such as the number of packets sent and received, the number of packets that are discarded, as well as the overall link utilization.

NetFlow is a Cisco IOS software feature and also the name of a Cisco protocol for collecting IP traffic information. Within NetFlow, a network flow is defined as a unidirectional sequence of packets between a given source and destination. The branch office router outputs a flow record after it determines that the flow is finished. This record contains information, such as timestamps for the flow start and finish time, the volume of traffic in the flow, and its source and destination IP addresses and source and destination port numbers.

NetFlow represents a more advanced source of management data than SNMP MIBs. For example, whereas data from standard SNMP MIB monitoring can be used to quantify overall link utilization, this class of management data can be used to identify which network users or applications are consuming the bandwidth.

The IETF is in the final stages of approving a standard (RFC 3917) for logging IP packets as they flow through a router, switch or other networking device and reporting that information to network management and accounting systems. This new standard, which is referred to as IPFIX (IP Flow Information EXport), is based on NetFlow Version 9.

An important consideration for IT organizations is whether they should deploy vendor-specific, packet inspection-based dedicated instrumentation. The advantage of deploying dedicated instrumentation is that it enables a more detailed view into application performance. The disadvantage of this approach is that it increases the cost of the solution. A compromise is to rely on data from SNMP

MIBs and NetFlow in small sites and to augment this with dedicated instrumentation in larger, more strategic sites.

Whereas gaining access to management data is relatively easy, collecting and analyzing details on every application in the network is challenging. It is difficult, for example, to identify every IP application, host and conversation on the network as well as applications that use protocols such as IPX or DECnet. It is also difficult to quantify application response time and to identify the individual sources of delay; i.e., network, application server, database. One of the most challenging components of this activity is to unify this information so the organization can leverage it to support myriad activities associated with managing application delivery.

6.0 Network and Application Optimization

Introduction

The phrase *network and application optimization* refers to an extensive set of techniques that organizations have developed in an attempt to optimize the performance of networks and applications as part of assuring acceptable application performance. The primary role that these techniques play is to:

- Reduce the amount of data that is sent over the WAN.
- Ensure that the WAN link is never idle if there is data to send.
- Reduce the number of round trips (a.k.a., transport layer or application turns) that are necessary for a given transaction.
- Offload computationally intensive tasks from client systems and servers

There are two principal categories of network and application optimization. One category focuses on the nega-

tive effect of the WAN on application performance. This category will be referred to in this report as Branch Office Optimization Solutions. In most cases, the Branch Office Optimization Solutions are referred to as *symmetric solutions* because they require an appliance in both the data center as well as the branch office. Some vendors, however, have implemented solutions that call for an appliance in the data center, but instead of requiring an appliance in the branch office only requires software on the user's computer. This class of solution is often referred to as a *software only solution* and is most appropriate for individual users or small offices.

The second category is often referred to as an Application Front End (AFE) or Application Device Controller (ADC). This solution is typically referred to as being an *asymmetric solution* because an appliance is only required in the data center and not the branch office. The genesis of this category of solution dates back to the IBM mainframe-computing model of the late 1960s and early 1970s. Part of that computing model was to have a Front End Processor (FEP) reside in front of the IBM mainframe. The primary role of the FEP was to free up processing power on the general purpose mainframe computer by performing communications processing tasks, such as terminating the 9600 baud multi-point private lines, in a device that was designed just for these tasks. The role of the AFE is somewhat similar to that of the FEP in that the AFE performs computationally intensive tasks, such as the processing of SSL traffic, and hence frees up server resources. However, another role of the AFE is to function as a Server Load Balancer (SLB) and, as the name implies, balance traffic over multiple servers.

Companies deploy Branch Office Optimization Solutions and AFEs in different ways. The typical company, for example, has many more branch offices than data centers. Hence, the question of whether to deploy a solution in a limited tactical manner vs. a broader strategic manner applies more to Branch Office Optimization Solutions than

it does to AFEs. Also, AFEs are based on open standards and as a result a company can deploy AFEs from different vendors and not be concerned about interoperability. In contrast, Branch Office Optimization Solutions are based on proprietary technologies and so a company would tend to choose a single vendor from which to acquire these solutions.

Kubernan recently surveyed IT professionals relative to whether they quantify the impact of optimization (i.e., caching, compression, protocol acceleration) on the application. We found that:

- 36% of IT organizations currently perform this task.
- 60% of the companies that perform this task claim that they do it well.

Alice in Wonderland Revisited

The previous section of this report dealt with the planning component of application delivery. That section began with a reference to *Alice in Wonderland* and discussed the need for IT organizations to set a direction for things such as application performance. That same reference to *Alice in Wonderland* applies to the network and application optimization component of application delivery. In particular, no network and application optimization solution on the market solves all possible application performance issues.

To deploy the appropriate network and application optimization solution, IT organizations need to understand the problem they are trying to solve.

Section 4 of this report described some of the characteristics of a generic application environment and pointed out that to choose an appropriate solution, IT organizations need to understand their unique application environment. In the context of network and application optimization, if the company either already has or plans to consolidate servers out of branch offices and into centralized data

centers, then as described later in this section, a WAFS (Wide Area File Services) solution might be appropriate. If the company is implementing VoIP, then any Branch Office Optimization Solution that it implements must be able to support traffic that is both real-time and meshed. Analogously, if the company is making heavy use of SSL, it might make sense to implement an AFE to relieve the servers of the burden of processing the SSL traffic.

In addition to high-level factors of the type the preceding paragraph mentioned, the company's actual traffic patterns also have a significant impact on how much value a network and application optimization solution will provide. To exemplify this, consider the types of advanced compression most solution providers offer. The effectiveness of advanced compression depends on two factors. One factor is the quality of the compression techniques that have been implemented in a solution. Since many compression techniques use the same fundamental and widely known mathematical and algorithmic foundations, the performance of many of the solutions available in the market will tend to be somewhat similar.

The second factor that influences the effectiveness of advanced compression solutions is the amount of redundancy of the traffic. Applications that transfer data with a lot of redundancy, such as text and html on web pages, will benefit significantly from advanced compression. Applications that transfer data that has already been compressed, such as the voice streams in VoIP or jpg-formatted images, will see little improvement in performance from implementing advanced compression and could possibly see performance degradation. However, other techniques, such as TCP optimization, can improve performance even when advanced compression cannot.

Because a network and optimization solution will provide varying degrees of benefit to a company based on the unique characteristics of its environment, third party tests of these solutions are helpful, but not conclusive.

In order to understand the performance gains of any network and application optimization solution, that solution must be tested in an environment that closely reflects the environment in which it will be deployed.

Packet Loss	Congestion Control Forward Error Correction (FEC)
Network Contention	Quality of Service (QoS)

Below is a brief description of some of the principal WAN optimization techniques.

Branch Office Optimization Solutions

Background

The goal of Branch Office Optimization Solutions is to improve the performance of applications delivered from the data center to the branch office or directly to the end user. If organizations can improve the performance of these applications enough, it might well be possible to centralize some or all of the existing branch office IT infrastructure; e.g., file servers, email servers, SMS servers, and local tape backup equipment.

Myriad techniques comprise branch office optimization solutions. Table 6.1 lists some of these techniques and indicates how organizations can use each of these techniques to overcome some characteristic of the WAN that impairs application performance.

Table 6.1: Techniques to Improve Application Performance

WAN Characteristic	WAN Optimization Techniques
Insufficient Bandwidth	Data Reduction: <ul style="list-style-type: none"> • Data Compression • Differencing (a.k.a., de-duplication) • Caching
High Latency	Protocol Acceleration: <ul style="list-style-type: none"> • TCP • HTTP • Mitigate Round-trip Time • Request Prediction • Response Spoofing • CIFS • NFS • MAPI Mitigate Round-trip Time <ul style="list-style-type: none"> • Request Prediction • Response Spoofing

Caching

This refers to keeping a local copy of information with the goal of either avoiding or minimizing the number of times that information must be accessed from a remote site. If caching is done at the object level (e.g., file, web object, or email attachment) then there is always a risk that the local copy is not identical to the actual object being requested from the central server.

Compression

The role of compression is to reduce the size of a file prior to transmitting that file over a WAN.

Congestion Control

The goal of congestion control is to ensure that the sending device does not transmit more data than the network can accommodate. To achieve this goal, the TCP congestion control mechanisms are based on a parameter referred to as the congestion window. TCP has multiple mechanisms to determine the *congestion window*.

Differencing; a.k.a., De-duplication

The goal of differencing is to avoid sending an entire file or data stream from origin to destination. In particular, the goal of differencing is to send only the changes that have been made to the file or data stream since the last time it was sent.

Forward Error Correction (FEC)

FEC is typically used at the physical layer (Layer 1) of the OSI stack. FEC can also be applied at the

network layer (Layer 3) whereby an extra packet is transmitted for every n packets sent. This extra packet is used to recover from an error and hence avoid having to retransmit packets.

TCP Acceleration

TCP acceleration involves a range of techniques, including simple steps such as increasing the TCP window size. The TCP window size refers to the number of packets that can be sent without receiving an acknowledgment. Other techniques include connection pooling, limited and fast re-transmits, and support for high speed TCP. The goal of these techniques is to make TCP perform better in a wide range of high latency environments.

HTTP Acceleration

This involves techniques such as object pre-fetching, HTTP pipelining as well as the caching of static Web pages to improve the performance of HTTP.

Quality of Service (QoS)

QoS refers to the ability of the network to use functionality such as queuing to provide preferential treatment to certain classes of traffic, such as voice traffic. Some QoS solutions allocate bandwidth, while others prioritize the packets in the queue. Some solutions do both independently.

Request Prediction

By understanding the semantics of specific protocols or applications, it is often possible to anticipate a request a user will make in the near future. Making this request in advance of it being needed eliminates virtually all of the delay when the user actually makes the request.

Many applications or application protocols have a wide range of request types that reflect different user actions or use cases. It is important to understand what a vendor means when it says it has a certain application level optimization. For example, in the CIFS (Windows file sharing) protocol, the simplest interactions that can be optimized involve *drag and drop*. But many other interactions are more complex. Not all vendors support the entire range of CIFS optimizations.

Request Spoofing

This refers to situations in which a client makes a request of a distant server, but the request is responded to locally.

In some cases, one of the techniques listed above might be the entire solution. For example, a company that ships large files between the United States and India could decide to put an appliance that only does compression on each end of that link. In many cases, however, techniques like the ones listed above are combined into a broader based solution. For example, as discussed in section 4.6, companies that consolidate servers out of branch offices and into a centralized data center end up running CIFS over their WAN. Since CIFS is a chatty protocol, this can result in poor performance. To compensate, many vendors have deployed multi-function WAN optimization solutions. These solutions typically implement myriad technologies, such as CIFS and NFS acceleration, caching, compression, differencing, request prediction and spoofing, as well as security functionality, such as encryption.

Tactical vs. Strategic Solutions

To put the question of tactical vs. strategic in context, refer again to the IT organization that Section 2 of this report referenced. For that company to identify the problem that it is trying to solve, it must answer questions such as: Is the problem just the performance of this one applica-

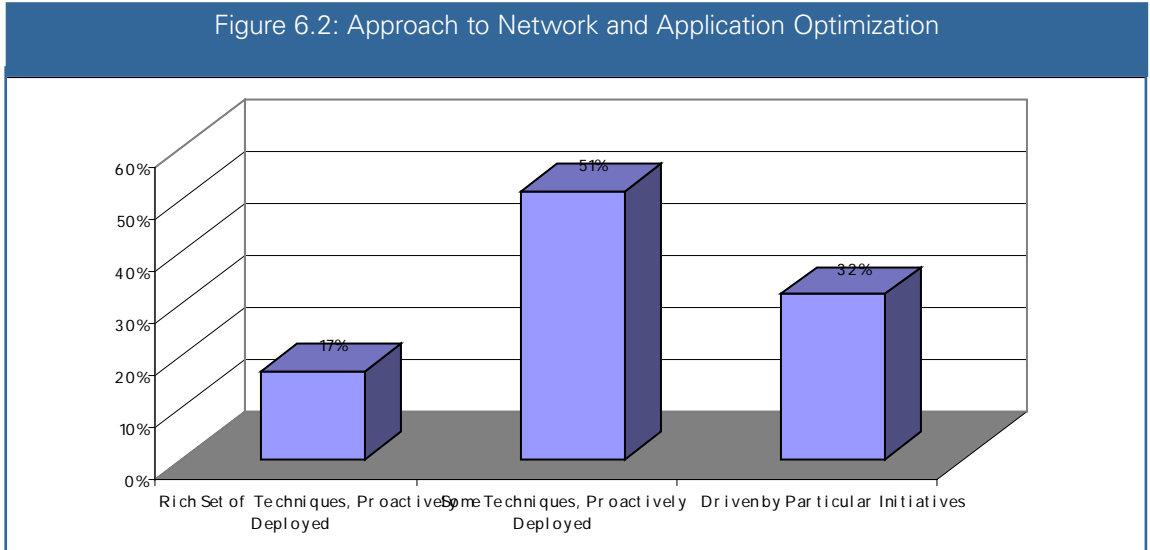
tion as used just by employees in the Pac Rim? If that is the problem statement, then the company is looking for a very tactical solution. However, the company might decide that the problem that it wants to solve is how can it guarantee the performance of all of their critical applications for all of its employees under as wide a range of circumstances as possible. In this case, the company needs a strategic solution.

Historically, Branch Office Optimization Solutions have been implemented in a tactical fashion. That means that companies have deployed the least amount of equipment possible to solve a specific problem. Kubernan recently asked several hundred IT professionals about the tactical vs. strategic nature of how they use these techniques. Their answers, which Figure 6.2 shows, indicate the deployment of these techniques is becoming a little more strategic.

The Electronics COO who noted that his company's initial deployment of network and application optimization techniques was to solve a particular problem supports that position. He also stated that his company is "absolutely becoming more proactive moving forward with deploying these techniques."

Similarly, The Motion Picture Architect commented that his organization has been looking at these technologies for a number of years, but has only deployed products to solve some specific problems, such as moving extremely large files over long distances. He noted that his organization now wants to deploy products proactively to solve a broader range of issues relative to application performance.

According to The Motion Picture Architect, "Even a well written application does not run well over long distances. In order to run well, the application needs to be very thin and it is very difficult to write a full featured application that is very thin."



Current Deployments

Table 6.3 depicts the deployment status of some of the primary branch office optimization techniques.

One conclusion that can be drawn from the data in Table 6.3 is that IT organizations have plans to deploy a wide range of branch office optimization techniques.

The Team Leader pointed out that his company has historically made significant use of satellite links and, as a result, they have been deploying some form of optimization appliance for about 10 years. One of the primary differences he said he sees in the current generation of optimization appliances is that they provide a broader range of functionality than previous generations of these appliances provided.

The Engineering CIO stated that his organization originally deployed a WAFS solution to alleviate redundant file copy. He said he has been pleasantly surprised by the

Table 6.3 : Deployment of Branch Office Optimization Techniques

	No plans to deploy	Have not deployed, but plan to deploy	Deployed in test mode	Limited production deployment	Broadly deployed
Compression	27%	22%	9%	25%	17%
Caching	31%	19%	11%	19%	20%
HTTP acceleration	37%	20%	8%	18%	17%
TCP acceleration	37%	24%	13%	17%	9%
Wide Area File Services (WAFS)	51%	23%	9%	11%	5%

additional benefits of using the solution. In addition, his organization plans on doing more backup of files over the network and he expects the WAFS solution they have already deployed will assist with this.

The points The Engineering CIO raised go back to the previous discussion of a tactical vs. a strategic solution. In particular, most IT organizations that deploy a network and application optimization solution do so tactically and later expand the use of that solution to be more strategic.

When choosing a network and application optimization solution it is important to ensure that the solution can scale to provide additional functionality over what is initially required.

Selection Criteria

Below is a set of criteria that IT organizations can use to select a Branch Office Optimization Solution.

Performance

Third party tests of a solution can be helpful. It is critical, however, to quantify the kind of performance gains that the solution will provide in the particular environment where it will be installed. As part of this quantification, it is important to identify if the performance degrades as either additional functionality within the solution is activated or if

the solution is deployed more broadly across the organization.

Transparency

It should be possible to deploy the solution and not have anything such as routing, security or QoS

break. The solution should also be transparent relative to both the existing server configurations and the existing Authentication, Authorization and Accounting (AAA) systems. In addition, the solution should not make troubleshooting any more difficult.

Solution Architecture

If the organization intends the solution to be able to support additional optimization functionality over time, it is important to determine if the hardware and software architecture can support new functionality without an unacceptable loss of performance.

OSI Layer

Organizations can apply many of these techniques at various layers of the OSI model. They can apply compression, for example, at the packet layer. The advantage of applying compression at this layer is that it supports all transport protocols and all applications. The disadvantage is that it cannot directly address any issues that occur higher in the stack.

Alternatively, having an understanding of the semantics of the application means that compression can also be applied to the application; e.g., SAP or Oracle. Applying compression, or other

techniques such as request prediction, in this manner has the potential to be more effective.

Capability to Perform Application Monitoring

Some solutions provide the ability to monitor the end-to-end response time of an n-tier application or the Mean Opinion Score for VoIP traffic. Section 7.7 will explain this capability in more detail. Ideally, these solutions also provide the capability to isolate the source of performance problems.

Alternatively, some solutions work well with third party monitoring tools meaning in part that the solution does not negate the ability of these third party tools to see critical data.

Scalable

One aspect of scalability is the size of the WAN link that can be terminated on the appliance. More important is how much throughput the box can actually support with the relevant and desired optimization functionality turned on. Other aspects of scalability include how many simultaneous TCP connections the appliance can support as well as how many branches or users a vendor's complete solution can support.

Downward scalability is also important. Downward scalability refers to the ability of the vendor to offer cost effective products for small branches or even individual laptops.

Cost effective

This criterion is related to scalability. In particular, it is important to understand what the initial solution costs. It is also important to understand how the cost of the solution changes as the scope and scale of the deployment increases.

Application Sub-classification

As mentioned in section 4.2, an application such as Citrix or SAP is comprised of multiple modules with varying characteristics. Some Branch Office Optimization Solutions can classify at the individual module level, while others can only classify at the application level.

Module vs. Application Optimization

In line with the previous criterion, some Branch Office Optimization Solutions treat each module of an application in the same fashion. Other solutions treat modules based both on the criticality and characteristics of that module. For example, some solutions apply the same optimization techniques to all of SAP, while other solutions would apply different techniques to the individual SAP modules based on factors such as their business importance and latency sensitivity.

Disk vs. RAM

Advanced compression solutions can be either disk- or RAM-based. Disk-based systems typically can store as much as 1,000 times the volume of patterns in their dictionaries as compared with RAM-based systems, and those dictionaries can persist across power failures. The data, however, is slower to access than it would be with the typical RAM-based implementations, although the performance gains of a disk-based system are likely to more than compensate for this extra delay. While disks are more cost-effective than a RAM-based solution on a per byte basis, given the size of these systems they do add to the overall cost and introduce additional points of failure to a solution. Standard techniques such as RAID can mitigate the risk associated with these points of failure.

Protocol support

Some solutions are specifically designed to support a given protocol (e.g., UDP, TCP, HTTP, CIFS, MAPI) while other solutions support that protocol generically. In either case, the critical issue is how much of an improvement in the performance of that protocol the solution can cause in the type of environment in which the solution will be deployed.

It is also important to understand if the solution makes any modifications to the protocol that could cause unwanted side effects.

Security

The solution must not break the current security environment, such as breaking firewall Access Control Lists (ACLs) by hiding TCP header information. In addition, the solution itself must not create any additional security vulnerabilities.

Easy of Deployment and Management

As part of deploying a Branch Office Optimization Solution, an appliance needs to be deployed in branch offices that will most likely not have any IT staff. As such, it is important that unskilled personnel can install the solution. In addition, the greater the number of appliances deployed, the more important it is that they are easy to configure and manage.

It's also important to consider what other systems will have to be touched to implement the Branch Office Optimization Solution. Some solutions, especially cache-based or WAFS solutions, require that every file server be accessed during implementation.

Change Management

Since most networks experience periodic changes such as the addition of new sites or new applica-

tions, it is important that the Branch Office Optimization Solution can adapt to these changes easily. It is preferable if the Branch Office Optimization Solution can adjust to these changes automatically.

Support of Meshed Traffic

As section 4.3 mentioned, a number of factors are causing a shift in the flow of WAN traffic away from a simple hub-and-spoke pattern and to more of a meshed flow. If a company is making this transition, it is important that the Branch Office Optimization Solution that they deploy can support meshed traffic flows and can support a range of features such as asymmetric routing.

Support for Real Time Traffic

As section 4.2 mentioned, many companies have deployed real-time applications. For these companies it is important that the Branch Office Optimization Solution can support real time traffic.

Some real time traffic like VOIP and live video can't be accelerated because it is real time and already highly compressed. Header compression might be helpful for VoIP traffic and most real time traffic will benefit from QoS.

Link by link vs. a Global Solution

The distinction between these two approaches is the amount of automated and intelligent coordination that exists between the individual components of the solution.

In a link-by-link solution, the coordination between the component parts is manual. A global solution implements extensive automation, with real-time measurements of application performance communicated among the system components. The goal of the global approach is to enable the solu-

tion to make continuous, dynamic adjustments to the performance of the company's applications.

Application Front Ends (AFEs)

Background

As section 6.1 mentions, an historical precedent exists to the current generation of AFEs (a.k.a., ADCs). That precedent is the Front End Processor (FEP) that was introduced in the late 1960s and was developed and deployed in order to support mainframe computing. From a more contemporary perspective, the current generation of AFEs evolved from the earlier generations of Server Load Balancers (SLBs) that were deployed in front of server farms.

While an AFE still functions as a SLB, the AFE has assumed, and will most likely continue to assume, additional roles. AFEs, for example, serve to both manage application traffic within the data center, as well as accelerate the delivery of applications asymmetrically from the data center to individual remote users. One of the primary new roles played by an AFE is to offload processing-intensive tasks that are network related and that consume considerable CPU cycles. As previously mentioned, an example of server offload is the SSL processing in the data center. SSL offload allows the intranet Web and Internet eCommerce servers to process more requests for content and handle more transactions. This technique provides a significant increase in the performance of these secure sites without adding additional server capacity. AFEs are also beginning to incorporate security services. This seems to be reasonable given that AFEs are deployed in the data center network behind the firewalls and in front of the company's servers.

An AFE provides more sophisticated functionality than a SLB does.

For example, to balance traffic across multiple servers an AFE or a traditional SLB merely has to inspect each packet and intelligently route the packet to the appropriate

server. However, in order to provide functionality such as SSL offload, an AFE has to terminate the SSL connection and act as a proxy for the server.

In summary, one or more of the following factors motivates IT organizations that are looking to deploy an AFE:

- I. Maximize the efficiency of the company's servers by offloading processing-intensive tasks.
- II. Maximize the efficiency of the company's servers through intelligent sending of application requests to the most appropriate server based on Layer 4 or Layer 7 factors, such as server load or application type.
- III. Maximize application availability and performance by sending requests to available servers and avoiding re-routing of the requests away from unavailable servers.
- IV. Accelerate the performance of applications delivered over the WAN by implementing techniques such as compressing, caching and TCP optimization.
- V. Provide incremental application-layer security to maximize uptime for servers and applications and to better ensure the integrity of the company's data.

To provide the functionality listed above, AFEs implement some of the optimization techniques section 6.2.1 described. AFEs, however, also implement functionality not found in a Branch Office Optimization Solution. An example of this additional functionality is reverse caching. Reverse caching refers to having a cache inside of the AFE, the role of which is to accelerate the delivery of Web pages. All Web pages sent from a server back to the user's browser move through and are stored in the reverse cache. If the next request for a Web page has already been stored in the cache, it is retrieved from the cache. This minimizes the involvement of the servers in rendering Web pages to users.

Another example of the additional functionality found within an AFE is TCP session management and multiplexing. This functionality is designed to deal with the complexity associated with the fact that each object on a Web page requires its own short-lived TCP connection. Processing all of these connections can consume an inordinate amount of the server's CPU resources. Instead, the AFE manages all of the short-lived TCP connects and presents to the server a single long-lived connection.

Current Deployments

Table 6.4 depicts the deployment status of both traditional SLBs as well as the more fully functional AFEs.

One conclusion that can be drawn from Table 6.4 is that that deployment of AFEs should increase significantly.

The architecture of the AFE is an important selection criterion, as it will dictate the ability of the AFE to support:

- Increased traffic volume.
- Additional servers.
- New applications.

The AFE has to be able to easily integrate with other components of the data center such as the firewalls and Layer 3 switches. In some companies, for example, it is important to integrate the Layer 3 switch, the AFE and firewalls so that all three products function as one traffic stream at a point in the data center

The AFE must also support the range of applications that the IT organization is likely to deploy. Hence, another key

differentiator is the breadth of applications that the AFE can accelerate. All AFEs can accelerate Web-based applications. Some AFEs, however, can also accelerate other classes of applications.

Table 6.4: Deployment of SLBs and AFEs

	No plans to deploy	Have not deployed, but plan to deploy	Deployed in test mode	Limited production deployment	Broadly deployed
SLB	19.0%	12.3%	9.8%	25.8%	33.1%
AFE	22.1%	24.8%	6.7%	26.8%	19.5%

Selection Criterion

Many of the selection criteria for an AFE are the same as for a Branch Office Optimization Solution. This includes transparency, cost effectiveness, security and ease of management. Performance is also an important selection criterion and refers to how the solution will perform in the particular environment in which it will be deployed.

A selection criterion that is likely to grow in importance is virtualization. This refers to the ability of the AFE to leverage any service (e.g., server load balancing, acceleration, security) across any application on a context-by-context basis. One benefit of virtualization is a lowering of the cost of ownership within the data center due to a reduced need for servers and the associated power and cooling capability.

Some overlap exists between the advantages that result from deploying an AFE and the advantages that result from deploying a Branch Office Optimization Solution. Each solution, for example, conserves server resources. The AFE conserves server resources by offloading computationally intensive tasks from the servers. Many Branch Office Optimization Solutions conserve server resources by caching information in the branch offices and minimizing the number of times that the data center servers have to respond to a request for information.

In most cases which type of network and application optimization solution best meets the requirements is fairly clear. The choice of solution depends on factors such as:

- The range of applications: purely Web based or fat client?

- The location and type of the end users: Employees in a branch? Traveling Employees? Third party users?
- The need to consolidate branch infrastructure or just to accelerate application response times?

It is worth noting that conserving servers, an advantage of both solutions, has a number of positive side effects, such as reducing the number of software licenses, reducing the amount of maintenance that is required, and freeing up real estate in the data centers. Another side effect growing in importance is that as an IT organization reduces the number of servers that it has in its data centers, it also reduces the corresponding power requirements. For some IT organizations, that benefit is critical.

7.0 Management

Introduction

The primary management tasks associated with application delivery are to:

- Discover the applications running over the network and identify how they are being used.
- Gather the appropriate management data on the performance of the applications and the infrastructure that supports them.
- Provide end-to-end visibility into the ongoing performance of the applications and the infrastructure.
- Identify the sources of delay in the performance of the applications and the infrastructure.
- Automatically identify performance issues and resolve them.
- Gain visibility into the operational architecture and dynamic behavior of the network.

As section 2, mentioned Kubernan asked more than 300 IT professionals: “If the performance of one of your company’s key applications is beginning to degrade who notices it first? The end user or the IT organization?” Three-quarters of the survey respondents indicated that it was the end user.

IT organizations will not be considered successful with application delivery as long as the end user, and not the IT organization, first notices application degradation.

The Consulting Architect commented that, within his company, the end user and not the IT organization usually first finds application-performance issues. He stated that once a problem has been reported that identifying the root cause of the problem bounces around within the IT organization and that “It’s always assumed to be the network. Most of my job is defending the network.”

As part of that survey, Kubernan also asked the survey respondents to indicate what component of IT was the biggest cause of application degradation. Figure 7.1 summarizes their answers. In Figure 7.1, the answer *shared equally* means that multiple components of IT are equally likely to cause application degradation.

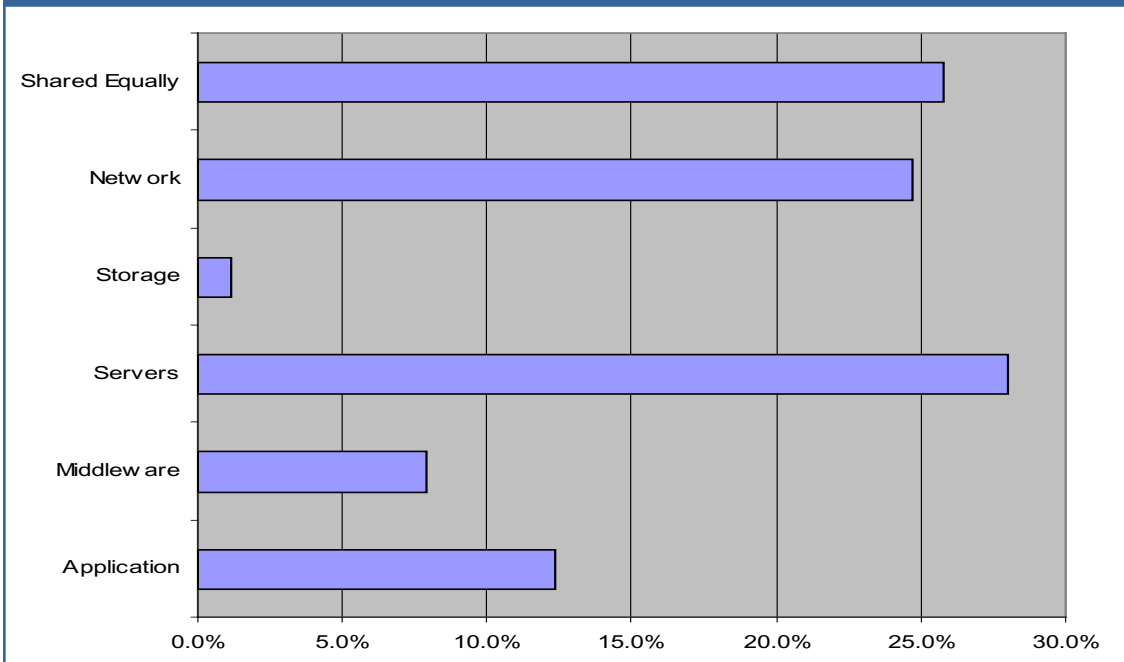
The data in Figure 7.1 speaks to the technical complexity associated with managing application performance.

When an application experiences degradation, virtually any component of IT could be the source of the problem.

The Organizational Complexity

To understand how IT organizations respond to application degradation, Kubernan asked several hundred IT professionals to identify which organization or organizations has responsibility for the ongoing performance of applications once they are in production. Table 7.1 contains their answers.

Figure 7.1: Causes of Application Degradation



Taken together with the data in Figure 7.1, managing application performance clearly is complex, both technically and organizationally.

The ASP Architect provided insight into the challenges of determining the source of an application-performance issue. He stated, “We used to have a real problem with identifying performance problems. We would have to run around with sniffers and other less friendly tools to trouble shoot problems. The finger pointing was often pretty

bad.” He went on to say that to do a better job of identifying performance problems the IT organization developed some of their own tools. The traditional IT infrastructure groups as well as by some of the application teams are using the tools that his organization developed. He went on to say that the reports generated by those tools helped to develop credibility for the networking organization with the applications-development organization.

Table 7.1: Organization Responsible for Application Performance

Group	Percentage of Respondents
Network Group – including the NOC	64.6%
Application development group	48.5%
Server group	45.1%
Storage group	20.9%
Application performance-management group	18.9%
Other	12.1%
No group	6.3%

The Team Leader pointed out that within his IT organization no single group has responsibility for application performance prior to the deployment of an application. Once the organization deploys the application, however, a group is responsible for the ongoing performance of that application.

The data in Table 7.1 speaks to the organizational complexity associated with managing application performance.

To be successful with application delivery, IT organizations need tools and processes that can identify the root cause of application degradation and which are accepted as valid by the entire IT organization.

The good news is that most IT organizations recognize the importance of managing application performance. In particular, research conducted by Kubernan indicates that in only 2% of IT organizations is managing application performance losing importance. In slightly over half of the IT

organizations, it is gaining in importance and keeping about the same importance in the rest of the IT organizations.

The Process Barriers

Kubernan also asked hundreds of IT professionals if their companies have a formalized set of processes for identifying and resolving application degradation. Table 7.2 contains their answers. The data in Table 7.2 clearly indicate that the majority of IT organizations either currently have processes, or soon will, to identify and resolve application degradation.

Table 7.2: Existence of Formalized Processes

Response	Percentage of Respondents
Yes, and we have had these processes for a while	22.4%
Yes, and we have recently developed these processes	13.3%
No, but we are in the process of developing these processes	31.0%
No	26.2%
Other	7.1%

The Infrastructure Engineering Manager said his IT organization does not have formalized processes for managing application performance, but that it is working on it. He explained it was motivated to develop these processes because application performance has become more of an issue recently in large part because the IT organization is increasingly hosting applications in a single data center, and having users from all of the world access those applications. As a result, the parameters of the WAN that impact application performance (i.e., delay, jitter, packet loss) are more pronounced than they would be if there were less distance between the user and the application.

Kubernan gave the same set of IT professionals a set of possible answers and asked them to choose the two most significant impediments to effective application delivery.

Table 7.3 shows the answers that received the highest percentage of responses.

Table 7.3: Impediments to Effective Application Delivery

Answer	Percentage of Companies
Our processes are inadequate	39.6%
The difficulty in explaining the causes of application degradation and getting any real buy-in	33.6%
Our tools are inadequate	31.5%
The application development group and the rest of IT have adversarial relations.	24.2%

The data in Table 7.3 indicates that three out of the top four impediments to effective application delivery have little to do with technology. The data in this table also provides additional insight to the data in Table 7.2. In particular, the data in Table 7.2 indicates that the vast majority of IT organization either have formalized processes for identifying and resolving application degradation, or are developing these processes. The data in Table 7.3, however, indicate that, in many cases, these processes are inadequate.

Organizational discord and ineffective processes are at least as much of an impediment to the successful management of application performance as are technology and tools.

The ASP Architect stated that the infrastructure component of the IT organization has worked hard to improve their processes in general, and improving its communications with the business units in particular. He pointed out that the infrastructure is now ISO certified and it is adopting an ITIL model for problem tracking. These improvements have greatly enhanced the reputation of the infrastructure organization, both within IT and between the infrastructure organization and the company's business units. It has reached the point that the applications-development groups have seen the benefits and are working,

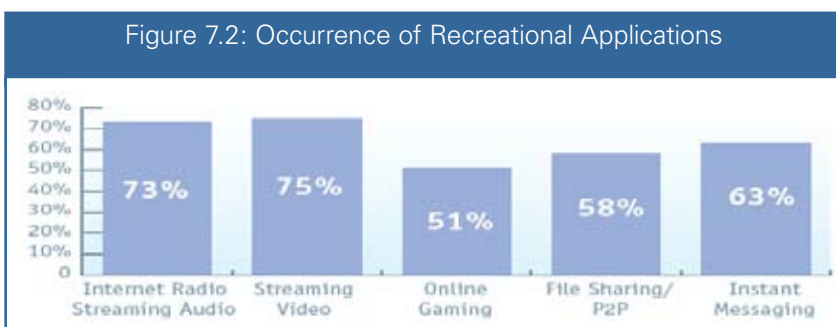
with the help of the infrastructure organization, to also become ISO certified.

Discovery

Section 5 of this report commented on the importance of identifying which applications are running on the network as part of performing a pre-deployment assessment. Due to the dynamic nature of IT, it is also important to identify which applications are running on the network on an ongoing basis.

Section 4 mentions one reason identifying which applications are running on the network on an ongoing basis is important: successful application delivery requires that IT organizations are able to eliminate the applications that are running on the network and have no business relevance.

To put this in context, Figure 7.2 shows a variety of recreational applications along with how prevalent they are. In a recent survey of IT professionals, 51 percent said they had seen unauthorized use of their company's network for applications such as Doom or online poker. Since IT professionals probably don't see all the instances of recreational traffic on their networks, the occurrence of recreational applications is likely higher than what Figure 7.2 reflects.



These recreational applications are typically not related to the ongoing operation of the enterprise and, in many cases, these applications consume a significant amount of bandwidth.

Kubernan recently surveyed IT professionals relative to both discovery in general, as well as identifying non-approved and inappropriate applications in particular. The results of that market research indicate:

- Just over half (55%) of IT organizations perform discovery.
- Only 42% of the IT organizations that perform discovery claim that they do it well.
- Only 41% of companies regularly attempt to identify non-approved and inappropriate applications.
- Less than two-thirds of the companies (61%) that regularly attempt to identify non-approved and inappropriate applications claim that they do it well.

The Team Leader stated his organization does not attempt to discover all of the applications that are running on its network nor does it try to identify all of the users of those applications. He pointed out that his organization feels that the arrival of a new application onto the network is important and deserves more attention than say, the arrival of new users onto the network. As a result, it has identified SNMP and NetFlow events that indicate when a new application is running over the network.

As the next section of this report will discuss, discovery is also an important component of traffic management.

End-to-End Visibility

Our industry uses the phrase *end-to-end visibility* in various ways. Given that one of this report's major themes is that IT organizations need to implement an application-delivery function that focuses directly on applications and not on the individual components of the IT infrastructure, this report will use the following definition of end-to-end visibility.

End-to-end visibility refers to the ability of the IT organization to examine every component of IT that impacts the communications once users hit ENTER or click the mouse to when they receive a response from an application.

End-to-end visibility is one of the cornerstones of assuring acceptable application performance. End-to-end visibility is important because it:

- Provides the information that allows IT organizations to notice application performance degradation before the end user does.
- Identifies the correct symptoms of the degradation and as a result enables the IT organization to reduce the amount of time it takes to remove the sources of the application degradation.
- Facilitates making intelligent decisions and getting buy-in from other impacted groups. For example, end-to-end visibility provides the hard data that enables an IT organization to know that it has to add bandwidth or redesign some of the components of the infrastructure because the volume of traffic associated with the company's sales order tracking application has increased dramatically. It also positions the IT organization to curb recreational use of the network.
- Allows the IT organization to measure the performance of critical applications before, during and after it makes changes. These changes could be infrastructure upgrades, configuration changes or the deployment of a new application. As a result, the IT organization is in a position both to determine if the change has had a negative impact and to isolate the source of the problem it can fix the problem quickly.
- Enables better cross-functional collaboration. As section 7.2 discussed, having all members of the IT

organization have access to the same set of tools that are detailed and accurate enough to identify the sources of application degradation facilitates cooperation.

The type of cross-functional collaboration the preceding bullet mentioned is difficult to achieve if each group within IT has a different view of the factors causing application degradation.

To enable cross-functional collaboration, it must be possible to view all relevant management data from one place.

Providing detailed end-to-end visibility is difficult due to the complexity and heterogeneity of the typical enterprise network. The typical enterprise network, for example, is comprised of switches and routers, firewalls, application front ends, optimization appliances, intrusion detection and intrusion prevention appliances as well as a virtualized network. An end-to-end monitoring solution must profile traffic in a manner that reflects not only the physical network but also the logical flows of applications, and must be able to do this regardless of the vendors who supply the components or the physical topology of the network.

As section 5.5 discussed, IT organizations typically have easy access to management data from both SNMP MIBs and from NetFlow. IT organizations also have the option of deploying dedicated instrumentation to gain a more detailed view into the types of applications listed below.

An end-to-end visibility solution should be able to identify:

- Well known applications; e.g., FTP, Telnet, Oracle, HTTPS and SSH.
- Complex applications; e.g., SAP and Citrix.
- Applications that are not based on IP; e.g., applications based on IPX or DECnet.
- Custom or homegrown applications.

- Web-based applications.
- Multimedia applications.

Other selection criteria include the ability to:

- Scale as the size of the network and the number of applications grows.
- Provide visibility into virtual networks such as ATM PVCs and Frame Relay DLCIs.
- Support a wide range of topologies both in the access, distribution and core components of the network as well as in the storage area networks.
- Provide visibility into encrypted networks.
- Support real-time and historical analysis.
- Support flexible aggregation of collected information.
- Provide visibility into complex network configurations such as load-balanced or fault-tolerant, multi-channel links.
- Support the monitoring of real traffic.
- Generate and monitor synthetic transactions.

Network and Application Alarming

Static Alarms

Historically, one of the ways that IT organizations attempted to manage performance was by setting static threshold performance-based alarms. In a recent survey, for example, roughly three-quarters (72.8%) of the respondents said they set such alarms. The survey respondents were then asked to indicate the network and application parameters against which they set the alarms. Table 7.4 contains their answers to that question. Survey Respondents were instructed to indicate as many parameters as applied to their situation.

Table 7.4: Percentage of Companies that Set Specific Thresholds

Parameter	Percentage
WAN Traffic Utilization	81.5%
Network Response Time (Ping, TCP Connect)	58.5%
LAN Traffic Utilization	47.8%
Application-Response Time (Synthetic Transaction Based)	30.2%
Application Utilization	12.2%
Other	5.9%

As Table 7.4 shows, the vast majority of IT organizations set thresholds against WAN traffic utilization or some other network parameter. Less than one-third of IT organizations set parameters against application-response time.

Many companies that set thresholds against WAN utilization use a rule of thumb that says network utilization should not exceed 70 or 80 percent. Companies that use this approach to managing network and application performance implicitly make two assumptions:

1. If the network is heavily utilized, the applications are performing poorly.
2. If the network is lightly utilized, the applications are performing well.

The first assumption is often true, but not always. For example, if the company is primarily supporting email or bulk file transfer applications, heavy network utilization is unlikely to cause unacceptable application performance.

The second assumption is often false. It is quite possible to have the network operating at relatively low utilization levels and still have the application perform poorly. An example of this is any application that uses a chatty protocol over the WAN. In this case, the application can perform badly because of the large number of application turns, even though the network is exhibiting low levels of delay, jitter and packet loss.

Application management should focus directly on the application and not just on factors that have the potential to influence application performance.

The Survey Respondents were also asked to indicate the approach that their companies take to setting performance thresholds. Table 7.5 contains their answers.

Table 7.5: Approach to Setting Thresholds

Approach	Percentage of Companies
We set the thresholds at a high-water mark so that we only see severe problems.	64.3%
We set the thresholds low because we want to know every single abnormality that occurs.	18.3%
Other (Please specify).	17.4%

Of the Survey Respondents that indicated *other*, their most common responses were that their companies set the thresholds at what they consider to be an average value.

One conclusion that can be drawn from Table 7.5 is that the vast majority of IT organizations set the thresholds high to minimize the number of alarms that they receive. While this approach makes sense from operationally, it leads to an obvious conclusion.

Most IT organizations ignore the majority of the performance alarms.

Proactive Alarms

As noted, most IT organizations implement static performance alarms by setting thresholds at the high water mark. This means that the use of static performance alarms is reactive. The problems static performance alarms identify are only identified once they have reached the point where they most likely impact users.

The use of static performance alarms has two other limitations. One is that the use of these alarms can result in a

lot of administrative overhead due to the effort required to initially configure the alarms, as well as the effort needed to keep up with tuning the settings in order to accommodate the constantly changing environment. Another limitation of the use of these alarms is accuracy. In particular, in many cases the use of static performance alarms can result in an unacceptable number of false positives and/or false negatives.

Proactive alarming is sometimes referred to as network analytics. The goal of proactive alarming is to automatically identify and report on possible problems in real time so that organizations can eliminate them before they impact users. One key concept of proactive alarming is that it takes the concepts of baselining, which section 5.4 describes, and applies these concepts to real-time operations.

A proactive alarming solution needs to be able to baseline the network to identify normal patterns and then identify in real time a variety of types of changes in network traffic. For example, the solution must be able to identify a spike in traffic, where a spike is characterized by a change that is both brief and distinct. A proactive alarming solution must also be able to identify a significant shift in traffic as well as the longer-term drift.

Some criteria organizations can use to select a proactive alarming solution include that the solution should:

- Operate off real-time feeds of performance metrics.
- Not require any threshold definitions.
- Integrate with any event console or enterprise-management platform.
- Self-learn normal behavior patterns, including hourly and daily variations based on the normal course of user community activities.
- Recognize spike, shift and drift conditions.
- Discriminate between individual applications and users.

- Discriminate between physical and virtual network elements.
- Collect and present supporting diagnostic data along with alarm.
- Eliminate both false positive and false negative alarms.

Measuring Application Performance

Evaluating application performance has been used in traditional voice communications for decades. In particular, evaluating the quality of voice communications by using a Mean Opinion Score (MOS) is common.

The Mean Opinion Score is defined in “Methods for Subjective Determination of Voice Quality (ITU-T P.800).” As that title suggests, a Mean Opinion Score is a result of subjective testing in which people listen to voice communications and place the call into one of five categories. Table 7.6 depicts those categories, and the numerical rating associated with each.

Table 7.6: Mean Opinion Scores and Speech Quality

MOS	Speech Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

A call with a MOS of 4.0 or higher is deemed to be of toll quality.

To increase objectivity, the ITU has developed another model of voice quality. Recommendation G.107 defines this model, referred to as the E-Model. The E-Model is intended to predict how an average user would rate the quality of a voice call. The E-Model calculates the transmission-rating factor R, based on transmission parameters such as delay and packet loss.

Table 7.7⁴ depicts the relationship between R-Values and Mean Opinion Scores.

Table 7.7: Comparison of R-Values and Mean Opinion Scores

R-Value	Characterization	MOS
90 - 100	Very Satisfied	4.3+
80 - 90	Satisfied	4.0 - 4.3
70 - 80	Some Users Dissatisfied	3.6 - 4.0
60 - 70	Many Users Dissatisfied	3.1 - 3.6
50 - 60	Nearly All Users Dissatisfied	2.6 - 3.1
0 - 60	Not Recommended	1.0 - 2.6

A number of vendors have begun to develop application-performance metrics based on a somewhat similar approach to the ITU E-Model. For example, the Apdex Alliance⁵ is a group of companies collaborating to promote an application-performance metric called Apdex (Application Performance Index) which the alliance states is an open standard that defines a standardized method to report, benchmark and track application performance.

Route Analytics

As section 4 mentions, many organizations have moved away from a simple hub-and-spoke network topology and have adopted either a some-to-many or a many-to-many topology. By the nature of networks that are large and which have complex network topologies, it is not uncommon for the underlying network infrastructure to change, experience instabilities, and to become mis-configured. In addition, the network itself is likely designed in a sub-optimum fashion. Any or all of these factors have a negative impact on application performance. As a result, an organization that has a large complex network needs visibility into the operational architecture and dynamic behavior of the network.

⁴ *Overcoming Barriers to High-Quality Voice over IP Deployments, Intel*

⁵ <http://www.apdex.org/index.html>

One of the many strengths of the Internet Protocol (IP) is its distributed intelligence. For example, routers exchange reachability information with each other via a routing protocol such as OSPF (Open Shortest Path First). Based on this information, each router makes its own decision about how to forward a packet. While this distributed intelligence is a strength of IP, it is also a weakness. In particular, while each router makes its own forwarding decision, there is no single repository of routing information in the network.

The lack of a single repository of routing information is an issue because routing tables are automatically updated and the path that traffic takes to go from point A to point B may change on a regular basis. These changes may be precipitated by a manual process such as adding a router to the network, the mis-configuration of a router or by an automated process such as automatically routing around a failure. In this latter case, the rate of change might be particularly difficult to diagnose if there is an intermittent problem causing a flurry of routing changes typically referred to as route flapping. Among the many problems created by route flapping is that it consumes a lot of the processing power of the routers and hence degrades their performance.

The variability of how the network delivers application traffic across its multiple paths over time can undermine the fundamental assumptions that organizations count on to support many other aspects of application delivery. For example, routing instabilities can cause packet loss, latency, and jitter on otherwise properly configured networks. In addition, alternative paths might not be properly configured for QoS. As a result, applications perform poorly after a failure. Most importantly, configuration errors that occur during routine network changes can cause a wide range of problems that impact application delivery. These configuration errors can be detected if planned network changes can be simulated against the production network.

Factors such as route flapping can be classified as logical as compared to a device specific factor such as a link outage. However, both logical and device-specific factors impact application performance. To quantify how often a logical factor vs. a device specific factor causes an application delivery issue, 200 IT professionals were given the following survey question:

“Some of the factors that impact application performance and availability are logical in nature. Examples of logical factors include sub-optimal routing, intermittent instability or slowdowns, and unanticipated network behavior. In contrast, some of the factors that impact application performance and availability are device specific. Examples of device specific factors include device or interface failures, device out of memory condition or a failed link. In your organization, what percentage of the time that an application is either unavailable or is exhibiting degraded performance is the cause logical? Is the cause device specific?”

The responses to that question are contained in the middle column of the following table.

Table 7.8: Impact of Logical vs. Device Specific Factors

	Percentage of Respondents	Percentage of Respondents - Removing "don't know"
Less than 10% logical vs. 90% device specific	19.5%	26.8%
Up to 30% logical vs. 70% device specific	22.1%	30.4%
50% logical, 50% device specific	10.5%	14.5%
70% logical, 30% device specific	11.6%	15.9%
90% logical, 10% device specific	8.9%	12.3%
Don't know	27.4%	

As Table 7.8 shows, a high percentage of survey respondents answered *don't know*. To compensate for this, the far right column of Table 7.8 reflects the responses of those survey respondents who provided an answer other than *don't know*.

Logical factors are almost as frequent a source of application performance and availability issues as are device-specific factors.

Section 7 gave many examples of the use of SNMP-based management data. SNMP-based management systems can discover and display the individual network elements and their physical or Layer 2 topology, however they cannot identify the actual routes packets take as they transit the network. As such, SNMP-based systems cannot easily identify problems such as route flaps or misconfigurations.

Section 7.6 used the phrase *network analytics* as part of the discussion of proactive alarming. Network analytics and route analytics have some similarities. For example, each of these techniques relies on continuous, real-time monitoring. Whereas the goal of network analytics is to overcome the limitation of setting static performance thresholds, the goal of route analytics is to provide visibility, analysis and diagnosis of the issues that occur at the routing layer. A route analytics solution achieves this goal by providing an understanding of precisely how IP networks deliver application traffic. This requires the creation and maintenance of a map of network-wide routes and of all of the IP traffic flows that traverse these routes. This in turn means that a route analytics solution must be able to record every change in the traffic paths as controlled and notified by IP routing protocols.

By integrating the information about the network routes and the traffic that flows over those routes, a route analytics solution can provide information about the volume, application composition and class of service (CoS) of traf-

fic on all routes and all individual links. This network-wide, routing and traffic intelligence serves as the basis for:

- Real-time monitoring of the network's Layer 3 operations from the network's point of view.
- Historical analysis of routing and traffic behavior as well as for performing a root causes analysis.
- Modeling of routing and traffic changes and simulating post-change behavior.

The key functional components in a route analytics solution are:

- Listening to and participating in the routing protocol exchanges between routers as they communicate with each other.
- Computing a real-time, network-wide routing map. This is similar in concept to the task performed by individual routers to create their forwarding tables. However, in this case it is computed for all routers.
- Mapping Netflow traffic data, including application composition, across all paths and links in the map.
- Monitoring and displaying routing topology and traffic flow changes as they happen.
- Detecting and alerting on routing events or failures as routers announce them, and reporting on correlated traffic impact.
- Correlating routing events with other information, such as performance data, to identify underlying cause and effect.
- Recording, analyzing and reporting on historical routing and traffic events and trends.
- Simulating the impact of routing or traffic changes on the production network.

One instance in which a route analytics solution has the potential to provide benefits to IT organizations occurs when the IT organization runs a complex private network. In this case, it might be of benefit to the IT organization to take what is likely to be a highly manual process of monitoring and managing routing and to replace it with a highly automated process. Another instance in which a route analytics solution has the potential to provide benefits to IT organizations is when those IT organizations use MPLS services provided by a carrier who uses a route analytics solution. One reason that a route analytics solution can provide value to MPLS networks is that based on the scale of a carrier's MPLS network, these networks tend to be very complex and hence difficult to monitor and manage. The complexity of these networks increases when the carrier uses BGP (Border Gateway Protocol) as BGP is itself a complex protocol. For example, a mis-configuration in BGP can result in poor service quality and reachability problems as the routing information is transferred between the users' CE (Customer Edge) routers to the service provider's PE (Provider Edge) routers.

Route analytics can also be useful in simulating and analyzing the network-wide routing and traffic impact of various failure scenarios as well as the impact of planned network changes such as consolidating servers out of branch offices, or implementing new WAN links or router hardware. The purpose of this simulation is to ensure that the planned and unplanned changes will not have a negative effect on the network.

One criterion that an IT organization should look at when selecting a route analytics solution is the breadth of routing protocol coverage. For example, based on the environment, the IT organization might need the solution to support of protocols such as OSPF, IS-IS, EIGRP, BGP and MPLS VPNs. Another criterion is that the solution should be able to collect data and correlate integrated routing and Netflow traffic flow data. Ideally, this data is collected and reported on in a continuous real-time fashion and is also

stored in such a way that it is possible to generate meaningful reports that provide an historical perspective on the performance of the network. The solution should also be aware of both application and CoS issues, and be able to integrate with other network management components. In particular, a route analytics solution should be capable of being integrated with network-agnostic application performance management tools that look at the endpoint computers that are clients of the network, as well as with traditional network management solutions that provide insight into specific points in the network; i.e., devices, interfaces, and links.

8.0 Control

Introduction

To effectively control how applications perform as well as who has access to which applications, IT organizations must be able to:

- Control the availability, security and performance of the desktop.
- Enforce company policy relative to what devices can access the network.
- Classify traffic based on myriad criteria.
- Prioritize traffic that is business critical and delay sensitive.
- Perform traffic management and dynamically allocate network resources.
- Authenticate traffic.
- Provide call control and signaling functionality.
- Affect the routing of traffic through the network.

Controlling the Desktop

For the sake of this report, the term *desktop* refers to the typical desktop or laptop computer that employees use for a variety of information-processing functions. It will

increasingly become reasonable, however, to think of the term *desktop* as including a variety of other devices, such as smart phones.

Traditionally the role of the desktop has been limited to supporting normal business operations. That is changing. As section 4.8, mentioned, the vast majority of employees work outside of a headquarters site; i.e., at a remote office or a home office. What many companies have discovered, often by accident, is that having a high percentage of employees work remotely can be a key component of a business-continuity strategy. For example, a disaster such as a hurricane could strike a company's headquarters site and make that site inoperative by knocking out the electricity and making it impossible for employees to get to work. Employees at remote or home offices can still be functioning. These employees would access some of their company's key applications off of their PC and would also use that PC to access other key applications, typically from the company's backup data center.

With proper planning, the desktop is a key component of a business-continuity plan.

Whether to support normal operations or to enable effective business continuity, two fundamental issues are associated with the desktop that IT organizations need to resolve. One issue is compliance. One aspect of compliance is the mandate that IT organizations have to ensure that the company can comply with the security and privacy requirements of a growing number of government regulations, including HIPAA (Health Insurance Portability and Accountability Act), Sarbanes-Oxley and Gramm-Leach-Bliley. The second aspect of compliance is to ensure that the desktop can comply with the company's own policy requirements. For example, the company might have a policy that states that no desktop can attach to the network if it is not running the latest anti-virus software.

The second key issue associated with the desktop is that once the desktop is placed into service it begins to accu-

mulate unnecessary applications and files. These unnecessary applications and files will be referred to in this report as *detrimental additions*. In most cases, these detrimental additions will affect the availability and performance of the desktop, and hence directly affect application delivery.

The detrimental additions that accumulate on computers come from numerous sources. In some cases the user is not directly involved in adding the detrimental additions that accumulate on their PC. Well-known examples of this phenomenon include worms, viruses and spyware that attach themselves to the user's desktop once a user unwittingly opens an email from a malicious source. There are, however, some less well-known examples of this phenomenon, such as registry creep and bad DLL files that can also have a significant long-term impact on desktop performance.

While in many instances the user is not involved in the accumulation of detrimental additions on their desktop, there are also many instances in which they are. One way this occurs is when users download applications or files, typically for their own use and virtually always over the Internet. The Manufacturing IT Manager pointed out why this behavior is so common. He said, "The P in PC stands for personal." He went on to explain that most of his company's employees believe they have a right to use a company PC for a certain amount of non-business related activities.

The Technical Services Officer and the Manufacturing IT Manager work in somewhat different environments. The Technical Services Officer does not have to be concerned about any of the 2,000 workstations that his organization supports being brought home and used by an employee's family members. In addition, his organization has a policy of controlling what users can add to their desktops so that detrimental additions do not accumulate on those workstations. The Technical Services Officer commented that prior to implementing this policy they had a serious issue with the accumulation of various forms of detrimental additions

on those workstations. He commented that they used to have a serious problem with spyware, in particular.

The Manufacturing IT Manager used the metaphor of “the frog in hot water” to describe how users respond to having the performance of their desktop degrade over time. The premise of the metaphor is that if you place a frog in hot water, it will jump out. However, if you place a frog in cold water and slowly heat the water, the frog will not jump out. Using this metaphor, the Manufacturing IT Manager stated that a user’s tolerance for bad performance grows slowly over time.

Ensuring compliance as well as the availability and performance of the desktop is challenging in large part because the desktop support model most IT organizations use is extremely labor intensive. Any labor-intensive process tends to be expensive, slow to react, and error prone. This is particularly true for desktop processes because of the sheer number of desktops that most IT organizations have to support.

In our industry, people overuse the phrase *next generation*. In particular, when many companies introduce some new feature into a product or service, they start to refer to that product or service as being next generation. In most cases, these are not next-generation products and services. To qualify as being next generation, the change to the product or service has to be truly fundamental.

One of the primary characteristics of the next generation of desktop support systems is automation.

The next-generation desktop support systems, for example, must automatically enforce policy and must automatically prevent or resolve a significant percentage of the issues that lead to either degraded performance or an outage.

In addition to automation, other criteria to select a next-generation desktop support system include the ability of the system to:

- Delete the detrimental additions that accumulate on desktops.
- Support different operating environments for individual work groups.
- Enable the IT organization to exert varying levels of control over the desktops based on organizational constraints.
- Provide detailed reporting.
- Update only the changed components of the operating environment.
- Integrate with other management tools.
- Continue to operate even if the operating system is not working.
- Enable the IT organization to better manage its software licenses.

Traffic Management and 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. One example of a latency-sensitive, business-critical application is VoIP. Another example is that some of the SAP 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 its sales organization as it wastes time waiting for the SD module to respond. In addition, if the SD module times out, this can irritate customers to the point where they hang up, taking their business elsewhere.

The focus of the organization’s traffic management processes must be the company’s applications, and not solely the megabytes of traffic traversing the network.

To ensure that an application receives the required amount of bandwidth, or alternatively 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, 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, which form the basis of traditional Quality of Service (QoS) functionality, 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.

In the preceding sections the Team Leader discussed the progress that his organization has made relative to implementing various components of the application-delivery framework; i.e., planning as well as network and application optimization. He stated that his organization was going to now focus their attention on implementing more control, starting with a company wide deployment of QoS.

Some of the key steps in a traffic management process include:

Discovering the Application

Application discovery should occur at Layer 7. In particular, information gathered at Layer 4 or lower allows a network manager to assign a lower prior-

ity to their Web traffic than to other WAN traffic. Without information gathered at Layer 7, however, network managers are not able manage the company's application to the degree that allows them to assign a higher priority to some Web traffic than to other Web traffic.

Profiling the Application

Once the application has been discovered, it is necessary to determine the key characteristics of that application.

Quantifying the Impact of the Application

Since many applications share that same WAN physical or virtual circuit, these applications will tend to interfere with each other. In this step of the process, the degree to which a given application interferes with other applications is identified.

Assigning Appropriate Bandwidth

Once the organization has determined the bandwidth requirements and identifies the degree to which a given application interferes with other applications, it may now assign bandwidth to an application. In some cases, it will do this to ensure that the application performs well. In other cases, it will do this primarily to ensure that the application does not interfere with the performance of other applications. Due to the dynamic nature of the network and application environment, it is highly desirable to have the bandwidth assignment be performed dynamically in real time as opposed to using pre-assigned static metrics. In some solutions, it is possible to assign bandwidth relative to a specific application such as SAP. For example, the IT organization might decide to allocate 256Kbps for SAP traffic. In some other solutions, it is possible to assign bandwidth to a given session. For example, the IT organization could

decide to allocate 50 Kbps to each SAP session. The advantage of the later approach is that it frees the IT organization from having to know how many simultaneous sessions will take place.

Bandwidth does not solve all application performance issues. In particular, adding bandwidth will have little impact on a chatty application.

Many IT organizations implement QoS via queuing functionality found in their routers.

Others implement QoS by deploying MPLS based services. The use of MPLS services is one more factor driving the need for IT organizations to understand the applications that transit the network. This is the case because virtually all carriers have a pricing structure for MPLS services that includes a cost for the access circuit and the port speed. Most carriers also charge for a variety of advanced services, including network based firewalls and IP multicast.

In addition, most carriers charge for the CoS (class of service) profile. While most carriers offer between five and eight service classes, for simplicity assume that a carrier only offers two service classes. One of the service classes is referred to as real time and is intended for applications such as voice and video. The other is called best effort and is intended for any traffic that is not placed in the real-time service class.

The CoS profile refers to how the capacity of the service is distributed over these two service classes. In most cases, if all of the traffic were assigned to the real-time traffic class that would cost more than a 50-50 split in which half the traffic were assigned to real time and half to best effort. A 50-50 split would cost more than if all of the traffic was assigned to best effort. Hence, assigning more capacity to the real-time class than is necessary will increase the cost of the service. However, since most carriers drop any traffic that exceeds what the real-time traffic class was configured to support, assigning less

capacity to this class than is needed will likely result in poor voice quality. Some solutions provide QoS mechanisms to independently prioritize packets based on traffic class and latency sensitivity, thereby avoiding the problem of over-provisioning WAN bandwidth in an effort to ensure high performance.

Kubernan recently surveyed IT professionals about their use of traffic management and QoS. We found:

- 44% of IT organizations perform these tasks.
- 56% of the companies that perform these tasks claim they do them well.

Route Control

One challenge facing IT organizations that run business-critical applications on IP networks is the variability in the performance of those networks. In particular, during the course of a day, both private and public IP networks exhibit a wide range of delay, packet loss and availability⁶. Another challenge is the way that routing protocols choose a path through a network. In particular, routing protocols choose the least-cost path through a network. The *least-cost* path through a network is often computed to be the path with the least number of hops between the transmitting and receiving devices. Some sophisticated routing protocols, such as OSPF, allow network administrators to assign cost metrics so that some paths through the network are given preference over others. However it is computed, the least-cost path through a network does not necessarily correspond to the path that enables the optimum performance of the company's applications.

A few years ago, organizations began to deploy functionality referred to as route control. The goal of route control is to make more intelligent decisions relative to how traffic is routed through an IP network. Route control achieves this goal by implementing a four-step process. Those steps are:

⁶ *Assessment of VoIP Quality over Internet Backbones, IEEE INFOCOM, 2002*

1. Measurement

Measure the performance (i.e., availability, delay, packet loss, and jitter) of each path through the network.

2. Analysis and Decision Making

Use the performance measurements to determine the best path. This analysis has to occur in real time.

3. Automatic Route Updates

Once the decision has been made to change paths, update the routers to reflect the change.

4. Reporting

Report on the performance of each path as well as the overall route optimization process.

Section 7.7 discussed measuring application performance. Both measuring application performance and route control have merit as stand-alone techniques. In addition, it is possible to link the process that measures application performance with the route-optimization process. In this way, once an organization determines it has an application-performance issue, it can automatically reroute that application's traffic to a more appropriate network path.

9.0 Conclusion

For the foreseeable future, the importance of application delivery is much more likely to increase than it is to decrease. Analogously, for the foreseeable future the impact of the factors that make application delivery difficult, many of which sections 4 and 7 discussed, is much more likely to increase than it is to decrease. To deal with these two forces, IT organizations need to develop a systematic approach to applications delivery. Given the complexity associated with application delivery, this approach cannot focus on just one component of the task such as network and application optimization. To be successful, IT organizations must implement an approach to application delivery that integrates the key components of planning,

network and application optimization, management and control.

This report identified a number of conclusions that IT organizations can use when formulating their approaches to ensuring acceptable application delivery. Those conclusions are:

- If you work in IT, you either develop applications or you deliver applications.
- In the vast majority of instances when a key business application is degrading, the end user, not the IT organization, first notices the degradation.
- The goal of application delivery is to help IT organizations develop the ability to minimize the occurrence of application performance issues and to both identify and quickly resolve issues when they occur.
- Application delivery is more complex than network and application acceleration.
- Application delivery needs to have top-down approach, with a focus on application performance.
- Companies that want to be successful with application delivery must understand their current and emerging application environments.
- In the majority of cases, there is at most a moderate emphasis during the design and development of an application on how well that application will run over a WAN.
- Successful application delivery requires IT organizations identify the applications running on the network and ensure the acceptable performance of the applications that are relevant to the business while controlling or eliminating irrelevant applications.
- Every component of an application-delivery solution has to be able to support the company's traffic pat-

terns, whether they are one-to-many, many-to-many or some-to-many.

- Just as WAN performance impacts n-tier applications more than monolithic applications; WAN performance impacts Web services-based applications more than are n-tier applications.
- The webification of application introduces chatty protocols into the network. In addition, some of these protocols (i.e., XML) tend greatly increase the amount of data that transits the network and is processed by the servers.
- While server consolidation produces many benefits, it can also produce some significant performance issues.
- One effect of data-center consolidation and single hosting is additional WAN latency for remote users.
- In the vast majority of situations, when people access an application they are accessing it over the WAN.
- To be successful, application delivery solutions must function in a highly dynamic environment. This drives the need for both the dynamic setting of parameters and automation.
- Only 14% of IT organizations claim to have aligned the application delivery with application development. Eight percent (8%) of IT organizations state they plan and holistically fund IT initiatives across all of the IT disciplines. Twelve percent (12%) of IT organizations state that troubleshooting IT operational issues occurs cooperatively across all IT disciplines.
- People use the CYA approach to application delivery to show it is not their fault that the application is performing badly. In contrast, the goal of the CIO approach is to identify and then fix the problem.

- It is extremely difficult to make effective network and application-design decisions if the IT organization does not have well-understood and adhered-to targets for application performance.
- Hope is not a strategy. Successful application delivery requires careful planning coupled with extensive measurements and effective proactive and reactive processes.
- The ability to understand how to optimally improve the performance of an application requires a complete profile of that application.
- The application-delivery function needs to be involved early in the applications development cycle.
- A primary way to balance the requirements and capabilities of the application development and the application-delivery functions is to create an effective architecture that integrates those two functions.
- IT organizations need to modify their baselining activities to focus directly on delay.
- Organizations should baseline their network by measuring 100% of the actual traffic from real users.

To deploy the appropriate network and application-optimization solution, IT organizations need to understand the problem that they are trying to solve.

- To understand the performance gains of any network and application-optimization solution, organizations must test that solution in an environment that closely reflects the environment in which they will deploy it.
- When choosing a network and application optimization solution, organizations must ensure the solu-

tion can scale to provide additional functionality over what they initially require.

- An AFE provides more sophisticated functionality than a SLB does.
- IT organizations will not be successful with application delivery as long as long as the end user, and not the IT organization, first notices application degradation.
- When an application experiences degradation, virtually any component of IT could be the source of the problem.
- To be successful with application delivery, IT organizations need tools and processes that can identify the root cause of application degradation. The entire IT organization must accept those tools and processes as valid.
- Organizational discord and ineffective processes are at least as much of an impediment to the successful management of application performance as are technology and tools.
- End-to-end visibility refers to the ability of the IT organization to examine every component of IT that impacts the communications once users hit ENTER or click the mouse to when they receive responses from an application.
- To enable cross-functional collaboration, it must be possible to view all relevant management data from one place.
- Application management should focus directly on the application and not just on factors that have the potential to influence application performance.
- In most IT organizations people ignore the majority of the performance alarms.

- Logical factors are almost as frequent a source of application performance and availability issues as are device-specific factors.
- With proper planning, the desktop is a key component of a business-continuity plan.
- One primary characteristic of the next generation of desktop-support systems is automation.
- The focus of the organization's traffic management processes must be the company's applications, and not merely the megabytes of traffic traversing the network.

10.0 Bibliography

Articles by Jim Metzler

Newsletters written for Network World

Morphing tactical solutions to becoming strategic ones
<http://www.networkworld.com/newsletters/frame/2006/1218wan1.html>

The benefits of thinking strategic when deploying network optimization
<http://www.networkworld.com/newsletters/frame/2006/1211wan2.html>

Identify and then test
<http://www.networkworld.com/newsletters/frame/2006/1127wan2.html>

WAN optimization tips
<http://www.networkworld.com/newsletters/frame/2006/1127wan1.html>

Insight from the road
<http://www.networkworld.com/newsletters/frame/2006/1120wan1.html>

What is termed network misuse in one company may not be so in another
<http://www.networkworld.com/newsletters/frame/2006/1113wan1.html>

Naïve users who hog (or bring down) the network
<http://www.networkworld.com/newsletters/frame/2006/1106wan2.html>

Network managers reveal extent of network misuse on their nets
<http://www.networkworld.com/newsletters/frame/2006/1106wan1.html>

Network managers plugged into the importance of application delivery
<http://www.networkworld.com/newsletters/frame/2006/1023wan1.html>

Cisco vs. Microsoft: The battle over the branch office, unified communications, and collaboration
<http://www.networkworld.com/newsletters/frame/2006/0925wan2.html>

Cisco gets serious about application delivery
<http://www.networkworld.com/newsletters/frame/2006/0925wan1.html>

Application Acceleration that Focuses on the Application, Part 1
<http://www.networkworld.com/newsletters/frame/2006/0821wan1.html>

Application Acceleration that Focuses on the Application, Part 2
<http://www.networkworld.com/newsletters/frame/2006/0821wan2.html>

CIOs don't take enough notice of application delivery issues
<http://www.networkworld.com/newsletters/frame/2006/0807wan2.html>

WAN-vicious apps are a net manager's worst nightmare
<http://www.networkworld.com/newsletters/frame/2006/0807wan1.html>

Who in your company first notices when apps performance starts to degrade?
<http://www.networkworld.com/newsletters/frame/2006/0731wan2.html>

When applications perform badly, is the CYA approach good enough?
<http://www.networkworld.com/newsletters/frame/2006/0529wan2.html>

Making sure the apps your senior managers care about work well over the WAN
<http://www.networkworld.com/newsletters/frame/2006/0403wan1.html>

Where best to implement network and application acceleration, Part 1
<http://www.networkworld.com/newsletters/frame/2006/0327wan1.html>

Where best to implement network and application acceleration, Part 2
<http://www.networkworld.com/newsletters/frame/2006/0327wan2.html>

A new convergence form brings together security and application acceleration
<http://www.networkworld.com/newsletters/frame/2006/0227wan2.html>

What makes for a next-generation application performance product?
<http://www.networkworld.com/newsletters/frame/2006/0220wan1.html>

The limitations of today's app acceleration products
<http://www.networkworld.com/newsletters/frame/2006/0213wan2.html>

What slows down app performance over WANs?

<http://www.networkworld.com/newsletters/frame/2006/0213wan1.html>

Automating application acceleration

<http://www.networkworld.com/newsletters/frame/2006/0130wan2.html>

Advancing the move to WAN management automation

<http://www.networkworld.com/newsletters/frame/2006/0130wan1.html>

Application benchmarking helps you to determine how apps will perform

<http://www.networkworld.com/newsletters/frame/2006/0403wan2.html>

How do you feel about one-box solutions?

<http://www.networkworld.com/newsletters/frame/2005/1212wan1.html>

Users don't want WAN optimization tools that are complex to manage

<http://www.networkworld.com/newsletters/frame/2005/1121wan1.html>

QoS, visibility and reporting are hot optimization techniques, users say

<http://www.networkworld.com/newsletters/frame/2005/1114wan2.html>

Survey finds users are becoming proactive with WAN mgmt.

<http://www.networkworld.com/newsletters/frame/2005/1114wan1.html>

Microsoft attempts to address CIFS' limitations in R2

<http://www.networkworld.com/newsletters/frame/2005/1107wan2.html>

WAFS could answer CIFS' limitations

<http://www.networkworld.com/newsletters/frame/2005/1107wan1.html>

Disgruntled users and the centralized data center

<http://www.networkworld.com/newsletters/frame/2005/1031wan2.html>

WAFS attempts to soothe the problems of running popular apps over WANs

<http://www.networkworld.com/newsletters/frame/2005/1031wan1.html>

WAN optimization helps speed up data replication for global benefits firm

<http://www.networkworld.com/newsletters/frame/2005/1017wan1.html>

Application accelerators take on various problems

<http://www.networkworld.com/newsletters/frame/2005/0829wan1.html>

The gap between networks and applications lingers

<http://www.networkworld.com/newsletters/frame/2005/0815wan2.html>

Is Cisco AON the new-age message broker?

<http://www.networkworld.com/newsletters/frame/2005/0718wan1.html>

Controlling TCP congestion

<http://www.networkworld.com/newsletters/frame/2005/0704wan1.html>

How TCP ensures smooth end-to-end performance

<http://www.networkworld.com/newsletters/frame/2005/0627wan2.html>

Mechanisms that directly influence network throughput

<http://www.networkworld.com/newsletters/frame/2005/0627wan1.html>

Increase bandwidth by controlling network misuse

<http://www.networkworld.com/newsletters/frame/2005/0620wan1.html>

Cisco's FineGround buy signals big change in the WAN optimization sector

<http://www.networkworld.com/newsletters/frame/2005/0530wan1.html>

The thorny problem of supporting delay-sensitive Web services

<http://www.networkworld.com/newsletters/frame/2005/0516wan2.html>

Uncovering the sources of WAN connectivity delays

<http://www.networkworld.com/newsletters/frame/2005/0502wan2.html>

Why adding bandwidth does nothing to improve application performance

<http://www.networkworld.com/newsletters/frame/2005/0502wan1.html>

Organizations are deploying MPLS and queuing for QoS, survey finds

<http://www.networkworld.com/newsletters/frame/2005/0425wan1.html>

The trick of assigning network priority to application suites

<http://www.networkworld.com/newsletters/frame/2005/0418wan2.html>

TCP acceleration and spoofing acknowledgements

<http://www.networkworld.com/newsletters/frame/2005/0321wan2.html>

How TCP acceleration could be used for WAN optimization

<http://www.networkworld.com/newsletters/frame/2005/0321wan1.html>

Antidote for 'chatty' protocols: WAFS

<http://www.networkworld.com/newsletters/frame/2005/0314wan1.html>

How are you optimizing your branch-office WAN?

<http://www.networkworld.com/newsletters/frame/2005/0307wan1.html>

What the next generation Web services mean to your WAN

<http://www.networkworld.com/newsletters/frame/2005/0221wan2.html>

Bandwidth vs. management: A careful balancing act

<http://www.networkworld.com/newsletters/frame/2005/0214wan2.html>

IT Impact Briefs

WAN Vicious Applications

<http://www.webtorials.com/main/resource/papers/netscout/briefs/brief-11-06.htm>

The Movement to Implement ITIL

<http://www.webtorials.com/main/resource/papers/netscout/briefs/brief-10-06.htm>

Business Process Redesign

<http://www.webtorials.com/main/resource/papers/netscout/briefs/brief-09-06.htm>

Network Misuse Revisited

<http://www.webtorials.com/main/gold/netscout/it-impact/briefs.htm>

Moving Past Static Performance Alarms

<http://www.webtorials.com/main/resource/papers/netscout/briefs/brief-07-06.htm>

Analyzing the Conventional Wisdom of Network Management Industry Trends

<http://www.webtorials.com/main/resource/papers/netscout/briefs/brief-06-06.htm>

The Cost and Management Challenges of MPLS Services

<http://www.webtorials.com/main/resource/papers/netscout/briefs/brief-05-06.htm>

The Movement to Deploy MPLS

http://www.webtorials.com/main/resource/papers/netscout/briefs/04-06/NetScout_iib_Metzler_0406_Deploy_MPLS.pdf

Managing VoIP Deployments

http://www.webtorials.com/main/resource/papers/netscout/briefs/03-06/NetScout_iib_Metzler_0306_Managing_VoIP_Deployments.pdf

Network and Application Performance Alarms: What's Really Going On?

http://www.webtorials.com/main/resource/papers/netscout/briefs/02-06/NetScout_iib_Metzler_0206_Network_Application_Alarms.pdf

Netflow – Gaining Application Awareness

http://www.webtorials.com/main/resource/papers/netscout/briefs/01-06/NetScout_iib_Metzler_0106_NetFlow_Application_Awareness.pdf

Management Issues in a Web Services Environment

http://www.webtorials.com/main/resource/papers/netscout/briefs/11-05/Management_Web_Services.pdf

The Movement to Deploy Web Services

http://www.webtorials.com/main/resource/papers/netscout/briefs/10-05/NetScout_iib_Metzler_1005_Web_Services_Deployment_SOA.pdf

The Rapidly Evolving Data Center

http://www.webtorials.com/main/resource/papers/netscout/briefs/09-05/Data_Center.pdf

What's Driving IT?

http://www.webtorials.com/main/resource/papers/netscout/briefs/08-05/Whats_Driving_IT.pdf

The Lack of Alignment in IT

http://www.webtorials.com/main/resource/papers/netscout/briefs/07-05/Lack_of_Alignment_in_IT.pdf

It's the Application, Stupid

<http://www.webtorials.com/main/resource/papers/netscout/briefs/06-05/0605Application.pdf>

Identifying Network Misuse

http://www.webtorials.com/main/resource/papers/netscout/briefs/05-05/0505_Network_Misuse.pdf

Why Performance Management Matters

<http://www.webtorials.com/abstracts/Why%20Performance%20Management%20Matters.htm>

Crafting SLAs for Private IP Services

<http://www.webtorials.com/abstracts/Crafting%20SLAs%20for%20Private%20IP%20Services.htm>

White Papers

Eliminating The Roadblocks to Effectively Managing Application Performance

www.kubernan.com

Closing the WAN Intelligence Gap

www.kubernan.com

Taking Control of Secure Application Delivery

www.kubernan.com

Innovation in MPLS-Based Services

www.kubernan.com

Supporting Server Consolidation Takes More than WAFS

www.kubernan.com

Proactive WAN Application Optimization – A Reality Check

www.kubernan.com

The Mandate to Implement Unified Performance Management

www.kubernan.com

The Three Components of Optimizing WAN Bandwidth

www.kubernan.com

Branch Office Networking

<http://www.webtorials.com/abstracts/ITBB-BON-2004.htm>

The Successful Deployment of VoIP

<http://www.webtorials.com/abstracts/The%20Successful%20Deployment%20of%20VoIP.htm>

The Challenges of Managing in a Web Services Environment

<http://www.webtorials.com/abstracts/NetworkPhysics6.htm>

Buyers Guide: Application Delivery Solutions

<http://www.webtorials.com/abstracts/SilverPeak3.htm>

Best Security Practices for a Converged Environment

<http://www.webtorials.com/abstracts/Avaya27.htm>

Articles Contributed by the Sponsors

Network Performance Management Buyers Guide

<http://www.netscout.com/library/buyersguide/default.asp?cpid=metzler-1206&pdf=bg>

Integrating NetFlow Data into Your Network and Application Performance Monitoring System

http://www.netscout.com/library/whitepapers/NetFlow_performance_management.asp?cpid=metzler-1206&pdf=wp_netflow <http://www.netscout.com/library/whitepapers/NetFlow_performance_management.asp?cpid=metzler-1206&pdf=wp_netflow>

VoIP Implementation Guide for Network Performance Management

http://www.netscout.com/library/whitepapers/voip_best_practices.asp?cpid=metzler-1206&pdf=wp_voip_guide <http://www.netscout.com/library/whitepapers/voip_best_practices.asp?cpid=metzler-1206&pdf=wp_voip_guide>

Big Picture in High Definition - Application Visualization for Enterprise Network Performance Management

http://www.netscout.com/library/whitepapers/hdpm_application_network_management.asp?cpid=metzler-1206&pdf=wp_app_visual <http://www.netscout.com/library/whitepapers/hdpm_application_network_management.asp?cpid=metzler-1206&pdf=wp_app_visual>

Cutting through complexity of monitoring MPLS Networks

http://www.netscout.com/library/whitepapers/MPLS_multi_protocol_label_switching.asp <http://www.netscout.com/library/whitepapers/MPLS_multi_protocol_label_switching.asp?cpid=metzler-1206&pdf=wp_mpls> <http://www.netscout.com/library/whitepapers/MPLS_multi_protocol_label_switching.asp?cpid=metzler-1206&pdf=wp_mpls>

Accelerating Response Times in Branch Offices

http://www.cisco.com/en/US/products/ps6870/products_white_paper0900aecd8051c07f.shtml

The Value of Network-Transparent Application Acceleration and WAN Optimization

http://www.cisco.com/en/US/products/ps6870/products_white_paper0900aecd8051d553.shtml

Testing Report: Veritest on WAFS Results for Microsoft Office Applications

http://www.cisco.com/application/pdf/en/us/guest/products/ps6870/c1031/cdcont_0900aecd8051c157.pdf

WAN Optimization for Centralized Email Applications

http://www.cisco.com/en/US/products/ps6870/products_white_paper0900aecd8051c11d.shtml

WAN Optimization for Microsoft Sharepoint

http://www.cisco.com/en/US/products/ps6870/products_white_paper0900aecd8051c13b.shtml

Cisco Solutions, Design Guides and Case Studies for Oracle

<http://www.cisco.com/go/oracle>

IP Route Analytics: A New Foundation for Modern Network Operations:

<http://www.packetdesign.com/documents/IP%20Route%20Analytics%20White%20Paper.pdf>

Network-Wide IP Routing and Traffic Analysis: An Introduction to Traffic Explorer

<http://www.packetdesign.com/documents/Tex-WP-v1.0.pdf>

Easing Data Center Migration with Traffic Explorer

http://www.packetdesign.com/documents/Easing_Data_Center_Migration_with_Traffic_Explorer.pdf

Network-Layer Management: Foundation for Predictable IP Service Delivery

<http://www.packetdesign.com/documents/Net-LayerMgmt-WP.pdf>

Route Analysis for Converged Networks: Filling the Layer 3 Gap in VoIP Management

<http://www.packetdesign.com/documents/Route-Analysis-VoIP-v1.0.pdf>

US Navy Case Study

<http://www.comnews.com/stories/articles/0106/0106coverstory.htm>

Application Delivery

http://www.riverbed.com/info/ty_wp_master.php?mtcCampaign=3064&mtcPromotion=nw_p>metz>wp <http://www.riverbed.com/info/ty_wp_master.php?mtcCampaign=3064&mtcPromotion=nw_p%3emetz%3ewp>

Instant Business Service Visibility

http://www.networkgeneral.com/Uploads/WhitePapers/20069205673807/WP_BusVisibility_0906.pdf

Unleash the Power of Performance Management, A Guide for selecting a Network Management Solution

http://www.networkgeneral.com/Uploads/WhitePapers/20069205628386/WP_PerfMgmt_0906.pdf

Next-Generation, High-Visibility Tools for Virtualization Management

http://www.networkgeneral.com/Uploads/WhitePapers/20069203398485/WP_Virtualization_0906.pdf

Integrated Visibility for VoIP Management

http://www.networkgeneral.com/Uploads/WhitePapers/20069203283916/WP_VoIP_0906.pdf

Strategies for Optimizing Applications on the WAN

<http://www.packeteer.com/resources/prod-sol/MngAppTraf.pdf>

iShared Architecture and Technology

<http://www.packeteer.com/resources/prod-sol/WANOptimization.pdf>

Configuring PacketShaper for Threat Containment

http://www.packeteer.com/resources/prod-sol/OrangeBook_WP.pdf

Best Practices Handbook for Ensuring Network Readiness for Voice and Video Over IP

http://www.packeteer.com/resources/prod-sol/Wainhouse_IPReadiness.pdf

Managing the Performance of Converged VoIP and Data Applications

<http://www.netqos.com/resourceroom/whitepapers/forms/voip.asp>

Performance First: Performance-Based Network Management Keeps Organizations Functioning at Optimum Levels

<http://www.netqos.com/resourceroom/whitepapers/forms/performancefirst.asp>

Best Practices for NetFlow/IPFIX Analysis and Reporting

<http://www.netqos.com/resourceroom/whitepapers/forms/netflownew.asp>

Improve Networked Application Performance Through SLAs
<http://www.netqos.com/resourceroom/whitepapers/forms/improvedSLAs.asp>

Solving Application Response Time Problems - Metrics that Matter
<http://www.netqos.com/resourceroom/whitepapers/forms/metrics.asp>

Enabling Productivity Through Application Acceleration
http://www.expand.com/products/WhitePapers/Expand_Enabling_Productivity_Application_Acceleration.pdf

LAN-like Performance from the WAN
<http://www.expand.com/products/WhitePapers/wanForLan.pdf>

Everything you always wanted to know about WAN optimization but were afraid to ask
http://www.ipanematech.com/New/DocUpload/Docupload/mc_wp_WAN_Optimization_en_0609.pdf

Beyond Class of Service- A user guide
http://www.ipanematech.com/New/DocUpload/Docupload/mc_wp_cos_user_guide_en_0607.pdf

Maximizing Network Application Performance (solution overview) :
http://www.ipanematech.com/New/DocUpload/Docupload/mc-ipa-solution_overview_en_0607.pdf

Investing for Success
<http://www.avaya.com/master-usa/en-us/resource/assets/whitepapers/svc3234.pdf>

The Power of Remote Monitoring and Remote Diagnostics: EXPERT Systems Reports from Avaya
<http://www1.avaya.com/campaign/demo/expert/index.html>

11.0 Interviewees

In order to gain additional insight application delivery from the perspective of IT organizations, a number of IT professionals were interviewed. The following table depicts the title of each of the interviewees, the type of industry that they work in, as well as how they are referred to in this report.

Job Title	Industry	Reference
COO	Electronic Records Management Company	The Electronics COO
Chief Architect	Entertainment	The Motion Picture Architect
CIO	Diverse Industrial	The Industrial CIO
IT Network Director	Manufacturing	The Manufacturing Director
Chief Technical Services Officer	Government Agency	The Technical Services Officer
IT Manager	Manufacturing	The Manufacturing IT Manager
Network Engineer	Automotive	The Automotive Network Engineer
Global Network Architect	Consulting	The Consulting Architect
Global Infrastructure Engineering Manager	Automotive	The Infrastructure Engineering Manager
Enterprise Architect	Application Service Provider (ASP)	The ASP Architect
Manager of Network Services and Operations	Manufacturing	The Manufacturing Manager
Team Leader, Network Architecture, Strategy and Performance	Energy	The Team Leader
CIO	Engineering	The Engineering CIO

The End of Geography

How senior executives can re-think global business operations, provide better services, do more with less, and reduce IT expenses – All at the same time.

Strategic Imperatives in 2007 – Eliminating Geography as a Decision Input

For CEOs, CFOs and CIOs, there are some key strategic issues that are looming large in the latter half of this decade. Certainly globalization and all its implications are near the top of the list. What does globalization mean in practical terms? Aside from facing down low cost competition from overseas, it means business leaders have to respond, sometimes in kind; they have to expand their companies to all corners of the Earth. They have to find their own low-cost suppliers; they need to get closer to customers; then need to expand their recruiting horizons and find the best people in every field, no matter where they live.

A component of “globalization” is certainly out-sourcing – finding partners around the world to do work that isn’t within your core competency is now standard operating procedure in many industries. Despite the sometimes politically incorrect aspect of outsourcing, it’s a fact of life, and has been embraced by most large corporations and many small ones.

With globalization and outsourcing comes lower costs in some areas (labor for instance), but higher costs in others (IT, distributed/replicated systems, bandwidth, travel and so forth), and higher risks (data protection, disaster recovery, security, attacks from malicious insiders or outsiders and the like). While some may say that “The World is Flat”, there are plenty of potholes in the roads that connect all of us together in this flattening world.

This paper summarizes some of the specific costs of “geography” – that is, the costs of decentralized busi-

ness, and offers a solution that can mitigate many of them, and completely eliminate others. A new class of IT infrastructure is emerging that allows business leaders to take their companies where they must go, yet allows them to minimize the new threats and risks that would otherwise be introduced.

The Impact of Geography

Geography – and its impact on the distribution of people & information - costs enterprises around the world billions of dollars in unnecessary spending and lost opportunity every year. While geographic expansion and decentralization has led to lower costs in many areas, there are other effects of our expanding geographical reach are like a plaque, inefficiency building up in the arteries of companies, slowing them down, costing them money and leading to serious problems, sometimes potentially fatal ones. Yet, it’s not a topic that gets discussed much by business executives. It’s taken as gospel that a “local area” network is different from a “wide area” network, and that certain things are just too difficult to do over a wide area network (“WAN”).

How Geography Affects Growth and Profitability

Every company wants to be flexible, to respond quickly to new opportunities and to adapt quickly to change. Yet flexibility is a big challenge when the opportunities are far away. Companies spend incredible resources in an attempt to capitalize on global opportunity efficiently, and unfortunately they over-spend by billions every year. In fact, according to a recent Gartner report, enterprises will waste \$100 billion on network overspending over the next

five years¹. The reason is that alternatives exist today that give CIOs new tools to deliver much better services at far lower costs. Firms that embrace the new solutions will benefit from huge savings on IT, more flexibility, faster time to market, and less risk; those who don't will incur additional expenses and suffer from reduced growth.

To illustrate the severity of geography on the operations of a global business, we'll look at an example of a global manufacturing firm based in Illinois in the United States, and focus on a scenario where the firm wins a major new contract in Shanghai. This scenario would be the same for virtually any other business; virtually all the technologies discussed are widely used across all industries.

This company has tens of thousands of employees around the world, and already has fairly extensive operations in China, but they are running at capacity due to the sustained growth there. Worldwide, the company has almost a hundred offices; they are growing fast and operate in fifty-seven countries around the world, with major facilities in Asia, North America, Europe and the Middle East. Last year they won the new contract in China, and we'll examine it because it has exposed the serious issues caused by geography.

The new project, effectively the design and construction of a new communications infrastructure in a large part of the country, will require hiring almost 1,000 additional people over the expected seven year life of the project: Engineers, manufacturing engineers, factory works designers, project managers, accountants and support personnel will all need to be added to the project as quickly as possible.

Deploying Steelhead appliances was one of the three things that have made a real difference to the way we work in the last twenty years. The other two were the web and email.

- CTO Global Engineering Firm

Why Can't Under-Utilized Staff In the U.S. or Europe be Re-Assigned to China?

It seems like a no-brainer: Just take the under-utilized engineers in the US and Europe, and put them on the over-taxed China program. There are several reasons why this is harder than it seems. First, collaboration over long distances is very challenging, and it's not just the language or different time zones. Modern design & engineering (and virtually every other industry, from semiconductor design to pharmaceuticals, from software to call centers) is heavily dependent on information technology, data and sophisticated applications used to manage projects, design equipment, store files, share documents and ensure the project is finished on time and on budget.

Those files involved can be enormous; moving them across Wide Area Networks (WANs) can take many minutes every time they are opened, saved or copied (and for large files, the time can be hours). In fact in many cases it is faster to burn a CD or DVD and send it to Shanghai by FedEx than it would be to try to send the data over the WAN. Obviously waiting hours to open a file that you're working on with a colleague is not acceptable, hence the preference to have people to work on local projects. Even standard business applications rely on having local IT infrastructure to work properly and at acceptable performance.

The most serious issue though is lost opportunity. Because the firm has so much trouble finding the staff in Shanghai required to expand quickly, they are simply not bidding on other new projects in China for which their core competence are ideally suited. They are literally leaving millions of dollars of potential business on the table because they cannot handle any more growth.

¹ <http://www.networkworld.com/news/2006/101206-gartner-network-overspending-waste.html>

Data management becomes a serious burden

Like companies in any industry, the firm has found that most applications absolutely require local IT infrastructure in order to work properly. File servers and filers must be deployed in virtually every office around the world; of course those servers and filers need to be backed up, so there are local tape backup systems as well. The Shanghai office, like all their offices of twenty or more people, has their own Exchange server in order to provide acceptable email performance. Overall, the company maintains eighteen Exchange servers around the world.

Data Protection

The tapes from the backup systems required to protect the Shanghai operations are stored off-site by a data-protection / records management firm who picks them up every day. The company also maintains a local content “vaults” for all their design drawings, which can easily be hundreds of megabytes each. Those design files have to be replicated every night to every other office around the world so that their colleagues can collaborate.

Of course, there’s no absolute guarantee that the data is being stored or backed up properly. Similarly, there’s no telling who really has access to the data – is it truly protected and secured in China? Who has access to the servers? To the backup tapes? Where are they stored anyway? What if someone engineered a wholesale theft of the company’s proprietary product designs? Who would ever know? These are all questions that are nagging at senior management – data protection is something people used to just set aside; obviously in today’s climate it’s become much more important.

Distributed Expertise

In any complex design or manufacturing project, there are always specialist and experts on particular topics who are not based in the same place as the primary project staff, and who need access to the current project files. Because of this challenge, the IT staff has devised a

complex replication process that ensures most files are copied to all other offices every night. The whole replication process currently takes anywhere from three to ten hours for a partial job; a complete replication of everything is not possible on a daily basis because it simply takes too long – the backup window available each day is only a few hours long, and if the backup takes eight to ten hours, it simply can’t be done. So the IT staff meets every morning with the project managers to decide what needs to be replicated and to where.

IT Management

All of this IT infrastructure must be managed, so the Shanghai office has a full-time IT staff of five, and contracts out server maintenance to a local company. They also subscribe to the services of an off-site media management company for off-site tape storage. In addition, since the company operates in other cities around China, all of the challenges described here exist elsewhere too. In order to ensure their Asia-Pacific operations have a sound disaster-recovery plan, the company has also invested in their own regional data center that serves the offices in China, Korea, Singapore and Australia.

Despite all of that money, infrastructure and staff, the Shanghai office is struggling to take advantage of the vast opportunity before it. The cost, complexity and human factors are compounded by their remote location ten thousand miles from the company’s main office.

Eliminate Geography from the Equation

If geography was not a factor, how different would things be? That is, if an engineer in Shanghai could connect to the people and servers in Cincinnati or Dortmund as easily as with those down the hall, what would it mean?

Well, things would be completely different. The company would be able to grow faster while investing less in IT; it would be able to provide better and more uniformly high quality services, and ensure that their data is protected,

better managed more soundly protected. In this section of the paper we'll examine what forward-looking companies are doing to eliminate geography from their thinking.

Let's look at the company after incorporating Riverbed's Wide-area Data Services (WDS) appliances.

By deploying Riverbed's Steelhead® appliances in each remote branch office, including Shanghai, a few things change immediately. Engineers and designers around the world suddenly find that opening a file or sending a drawing to a colleague in Ohio (or Germany, or anywhere else in the world) takes a tiny fraction of the time it used to take. Instead of taking 30 to 90 minutes for a file to arrive (or to be saved to a remote file share), huge files can be moved in under a minute. More common Office files (like Word or Excel) can be moved in seconds instead of many minutes. In effect, people connected by the WAN get an experience much closer to the LAN-like experience that they are used to.

Better Collaboration.

Improving the speed or response time dramatically has an immediate impact on the way people collaborate around the world. Suddenly it is possible for an engineer in Shanghai to work with their colleagues in Europe and North America as easily as they could with their colleagues down the hall. Users will adopt the normal practices of using file shares as primary storage rather than relying on emailing files or storing things locally on their laptops.

"In-Sourcing".

The firm can now put people on the Shanghai project (or any other) from under-utilized offices around the world which allows them to load-balance their work in a more intelligent way; colleagues in those offices can begin contributing immediately to over-loaded projects without having to move or take repeated extended trips to get involved.

Faster Growth.

Because of this new in-sourcing capability, the firm has added over a hundred new clients around the globe in the last year, a much faster rate than would have been possible before because of the problems described above. Because they can put people on projects anywhere in the world regardless of the actual location of the people, the firm is able to grow much much faster than before.

Vastly Simplified Infrastructure.

Instead of twenty servers and filers in the Shanghai office, the firm has eliminated virtually all of them. All file servers, have been centralized back to California; the Exchange server has been removed; the SMS servers has been eliminated; the specialized file "vaults" (additional servers for CAD files) have also been centralized. Of course, along with the elimination of those servers, the associated tape-backup systems have also been eliminated, and the contract to store backup tapes off site has been cancelled because there are no more tapes.

Inter-office collaboration now occurs on 50% of our projects, up from about 5% two years ago. We've also been able to take on over 100 new clients in the last year – there's no way we could have absorbed that kind of growth before.

- CTO Global Engineering Firm

Tremendous Savings on IT.

The Return on Investment (ROI) is impressive and immediate. The company saves money immediately on the reduction in server maintenance (and of course replacement costs since they are not longer present). They avoid the costs of upgrading the WAN in an effort to increase performance (see WAN savings below) – overseas, the cost of a WAN link to a branch can easily be thousands of dollars per month. Even the costs of backup

and associated tapes and off-site media storage can be very significant.

Shut-down of Asia Pacific Data Center.

There is no need to maintain a special data center in Singapore any longer because all the Asia Pacific offices are relying on redundant trans-Pacific WAN links connecting them to infrastructure in California.

Faster Restores After Failures.

The firm did leave a couple of filers in the local branch to support Shanghai-centric projects (non-collaborative projects). Those servers are backed up every night over the network (in about an hour) to the existing primary data center in San Francisco (no tapes). In the past, if a filer went down and had to be restored, that restoration process would take a day or even several days. First the IT staff had to diagnose and fix the problem, then the latest tape had to be identified and retrieved from the off-site storage facility, and the whole process would be disruptive. Now, since the WDS implementation is symmetric, servers or filers can be restored just as quickly as they are backed up over the WAN, in this case in about an hour.

Lower WAN Expenses.

After deploying Riverbed's WDS systems, the WAN traffic from file sharing, backup and document management dropped by 90%, allowing them to rollout VOIP and keep their network as it was, avoiding hundreds of thousands of dollars in annual bandwidth expenses.

The End of Geography. It's Within Reach for Your Company Too.

Overall the firm now operates globally in much the same way as it would if it had a single office. They have very little local infrastructure in branch offices because it's no longer necessary. When they open a new office to take advantage of new business opportunities, the only IT infrastructure that gets deployed is a router with an integrated

VPN, the workstations & phones to support the local users, and a Steelhead appliance from Riverbed.

About Riverbed

Riverbed is the performance leader in wide-area data services (WDS) solutions for customers worldwide. By enabling application performance over the wide area network (WAN) that is orders of magnitude faster than what users experience today, Riverbed is changing the way people work, and enabling a distributed workforce that can collaborate as if they were local. Additional information about Riverbed (Nasdaq: RVBD) is available at www.riverbed.com.