The 2009 Handbook of Application Delivery

A Guide to Decision Making in Challenging Economic Times

By Dr. Jim Metzler

SPONSORED IN PART BY

f5

Webtorials

www.webtorials.com

The Handbook of Application Delivery

Contents

IT Innovation Report

Published By
Webtorials
www.webtorials.com

Founder
Jim Metzler
jim@ashtonmetzler.com

Design/Layout Artist
Debi Vozikis

For Editorial and Sponsorship Information
Contact 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 judgment 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

Throughout this handbook, the phrase *application delivery* will refer to the task of ensuring that the applications that an enterprise uses:

- Can be effectively managed
- Exhibit acceptable performance
- Incorporate appropriate levels of security
- Are cost effective

Over the last few years application delivery has become a priority for virtually all IT organizations. However, while many IT organizations have become better at application delivery, the majority of IT organizations struggle with this highly complex task. One of the primary goals of this handbook is to help IT organizations become better with application delivery. One of the ways that the handbook achieves that goal is by creating a framework that IT organizations can customize for use in their environment. The four primary components of the framework are:

- Planning
- Network and application optimization
- Management
- Control

Another way that the handbook achieves that goal is by identifying criteria that IT organizations can use when evaluating alternative solutions.

The Worldwide Economic Environment

In 2009 the economy is politely referred to as being challenging. While this *challenging* economic environment will certainly put pressure on IT budgets, IT organizations need to continue to invest in application delivery solutions. That follows in large part because IT organizations have made, and continue to make, significant investments in enterprise applications to support key business processes. IT organizations need to protect these investments, and the business processes that they enable, by ensuring that these applications can be effectively delivered. It also follows because many of the optimization techniques described in chapter 6 will reduce cost and many of the techniques described in the rest of the handbook also contribute to the IT organization's ability to manage cost. For example, the majority of IT organizations are deploying virtualized servers as a way to reduce cost. As discussed in chapter 4, these initiatives will not be successful if IT organizations do not overcome the management challenges that are associated with virtualized servers.

Application Delivery Challenges

One of the key challenges associated with ensuring acceptable application delivery is that in the vast majority of instances it is the end user, and not the IT organization, that first notices application degradation. Making matters worse, when the IT organization is made aware that the performance of an application is degrading, the organization is often unsure of the cause in part because any component of IT could be the cause of that degradation.

In addition, in most instances:

- IT organizations tend to plan and manage each component of IT in isolation from each other.
- In many instance, each component of the IT infrastructure can perform well, yet the overall performance of the application is unacceptable.
- White papers and other documents that are intended to help IT organizations get better at application delivery tend to focus on just one component of IT. As such, these documents provide little guidance relative to the end-to-end issues caused by complex applications running over a complex infrastructure.

This handbook is designed to provide the end-to-end guidance that the typical white paper cannot provide. An unfortunate side affect of providing that guidance is that the handbook is lengthy. To compensate for that, this executive summary is intended to provide a summary of the key topics found in the handbook and to provide the online reader with a hyperlink to specific sections of the handbook that provide greater detail.

As noted, one of the key challenges associated with ensuring acceptable application performance is that the individual groups within the IT organization function in a siloed or stove-piped fashion. That means that these groups typically do not share:

- Terminology[1]
- Goals
- Tools
- Processes

While they can make some progress on their own, there is a limit to how much success the rank and file of the IT organization can make relative to eliminating organizational and technological silos. This creates a CIO mandate whereby the CIO must drive a transformation to where the IT infrastructure organization focuses not on individual technologies, but on a systematic approach to application delivery. For example, the role of the network operations center (NOC) has changed significantly. In most cases, this change has occurred without the support of senior IT management. In large part due to the lack of that support, many NOCs today are only moderately successful. Using the NOC as just one example of the CIO mandate, what is needed is for senior IT managers to drive the evolution to an Integrated Operations Center (IOC) that has effective tools and processes to support all components of IT, both individually and as an end-to-end system.

The complexity of application delivery is driven in part by the evolving applications environment. Some of the components of that environment that drive complexity include:

- An application development process that largely ignores the impact of the WAN.
- The webification of applications that introduces protocols that are chatty and dense.
- Server consolidation that results in more users accessing applications over a WAN.
- Data center consolidation that results in the WAN link between the user and the application being lengthy and hence exhibiting high levels of latency and packet loss.

[1] The word *service* is a good example of this lack of common terminology. The various subgroups within an IT organization typically use the word service to refer to different things.

- The decentralization of employees that results in even more employees accessing applications over the WAN.
- The use of Software as a Service that results in less IT control over the management and performance of applications.
- The deployment of distributed applications which introduces additional sources of delay.
- The use of Web 2.0 applications and application development techniques both of which result in less IT control over the management and performance of applications.

Another factor that increases the complexity of application delivery is virtualization. There can be compelling reasons to virtualize servers, desktops and storage. However, server virtualization, desktop virtualization and storage virtualization all present distinct management and optimization challenges.

Planning

There are a number of planning functions that are critical to successful application delivery. One of these functions is WAN emulation. One of the goals of WAN emulation is to enable and encourage software engineers to develop applications that perform well over a WAN. Another key function is baselining. Baselining provides a reference from which service quality and application delivery effectiveness can be measured. It does so by quantifying the key characteristics (e.g., response time, utilization and delay) of applications and various IT resources including servers, WAN links and routers. IT organizations that are looking to deploy solutions to baseline their networks should evaluate these solutions using a broad set of selection criteria.

An important task for all IT organizations is to integrate planning and operations. One of the reasons to integrate planning and operations is that it results in the reduction in the number of management tools that must be acquired and supported. This reduces cost, which is particularly important in this challenging economic environment. Another reason to integrate planning and operations is because it also increases the communications within the IT organization. This follows because fewer tools result in less disagreement over the health of the IT infrastructure and the applications that use that infrastructure. One of the technologies that can be used to better integrate planning and operations is route analytics.

Network and Application Optimization

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

- Reduce the amount of data 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) necessary for a given transaction;
- Mitigate the inefficiencies of older protocols;
- Offload computationally intensive tasks from client systems and servers.

Some of the basic tasks of network and application optimization can be gained by deploying devices that function within the *packet delivery network*. By packet delivery network is meant the packet payload and the transport, network and data link layers of the Internet protocol suite. However, more sophisticated techniques require an application delivery network (ADN). ADN solutions leverage functionality that resides higher in the OSI protocol stack and can improve the effectiveness of application delivery based on the ability of these solutions to recognize application layer signatures and to then differentiate among the various applications that share and contend for common transport resources. Some of the primary ADN characteristics include optimization, management and control.

There are two principal categories of network and application optimization products. One category is typically referred to as a WAN Optimization Controller (WOC). WOCs are often referred to as *symmetric solutions* because they typically require an appliance in both the data center as well as the branch office. In most cases, WOCs are implemented as an appliance. As is described below, however, some vendors have implemented virtualized WOCs. This class of solution is often referred to as a *software only solution* or as a *soft WOC*.

The goal of a WOC is to improve the performance of applications delivered from the data center to the branch office or directly to the end user over networks such as Frame Relay, ATM or MPLS. WOCs implement a wide variety of technologies, including caching, compression, congestion control, forward error correction, protocol acceleration, as well as request prediction and spoofing. IT organizations that are evaluating these products should conduct that evaluation based on a broad set of WOC evaluation criteria.

The second category of network and application optimization products is typically referred to as Application Delivery Controllers (ADCs). ADCs began as simple layer 4 load balancers but now provide a wide range of functionality including SSL offload, application firewall, global traffic distribution, rate shaping, DDoS/DoS protection, asymmetrical application acceleration and response time monitoring. Similar to the situation with WOCs, IT organizations that are evaluating ADCs should conduct that evaluation based on a broad set of ADC evaluation criteria.

Just as devices such as servers can be virtualized, so can appliances such as WOCs. A *Virtual Appliance* is based on network appliance software, together with its operating system, running in a virtual machine in a virtualized server. Virtual appliances can include WOCs, firewalls, and performance monitoring solutions among others. A virtual appliance offers the potential to alleviate some of the management burdens in branch offices because most of the provisioning, software updates, configuration, and other management tasks can be automated and centralized at the data center.

Another form of virtualization is clustering. For example, it is possible to cluster a number of ADCs and have the cluster perform as a single ADC. Another option is to implement a cluster of physical appliances with an ADC providing the load balancing across the individual appliance platforms.

Several types of appliances such as ADCs can support yet another form of virtualization, where the system's hardware platform supports a number of independent software partitions. A partitioned appliance can be configured to dedicate a separate partition to each application or service being delivered. This allows the configuration of each partition to be optimized for the specific type of application traffic being processed.

Managed Service Providers (MSP)

Managed Service Providers (MSPs) are not a new phenomena. The last few years, however, have seen the development of a new class of MSP – the Application Delivery MSP (ADMSP). Two of the many benefits of using an ADMSP are the ability to leverage both the ADMSP's expertise and their technology.

There are two primary categories of managed application delivery services provided by ADMSPs: site-based services and Internet-based services. Site-based services are comprised of managed WOCs and/or ADCs installed at participating enterprise sites. The application optimization service may be offered as an optional add-on to a WAN service or as a standalone service that can run over WAN services provided by a third party. Where the application delivery service is bundled with a managed router and WAN service, both the WOC and the WAN router would be deployed and managed by the same MSP.

Whether implemented in a do-it-yourself (DIY) manner or via site-based services, the traditional classes of application delivery solutions (ADC, WOC, soft WOC) were designed to address application performance issues at both the client and server endpoints. These solutions make the assumption that performance characteristics within the WAN itself are not optimizable because they are determined by the relatively static service parameters controlled by the WAN service provider. This assumption is reasonable in the case of private WAN services. However, this assumption does not apply to enterprise application traffic that transits the Internet because there are significant opportunities to optimize performance within the Internet itself based on Application Delivery Services (ADSs).

An ADS is an Internet-based services that focuses on the acceleration of the increasing number of applications that traverse the Internet. Ensuring acceptable application performance over the Internet is difficult because the Internet is a network of networks and the only service providers that get paid to carry Internet traffic are the providers of the first and last mile services. All of the service providers that carry traffic between the first and last mile do so without compensation. One of the affects of this business model is that there tend to be availability and performance bottlenecks at the peering points. Another affect is that since there is not a single, end-to-end provider, service level agreements (SLAs) for the availability and performance of the Internet are not available.

An ADS leverages service provider resources that are distributed throughout the Internet in order to optimize the performance, security, reliability, and visibility of the enterprise's Internet traffic. All client requests to the application's origin server in the data center are redirected via DNS to an ADS server in a nearby point of presence (PoP). This edge server then optimizes the traffic flow to the ADS server closest to the data center's origin server.

Management

Part of the challenge facing IT organizations, and another reason for the mandate to have CIOs drive a transformation of the IT organization, is the organizational dynamic that exists inside of many IT organizations. Part of that organizational dynamic is that less than half of IT organizations indicate that there is a cooperative relationship between their application development organization and their network organization. Another part of the dynamic is that ineffective management processes are one of the biggest impediments to effective application delivery.

There are a number of management tasks that are essential to successful application delivery. As noted, one of the key challenges associated with ensuring acceptable application delivery is that in the vast majority of instances it is the end user, and not the IT organization, the first notices application degradation. As such, the ability to have End-to-End Visibility is a minimum management requirement. In this context, *end-to-end visibility* refers to the ability of the IT organization to examine every component of IT that impacts communications once users hit ENTER or click the mouse button when they receive a response from an application. IT organizations that are looking to deploy a tool to provide end-to-end visibility should evaluate these solutions using a broad set of selection criteria.

The port 80 black hole creates a management and a control challenge. Port 80 is the port that servers listen to while expecting to receive data from Web clients. As a result, a firewall can't block port 80 without eliminating much of the traffic on which a business may depend. Taking advantage of this fact, many applications will port-hop to port 80 when their normally assigned ports are blocked by a firewall. This behavior creates what is referred to as the *port 80 black hole*.

Well-known applications that do port hoping include AOL's instant messaging (AIM), Skype and applications based on the Financial Information eXchange (FIX) protocol. Just looking at these three applications, the port 80 black hole creates issues relative to:

- Security – AIM can carry viruses and worms
- Compliance – In some instances, regulations require that IMs must be archived
- Legal – The file sharing enabled by Skype can be against the law
- Performance – FIX based applications can be very time sensitive

The traditional approach that most IT organizations have taken relative to network and application alarming is to set static threshold alarms. One of the problems with static threshold alarms is that most IT organizations set them at a high value. As a result, most IT organizations miss the majority of alarms. An alternative approach is referred to as proactive alarms or analytics. The goal of analytics is to automatically identify and report on possible problems in real time so that organizations can eliminate the problems before they impact users. One key concept of proactive alarming is that it takes the concepts of baselining and applies these concepts to real-time operations. One of the key selection criteria for an analytics solution is that the 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.

As mentioned, one of the technologies that can be used to better integrate planning and operations is route analytics. From an ongoing management and operations perspective, 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. One of the key selection criteria that an IT organization should look at when selecting a route analytics solution is the breadth of routing protocol coverage that the solution provides.

Application Performance Management

Application performance management (APM) is a relatively new management discipline. The newness of APM is attested to by the fact that ITIL has yet to create a framework for APM. Successful APM requires a holistic approach based on integrated management of both the application itself as well as the end-to-end IT infrastructure. This approach must focus on the experience of the end user of the application or service and must address most, if not all, of the following aspects of management:

- Adoption of service level agreements (SLAs) for at least a handful of key applications and services.
- End-to-end monitoring of all end user transactions.
- Automatic discovery of all the elements in the IT infrastructure that support the key applications and services. These are referred to as The Key Elements.
- Proactive and predictive monitoring of The Key Elements.
- Prioritizing outages and other incidents based on potential business impact.

- Triage and root cause analysis applied at both the application and infrastructure levels.
- Automated generation of performance dashboards and historical reports to allow both IT and business managers to gain insight into SLA compliance and performance trends.

Automation

The automation of management tasks is a critical topic for multiple reasons. One reason is that the majority of IT capital and personnel resources are consumed maintaining the status quo and this percentage increases every year as more functionality is added to the IT infrastructure. The second reason is that performing repetitive, time-consuming tasks is error prone. Automation has the potential to reduce the amount of resources consumed by management tasks and simultaneously to improve the quality associated with those tasks.

Some of the tasks that it makes the most sense to automate include:

- Configuration Management
- Event Management
- Service Level Management
- Security Management

Control

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

Route optimization

The goal of route optimization is to make intelligent decisions relative to how traffic is routed through an IP network. Route optimization is an integral part of an Internet-based service from an ADMSP.   Route optimization can also be applied by IT organizations whenever there is the possibility of multiple paths from between end points.

SSL VPN Gateways

One of the purposes of an SSL VPN gateway is to communicate directly with both the user's browser and the target applications and enable communications between the two. Another purpose of the SSL VPN gateway is to control both access and actions based on the user and the endpoint device. IT organizations should evaluate SSL VPN gateways based on a variety of selection criteria.

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 where there are one or more delay-sensitive, business-critical applications.

Next Generation WAN Firewall

In order to overcome challenges such as the previously mentioned Port 80 Black Hole, a new generation of WAN firewall is required. IT organizations should evaluate next generation firewalls based on a variety of selection criteria.

2.0 Introduction

Background and Goal

Throughout this handbook, the phrase *application delivery* will refer to the task of ensuring that the applications that an enterprise uses:

- Can be effectively managed
- Exhibit acceptable performance
- Incorporate appropriate levels of security
- Are cost effective

Over the last few years application delivery has become a priority for virtually all IT organizations. However, while many IT organizations have become better at application delivery, the majority of IT organizations struggle with the task. For example, Webtorials recently asked 345 IT professionals the following question. "If the performance of one of your company's key applications starts to degrade, who is the most likely to notice it first – the IT organization or the end user?" Seventy three percent (73%) 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 situations in which the end user is typically the first to notice application degradation, the reputation of the IT organization is tarnished.

A goal of this handbook is to help IT organizations develop the ability to minimize the occurrence of application performance issues and to identify and quickly resolve issues when they do occur. To achieve that goal, this handbook develops a framework for application delivery that can be customized by IT organizations for use in their environment.

It is important to note that most times when the industry uses the phrase application delivery, it simply means network and application optimization, which are important; however, achieving this handbook's goal requires a broader view of the factors impacting the ability of the IT organization to assure acceptable application performance.

Application delivery must have a top-down approach, with a focus on application performance as seen by the user of the application.

With these factors in mind, the application delivery framework this handbook describes comprises four primary components.

Successful application delivery requires the integration of:

- ***Planning***
- ***Network and application optimization***
- ***Management and***
- ***Control.***

Some overlap exists in the framework, as a number of common IT processes are part of multiple components of the framework. This includes processes such as discovery (what applications are running on the network and how are they being used), baselining, visibility and reporting.

This handbook details many factors that currently complicate application delivery. This includes the centralization of IT resources, the decentralization of employees and the complexity associated with the current generation of n-tier applications. However, IT organizations that are attempting to become more proficient at application delivery need to account for not just the existing challenges, but also the emerging challenges.

The complexity associated with application delivery will increase over the next few years.

That follows in part because as explained in this handbook, the deployment of new application architectures such as Services Oriented Architecture (SOA), Rich Internet Architecture and Web 2.0 will dramatically increase the difficulty of ensuring acceptable application performance. It also follows because of the increasing management complexity associated with the burgeoning deployment of the virtualization of IT resources (i.e., desktops, servers, storage and applications), the growing impact of wireless communications, the need to provide increasing levels of security as well as emerging trends such as storage optimization.

Instead of reaching a point where the challenges associated with application delivery are going away, we are in fact merely ending the first phase of a fundamental transformation of the IT organization. At the beginning of this transformation, virtually all IT organizations were composed of myriad stove piped functions. By stove piped we mean that these functions were relatively isolated, with few common goals, terminology, tools or processes. A major component of the transformation is that leading edge IT organizations are now creating an environment characterized by the understanding that:

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

Put another way, leading edge companies are evolving their IT organizations to be comprised of two functions: application development and application delivery. Both of these functions must work holistically and cooperatively in order to ensure acceptable application performance.

This view of IT affects everything – including the organizational structure, the management metrics, the requisite processes, technologies and tools. While the transformation is indeed fundamental, it will not happen quickly. We have spent the last few years coming to understand the importance and difficulty associated with application delivery. As a result of this understanding, many IT organizations have deployed a first generation of tools typically in a stand-alone, tactical fashion, and have also utilized frameworks such as the IT Infrastructure Library (ITIL) in an attempt to implement more effective processes.

As we enter the next phase of application delivery, leading edge IT organizations will develop plans for evolving from a stove-piped IT infrastructure function to an integrated application delivery function.

Senior IT management needs to ensure that their organization evolves to where it looks at application delivery holistically and not just as an increasing number of stove-piped functions

This transformation will not be easy, in part because it crosses myriad organizational boundaries and involves rapidly changing technologies never before developed by vendors, nor planned, designed, implemented and managed by IT organizations in a holistic fashion. Another one of the goals of this handbook is to help IT organizations plan for that transformation in spite of the challenges created by the difficult economic environment. Hence, the 2009 application delivery handbook is subtitled: A Guide to Decision Making in Challenging Economic Times.

Foreword to the 2009 Edition

This handbook builds on the 2008 edition of the application delivery handbook. However, every chapter of the 2008 edition of the handbook has been modified prior to inclusion into the 2009 edition. First, information that was contained in the 2008 edition that is no longer relevant was deleted from this edition. This included anecdotal input from IT organizations that focused on the importance of application delivery. That material was eliminated, as most IT organizations currently understand the importance of the topic.

Second, information was added to increase both the breadth and depth of this edition. For example, later in this chapter of the handbook there is a section (The Perfect Storm) that crystallizes how some emerging technologies will complicate the task of application delivery. The end of this chapter of the handbook also contains a section (The Challenging Economy) that discusses what the economic environment is likely to mean to the priority that IT organizations place on application delivery in 2009.

Because of the need for senior IT managers to ensure that their IT organization approaches application development and application delivery holistically, there is a new chapter (Chapter 3) that discusses the role of the CIO. Also, given the growing importance of virtualization:

- Chapter 4 contains a section that highlights the advantages and challenges associated with virtualized servers, desktops and storage.
- Chapter 6 contains a section that discusses the virtualization of network and application optimization appliances.
- Chapter 8 contains a section that describes the management issues associated with virtualized servers.

Chapter 5 makes the case that most IT organizations would be more effective at application delivery if they had fewer tools. In particular, chapter 5 makes the case that it is important that the IT professionals that are planning the enterprise network use at least some of the same tools as do the IT professionals that manage the network. Chapter 5 also contains a framework for ensuring that applications that are developed over a LAN, will perform well when run over a WAN.

Chapter 6 and Chapter 11 discuss the fact that the functionality necessary to ensure packet delivery is not sufficient to ensure application delivery. Towards that end, Chapter 6 defines what is meant by an application delivery network (ADN). Chapter 11 applies the application delivery framework in order to describe the key characteristics of an ADN.

Chapter 7 contains a discussion of the fact that when an IT organization develops a strategy for application delivery, it is not sufficient to just consider applications that are delivered to the enterprise's branch office employees over network technologies such as private lines, frame relay, ATM or MPLS. As part of their strategy, IT organizations must also consider how they will extend their application delivery strategy to myriad constituencies (e.g., employees, customers, suppliers, distributors) who access applications over the Internet.

In addition to describing the management issues associated with virtualized servers, Chapter 8 also presents a framework for successful application performance management and discusses the importance of automation. Chapter 9 expands on the discussion of an Integrated Operations Center that was included in the 2008 edition of the handbook.

Chapter 11 is new. Chapter 11 uses the application delivery framework to describe both the key characteristics of an application delivery network as well as how to ensure the successful delivery of a key application - voice over IP.

In addition, in order to avoid just presenting a theoretical framework, the 2009 application delivery handbook also contains a number of case studies. These case studies were written by IT organizations and are intended to highlight specific steps that these IT organizations have taken to become better at application delivery.

Unfortunately, the handbook is somewhat lengthy. It does not, however, require linear, cover-to-cover reading. A reader might start reading this handbook in the middle and use the hyperlink references embedded in the text as forward and backward pointers to related information.

The Perfect Storm

The goal of this section is to describe how the deployment of some emerging technologies, each of which adds significant value, will greatly complicate the task of application delivery. This section is not intended to be unduly pessimistic, but is intended to guide the types of application delivery solutions[2] that IT organizations implement.

The application delivery solutions that IT organizations deploy must be able to scale to support clearly discernable emerging requirements.

As is discussed in Chapter 4, IT organizations have already made significant deployment of varying forms of virtualization and intend to make further deployment in 2009. As such, in the not too distant future it will be common for a user in a branch office to utilize a virtualized desktop. That user will likely access the branch office router over a virtual LAN (VLAN). The branch office router may well change in the near term. In addition to routing, the router may also host virtual machines that support a variety of applications and/or Web services. In addition, since the deployment of WAN Optimization Controllers (WOCs) is increasing, in the near future it will be much more likely than it is today that the data flow within the branch office transits a WOC before it hits the WAN. However, this may not be the traditional WOC. For example, in addition to providing standard WOC functions such as caching, compression and protocol acceleration, this WOC will also host virtual machines that provide network services such as DNS and DHCP. Given the ever-increasing concern about security, in the near future it will be even more likely than it is today that there will also be a firewall in the branch office. This may be a traditional firewall, or firewall software running on a virtualized machine, possibly inside either the router or the WOC.

[2] Application delivery solutions refers to a combination of the people, tools and processes necessary to ensure successful application delivery.

The data flow next transits a WAN link that today is almost always a terrestrial link. However, for both backup and performance reasons, we will see the deployment of 3G links that have delay characteristics that will further exacerbate the WAN performance issues. Upon entering the data center, the traffic hits a virtualized application delivery controller (ADC). After transiting the ADC, the next step for the traffic is to transit a number of virtualized web servers, application servers and database server, each of which may or may not be isolated from each other by virtualized firewalls. When the application requires data it gets it from a pool of virtualized storage.

Virtualization, however, is not the only emerging technology that will complicate application delivery. New application architectures such as a Service-Oriented Architecture (SOA) with Web services and Web 2.0 will also complicate application delivery.

In a Web services-based application, the Web services that comprise the application typically run on servers that are housed within multiple data centers. In many instances, at least some of these Web services reside in data centers owned by a company's partners, customers and suppliers. As a result, the IT organization has little insight into, or control over, what is happening in those data centers. In addition, the WAN impacts multiple traffic flows and hence has a greater overall impact on the performance of a Web services-based application that it does on the performance of an n-tier application.

Many IT professionals associate the phrase Web 2.0 with social networking sites such as MySpace. While that is reasonable, one of the most concrete aspects of Web 2.0 is not what it does, but the fact that Web 2.0 applications are typically constructed by aggregating other applications together. This has become such a common concept that a new term, mashup, has been coined to describe it. According to Wikipedia, a mashup is a web application that combines data from more than one source into a single integrated tool.

Mashups are powerful but challenging. When you have an application that calls on another application that is designed, controlled and operated by another organization, you have given up virtually all visibility and control over that piece of your overall application. If there is an availability or performance problem, in many cases you have little recourse other than to wait for the problem to go away.

The Challenging Economy

The worldwide economy is mercurial at best. For example, in the last year, the Dow Jones Industrial average (DJIA) hit a high of 13,990 and a low of 7,882. In part due to the volatile nature of the economy, virtually all companies are reevaluating their strategies and are trying to decide how aggressive or conservative they should be in 2009. These decisions have a direct impact on IT organizations.

Given all of the uncertainty relative to the economy, predicting what IT budgets will look like in 2009 is extremely difficult. In November of 2008 Webtorials surveyed three hundred IT professionals about a number of topics, including the impact of the economy on their IT organization. Two thirds of the survey respondents indicated that they anticipated that the global economic conditions would have either a moderate or a severely negative impact on their IT operations in 2009. In particular, forty percent of the survey respondents stated that they expected at least some decrease in their IT budget in 2009. In addition, roughly 50% of the survey respondents indicated that IT would be more important in 2009. The most viable interpretation of those apparently contradictory data points is that business unit managers will be under heavy pressure to meet their goals in 2009 and will look to the IT organization for additional help.

At any point in time, the typical IT organization has a number of new initiatives underway. Typically some of these initiatives are strategic in orientation while others are more tactical. In challenging economic times there is usually more attention paid to tactical initiatives in general, and to cost cutting initiatives in particular. Many of the technologies and services discussed in this handbook can result in significant cost savings. For example, the virtualization initiatives that are discussed in Chapter 4 are likely to get a lot of attention in 2009. Unfortunately, this will result in some significant management challenges. Because of their ability to both mitigate the need for more bandwidth and improve server performance, there is likely to be heightened interest in deploying both WOCs and ADCs in 2009. IT organizations have been automating network management functions for decades. Due in part to the resultant cost savings; IT organizations will likely take a hard look at stepping up their use of automation in 2009.

Managed Service Providers (MSPs) are nothing new. The survey respondents indicated that over 80% of enterprises currently use an MSP. In addition, almost 45% of the survey respondents indicted that they expected that their organization would make additional use of MSPs in the future. As noted, many of the survey respondents indicated that their company might cut the IT headcount and the capital budget and yet still demand new functionality. These conflicting goals would seem to be further evidence that the use of MSPs will increase in 2009.

The bottom line is that the demanding economic environment will exacerbate the challenges associated with successful application delivery while simultaneously increasing its importance.

3.0 The Role of the CIO

Successful CIOs spend three quarters of their time focused outside the IT organization, managing relationships with the company's business and functional managers as well as with customers, partners, vendors and other important stakeholders. That is a critical role for the CIO to play for several reasons, including the fact that in many cases it is up to the CIO to sell the company's senior managers on the advantages of making a major shift in technology and to demonstrate how technology can be leveraged to meet aggressive business goals.

In addition, in many organizations CIOs are part of the Executive Team and report to the CEO. As such, they play a critical role in helping to shape and manage company strategies and business directions. For example, in the case study entitled Selling the Benefits of SOA/BPM to Senior Business Leaders, Frederic Kunzi talks about the approach he developed to get management agreement to move forward with implementing a Service Oriented Architecture combined with advanced Business Process Management capabilities

CASE STUDY: Selling the Benefits of SOA/BPM to Senior Business Leaders

By Frederic Kunzi, CIO at LCEC
September 12, 2008

Over the course of my IT career I had the privilege to evolve with many innovations and take part in implementations that have produced tremendous value and business benefits for companies around the globe. Planning and deploying multi-tier architectures with middleware concepts in the 1990s to Service Oriented Architectures combined with Business Process Management provided me with the ability to build knowledge on how to introduce, validate and sell technology based solutions to senior Executives. The objective of this brief case study is to share some of my experiences in positioning SOA/BPM and leveraging technology for better business results.

There are many publications and discussion forums that cover SOA though most of the times they are focusing on technology. Selling the concept of SOA/BPM requires a well defined set of business objectives. Consider SOA/BPM as enablers for business process automation. Picture a place where processes are the real drivers invoking services that are exposed by core applications or by in-house developed functional service modules; or presented by a partner/supplier through a B2B server, or from employees or customers through Web Portals! Picture how end-to-end process automation can add value inside and outside the boundaries of your Enterprise fabric including customers and partners!

The vision I shared with many leaders during an initial phase, is that business and operation teams should have full management control on all the processes they own. A business team should be able to come together in a meeting room, project a living process on a screen and see how a given process is working and performing in real time mode. Team members may detect a problem in a sub process that is highlighted in red with access to life data indicating how the sub process is performing or mal performing. At that time they may elect to take the process out of the production context and run a simulation based on a set of newly defined assumptions. They may run several iterations before they find the right modus operandi! Once they agree on a scenario they can go back and post the process for modification and redeployment. Team members will further receive follow-up analysis and reports to validate changes and show return in productivity enhancements. Notifications will follow when a process is operating outside agreed upon thresholds for proactive actions to be initiated.

Some of the real benefits of deploying SOA/BPM can be summarized in three domains of interest. The first is increased efficiency in the way a company is doing work. In other words deliver more in a better way a shorter time and at less cost. The second refers to agility and flexibility and the ability to have a positive impact on Change Management. The third and perhaps the most important one is to enable control of your processes understand the state of any given process at any given time; understand how they perform and where they don't. All three domains need to be matched with metrics and control points linked to business intelligence including analytics and reporting capabilities.

In regard with a return on investment analysis, SOA/BPM returns will emerge from taking multiple processes from their initial stage to an automated stage. Taking a step by step approach starting with simple processes and evolving as your organization gets skilled to tackle more complex processes. Several leading consulting firms issued SOA maturity models that are great tools to explain how a gradual deployment should take place and how business value is building up step by step.

The road to success depends on many aspects; here is what you may consider:

- Strategic alignment
- Senior management Support
- Planning and budgeting for a multi-year initiative
- SOA Governance linked with PMO and a strong Process Practice on board
- Establish an Enterprise Business Process Architecture
- Standardize on one process methodology, one process repository, process simulation and process publishing capabilities.
- Earmark cross functional process improvement initiatives Order to Cash, Inventory, etc.
- Enable Business Intelligence if not available in your enterprise
- Take a gradual and controlled approach
- Report back on all winning steps to keep the momentum going

To summarize, keep the focus on business aspects and benefits, solving business problems should be your primary objective. Keep in mind that the deployment of an SOA/BPM initiative may take two to three years to show a real return. The most attractive outcome is that the benefits will cumulate as you are bringing more and more processes to automation. There are several levels of benefits that you will be able to track. Level one benefits directly related to changes brought from an "as is" to an "as to be" process while secondary and perhaps third level benefits are not always visible at first hand they will emerge later in the post deployment phase.

From an IT perspective implementing SOA/BPM will yield benefits in many aspects as more applications are getting service enabled. You will also have the opportunity to gradually build a rich library of service modules that you will be able to use to create flexibility, improve productivity, reduce costs, shorter development time for new business requirements and improve time to deliver. Make sure that you have a good handle on all processes that you are planning for automation such that you are able to measure and benchmark progress. Succeeding in supporting business transformation by providing stronger linkages between businesses and enabling IT should be a main objective of your strategy.

It is critical that CIOs continue in their traditional role of interfacing with key stakeholders; however, CIOs must find time to create fundamental change in the IT organization.CIOs must lead the transformation of the IT organization to where it has a focus on application delivery that it tightly integrated with its approach to application development.

The CIO Mandate

Much of the impetus for transforming the IT organization to date has come from the bottom-up efforts of IT professionals who see the technological and organizational stovepipes and work to reduce their impact. There are, however, limitations to how broad a transformation of the IT function can take place by only a bottom-up approach.

In order to transform the IT organization such that it can effectively deliver applications, CIOs must get involved in setting the direction and monitoring progress.

CIOs must get involved in setting the direction and monitoring progress in every aspect of application delivery from planning to ongoing operations. For example, Chapter 4 of the handbook discusses the flawed application development process. As we point out in that chapter, in the typical IT organization 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. CIOs must drive change within the application development organization to where the application's performance over the WAN is a key component of the application development process.

Chapter 9 describes the changing role of the Network Operations Center (NOC). As recently as a few years ago, NOC personnel spent the majority of their time managing the availability of *networks*. However, a recent survey of IT professionals indicates that the role of the NOC is shifting and that NOC personnel now spend the bulk of their time on the availability and performance of *applications*.

The majority of the survey respondents indicated that, in an attempt to facilitate change, their organizations have used a framework such as IT Infrastructure Library (ITIL) to develop more effective IT processes. Of those who have not, the majority stated that their organization would use such a framework within the next 12 months. While the significant efforts NOCs have undergone to improve their processes have had some impact, they still have a long way to go. Almost a third of the survey respondents indicated their NOC is not meeting the organization's current needs.

The primary reason the NOC is only moderately successful at improving its application delivery processes is that the task is complex and cuts across multiple organizational boundaries. Without strong support from their CIO, few NOCs will be successful implementing processes that cut across multiple groups inside the IT organization. What is needed is an Integrated Operations Centre (IOC), whose goal is to ensure acceptable application performance by both proactively eliminating problems before they impact the end user, and responding more quickly to problems when they do impact end users.

CIOs must drive the vision of an Integrated Operations Centre (IOC)

In addition to driving the vision for the IOC, the CIO should appoint a champion whose role is to:

- Create a roadmap leading to the implementation of an effective IOC
- Manage the transition from a NOC to an IOC, and
- Market the value of the IOC both within the IT organization and more broadly within the company.

One example of an organization that is moving in this direction is PSS World Medical, Inc. The case study entitled *Creating an Application Delivery Function,* discusses the process that PSS World Medical, Inc. went through to shift from having an infrastructure organization that focused on stove piped technologies to one that focused on delivering applications to customers.

CASE STUDY: Creating an Application Delivery Organization

In the world of IT infrastructure, the search for performance improvements typically focuses on hardcore technology principles – uptime, change management, service packs, chip speeds, etc. After all, the NOC is the last bastion of "old-school" IT. While analysts and developers are out researching Open Source and SOA, the folks in the infrastructure team are still lifting servers into racks and plugging them in by hand. With the exception of virtual server "voodoo", it's generally business as usual in the NOC.

At PSS World Medical, we found that a subtle, psychological approach was effective in promoting change in our IT Infrastructure team. The change was unexpected, even accidental at times, but we have infused a holistic, customer-centric mentality into a group once focused on power cables, KVM switches and temperature sensors.

The pathway to progress began with the acquisition of a network management tool. The network guys would typically have reviewed the possible options and made the final selection on their own. But in an effort to break down some of the silos we knew existed in our shop, we brought the server and storage folks to one of the demos, as well as some senior IT management. Not only did the rest of the team feel like they had some input on the overall decision, but they found that the tool highlighted improvement opportunities in servers and applications – not just the network.

The result was that the entire infrastructure team embraced the tool and its use, and they began to work together to interpret the data it provided. Within weeks of the implementation, we were scheduling "war rooms" on Fridays, where the entire team would work together in a conference room all day while viewing the live data from our network and servers. The tool had provided the common bond that the team needed, and the change was underway.

About six months later, I was at an industry conference, and was presented with the idea that the essential purpose of an infrastructure teams is to deliver applications. Having made some progress in breaking down silos, the next step was to focus the group's attention on our end customer. With that in mind, we changed the name of the IT Operations team to the Application Delivery Group. The idea was simple – transform the mentality of the server jockeys from 'keeping a bunch of hardware running' to 'keeping a bunch of customers in business.' The complexity and interdependencies of today's IT architectures are mandating that infrastructure teams focus on their end product, not just the components within the data center.

The psychological effect of the change was that it prompted the group to open dialogue with our users about how they actually use the applications. As they internalized the impact of downtime or outages, the importance of SLAs became clearer. And as they responded more quickly to our customers' needs, the relationships with the project and support teams within IT improved. Speaking months later about the name change, our Senior Network Engineer said it best: "IT Operations sounds cold, sterile. But the Application Delivery Group clearly has customers." This is exactly the result we were hoping for.

I took several important lessons from this experience. First, basic leadership principles DO apply to IT infrastructure teams – regardless of the current structure of your NOC, there are fundamental frameworks through which any measure of change can be achieved. Second, we must all challenge the conventional view of data center management. The idea that all is well as long as all lights are green is a thing of the past. If your infrastructure teams don't understand their customers' needs, they can't be expected to satisfy them. Third, it is important to remember that "if you expect it, inspect it." We kept a close eye on our progress through this change, but we didn't track specific metrics like resolution times or problems avoided. Had we spent more time measuring the impact of our change, we could report results with numbers instead of anecdotes.

Our focus on Application Delivery has yielded some great results. I wish you luck if you decide to go down the same path.

The Positioning of the CIO

Surveyed IT professionals were asked three questions about their CIO. Throughout this section, that survey will be referred to as The CIO Survey. To gain further insight, Stuart McGuigan, the CIO of Liberty Mutual, was interviewed. The responses to the survey question are shown in Tables 3.1 – 3.3. Note that while some members of the community believe their CIO is doing a great job, the overall impression is that there is some room for improvement.

Aligning IT with Business

The CIO Survey asked first about the alignment of the CIO with the company's business and functional managers. Responses are shown in Table 3.1.

Survey Question		Disagree			Agree	Agree	Agree Strongly
Our CIO is an innovative, strategic thinker who is closely aligned with our company's senior business and functional managers	2.5%	3.5%	4.9%	23.5%	13.7%	21.4%	30.5%

Table 3.1: The CIO as Strategic Thinker

McGuigan stated that his IT organization was very well aligned with the company's business and functional managers and that this alignment was the result of a lot of hard work. He pointed out that one of the ways that they developed better links with the company's business managers was to adopt an organizational model similar to that of the overall business. In particular, Liberty Mutual is a fairly decentralized organization. In response to this business model, when McGuigan started at Liberty Mutual four years ago he created the position of CIO for each of the business units. While these CIOs report to McGuigan, they function as though they report to the head of the business unit.

The organizational model for the IT function needs to be similar to the organizational model for the enterprise.

Another initiative that McGuigan credits with improving the alignment of IT with the business managers is that where appropriate they have implemented common technology. For example, McGuigan mentioned that they have a highly standardized desktop, allowing them to cut cost while improving quality metrics as there are fewer technologies to support. He did add that they do have an exception process to deal with situations in which the standardized solution does not meet the business need.

Operational Focus

The CIO Survey respondents were asked if their CIO had a strong operational focus. Their responses are shown in Table 3.2.

Survey Question		Disagree			Agree	Agree	Agree Strongly
Our CIO has a strong focus on operational issues such as assuring availability and minimize cost	1.4%	2.8%	3.5%	18.6%	17.9%	22.1%	33.7%

Table 3.2: The CIO's Focus on Operations

McGuigan said that when he first arrived at Liberty Mutual he spent 80% of his time on operational issues. This approach was necessitated by the fact that they were experiencing an unacceptable level of outages in some of their core applications. As availability issues were resolved, McGuigan spent less and less time on operations, now spending only 20% of his time in that arena. He added that the percentage of time that he spends on operational issues varies by business unit based on their level of business and IT maturity.

Leveraging Technology for Business Value

The CIO survey respondents were next asked if their CIO was a leader in terms of leveraging technology for business value. Their responses are shown in Table 3.3.

Survey Question		Disagree			Agree	Agree	Agree Strongly
Our CIO has a strong grasp of technology and how technology can be used to drive business value	2.5%	3.2%	7.0%	22.2%	15.5%	22.2%	27.5%

Table 3.3: The CIO's grasp on technology and its use

McGuigan stated that he regards himself as a business leader who manages technology – and his approach is that any IT expenditure must have business value. He agreed with the conventional wisdom that, after the dot com implosion, IT was seen as having little direct business value and that view is now changing. According to McGuigan, we are at a turning point in terms of how the customer gets supported. As he asserts, this type of work used to be very people intensive but is being increasingly automated -- and allowable business processes are determined by business rules inside relevant enterprise applications. As such, business and IT professionals need a higher level of understanding of the applications than was previously necessary. McGuigan also stated that he has adopted an approach that was first put forth by one of Liberty Mutual's senior managers, which is that "there is no such thing as an IT project, just business projects that have an IT component."

CIO Priorities

Table 3.4 highlights the top five CIO initiatives from three different perspectives. The first column reflects what the author believes should be the CIO's priorities (Jim Metzler's Priorities). The second column reflects what the respondents to The CIO Survey[3] stated their CIO's priorities are (CIO Priorities) and the third column is what the surveyed community believed the CIO's priorities should be (Survey Respondent's Priorities).

Jim Metzler's Priorities	CIO Priorities	Survey Respondent's Priorities
Develop a strong application delivery function	Cost Control	Upgrade significant portions of the IT infrastructure
Develop and acquire more WAN friendly applications	Enhance security	Break down organizational and technological stovepipes
Break down the organizational and technological stovepipes	Align goals across all of IT	Develop and acquire more WAN friendly applications
Implement more effective processes both within IT and between IT and the rest of the business	Demonstrate the business value of IT	Implement more effective processes both within IT and between IT and the rest of the business
Cost Control	Develop a strong application delivery function	Simplify IT – both applications and infrastructure

Table 3.4: Summary of CIO Priorities

One immediate conclusion is that two of Jim's Priorities are also two of the CIO's Priorities: cost control and develop a strong application delivery function. In addition, the other three of Jim's Priorities are also part of the Survey Respondent's Priorities: develop and acquire more WAN friendly applications, break down the organizational and technological stovepipes, and implement more effective processes both within IT and between IT and the rest of the business. Notably – and unfortunately – there was no match at all between the CIO's Priorities and the Survey Respondent's Priorities.

In many instances, there is little overlap between a CIO's priorities and what the IT organization believes those priorities should be.

McGuigan stated that one of his top priorities for the next year is situational. Liberty Mutual recently acquired a new company. Therefore, one of McGuigan's top priorities is the integration of that company. He added that another of his top priorities is continually to improve their processes. To exemplify the progress Liberty Mutual has already made relative to process improvement, McGuigan told us that, as recently as a year and a half ago, his organization took one hundred days from the time they got a request for a new UNIX server until that server was actually in production. This lengthy implementation interval was because a) a large number of groups were involved in the implementation and b) the groups

[3] The CIO Survey was conducted in the summer of 2008, shortly before the economic situation worsened.

worked in a serial fashion. As a result of changing their processes and breaking down organizational silos, they now implement a new UNIX server in under ten days.

Additional insight into how IT organizations have successfully removed organizational silos in order to provide better service can be found in the case study entitled Breaking Down Stovepipes, by Michael Fergang, the CIO of Grange Insurance.

CASE STUDY: Breaking Down Stovepipes

Michael Fergang
CIO
Grange Insurance

At Grange, our mantra is Ease of Doing Business®. See? We've even registered it as a trademark. When applied to agents, Ease Of Doing Business –or EODB®–means the agent knows that when he has a customer in his office he will be able to connect to our portal easily and quickly—every agent, every time. When applied to policyholders, again, EODB means that they know they can come to our website and get what they need quickly and efficiently.

To achieve these goals in a high-growth organization such as Grange—we doubled our size from $500 million to $1 billion and extended our geographical reach from six to 13 states in the past seven yearse—you simply can't allow stovepipes. Knowledge hoarding is unacceptable. There are massive projects – ranging from networks to storage to servers and applications and securitye— that impact multiple teams. Success is predicated on a mature process. The mature process is what breaks down the silos. There are effectively no silos if there is a process that people follow and to which they are held accountable.

At Grange, we have cultivated a mature development lifecycle—we go through formal requirements gathering and we then go through an iterative process with IT and business owners, as well as quality assurance officers, to make sure that everyone who is impacted or could be impacted is on the same page.

We are presently applying this process as we extend our Ease Of Doing Business model directly to our more than one million policyholders to offer enhanced self-service via our website. These policyholders want to visit our site to do business and transact with us. We need to be competitive in servicing the customer in the way in which they want to be serviced, as this then becomes a competitive advantage for our agents. Phase One will go live in January 2009. This will enable policyholders to do basic document lookup and querying, such as being able to view and pay their bill online, look up current claim status, send an email to their agent or even conduct an insurance "tune up." Perhaps most importantly, they will be able to print that very necessary–but easily lost–proof of insurance card.

Even with a crystal-clear process, you need a person in the driver's seat so that each project has a "go-to guy"—the one person who is accountable even if it is not their core competency. This is especially important when we are working with business departments that don't necessarily understand IT and all its intricacies. We have found that one face of the IT organization is critical to success.

Technology is a big piece of the puzzle as well. We invest in appropriate technologies that enable collaboration—for instance, tools that allow the network and application teams to resolve issues quickly and tools that are not exclusive to IT but are extremely important from a workflow perspective.

We are continually expanding our market and breadth of services while, at the same time, continuing to add customers and agents. The level of sophistication of our applications and the demands on the network will continue to grow (e.g. Flash technology and Web 2.0 – this is our future). We need to continually provide better user experience while being sensitive to the implications on the network.

Our process is constantly put to the test. Stay tuned!

4.0 The Applications Environment

This section of the handbook discusses some of the primary dynamics of the applications environment impacting 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 single product or service in the marketplace provides a best in class solution for every component of the application delivery framework. As a result, companies must 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 most situations application development focuses on ensuring that applications are developed on time, on budget, and with few security vulnerabilities. Such narrow focus, combined with the fact that application development has historically been done over a high-speed, low-latency LAN, means that the impact of the WAN on the performance of the application often remains unknown until after the application is fully developed and deployed.

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 an application's performance over the WAN often results in the deployment of "chatty" applications as illustrated in Figure 4.1.

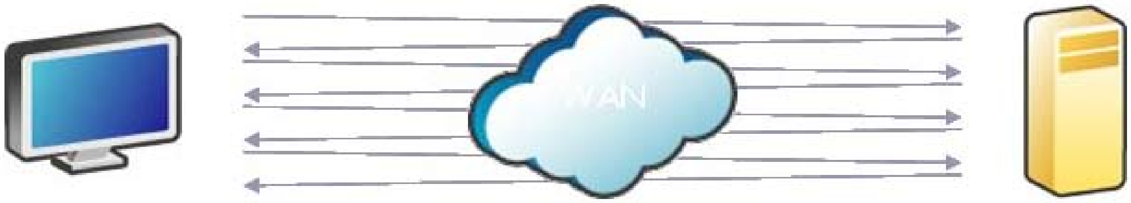


Figure 4.1: Chatty Application

A chatty application requires hundreds of application turns to complete a transaction. To exemplify the impact of a chatty protocol, let's assume that a given transaction requires 200 application turns. Further assume that the latency on the LAN on which the application was developed was 1 millisecond, but that the round trip delay of the WAN on which the application will be deployed is 100 milliseconds. For simplicity, the delay associated with the data transfer will be ignored and only the delay associated with the application turns will be calculated. In this case, the delay over the LAN is 200 milliseconds, which is generally not noticeable. However, the delay over the WAN is 20 seconds, which is very noticeable.

In such a case, it's obvious that developers need to be cognizant of the impact of the WAN on application performance *during* the application development lifecycle. In particular, it is important during application development to identify and eliminate any factor that could have a negative impact on application performance. This approach is far more effective than trying to implement a work-around *after* an application has been fully developed and deployed.

The preceding example also demonstrates the relationship between *network* delay and *application* delay.

A relatively small increase in network delay can result a significant increase in application delay.

Taxonomy of Applications

The typical enterprise has tens and often hundreds of applications that transit the WAN. These applications can be categorized as follows:

1. Business Critical

   A company typically runs the bulk of its key business functions utilizing a handful of applications. It can develop these applications internally, buy them from a vendor such as Oracle or SAP, or acquire them from a Software-as-a-Service (SaaS) provider such as Salesforce.com.

2. Communicative and Collaborative

   This includes delay sensitive applications such as Voice over IP (VoIP), telepresence and traditional conferencing, as well as applications that are less delay sensitive such as email.

3. Other Data Applications

   This category includes 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 nevertheless 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 that 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, music downloading and other media sharing.

6. Malicious

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

IT organizations may need to a) optimize the performance of some of their business critical applications, b) control the performance of all applications so that they do not interfere with the performance of applications such as VoIP and c) eliminate any malicious traffic.

Application delivery is more complex than merely optimizing the performance of all applications.

Because 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 that applications such as Citrix's XenApp[4] or SAP comprise multiple modules with varying characteristics. Thus, it is not particularly meaningful to say that Citrix XenApp traffic 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.

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. 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.

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

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

As we discuss in Chapter 10, the dense nature of XML also creates some security vulnerabilities.

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 Common Internet File System (CIFS), Exchange or Network File System (NFS) -- which were designed to run over the LAN -- running over the WAN. CIFS works by decomposing all files into smaller blocks prior to transmitting them. Assuming 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

[4] Citrix XenApp was formerly Citrix Presentation Server

back to the server. The server must wait for an acknowledgement prior to sending the next data block. As a result, the file may take several – noticeable – seconds to open.

Data Center Consolidation and Single Hosting

In addition to consolidating servers, many companies are also reducing the number of data centers they support worldwide. 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 a single data center.

> ***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 instead of the LAN.***

Software as a Service

According to Wikipedia[5], Software as a Service (SaaS) is a software application delivery model in which a software vendor develops a web-native software application and hosts and operates (either independently or through a third-party) the application for use by its customers over the Internet. Customers do not pay to own the software itself but instead pay to use it. They use it through an application programming interface (API) accessible over the Web and often written using Web Services.

There are many challenges associated with SaaS. For example, by definition, the user accesses the application over the Internet and hence incurs all of the issues associated therewith. (See Chapter 7 for a discussion of the use of managed service providers as a way to mitigate some of the impact of the Internet.) In addition, because the company that uses the software does not own the software, they cannot make changes to the software to improve performance.

Fractured IT Organizations

The application delivery function consists of myriad sub-specialties such as devices (e.g., desktops, laptops, point of sale devices, smart phones), networks, servers, storage, servers, security, operating systems, etc. The planning and management of these sub-specialties are typically not well coordinated within the application delivery function. In addition, market research indicates that typically little coordination exists between the application delivery function and the application development function.

[5] http://en.wikipedia.org/wiki/Software_as_a_Service

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 IT operational issues occurs cooperatively across all IT disciplines.

In order to be successful, IT organizations need to evolve from a fractured (a.k.a., CYA) approach to application delivery to a proactive (a.k.a., CIO) approach.

The CYA approach to application delivery focuses on deflecting fault when the application performs badly. The goal of the CIO approach is to rapidly identify and fix the problem without assigning blame.

Application Complexity

Companies began deploying mainframe computers in the late 1960s and mainframes became the dominant style of computing in the 1970s. The applications 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 often referred to as *n-tier applications*. Because these tiers are implemented on separate systems, WAN performance impacts *n*-tier applications more than monolithic applications. For example, the typical 4-tier application (Figure 4.2) is comprised of a Web browser, a Web server(s), an application server(s) and a database server(s). The information flow in a 4-tier application is from the Web browser to the application server(s) to the database, and then back again over the Internet using standard protocols such as HTTP or HTTPS.

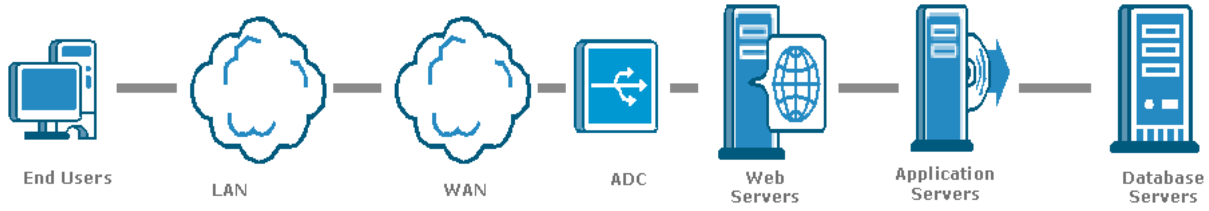


Figure 4.2: A 4-Tier Application

One of the primary reasons why an n-tier application architecture complicates the task of assuring IT service quality is because n-tier applications are inherently complex. For example, the typical n-tier application is comprised of myriad browsers and servers and often includes one or more application delivery controllers (ADCs). As is discussed in Chapter 6, at a minimum an ADC balances the load over multiple servers. In most cases, the ADC also offloads computationally intensive processes from the servers. In addition, the networks that support n-tier applications are comprised of switches, routers, access points, WAN optimization controllers, firewalls, intrusion detection systems and intrusion protection systems. If any of these components are not available or are not performing well, the performance of the overall service is impacted. In some instances, each component of the application architecture is performing well, but due to the sheer number of components, the overall delay builds up to a point where some function, such as a database query, fails. Another type of performance problem arises when the resources provisioned for each component of a complex application are not well matched, causing service performance issues because of one or more bottlenecks in the environment.

As the complexity of the environment increases, the number or sources of delay increases and the probability of application degradation increases in a non-linear way.

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 significantly more than WAN performance impacts n-tier applications.

To understand why the movement to Web services-based applications will drastically complicate the task of ensuring acceptable application performance, consider the 4-tier application architecture previously discussed. In a 4-tier application the Web server(s), 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 housed within multiple data centers. As a result, the WAN impacts multiple traffic flows and hence has a greater overall impact on the performance of a Web services-based application that it does on the performance of an *n*-tier application.

Web Services and Security

The expanding use of Web services creates some new security challenges. Part of this challenge stems from the fact that in most instances, the blueprint for Web services communication is outlined in Web Services Description Language (WSDL) documents. These documents are intended to serve as a guide to an IT organization's Web services. Unfortunately, they can also serve to guide security attacks against the organization.

Assuming that a hacker has gained access to an organization's WSDL document, the hacker can then begin to look for vulnerabilities in the system. For example, by seeing how the system reacts to invalid data that the hacker has intentionally submitted, the hacker can learn a great deal about the underlying technology and can use this knowledge to further exploit the system. If the goal of the hacker is to create a denial of service attack or degrade application performance, the hacker could exploit the verbose nature of both XML and SOAP [6]. When a Web services message is received, the first step the system takes is to read through, or parse, the elements of the message. As part of parsing the message, parameters are extracted and content is inserted into databases. The amount of work required by XML parsing is directly affected by the size of the SOAP message. Because of this, the hacker could submit excessively large payloads that would consume an inordinate amount of system resources and hence severely degrade application performance.

[6] Simple Object Access Protocol (SOAP) is the Web Services specification used for invoking methods on remote software components, using an XML vocabulary.

Chapter 10 discusses some of the limitations of the current generation of firewalls. One of these limitations is that the current generation of firewalls is largely incapable of parsing XML. As such, these firewalls are blind to XML traffic. As part of providing security for Web services, IT organizations must be able to inspect XML and SOAP messages and make intelligent decisions based on the content of these messages. For example, IT organizations must be able to perform anomaly detection in order to distinguish valid messages from invalid messages. In addition, they must be able to perform signature detection to detect the signature of known attacks.

Web 2.0

Defining Web 2.0

As we noted in the preceding section, the movement to a Service-Oriented Architecture (SOA) built on the use of Web services-based applications will drastically complicate the task of ensuring acceptable application performance. The same is true for the movement to Web 2.0. In the case of Web 2.0, however, the problem is further exacerbated because most IT organizations are not aware of the performance issues associated with Web 2.0.

> ***Many IT professionals view the phrase Web 2.0 as either just marketing hype devoid of any meaning or they associate it exclusively with social networking sites such as MySpace.***

While it is reasonable to associate Web 2.0 at least partly with social networking, it is not reasonable to dismiss social networking as not having a place inside the enterprise. In the case study entitled *Is Web 2.0 a viable technology platform in the enterprise,* Josh Hinkle demonstrates his belief that social media is a viable enterprise technology.

CASE STUDY: Is Web 2.0 a viable technology platform in the enterprise?

Josh Hinkle
Mgr, Network Mgt & Security, Department
American Heart Association

Since the term Web 2.0 was coined at the first Web 2.0 conference in 2004 by Tim O'Reilly many technologists have struggled to grasp all that Web 2.0 is and what it means to current and future business models. Are we really talking about a second Internet, or a new and improved Internet built on an existing platform?

The following case study intends to break down the components of the term "web 2.0" as well as confirm Web 2.0's viability in the enterprise environment. The supporting objectives of this case study will

1) Break down the components of web 2.0 as defined at the American Heart Association

2) Discuss the integration to the business model of the American Heart Association and

3) How the network support strategy from the AHA IT perspective has changed to adapt to this emerging technology.

Break down: Web 2.0 is Social Media

The American Heart Association has adopted the idea that the second generation of Internet applications considered Web 2.0 actually refers to Social Media. Social Media describes a way in which information is socialized to communities. That socialization serves as a combination of discrete channels to achieve a desired outcome.

Social media is an umbrella term that defines the various activities that integrate technology, social interaction, and the construction of words, pictures, videos and audio. This interaction, and the manner in which information is presented, depends on the varied perspectives and "building" of shared meaning among communities, as people share their stories, and understandings.

- Wikipedia

The first generation of the Internet was built on HTML and the web pages were full of static information. The second generation of Internet applications is powered by programming languages such as Java, AJAX, XML and flash creating a much more interactive experience for users. This improved experience has led to the rise of mainstream social media in the form of social-networking, Blogs, Wikis, Mashups, and RSS feeds. Social networking sites build communities and provide one on one interaction. Blogs, wikis, mashups and RSS feeds are technologies that connect people and communities with information.

Is Web 2.0 (Social Media) viable in the enterprise network?

Social Media's low cost of entry has made it a very attractive technology to businesses; as a result companies are quickly adapting their business models with the hope of increasing market penetration while lowering operation costs. Much of the focus in social media is around the building of corporate and customer communities using social networking sites such as Facebook and LinkedIn. Mashups, RSS feeds and Blogs are seen as ways for companies to create, maintain and share dynamic information internally as well as externally with customer bases.

The American Heart Association (AHA) is a national voluntary health agency whose mission is: "Building healthier lives, free of cardiovascular diseases and stroke." Over simplified, the AHA achieves its mission by funding life saving research, building awareness and providing information; the AHA believes social media is a tool by which the organization's broader goals can be realized. These activities include, but are not limited to, organization of events, acquisition of volunteers and movement members, cultivation of relationships, facilitation of research, sharing and general networking activities. Three key attributes of social media for AHA include: Networking: Facilitating connections between people; not just one to one but exponential. Participatory: Get people involved, comment, interact, recruit others, agree, disagree, sympathize, empathize and relate. Scalable: Social Networking sites and tools tend to be able to expand to absorb all who wish to be involved.

AHA was an early adopter of Facebook.com serving as one of its "Causes". Facebook was an opportunity for the AHA to start reaching new communities at little to no cost growing its volunteer network "You're the Cure Network". At the time of this case study the "You're the Cure" Facebook page has recruited almost 9000 members. The AHA has also launched a page to network audiences and increase participation for our Start! Walk Campaign. In another effort to increase collaboration AHA's Technology and Customer Strategies Department (TCS) has created a blog encouraging collaboration outside everyday operations and is often used to highlight technology driven company successes. The AHA has launched a Social Media project to identify opportunities and create a long-term strategy.

What is the impact of Social Media / Web 2.0 to the traditional Enterprise Network?

Traditional enterprise network management focuses on enterprise application delivery whereas the applications would fully reside inside the Wide Area Network (WAN) and the ownership and auditing of data and resources could be controlled; the customer experience of these applications is fully measurable and manageable. On the other hand social media is an Internet based platform full of dynamic information contributed from endless communities; access from corporate network s only requires bandwidth to the Internet.

Unlike traditional network management social media applications performance is not as manageable and may vary greatly with the general performance of the World Wide Web. User experience is not as measurable and the content is much richer. As a result the applications often require more resources, most notably bandwidth.

The adoption of social media into the AHA's business model has caused the proposition of 2 separate networks, either virtual or physical. The first network would be a traditional enterprise network transporting enterprise applications. Those applications would be fully managed for Class of Service (CoS) and delivery because all the resources reside within the existing infrastructure. Guaranteeing this type of performance allows the AHA to realize its existing investment in traditional enterprise applications.

The second network would be dedicated to all Internet bound traffic and include social media. The separation of enterprise applications from Internet bound applications (social media) allows the AHA to see the greatest gains in efficient utilization for all network traffic; enterprise applications are in a fully managed environment and see no competition from content rich Internet applications that can not be managed in an enterprise environment.

Conclusion: Web 2.0 is viable

The AHA believes that Social Media is a viable enterprise technology. Like most technologies there is a maturation period and this technology is still young; the AHA believes social media will continue to mature and gain market adoption as a leading technology."

The AHA is adapting its business model to integrate these technologies and the network team is changing its support strategies to be better aligned with business needs.

A key component of Web 2.0 is that the content is very dynamic and alive and that as a result people keep coming back to the website. The concept of an application that is itself the result of aggregating other applications has become so common that a new term, *mashup*, has been coined to describe it. According to Wikipedia [7] a mashup is a web application that combines data from more than one source into a single integrated tool.

Another industry movement often associated with Web 2.0 is the deployment of Rich Internet Applications (RIA). In a traditional Web application all processing is done on the server, and a new Web page is downloaded each time the user clicks. In contrast, an RIA can be viewed as "a cross between Web applications and traditional desktop applications, transferring some of the processing to a Web client and keeping (some of) the processing on the application server." [8] RIAs are created using technologies such as Adobe Flash Player, Flex, AJAX and Microsoft's Silverlight .

Webtorials recently presented over 200 IT professionals with the following question: "Which of the following best describes your company's approach to using new application architectures such as Services Oriented Architecture (SOA), Rich Internet Applications (RIA), or Web 2.0 applications including the use of mashups?" Their responses are shown in Table 4.1.

[7] http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid)

[8] Wikipedia on Rich Internet Applications:http://en.wikipedia.org/wiki/Rich_Internet_Application

Response	Percentage of Respondents
Don't use them	24.4%
Make modest use of them	37.2%
Make significant use of them	11.7%
N/A or Don't Know	24.4%
Other	2.2%

Table 4.1: Current Use of New Application Architectures

The same group of IT professionals was then asked to indicate how their company's use of those application architectures would change over the next year. Their responses are shown in Table 4.2.

Response	Percentage of Respondents
No change is expected	23.3%
We will reduce our use of these architectures	1.7%
We will increase our use of these architectures	46.7%
N/A or Don't Know	27.8%
Other	0.6%

Table 4.2: Increased Use of New Application Architectures

Emerging application architectures (SOA, RIA, Web 2.0) have already begun to impact IT organizations and this impact will increase over the next year.

Quantifying Application Response Time

As noted, Web 2.0 has some unique characteristics.

In addition to a services focus, Web 2.0 characteristics include dynamic, rich and in many cases, user created content.

A model is helpful to illustrate the potential performance bottlenecks in any application environment in general, as well as in a Web 2.0 environment in particular. The following model (Figure 4.3) is a variation of the application response time model created by Sevcik and Wetzel[9]. Like all models, the following is only an approximation and as a result is not intended to provide results that are accurate to the millisecond level. It is, however, intended to provide insight into the key factors impacting application response time. As shown below, the application response time (R) is impacted by amount of data being transmitted (Payload), the WAN bandwidth, the network round trip time (RTT), the number of application turns (AppTurns), the number of simultaneous TCP sessions (concurrent requests), the server side delay (Cs) and the client side delay (Cc).

[9] Why SAP Performance Needs Help, NetForecast Report 5084, http://www.netforecast.com/ReportsFrameset.htm

$$R \approx \frac{Payload}{Goodput} + \frac{(\#\ of\ AppsTurns\ *\ RTT)}{\text{Concurrent Requests}} + Cs + Cc$$

Figure 4.3: Application Response Time Model

The WOCs described in Chapter 6 were designed primarily to reduce the effective size of the payload and the number of application turns, and to increase goodput and mitigate the impact of chatty applications and protocols. The Application Delivery Controllers described in Chapter 6 were designed primarily to offload communications processing from servers; they were not designed to offload any backend processing. Neither of these solutions impacts the client side delay.

The Web 2.0 Performance Issues

As noted, existing network and application optimization solutions were designed to mitigate many of the factors highlighted in Figure 4.3. Microprocessor vendors such as Intel and AMD continually deliver products that increase the computing power available on the desktop. As a result, these products minimize the delays associated with client processing (Cc). This leaves just one element of the preceding model that must be more fully accounted for – server side delay. This is the critical performance bottleneck needing to be addressed in order for Web 2.0 applications to perform well.

> ***The existing generation of network and application optimization solutions does not deal with a key requirement of Web 2.0 applications: the need to massively scale server performance.***

The reason this is so critical is that unlike clients, servers suffer from scalability issues. In particular, servers must support multiple users and each concurrent user consumes some amount of server resources: CPU, memory, I/O. Chris Loosley[10] highlighted the scalability issues associated with servers. He pointed out that activities such as catalog browsing are relatively fast and efficient activities that do not consume a lot of server resources. He contrasted that with an activity that required the server to update something, such as clicking a button to add an item to a shopping cart. His points out that activities such as updating consumes significant server resources, so the number of concurrent transactions -- server interactions that update a customer's stored information -- plays a critical role in determining server performance.

Virtualization in the Application Environment[11]

Another factor impacting application delivery is that it is now possible to implement most components of the IT infrastructure in a virtualized fashion. As we discuss below, there are many advantages to virtualizing the IT infrastructure. However, the virtualization of IT resources introduces management, security and performance issues that can significantly impact the ability of the IT organization to ensure acceptable application performance.

In general, virtualization involves a logical abstraction of physical systems that allows one of the following:

- A single physical system to be partitioned to appear as multiple independent logical systems; e.g., multiple VLANs defined on a single physical LAN.
- Multiple physical systems to appear as single logical system; e.g., a compute cluster with a single system image or RAID disk array appearing to be a single large, reliable disk.

[10] Rich Internet Applications: Design, Measurement and Management Challenges, Chris Loosley, http://www.keynote.com/docs/whitepapers/RichInternet_5.pdf

[11] This section was written in cooperation with Rolf McClellan

The Application Environment of the data center can potentially support a wide variety of virtualization technologies including, but not limited to:

- Virtual Servers and Virtual Appliances
- Virtual Desktops
- Virtual Storage
- Virtually Partitioned Appliances; e.g. a firewall or server load balancer with multiple logical partitions
- Clustered Appliances or Servers; e.g., multiple WAN Optimization Controllers or Application Delivery Controllers clustered for higher availability and performance

In a survey completed in August 2008, 205 IT professionals were asked to indicate how much deployment their organization will have made of desktop, server and storage virtualization by the end of next year. Their answers are shown in Table 4.3.

	None	Some	Moderate Amount	Significant Amount	Very Significant Amount
Server Virtualization	7.0%	27.6%	24.4%	22.4%	18.6%
Storage Virtualization	16.9%	19.7%	26.1%	22.5%	14.8%
Desktop Virtualization	34.7%	27.2%	21.8%	12.2%	4.1%

Table 4.3: Anticipated Deployment of Virtualization

As the data in Table 4.3 indicates, IT organizations have significant interest in deploying a wide range of virtualization forms.

Server and storage virtualization have crossed the chasm and are now mainstream technologies.

The deployment of desktop virtualization lags behind that of server and storage virtualization.

The remainder of this section of the handbook focuses on the potential benefits, as well as the challenges, of deploying server, desktop, and storage virtualization in the application environment. A more detailed discussion of Virtual Appliances, Partitioned Appliances, and Clustered Appliances is provided in Chapter 6

Virtual Servers and Virtual Appliances

Server virtualization based on Virtual Machine (VM) software is becoming a popular solution for consolidation of data center servers. With VM software, a single physical machine can support a number of guest operating systems (OSs), each of which runs on its own complete virtual instance of the underlying physical machine, as shown in Figure 4.4. The guest OSs can be instances of a single version of one OS, different releases of the same OS, or completely different OSs; e.g., Linux, Windows, Mac OS-X, or Solaris.

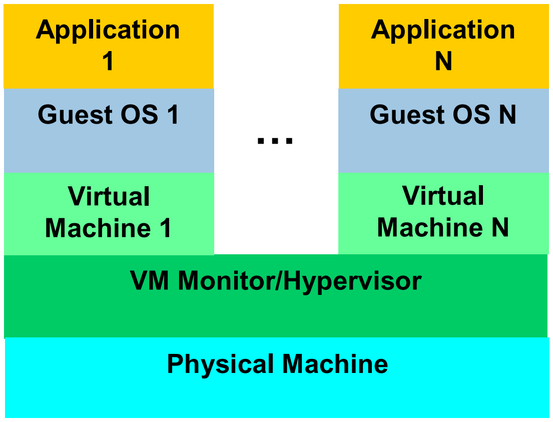


Figure 4.4: Simplified View of Virtual Machine Technology

A thin software layer called a *Virtual Machine Monitor* (VMM*)* or *hypervisor* creates and controls the virtual machine's other virtual subsystems. The VMM also takes complete control of the physical machine, and provides resource guarantees for CPU, memory, storage space, and I/O bandwidth for each guest OS. The VMM can provide a management interface that allows server resources to be dynamically allocated to match temporal changes in user demand for different applications.

Figure 4.5 shows how Ethernet Networking I/O is typically virtualized by VMM software. The VMs within a virtualized server typically share a conventional physical Ethernet NIC (PNIC) to connect to a data center LAN. The VMM provides each VM with a virtual NIC (VNIC) instance complete with MAC and IP addresses and creates a virtual switched network to provide the connectivity between the VNICs and the PNIC. The virtual switched network performs the I/O transfers using shared memory and asynchronous buffer descriptors in similar fashion to a shared memory Ethernet switch. With this software-based I/O Virtualization, the VMM provides the data path for Inter-VM traffic and to the external LAN.

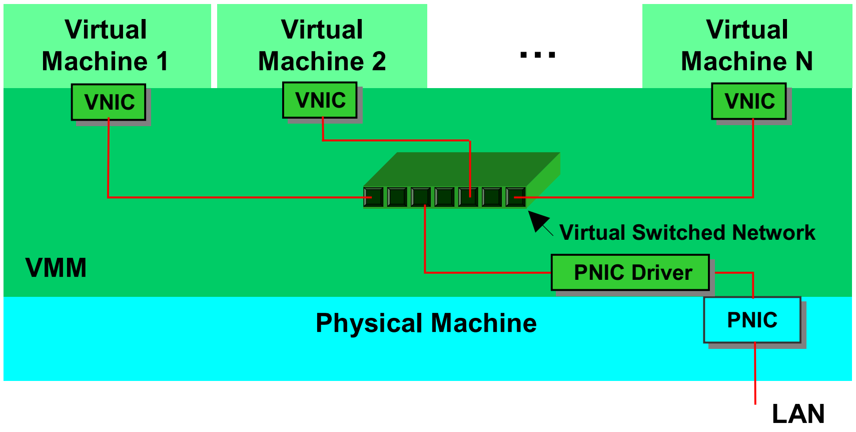


Figure 4.5: Software-based Virtual Ethernet I/O

In the data center, it is conceivable that an entire n-tier application architecture could be implemented on a single high performance, multi-core virtualized server. A group of VMs can be allocated to each server function, and additional VMs could be dedicated to *virtual appliances* consisting of a firewall, an Application Delivery Controller, or WAN Optimization Controller software running on a guest OS supported by the hypervisor. In instances such as this, where the network is partially subsumed by highly integrated virtualized servers, inter-VM traffic stays within the server and does not traverse the physical network. As we discuss in Chapter 8, the fact that inter-VM traffic does not traverse the physical network can result in IT organizations being unable to monitor this traffic.

Benefits of Server Virtualization

There are numerous benefits that can be derived from server virtualization, including:

- Server Consolidation

  A single physical server can easily support 10 to 100 VMs, allowing numerous applications normally requiring dedicated servers to share a single physical server. This allows the number of servers in the data center to be reduced while increasing average utilization from as low as 5-10% up to 50-60%.

- Flexible Server/Application Provisioning

  A production virtual machine can be transferred to a different physical server connected to the same storage area network (SAN) without service interruption. This enables workload management/optimization across the virtual infrastructure as well as zero-downtime maintenance. VMs also help to streamline the provisioning of new applications and backup/restore operations.

- Lower TCO

  Server virtualization allows significant savings in the both CAPEX (i.e., costs of server hardware, SAN Host bus adapters, and Ethernet NICs) and OPEX; i.e., server management labor expense, plus facility costs such as power, cooling, and floor space.

- Green IT

  In some instances, part of the motivation to deploy virtualized servers is to reduce the number of servers, not just to reduce the cost of those servers, but to also reduce the associated carbon footprint as part of a Green IT initiative.

Deployment of virtualized servers can result in significant cost savings.

Challenges of Server Virtualization

Along with the benefits, there are a number of challenges inhibiting broader adoption of server virtualization in the enterprise data center, including the following:

- The fact that it is easy to move VMs from one physical server to another contributes greatly to resource agility, high availability, and energy efficiency; however, it can be challenging to ensure the migrated VM retains the same security, storage access, and QoS configurations and policies. Keeping all the required configurations synchronized requires linkages among the management systems for physical and virtual servers, network devices, and storage. It often happens that relatively autonomous groups within the IT department manage these three resource domains.
- The flexibility of the virtual environment can make it difficult to properly plan and allocate computing and storage resources.

- Lack of visibility into the traffic flows between and among the VMs on the same physical platform affects security, performance monitoring, and troubleshooting.
- Over-consolidation of servers and/or appliances can cause performance problems because of limited CPU cycles or because of I/O bottlenecks due to over-subscribed physical NICs and Host Bus Adapters (HBAs).

Deployment of virtualized servers can result in significant management, performance and security issues.

Virtual Desktops

With virtual desktops, a VM on a data center server hosts a complete user desktop including all its applications, configurations, and privileges. The PC client then accesses the applications via the network with the desktop/application objects delivered on demand over the network from the virtual desktop servers via a remote display protocol, such as Microsoft Remote Desktop Protocol (RDP) and/or Citrix ICA. On the client device, the enterprise desktop is isolated from whatever else is running on the PC. Another less dynamic approach is to load the entire enterprise desktop as a guest partition on the client PC. The desktop can be downloaded over the network or from a dedicated mobile storage device such as a cell phone or portable disk.

Benefits of Virtual Desktops

With a desktop virtualization solution, there are a number of benefits that contribute to a significantly lower Total Cost of Ownership (TCO) by simplifying management, improving security, and increasing reliability of desktop services. These benefits include the following:

- Because desktops are managed from within the data center, there is no requirement to visit individual offices to perform routine management tasks, such as configuration of applications and operating systems.
- Security is improved because it is easier to control access to confidential data. As the desktop is isolated from the user’s PC, malicious code intrusion is less likely and the burdens of regulatory compliance are reduced.
- Computing resources are better utilized, reducing requirements for frequent upgrades to desktop PCs and reducing overall requirements for power and cooling.
- New desktops and desktop upgrades can be provisioned more rapidly and reliably.
- Through virtualization, desktop environments gain access to the same reliability, data protection and disaster recovery services that have been implemented for data center virtualized application servers. Automated VM failover helps to ensure high availability for virtual desktops and backups of desktop data can leverage the data center SAN or other shared storage.

A major component of the business case for deploying virtualized desktops is the resulting additional control and agility.

Challenges of Desktop Virtualization

Some of the challenges associated with desktop virtualization include the following:

- Virtual desktop systems need to be tightly integrated with policy management systems and user authorization to ensure that the desktop characteristics match the users' needs and permission profiles.
- Delivering acceptable application performance to the virtual desktop over the WAN can be challenging particularly when server I/O bottlenecks exist. The inefficiencies associated with LAN-centric remote display protocols, such as RDP, limit the performance of these protocols over the WAN. These performance challenges exist even when fairly sophisticated WAN Optimization Controllers are deployed to provide functionality such as TCP optimization, caching, and compression. The challenges of WAN performance have resulted in collaboration between WAN Optimization Controller vendors and desktop virtualization vendors to develop bundled client software. These collaborations may result in improved performance, plus new remote display protocols that provide a better user experience for media-rich virtual desktops delivered over the WAN.

There can be significant performance issues associated with the deployment of virtualized desktops.

Virtual Storage

Virtualization of storage helps achieve location independence by abstracting the physical location of the data. The virtualization system presents the user with a logical space for data storage and manages the mapping of logical addresses to physical addresses. The virtualization software or device is responsible for maintaining the mapping tables as *meta-data*. The most common form of virtualized storage is based on a SAN where a single block of data is addressed using a logical unit identifier (LUN) and an offset within that LUN. The virtualization entity within the storage area network (SAN) uses its meta-data to map a block's LUN /offset on a virtual disk to another LUN/offset on a physical disk.

Benefits of Storage Virtualization

The benefits associated with storage virtualization include the following:

- Storage virtualization has great potential for simplifying storage administration and reducing costs for managing widely-diverse storage assets.
- A SAN supporting storage virtualization is a powerful adjunct to server virtualization because it simplifies maintaining data linkages when virtual machines are migrated among physical platforms.

Challenges of Storage Virtualization

Storage virtualization is complex in its own right. The virtualization function can be provided on the host, or in in-band or out-of-band controllers within storage arrays, storage fabric switches or SAN-attached appliances. Complicating the situation is the fact that there are no standards for storage virtualization, preventing multi-vendor interoperability and limiting migration options.

Moving to virtualized storage as an adjunct to virtualized servers generally implies a move to some form of SAN, or possibly NAS. Several market research studies have concluded that roughly half of virtualized server environments use Fibre Channel SANs as a networked storage solution. Especially for small to medium sized enterprises, Fibre Channel can be a challenging and expensive technology to assimilate. iSCSI is another SAN alternative, but that technology over Gigabit Ethernet currently lacks the robustness and performance of Fibre Channel. iSCSI will probably become a more viable option in the future when "lossless" Ethernet (a.k.a., Data Center Ethernet) is available and 10 GbE has come down in price.

5.0 Planning

In the classic novel *Alice in Wonderland,* English mathematician Lewis Carroll first explained part of the need for the planning component of the application delivery framework (though he may not have known it at the time). Alice asks the Cheshire cat, "Which way should I go?" The cat replies, "Where do you want to get to?" Alice responds, "I don't know," to which the cat says, "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 a minority 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 observed.***

One primary factor driving the planning component of application delivery is 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 proactive steps in order to protect both the company's investment in the application as well as the political capital of the application delivery function itself.

> ***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:

- Establishing SLAs for at least a core set of business applications.
- Identifying the key elements (e.g., specific switches and routers) for each component of the IT infrastructure (e.g., servers, databases, networks) that support each application.
- Establishing SLAs for the key elements for each component of the IT infrastructure that support each application.
- Baselining the performance of the key applications.
- Baselining the performance of the key elements for each component of the IT infrastructure that support each application.
- Establishing application design guidelines that ensure that applications will perform optimally when run over a WAN.
- Profiling an application prior to deploying it, including running it in conjunction with a WAN emulator, to quantify the performance that can be expected.
- Performing an assessment of the IT infrastructure to support a new application prior to deploying it.

- Identifying in advance the impact of a change to the network, the servers, or to an application.
- Creating a network design that maximizes availability and minimizes latency.
- Creating a data center architecture that maximizes the performance of all of the resources in the data center.
- Determining what functionality to perform internally and what functionality to acquire from a third party. This topic will be expanded upon in Chapter 7.

WAN Emulation

Chapter 4 outlined many of the factors that complicate the task of ensuring acceptable application performance. One of these factors is the fact that in the vast majority of situations, the application development process does not take into account how the application runs over a WAN.

One class of tools that can be used to test and profile application performance throughout the application lifecycle is the WAN emulator. Used during application development and quality assurance (QA), these tools mimic the performance characteristics of the WAN, e.g., delay, jitter, packet loss. One of the primary benefits is that application developers and QA engineers can use them to quantify the impact of the WAN on the performance of the application under development, ideally while there is still time to modify the application. One of the secondary benefits of using WAN emulation tools is that over time the application development groups come to understand how to write applications that perform well over the WAN.

As an example, Table 5.1 depicts the results of a lab test done using a WAN emulator to quantify the affect that WAN latency would have 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.

Network Latency	Measured Response Time
0 ms	2 seconds
25 ms	2 seconds
50 ms	2 seconds
75 ms	2 seconds
100 ms	4 seconds
125 ms	4 seconds
150 ms	12 seconds

Table 5.1: Impact of Latency on Application Performance

As Table 5.1 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 represents the time spent in the application server and the database server. As network latency is increased up to 75 ms., it still has little impact on the application's response time, but if network latency goes above 75 ms, the response time of the application degrades rapidly and is quickly well above the target response time.

In a recent survey, over 200 IT professionals were asked "Which of the following describes your company's interest in a tool that can be used to test application performance throughout the application lifecycle – from application design through ongoing management?" The survey respondents were allowed to indicate multiple answers. Their responses are depicted in Table 5.2.

Response	Percentage of Respondents
If the tool worked well it would make a significant improvement to our ability to manage application performance	71%
The output of tools like this is generally not that helpful	9%
Tools like this tend to be too difficult to use, particularly during application development	13%
Our applications developers would be resistant to using such a tool	11%
Our operations groups lack the application specific skills to use a tool like this	17%

Table 5.2: Interest in an Application Lifecycle Management Tool

One obvious conclusion that can be drawn from table 5.2 is:

> ***The vast majority of IT organizations see significant value from a tool that can be used to test application performance throughout the application lifecycle.***

Given the complex and dynamic nature of the IT environment, a valid use of a WAN emulation tool is to provide insight into what happens if WAN delay increases from 70 ms to 100 ms. For example, would it increase the application delay by a second? By two seconds? By five seconds? The 80/20 rule applies here: 80% of the insight into application performance can be provided while only incurring 20% of the complexity. However, gaining additional insight requires that the tool become very complex, and typically requires a level of granular input that either does not exist or is unnecessarily time consuming to create.

The data in Table 5.2 indicates that IT professionals are well aware of the fact that many of these tools are unacceptably complex. In particular, while the survey respondents indicated a strong interest in these tools, thirty percent of the survey respondents indicated either that tools like this tend to be difficult or that their operations group would not have the skills necessary to use such a tool.

> ***In the vast majority of cases, a tool that is unduly complex is of no use to an IT organization.***

Using a WAN emulator as described above, to either develop more efficient applications or to quantify the impact of a change such as a data center initiative, is a *proactive* use of the tool. In many cases, IT organizations profile an application in a *reactive* fashion, which means the organization profiled the application only after users complained about its performance.

Alternatively, some IT organizations only 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 to 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 user expectation.
- Learn about the factors influencing how well an application will run over a WAN.

However, because companies perform these tests just before the application goes into production, it is usually too late to make any major changes.

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

In order to be deeply involved in the application development cycle, the application delivery function needs to either implement, or somehow have access to, an application performance laboratory. As part of pre-deployment application performance testing, the application delivery function should then:

- Establish goals for the performance of the application.
- Import detailed information on the actual performance of the production network.
- Run scripts that model the performance of the application under development, running over the production network, using a wide variety of scenarios.
- Identify the performance bottlenecks, if any.
- Either remove the bottlenecks or implement a work-around.
- Repeat the last three steps until the performance goals have been met.

Baselining

Baselining provides a reference from which service quality and application delivery effectiveness can be measured. It does so by quantifying the key characteristics (e.g., response time, utilization and 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. It is a component of several key processes, such as performing a pre-assessment of the network prior to deploying an application or performing proactive alarming.

The Key Steps

Four principal steps comprise baselining:

I. Identify the Key Resources

Most IT organizations do not have the ability to baseline all of their resources. They must therefore determine which are the most important resources and baseline those. One way to determine which resources are the most important is to identify the company's key business applications.   The IT resources that support these applications are the most important resources and should be baselined.

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

Organizations must compute the baseline over a normal business cycle. For example, the activity and response 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, or during holidays.

In most cases, baselining focuses on measuring the utilization of resources, such as WAN links. However, application performance is only indirectly 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 with insight into how their organizations are changing based on how their use of key applications is changing.

IV. Use the Information

The information gained from baselining has many uses, including 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 it gets all of the benefits it expects from that upgrade, the 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.

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 potential to miss important behavior that is infrequent and/or anomalous.

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

Baseline Solution 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, XenApp?
- 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 is: Can the network provide the necessary levels of availability and performance? 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 upheld.

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 key components of a pre-deployment network assessment are:

Create an inventory of the applications running on the network

This includes discovering all applications running on the network. Chapter 8 will discuss this task in greater detail.

It is also important to categorize those applications using an approach similar to that described in Chapter 4. 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 recreational use can free up additional WAN bandwidth. Chapter 8 quantifies the extent to which most 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 releases additional bandwidth.

Evaluate bandwidth to ensure available capacity for new applications

This activity involves baselining the network as previously described, with the goal of using the information about relevant network resource utilization trends to identify any parts of the network in need of upgrading to support the new application.

Baselining typically refers to measuring the utilization of key IT resources. Again, companies should 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 factors.

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 information 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 creates 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, 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 cannot 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 IP Flow Information Export (IPFIX), 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.

Another consideration is whether or not IT organizations should deploy software agents on end systems. One of the architectural advantages of this approach is that it monitors performance and events closer to the user's actual experience. A potential disadvantage of this approach is that there can be organizational barriers that limit the ability of the IT organization to put software on each end system. In addition, for an agent-based approach to be successful, it must not introduce any appreciable management overhead.

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 greatest challenges of this activity is unifying this information so the organization can leverage it to support myriad activities associated with managing application delivery.

Integrating Network Planning and Network Operations

With a life cycle approach to planning and managing application performance, a critical requirement is to consider not only whether the existing network can provide the necessary levels of availability and response time, but also to anticipate the impact that various changes in the infrastructure will have on targeted application service levels.

Addressing performance issues throughout the application lifecycle is greatly simplified if there are tight linkages between the IT personnel responsible for the planning and operational functions. The degree of integration between these IT functions can be significantly enhanced by a common tool set that:

1. Provides estimates of the impact on both network and application performance that would result from proposed changes in either the infrastructure or in application traffic patterns.

2. Verifies and ensures consistency of configuration changes to ensure error-free network operations and satisfactory levels of service

A common tool set that spans planning and operational functions also supports initiatives aimed at consolidation of network management tools in order to reduce complexity and maximize productivity of the IT staff.

The Gap Between Network Planning and Network Operations

For those organizations that run a large, meshed network there often is a significant gap between network planning and network operations. One of the reasons for this gap is that due to the complex nature of the network there tends to be a high degree of specialization amongst the members of the IT function. Put simply, the members of the organization who do planning understand planning, but typically do not understand operations. Conversely, the members of the organization who do operations understand operations, but typically do not understand planning.

Another reason for this gap is that historically it has been very difficult to integrate planning into the ongoing change management processes. For example, many IT organizations use a change management solution to validate changes before they are implemented. These solutions are valuable because they identify syntax errors that could lead to an outage. However, these solutions cannot identify how the intended changes would impact the overall performance of the network.

As is discussed in Chapter 9, within the majority of enterprise IT organizations the operations group is involved in what has traditionally been planning functions. In particular, that research showed that in the majority of IT organizations, the operations group is involved in:

- Network design
- Selection of new technologies; i.e., MPLS
- Selection of network service providers

The fact that network planning and network operations are working together on tasks such as network design is encouraging because that cooperation is likely to result in networks that are more highly available. However, as was pointed out, system complexity with multiple components and many types of interactions creates an environment where the relationship between actions and outcomes is not always obvious. As such, in order to design high availability networks and ensure that changes made to those networks do not negatively impact availability or performance, IT organizations need tools that can accurately predict the impact of change.

Predicting the Impact of Change

In order to be able to predict how a planned change will impact the performance of the network, some large IT organizations incur the cost of pre-testing a change in a lab environment prior to implementation. However, it is not possible to accurately represent a complex network in a lab. As a result, lab testing can only provide some insight into how a planned change will impact network performance. It has, however, the potential to miss some of the most significant components of how performance will be affected.

To overcome the limitations of lab testing, some IT organizations have deployed tools that model the performance of the network. Unfortunately, in many cases, these tools are very expensive, not only in terms of the cost of the software itself, but in terms of the personnel, training, and time needed to manually update the tools. Most IT organizations simply can’t afford the software or the personnel to run these tools. In addition, most network design modeling tools do not precisely mirror the real world network implementation because they cannot account for changes in the network that have occurred since the last time the model’s network configuration data was updated.

Another class of management tool that can facilitate the integration of planning and operations is typified by an IP route analytics solution. The goal of route analytics is to provide visibility, analysis, and diagnosis of the issues that occur at the routing layer in complex, meshed networks. A route analytics appliance draws its primary data directly from the network in real time by participating in the IP routing protocol exchanges. This allows the route analytics device to compute a real-time Layer 3 topology of the end-end network, detect routing events in real time, and correlate routing events or topology changes with other information, including application performance metrics. As a result, route analytics can help determine the impact on performance of both planned and actual changes in the Layer 3 network.

Route analytics is gaining in popularity because the only alternative for resolving logical issues involves a very time-consuming investigation of the configuration and log files of numerous individual devices. Route analytics is also valuable because it can be used to eliminate problems stemming from human errors in router configuration by allowing the effect of a configuration change to be previewed before the change is actually implemented (according to sources at Juniper Networks, between fifty and eighty percent of network outages are caused by human error[12]). From an application delivery perspective, route analytics allows the path that application traffic takes through the network to be predetermined before changes are implemented and then allows the application traffic to be tracked in real-time after the application has gone into production.

Route analytics is effective as a planning tool because it:

- Records an always-updated model of the network based on real-time routing and traffic changes.
- Is operationally accurate enough to be able to move back in time and perform simulated network changes using a peak traffic period or other important phenomena as a baseline.
- Incurs low network overhead.
- Is completely accurate in the way that it displays both current, historical behavior and modeled network behavior. This accuracy over an extended time frame makes route analytics a valuable tool for identifying trends in network behavior and for doing capacity planning.
- Enables IT organizations to easily and accurately simulate changing one piece of the network's routing or traffic, and calculate the effect on the rest of the network in a holistic fashion.

Route analytics can be used to help integrate network planning and operational functions by:

- Making sure that nothing bad will happen when a router is upgraded.
- Demonstrating how to tune routing metrics to spread traffic away from congested links to underutilized links.
- Ensuring that the network will behave as desired in a disaster recovery scenario.
- Making sure that design assumptions still hold true when changing, adding to, or upgrading the network in some fashion.
- Ensuring that service levels will be maintained after a initiative such as consolidating data centers.

[12] What's Behind Network Downtime?, www.juniper.com

6.0 Network and Application Optimization

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

- Reduce the amount of data 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) necessary for a given transaction;
- Mitigate the inefficiencies of older protocols;
- Offload computationally intensive tasks from client systems and servers.

There are two principal categories of network and application optimization products. One category focuses on the negative effect of the WAN on application performance. This category is often referred to as a WAN optimization controller (WOC) but will also be referred to in this handbook as Branch Office Optimization Solutions. Branch Office Optimization Solutions are often referred to as *symmetric solutions* because they typically 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 not in the branch office. This class of solution is often referred to as a *software only solution* or as a *soft WOC.*

The trade-off between a traditional symmetric solution based on an appliance and a software only solution is straightforward. Because the traditional symmetric solution involves an appliance in each branch office, it has the dedicated hardware allowing it to service a large user base. However, the number of hardware appliances required tends to make the traditional symmetric solution more expensive. The software-only solution is most appropriate for individual users or small offices.

The second category of products discussed in this Chapter is often referred to as an Application Front End (AFE) or Application Delivery 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 specifically for these tasks. The role of the ADC is somewhat similar to that of the FEP in that it performs computationally intensive tasks, such as the processing of Secure Sockets Layer (SSL) traffic, hence freeing up server resources. However, another role of the ADC that the FEP did not provide is that of Server Load Balancer (SLB) which, as the name implies, balances traffic over multiple servers.

Application Delivery Network Defined

The basic business benefits of WAN Optimization and Application Acceleration over the WAN are to:

- Reduce WAN bandwidth expenses
- Reduce congestion on WAN ports
- Reduce OPEX and CAPEX through the facilitation of consolidation and centralization of servers, applications, and storage resources
- Improve remote employee productivity through reduced application response time

Some of these basic benefits can be gained by deploying devices that are focused on optimizations within the packet delivery network. By *packet delivery network* is meant the packet payload and the transport, network and data link layers of the Internet protocol suite, as shown in Figure 6.1.

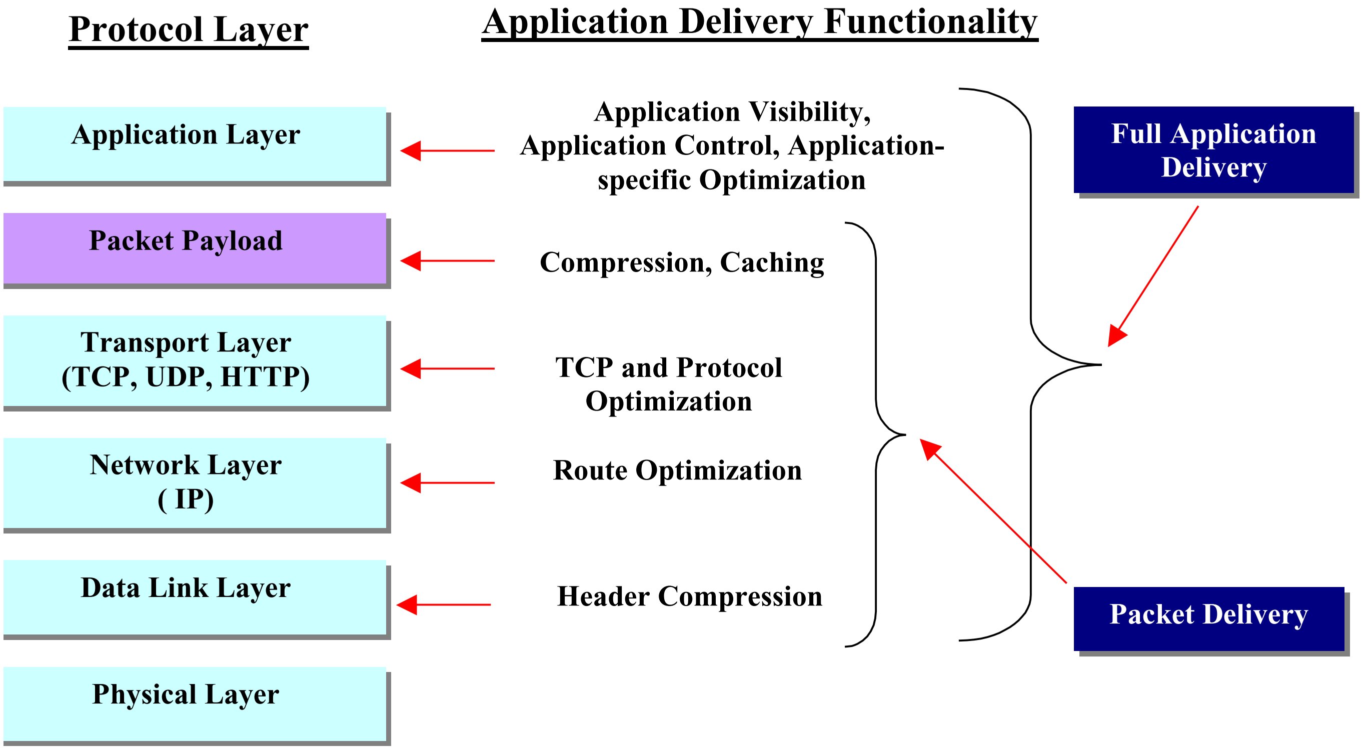

Protocol Layer
Application Delivery Functionality
Application Layer
Packet Payload
Transport Layer (TCP, UDP, HTTP)
Network Layer (IP)
Data Link Layer
Physical Layer
Application Visibility, Application Control, Application-specific Optimization
Compression, Caching
TCP and Protocol Optimization
Route Optimization
Header Compression
Full Application Delivery
Packet Delivery

Figure 6.1: Application Delivery vs. Packet Delivery

As Application Delivery technology continues to evolve, much more attention is being paid to the Application Layer. Solutions that leverage functionality that resides higher in the OSI protocol stack can improve the effectiveness of application delivery based on the ability of these solutions to recognize application layer signatures and to then differentiate among the various applications that share and contend for common transport resources. This is important because as mentioned in Chapter 4, 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 that are relevant to the business while controlling or eliminating applications that are not relevant.

Virtual Appliances, Clustered Appliances, and Partitioned Appliances

Chapter 4 of this handbook provided a discussion of virtualization in the application environment that focused on virtualization of servers, desktops, and storage. As previously noted, the general concept of virtualization can also be applied to many of the ancillary devices deployed in the data center, including WOCs, firewalls, ADCs, and other appliances that provide security or application services. The remainder of this section provides a description of three different ways virtualization can be implemented in these devices.

Virtual Appliances

A Virtual Appliance is based on network appliance software, together with its operating system, running in a virtual machine (VM) over the hypervisor in a virtualized server. Virtual appliances can include WOCs, firewalls, and performance monitoring solutions among others. A virtual appliance offers the potential to alleviate some of the management burdens in branch offices because most of the provisioning, software updates, configuration, and other management tasks can be automated and centralized at the data center.

Virtual appliances, such as virtual WOCs, would be of particular interest where server virtualization technology has already been disseminated to branch offices, as well as being implemented in the data center. In the branch office, a suitably placed virtualized server could potentially host a virtual WOC appliance, as well as other virtual appliances that would normally require a separate hardware platform. When server virtualization pervades the enterprise, application acceleration with virtual WOCs can potentially be implemented wherever it is needed, without the installation of additional hardware.

WOCs and virtual WOCs can also play a role in helping to distribute virtualization technologies to the branch office. For example, WOCs can greatly accelerate transfers of virtual images from the data center out to virtualized servers in the branch offices. WOCs can also potentially provide the boost in performance and scalability that would make virtual desktops a viable solution for provisioning and managing branch office desktop environments.

WOC virtualization can also be leveraged to provide “acceleration as a service” to facilitate and improve performance in deployments of service-oriented environments, including SOA and SaaS. In the case of SOA, WOC images can be easily deployed to be co-resident on the virtual servers that host the various components of a geographically-distributed SOA application. In the SaaS space, virtual WOCs can be provided as a standalone managed software service or bundled with other managed software services to increase their performance.

Clustered Appliances

While the performance of a virtual appliance can be well suited to the branch office, large data centers often need more appliance throughput than even a dedicated high performance physical appliance platform can provide. One option is to cluster a number of ADCs and have the cluster perform as a single ADC. Another option is to implement a cluster of physical appliances with an ADC providing the load balancing across the individual appliance platforms. *Clustered Appliances* such a clustered WOCs can be viewed as an additional example of virtualization that increases the scalability of appliance throughput capacity and provides redundancy that increases system availability.

Partitioned Appliances

Several types of appliances can support yet another form of virtualization, where the system’s hardware platform supports a number of independent software partitions, somewhat analogous to the independent VMs in a virtualized server. This type of *Partitioned Appliance* (e.g., ADC or firewall) can be configured to dedicate a separate partition to each application or service being delivered. This allows the configuration of each partition to be optimized for the specific type of application traffic being processed. Another benefit of partitioned appliances is that the technology is complementary to server virtualization, for example where a small cluster of physical servers, front-ended by only one or two WOCs, firewalls, and ADCs, could host a fairly large number of applications.

Alice in Wonderland Revisited

Chapter 5 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. The Cheshire Cat's assertion also 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.***

Chapter 4 of this handbook 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 plans to consolidate servers out of branch offices and into centralized data centers, or has already done so, then a WAFS (Wide Area File Services) solution might be appropriate. If the company is implementing VoIP, then any WOC it implements must be able to support traffic that is both real-time and meshed, and have strong QoS functionality. If, as another example, the company is making heavy use of SSL, it might make sense to implement an ADC to relieve the servers of the burden of processing the SSL traffic.

In addition to such high-level factors, the company's actual traffic patterns also have a significant impact on how much value a network and application optimization solution will provide. To illustrate this, consider the types of advanced compression most solution providers offer. The effectiveness of advanced compression depends on two factors. One is the quality of the compression techniques implemented in a solution. As 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 influencing the effectiveness of advanced compression solutions is the amount of redundancy of the traffic. Applications that transfer highly redundant data, 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.

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.

> ***Understanding the performance gains of any network and application optimization solution requires testing in an environment that closely reflects the live environment.***

WAN Optimization Controllers

The goal of a WOC is to improve the performance of applications delivered from the data center to the branch office or directly to the end user. 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.

WAN Characteristic	WAN Optimization Techniques
Insufficient Bandwidth	Data Reduction: Data Compression Differencing (a.k.a., de-duplication) Caching
High Latency	Protocol Acceleration: TCP HTTP CIFS NFS MAPI Mitigate Round-trip Time Request Prediction Response Spoofing
Packet Loss	Congestion Control Forward Error Correction (FEC)
Network Contention	Quality of Service (QoS)

Table 6.1: Techniques to Improve Application Performance

Below are some of the key techniques used in WAN optimization:

Caching

A copy of information is kept locally, with the goal of either avoiding or minimizing the number of times that information must be accessed from a remote site. Caching can take multiple forms:

Byte Caching

With byte caching the sender and the receiver maintain large disk-based caches of byte strings previously sent and received over the WAN link. As data is queued for the WAN, it is scanned for byte strings already in the cache. Any strings resulting in *cache hits* are replaced with a short token that refers to its cache location, allowing the receiver to reconstruct the file from its copy of the cache. With byte caching, the data dictionary can span numerous TCP applications and information flows rather than being constrained to a single file or single application type.

Object Caching

Object caching stores copies of remote application objects in a local cache server, which is generally on the same LAN as the requesting system. With object caching, the cache server acts as a proxy for a remote application server. For example, in Web object caching, the client browsers are configured to connect to the proxy server rather than directly to the remote server. When the request for a remote object is made, the local cache is queried first. If the cache contains a current version of the object, the request can be satisfied locally at LAN speed and with minimal latency. Most of the latency involved in a cache hit results from the cache querying the remote source server to ensure that the cached object is up to date.

If the local proxy does not contain a current version of the remote object, it must be fetched, cached, and then forwarded to the requester. Loading the remote object into the cache can potentially be facilitated by either data compression or byte caching.

Compression

The role of compression is to reduce the size of a file prior to transmission over a WAN. Compression also takes various forms.

Static Data Compression

Static data compression algorithms find redundancy in a data stream and use encoding techniques to remove the redundancy, creating a smaller file. A number of familiar lossless compression tools for binary data are based on Lempel-Ziv (LZ) compression. This includes zip, PKZIP and gzip algorithms.

LZ develops a codebook or dictionary as it processes the data stream and builds short codes corresponding to sequences of data. Repeated occurrences of the sequences of data are then replaced with the codes. The LZ codebook is optimized for each specific data stream and the decoding program extracts the codebook directly from the compressed data stream. LZ compression can often reduce text files by as much as 60-70%. However, for data with many possible data values LZ generally proves to be quite ineffective because repeated sequences are fairly uncommon.

Differential Compression; a.k.a., Differencing or De-duplication

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

While differential compression is restricted to those cases where the receiver has stored an earlier version of the file, the degree of compression is very high. As a result, differential compression can greatly reduce bandwidth requirements for functions such as software distribution, replication of distributed file systems, and file system backup and restore.

Real Time Dictionary Compression

The same basic LZ data compression algorithms discussed earlier can also be applied to individual blocks of data rather than entire files, which results in smaller dynamic dictionaries that can reside in memory rather than on disk. As a result, the processing required for compression and decompression introduces only a small amount of delay, allowing the technique to be applied to real-time, streaming data.

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.

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.

Later in this chapter, we will discuss some of the technical challenges associated with data replication and will describe how FEC mitigates some of those challenges.

Protocol Acceleration

Protocol acceleration refers to a class of techniques that improves application performance by circumventing the shortcomings of various communication protocols. Protocol acceleration is typically based on per-session packet processing by appliances at each end of the WAN link, as shown in Figure 6.2. The appliances at each end of the link act as a local proxy for the remote system by providing local termination of the session. Therefore, the end systems communicate with the appliances using the native protocol, and the sessions are relayed between the appliances across the WAN using the accelerated version of the protocol or using a special protocol designed to address the WAN performance issues of the native protocol. As described below, there are many forms of protocol acceleration.

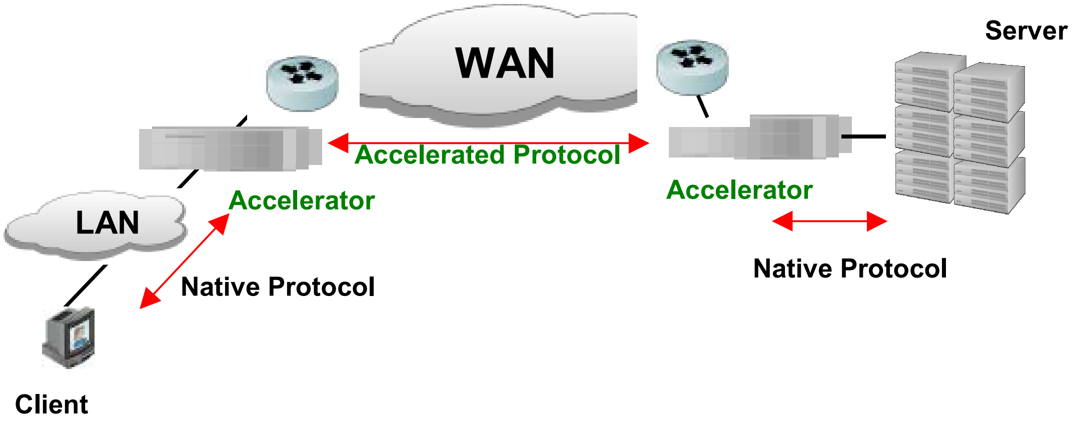


Figure 6.2: Protocol Acceleration Appliances

TCP Acceleration

TCP can be accelerated between appliances with a variety of techniques that increase a session's ability to more fully utilize link bandwidth. Some of these techniques include dynamic scaling of the window size, packet aggregation, selective acknowledgement, and TCP Fast Start. Increasing the window size for large transfers allows more packets to be sent simultaneously, thereby boosting bandwidth utilization. With packet aggregation, a number of smaller packets are aggregated into a single larger packet, reducing the overhead associated with numerous small packets. TCP selective acknowledgment (SACK) improves performance in the event that multiple packets are lost from one TCP window of data. With SACK, the receiver tells the sender which packets in the window were received, allowing the sender to retransmit only the missing data segments instead of all segments sent since the first lost packet. TCP slow start and congestion avoidance lower the data throughput drastically when loss is detected. TCP Fast Start remedies this by accelerating the growth of the TCP window size to quickly take advantage of link bandwidth.

Other performance challenges associated with TCP are discussed in the section of Chapter 7 entitled "The Limitations of the Internet".

CIFS and NFS Acceleration

CIFS and NFS use numerous Remote Procedure Calls (RPCs) for each file sharing operation. NFS and CIFS suffer from poor performance over the WAN because each small data block must be acknowledged before the next one is sent. This results in an inefficient ping-pong effect that amplifies the effect of WAN latency. CIFS and NFS file access can be greatly accelerated by using a WAFS transport protocol between the acceleration appliances. With the WAFS protocol, when a remote file is accessed, the entire file can be moved or pre-fetched from the remote server to the local appliance's cache. This technique eliminates numerous round trips over the WAN. As a result, it can appear to the user that the file server is local rather than remote. If a file is being updated, CIFS and NFS acceleration can use differential compression and block level compression to further increase WAN efficiency.

HTTP Acceleration

Web pages are often composed of many separate objects, each of which must be requested and retrieved sequentially. Typically a browser will wait for a requested object to be returned before requesting the next one. This results in the familiar ping-pong behavior that amplifies the effects of latency. HTTP can be accelerated by appliances that use pipelining to overlap fetches of Web objects rather than fetching them sequentially. In addition, the appliance can use object caching to maintain local storage of frequently accessed web objects. Web accesses can be further accelerated if the appliance continually updates objects in the cache instead of waiting for the object to be requested by a local browser before checking for updates.

Microsoft Exchange Acceleration

Most of the storage and bandwidth requirements of email programs, such as Microsoft Exchange, are due to the attachment of large files to mail messages. Downloading email attachments from remote Microsoft Exchange Servers is slow and wasteful of WAN bandwidth because the same attachment may be downloaded by a large number of email clients on the same remote site LAN. Microsoft Exchange acceleration can be accomplished with a local appliance that caches email attachments as they are downloaded. This means that all subsequent downloads of the same attachment can be satisfied from the local application server. If an attachment is edited locally and then returned to via the remote mail server, the appliances can use differential file compression to conserve WAN bandwidth.

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.

The benefits of WOC deployment are exemplified in the case study entitled "Optimizing Content and Applications over the WAN with Caching, Web Security and Control" by Chris Bress. Bress is the CIO of the Charlotte, FL county public schools.

CASE STUDY: Optimizing Content and Applications over the WAN with Caching, Web Security and Control

Chris Bress
Chief Information Officer
Charlotte County Public Schools.

At Charlotte County Public Schools (CCPS) we serve a mainly rural area of southwest Florida. Even so, we want to offer our 18,000 students the same advantages as a major metropolitan school system, including access to Web technology and streaming media. At the same time, this commitment to technology has created some IT challenges.

Because of our rural location, bandwidth capacity is an issue, and we don't have many options. Initially, we had T1 lines to connect each school to the district data center using a star topology. We then upgraded some sites to a wireless WAN infrastructure to get beyond the limitations of the area's local telecommunications infrastructure. While the upgrade has helped, streaming and other bandwidth-intensive applications can easily saturate each WAN link. The long and short of it is that Internet bandwidth remains constrained.

To give you an idea of what we're facing, the application mix within Charlotte County Public Schools includes:

- **Web Applications.** One of our more critical web applications is Compass Learning Odyssey, an application delivering Florida State standard curriculum (English, math, social studies, etc.) and preparing students for the FCAT (the Florida standardized test). Many of the district's other web-based applications are Internet or Application Service Provider (ASP)-based, such as Grollier's On-Line, Thompson Gale, Facts on File, and Electric Library.

- **Streaming.** CCPS also subscribes to United Streaming, which is a service that takes Discovery Channel content and makes it available over streaming video. A central 1.5TB server houses some of the United Streaming content at the district offices and is populated via their Internet connection. Each classroom makes heavy educational use of this content but because of bandwidth limitations, only several students at each site could watch United Streaming content.

- **HTTPS and Other Applications.** We also have several SSL-encrypted web applications including one for managing our individualized education plans for students above or below norms. Because these applications deal with confidential and sensitive information, we encrypt all user-application communication with SSL.

Given this mix of bandwidth-hungry applications - and despite bandwidth upgrades - we struggled to keep critical applications available and performing well on the network. The number of applications has also continued to grow, and the student appetite for bandwidth has made it difficult to deliver applications in a high-performance, prioritized way. While the distances across the county introduced some latency, much of CCPS' latency was caused by congestion. For example, my networking team noticed that after 15 students were using Compass Learning, the pipe was full.

Since my staff and I have a lot of experience with caching technology, we initially looked at bandwidth-expanding options that utilized caching servers. It soon became clear, however, that what we really needed was an application optimization solution that integrated compression and byte caching, content filtering and bandwidth management capabilities. The appliances we ultimately chose can cache video locally and improve the quality and capacity to serve large numbers of students. These appliances also provide visibility, acceleration, and control of SSL-encrypted traffic.

Once we had these appliances installed, our testing confirmed that the available bandwidth on our network pipes really increased by as much as 3X. One of the interesting things that the testing highlighted was that the installed appliances' were also able to byte cache Novell shares as well as CIFS (Microsoft). Delivery of United Streaming content, the key streaming media application, also saw a peak reduction in bandwidth of 100x.

Here are a couple of other ways this solution has helped our best efforts to optimize applications and content:

- This past summer, we did migration work for the ninth grade academy (one of our magnet schools on the far side of Charlotte County), moving roughly 500 students, their classrooms, and their computing infrastructure to a new location. Performing the desktop imaging over the network, we saw imaging time drop by 50-60% because of byte caching. Login times for this migration also went from 45-60 seconds, to 15 seconds – an improvement of 3x-4x.
- We've also been able to create bandwidth management policies around student activity to ensure that critical applications have the room they need to run. For example – we set a policy to ensure that Compass Learning Odyssey has 60% of the overall bandwidth available to it as needed at any given time. This prevents surges in acceptable (but non-critical) sites that can easily saturate the CCPS network. We can also perform bandwidth management and application control locally at each school as well and in the central district office. At the same time, we're looking at the potential of this Application Optimization solution to manage or block P2P, IM, and Skype traffic – replacing a legacy signature-based bandwidth management system in place today.

At Charlotte County Public Schools we have infrastructure limitations because of our geography, application mix, and user population. This Application Optimization solution we installed hasn't solved all our network problems – we still suffer from a daily struggle to keep newly identified unwanted applications at bay and at some point we will need to upgrade our bandwidth. However we have been able to optimize and secure bandwidth-intensive applications, prioritize the applications running on our network, and increase our network capacity. From a business perspective, this has enabled CCPS to ensure delivery of a variety of educational applications and content to a bandwidth-hungry group without breaking our budget.

WOC Selection Criteria

The recommended criteria for evaluating WAN Optimization Controllers are listed in Table 6.3. This list is intended as a fairly complete compilation of all possible criteria, so a given organization may apply only a subset of these criteria for a given purchase decision. In addition, individual organizations are expected to ascribe different weights to each of the criteria because of differences in WAN architecture, branch office network design, and application mix. Assigning weights to the criteria and relative scores for each solution provides a simple method for comparing competing solutions.

There are many techniques IT organizations can use to complete Table 6.3 and then use its contents to compare solutions. For example, the weights can range from 10 points to 50 points, with 10 points meaning not important, 30 points meaning average importance, and 50 points meaning critically important. The score for each criteria can range from 1 to 5, with a 1 meaning fails to meet minimum needs, 3 meaning acceptable, and 5 meaning significantly exceeds requirements.

As an example, consider solution A. For this solution, the weighted score for each criterion (WiAi) is found by multiplying the weight (Wi) of each criteria, by the score of each criteria (Ai). The weighted score for each criterion are then summed (WiAi) to get the total score for the solution. This process can then be repeated for additional solutions and the total scores of the solutions can be compared.

		Score for Solution “A”	Score for Solution “B”
Criterion		Ai	Bi
Performance			
Transparency			
Solution Architecture			
OSI Layer			
Capability to Perform Application Monitoring			
Scalability			
Cost-Effectiveness			
Application Sub-classification			
Module vs. Application Optimization			
Disk vs. RAM-based Compression			
Protocol Support			
Security			
Ease of Deployment and Management			
Change Management			
Support for Meshed Traffic			
Support for Real Time Traffic			
Total Score		ΣWiAi	ΣWiBi

Table 6.3: Criteria for WAN Optimization Solutions

Each of the criteria is explained below.

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. For example, if the IT organization is in the process of consolidating servers out of branch offices and into centralized data centers, or has already done so, then it needs to test how well the WAN optimization solution supports CIFS.

As part of this quantification, it is important to identify whether the performance degrades as additional functionality within the solution is activated, or as the solution is deployed more broadly across the organization.

Transparency

The first rule of networking is not to implement anything that causes the network to break. Therefore, an important criterion when choosing a WOC is that it should be possible to deploy the solution without breaking things such as routing, security, or QoS. The solution should also be transparent relative to both the existing server configurations and the existing Authentication, Authorization and Accounting (AAA) systems, and should not make troubleshooting any more difficult.

Solution Architecture

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

OSI Layer

Organizations can apply many of the optimization techniques discussed in this handbook 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 but is by definition application specific.

Capability to Perform or Support Application Monitoring

Many network performance tools rely on network-based traffic statistics gathered from network infrastructure elements at specific points in the network to perform their reporting. By design, all WAN optimization devices apply various optimization techniques on the application packets and hence affect these network-based traffic statistics to varying degrees. One of the important factors that determine the degree of these effects is based on the amount of the original TCP/IP header information retained in the optimized packets. This topic will be expanded in the subsequent section on transparency.

Scalability

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 Effectiveness

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

Application Sub-classification

An application such as XenApp or SAP is composed 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 WOCs 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, of have the ability to provide both options. Disk-based systems can typically 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, Microsoft Print Services, CIFS, MAPI) while other solutions support that protocol generically. In either case, the critical issue is how much of an improvement the solution can offer in the performance of that protocol, 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 be compatible with the current security environment. It must not, for instance, break 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 WAN optimization solution, an appliance will 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 modified in order to implement the WAN optimization solution. Some solutions, especially cache-based or WAFS solutions, require that every file server be accessed during implementation.

Change Management

As most networks experience periodic changes such as the addition of new sites or new applications, it is important that the WAN optimization solution can adapt to these changes easily – preferably automatically.

Support of Meshed Traffic

A number of factors are causing a shift in the flow of WAN traffic away from a simple hub-and-spoke pattern to more of a meshed flow. If a company is making this transition, it is important that the WAN optimization solution it deploys can support meshed traffic flows and can support a range of features such as asymmetric routing.

Support for Real Time Traffic

Many companies have deployed real-time applications. For these companies it is important that the WAN optimization solution can support real time traffic. Traffic such as VOIP and live video typically 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.

The Data Replication Bottleneck

While packet loss and out of order packets are merely a nuisance for a network that supports typical data applications[13] such as file transfer and email, it is a very serious problem when performing data replication and backup across the WAN. The former involves thousands of short-lived sessions made up of a small number of packets typically sent over low bandwidth connections. The latter involves continuous sessions with many packets sent over high capacity WAN links. Data applications can typically recover from lost or out of order packets by retransmitting the lost data. Performance might suffer, but the results are not catastrophic. Data replication applications, however, do not have the same luxury. If packets are lost, throughput can be decreased so significantly that the replication process cannot be completed in a reasonable timeframe.

Key WAN Characteristics: Loss and Out of Order Packets

Many IT organizations are moving away from a hub and spoke network and are adopting WAN services such as MPLS and IP VPNs. While there are significant advantages to MPLS and IP VPN services, there are drawbacks, a major one being high levels of packet loss and out of order packets. This is due to routers being oversubscribed in a shared network, resulting in dropped or delayed packet delivery.

The affect of packet loss on TCP has been widely analyzed[14]. Mathis et al. provide a simple formula that offers insight into the maximum TCP throughput on a single session when there is packet loss. That formula is:

$$Throughput <= (MSS/RTT)*(1 / sqrt\{p\})$$

where:

MSS: maximum segment size
RTT: round trip time
p: packet loss rate.

The preceding equation shows that throughput decreases as either RTT or p increases. To illustrate the impact of packet loss, assume that MSS is 1,420 bytes, RTT is 100 ms. and p is 0.01%. Based on the formula, the maximum throughput is 1,420 Kbytes/second. If however, the loss were to increase to 0.1%, the maximum throughput drops to 449 Kbytes/second. Figure 6.3 depicts the impact that packet loss has on the throughput of a single TCP stream with a maximum segment size of 1,420 bytes and varying values of RTT.

[13] The phrase *typical data application* refers to applications that involve inquiries and responses where moderate amounts of information are transferred for brief periods of time. Examples include file transfer, email, web and VoIP. This is in contrast to a data replication application that transfers large amounts of information for a continuous period of time.

[14] The macroscopic behavior of the TCP congestion avoidance algorithm by Mathis, Semke, Mahdavi & Ott in Computer Communication Review, 27(3), July 1997

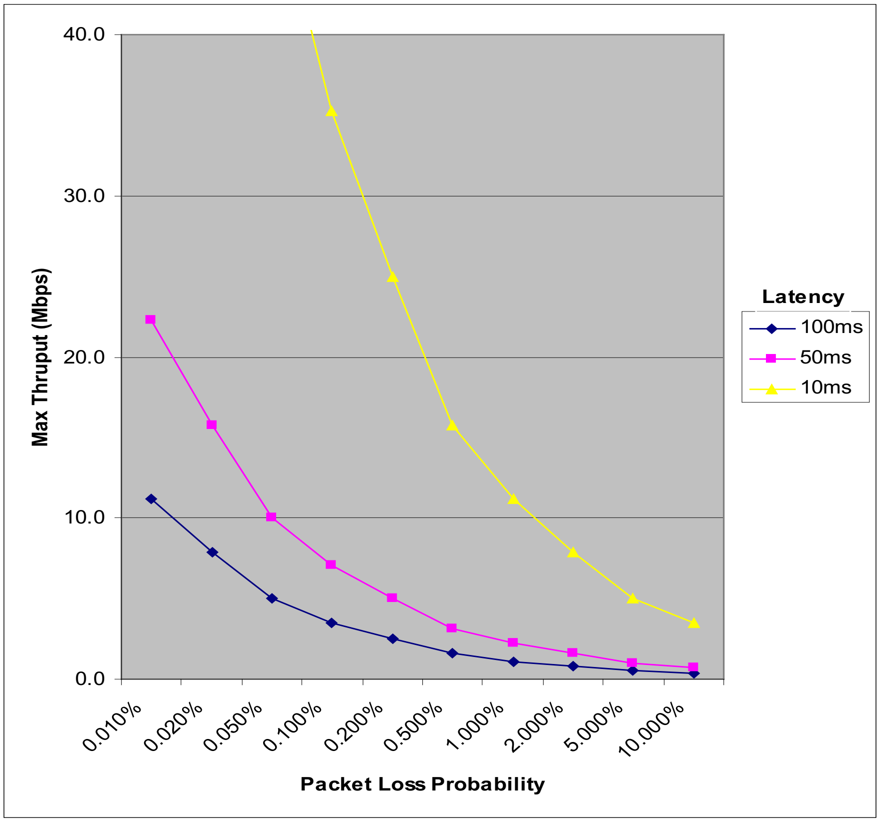


Figure 6.3: Impact of Packet Loss on Throughput

One conclusion we can draw from Figure 6.3 is:

Small amounts of packet loss can significantly reduce the maximum throughput of a single TCP session.

More specifically:

With a 1% packet loss and a round trip time of 50 ms or greater, the maximum throughput is roughly 3 megabits per second no matter how large the WAN link is.

Techniques for Coping with Loss and Out of Order Packets

The data in Figure 6.3 shows that while packet loss affects throughput for any TCP stream, it particularly affects throughput for high-speed streams, such as those associated with multi-media and data replication. As a result, numerous techniques, such as Forward Error Correction (FEC)[15], have been developed to mitigate the impact of packet loss.

FEC has long been used at the physical level to ensure error free transmission with a minimum of re-transmissions. Many enterprises have recently begun to use FEC at the network layer to improve the performance of applications such as data replication. The basic premise of FEC is that an additional error recovery packet is transmitted for every *n* packets sent. The additional packet enables the network equipment at the receiving end to reconstitute one of the *n* lost packets and hence negates the actual packet loss. The ability of the equipment at the receiving end to reconstitute the lost packets depends on how many packets were lost and how many extra packets were transmitted. In the case in which one extra packet is carried for every ten normal packets (1:10 FEC), a 1% packet loss can be reduced to less than 0.09%. If one extra packet is carried for every five normal packets (1:5 FEC), a 1% packet loss can be reduced to less than 0.04%. For example, assume that the MSS is 1,420, RTT is 100 ms, and the packet loss is 0.1%. Transmitting a 10 Mbyte file without FEC would take a minimum of 22.3 seconds. Using a 1:10 FEC algorithm would reduce this to 2.1 seconds and a 1:5 FEC algorithm would reduce this to 1.4 seconds.

The example demonstrates the value of FEC in a TCP environment; the technique applies equally well to any application regardless of transport protocol. FEC, however, introduces overhead which itself can reduce throughput. What is needed is a FEC algorithm that adapts to packet loss. For example, if a WAN link is not experiencing packet loss, no extra packets should be transmitted. When loss is detected, the algorithm should begin to carry extra packets and should increase the amount of extra packets as the amount of loss increases.

Application Delivery Controllers (ADCs)

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

While an ADC still functions as a SLB, the ADC has assumed, and will most likely continue to assume, a wider range of more sophisticated roles that enhance server efficiency and provide asymmetrical functionality to accelerate the delivery of applications from the data center to individual remote users.

An ADC provides more sophisticated functionality than a SLB does.

Among the functions users can expect from a modern ADC are the following:

- **Traditional SLB**
  ADCs can provide traditional load balancing across local servers or among geographically dispersed data centers based on Layer 4 through Layer 7 intelligence. SLB functionality maximizes the efficiency and availability of servers through intelligent allocation of application requests to the most appropriate server.

[15] RFC 2354, Options for Repair of Streaming Media, http://www.rfc-archive.org/getrfc.php?rfc=2354

- **SSL Offload**
  One of the primary new roles played by an ADC is to offload CPU-intensive tasks from data center servers. A prime example of this is SSL offload, where the ADC terminates the SSL session by assuming the role of an SSL Proxy for the servers. SSL offload can provide a significant increase in the performance of secure intranet or Internet Web sites. SSL offload frees up server resources, allowing existing servers to process more requests for content and handle more transactions.

- **XML Offload**
  Another function that can be provided by the ADC (as well as standalone devices) is to offload XML processing from the servers by serving as an XML gateway. As described in Chapter 4, Web services and Web 2.0 applications are XML based, and XML is a verbose protocol that is CPU-intensive. Hence, one of the roles of an XML gateway is to offload XML processing from the general-purpose servers and to perform this processing on hardware that was purpose-built for this task. Another role of an XML gateway is to provide additional security functionality to protect against the kinds of attacks that were described in Chapter 4.

- **Application Firewalls**
  ADCs may also provide an additional layer of security for Web applications by incorporating application firewall functionality. Application Firewalls are focused on blocking increasingly prevalent application-level attacks. As described in the Chapter 10, which discusses Next Generation firewalls in more detail, application firewalls are typically based on Deep Packet Inspection (DPI), coupled with session awareness and behavioral models of normal application interchange. For example, an application firewall would be able to detect and block Web sessions that violate rules defining the normal behavior of HTTP applications and HTML programming. Therefore, Application Firewalls complement traditional perimeter firewalls that are based on recognition of known network-level attack signatures and patterns. Application Firewalls also have the advantage of providing a measure of protection against *zero day* exploits by blocking the sessions of clients whose behaviors are outside the bounds of admissibility.

- **Asymmetrical Application Acceleration**
  ADCs can accelerate the performance of applications delivered over the WAN by implementing optimization techniques, such as reverse caching, asymmetrical TCP optimization, and compression. With reverse caching, new user requests for static or dynamic Web objects can often be delivered from the cache rather than having to be regenerated by the servers. Reverse caching therefore improves user response time and minimizes loading on Web servers, application servers, and database servers.

  Asymmetrical TCP optimization is based on the ADC serving as a proxy for TCP processing, minimizing the server overhead for fine-grained TCP session management. TCP proxy 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 about of the server's CPU resources, Acting as a proxy, the ADC terminates the client-side TCP sessions and multiplexes numerous short-lived network sessions initiated as client-side object requests into a single longer-lived session between the ADC and the Web servers.

  The ADC can also offload Web servers by performing compute-intensive HTTP compression operations. HTTP compression is a capability built into both Web servers and Web browsers. Moving HTTP compression from the Web server to the ADC is transparent to the client and so requires no client modifications. HTTP compression is asymmetrical in the sense that there is no requirement for additional client-side appliances or technology.

- **Response Time Monitoring**
  The application and session intelligence of the ADC also presents an opportunity to provide real-time and historical monitoring and reporting of the response time experienced by end users accessing Web applications. The ADC can provide the granularity to track performance for individual Web pages and to decompose overall response time into client-side delay, network delay, ADC delay, and server-side delay. The resulting data can be used to support SLAs for guaranteed user response times, guide remedial action, and plan additional capacity to maintain service levels.

ADC Selection Criteria

The ADC evaluation criteria are listed in Table 6.4. As was the case with WOCs, this list is intended as a fairly complete compilation of possible criteria. As a result, a given organization or enterprise might apply only a subset of these criteria for a given purchase decision.

Criterion		Score for Solution “A” Ai	Score for Solution “B” Bi
Features			
Performance			
Scalability			
Transparency and Integration			
Solution Architecture			
Functional Integration			
Virtualization			
Security			
Application Availability			
Cost-Effectiveness			
Ease of Deployment and Management			
Business Intelligence			
Total Score		ΣWiAi	ΣWiBi

Table 6.4: Criteria for Evaluating ADCs

Each of the criteria is described below.

Features

ADCs support a wide range of functionality including TCP optimization, HTTP multiplexing, caching, Web compression, image compression as well as bandwidth management and traffic shaping.

Performance

Performance is an important criterion for any piece of networking equipment, but it is critical for a device such as an ADC, because data centers are central points of aggregation. As such, the ADC needs to be able to support the extremely high volumes of traffic transmitted to and from servers in data centers.

A simple definition of performance is how many bits per second the device can support. While this is extremely important, in the case of ADCs other key measures of performance include how many Layer 4 connections can be supported as well as how many Layer 4 setups and teardowns can be supported.

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 application environment where it will be installed. As part of this quantification, it is important to identify if the performance of the solution degrades as either additional functionality within the solution is activated or if there are changes made to the application mix within the data center.

Transparency and Integration

Transparency is an important criterion for any piece of networking equipment. However, unlike proprietary branch office optimization solutions, ADCs are standards based, and thus inclined to be more transparent than other classes of networking equipment.

It is very important to be able to deploy an ADC solution and not break anything such as routing, security, or QoS. The solution should also be as transparent as possible relative to both the existing server configurations and the existing security domains, and should not make troubleshooting any more difficult.

The ADC also has to be able to easily integrate with other components of the data center, such as the firewalls, and other appliances that may be deployed to provide application services. In some data centers, it may be important to integrate the Layer 2 and Layer 3 access switches with the ADC and firewalls so that all that application intelligence, application acceleration, application security, and server offloading are applied at a single point in the data center network.

Scalability

Scalability of an ADC solution implies the availability of a range of products that span the performance and cost requirements of a variety of data center environments. Performance requirements for accessing data center applications and data resources are usually characterized in terms of both the aggregate throughput of the ADC and the number of simultaneous application sessions that can be supported. A related consideration is how device performance is affected as additional functionality is enabled.

Solution Architecture

Taken together, scalability and solution architecture identify the ability of the solution to support a range of implementations and to extend to support additional functionality. In particular, if the organization intends the ADC 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 and without unacceptable downtime.

Functional Integration

Many data center environments have begun programs to reduce overall complexity by consolidating both the servers and the network infrastructure. An ADC solution can contribute significantly to network consolidation by supporting a wide range of application-aware functions that transcend basic server load balancing and content switching. Extensive functional integration reduces the complexity of the network by minimizing the number of separate boxes and user interfaces that must be navigated by data center managers and administrators. Reduced complexity generally translates to lower TCO and higher availability.

Virtualization

Virtualization is becoming a key technology for realizing data center consolidation and its related benefits. For example, server virtualization supports data center consolidation by allowing a number of applications running on separate virtual machines to share a single physical server. Prior to virtualization, a common practice was to run only one application per server to maximize operating system stability. Not only was the extra hardware this approach required expensive, it also necessitated additional real estate and power, further increasing the cost.

ADCs can also be virtualized by partitioning a single physical ADC into a number of logical ADCs or ADC contexts. Each logical ADC can be configured individually to meet the server-load balancing, acceleration, and security requirements of a single application or a cluster of applications. Therefore, each virtualized ADC can consolidate the functionality of a number of physical ADCs dedicated to the support of single applications. Virtualization adds significantly to the flexibility of the data center by allowing applications to be easily moved from one physical server to another. For example, with a virtual ADC mapped to a virtual machine, the ADC would not need to be reconfigured when an application is moved or automatically fails-over to a new physical machine. Benefits of virtualization include lowering TCO through consolidation of ADC physical devices, higher availability when faced with a failover, plus the associated savings in management costs and power and cooling costs.

Security

The solution must be compatible with the current security environment, while also allowing the configuration of application-specific security features that complement general purpose security measures, such as firewalls and IDS and IPS appliances. In addition, the solution itself must not create any additional security vulnerabilities.

Security functionality that IT organizations should look for in an ADC includes protection against denial of service attacks, integrated intrusion protection, protection against SSL attacks and sophisticated reporting.

Application Availability

The availability of enterprise applications is typically a very high priority. Since the ADC is in line with the Web servers and other application servers, a traditional approach to defining application availability is to make sure that the ADC is capable of supporting redundant, high availability configurations that feature automated fail-over among the redundant devices. While this is clearly important, there are other dimensions to application availability. For example, as previously mentioned, an architecture that enables scalability through the use of software license upgrades tends to minimize the application downtime that is associated with hardware-centric capacity upgrades.

Cost Effectiveness

This criterion is related to scalability. In particular, it is important not only 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.

Ease of Deployment and Management

As with any component of the network or the data center, an ADC solution should be relatively easy to deploy and manage. It should also be relatively easy to deploy and manage new applications -- so ease of configuration management is a particularly important consideration where a wide diversity of applications is supported by the data center.

Business Intelligence

In addition to traditional network functionality, some ADCs also provide data that can be used to provide business level functionality. In particular, data gathered by an ADC can feed security information and event monitoring, fraud management, business intelligence, business process management and Web analytics.

7.0 Managed Service Providers

As previously noted, virtually all organizations are under increasing pressure to ensure acceptable performance for networked applications. Many IT organizations are responding to this challenge by enhancing their understanding of application performance issues and then implementing their own application delivery solutions based on the products discussed in the preceding chapter. Other IT organizations prefer to outsource all or part of application delivery to a Managed Service Provider (MSP).

Benefits of Using an MSP

There is a wide range of potential benefits that may be gained from outsourcing to an Application Delivery MSP (ADMSP), including:

Reduce Capital Expenditure

In cases where the ADMSP provides the equipment as CPE bundled with the service, the need for capital expenditure to deploy application optimization solutions can be avoided.

Lower the Total Cost of Ownership (TCO)

In addition to reducing capital expenditure, managed application delivery services can also reduce operational expense (OPEX) related to technical training of existing employees in application optimization or hiring of additional personnel with this expertise. In terms of OPEX, the customer of managed services can also benefit from the lower cost structure of ADMSP operations, which can leverage economies of scale by supplying the same type of service to numerous customers.

Leverage the MSP's Management Processes

The ADMSP should also be able to leverage sophisticated processes in all phases of application delivery, including application assessment, planning, optimization, management, and control. In particular, the ADMSP's scale of operations justifies their investment in highly automated management tools and more sophisticated management processes that can greatly enhance the productivity of operational staff. The efficiency of all these processes can further reduce the OPEX cost component underlying the service.

The ability to leverage the MSP's management processes is a factor that could cause an IT organization to use an MSP for a variety of services, including the provision of basic transport services. This criterion, however, is particularly important in the case of application delivery because as will be shown in the next Chapter, ineffective processes is one of the most significant impediments to successful application delivery.

Leverage the MSP's Expertise

In most cases, ADMSPs will have broader and deeper application-oriented technical expertise than an enterprise IT organization can afford to accumulate. This higher level of expertise can result in full exploitation of all available technologies and optimal service implementations and configurations that can increase performance, improve reliability, and further reduce TCO.

Similar to the discussion in the preceding paragraph, the ability to be able to leverage the MSP's expertise is a factor that could cause an IT organization to use an MSP for a variety of services. This criterion, however, is particularly important in the case of application delivery because the typical IT organization does not have personnel who have a thorough understanding of both applications and networks, as well as the interaction between them.

Leverage the MSP's Technology

Because of economies of scale, ADMSP facilities can take full advantage of the most advanced technologies in building their facilities to support service delivery. This allows the customer of managed application delivery services to gain the benefits of technologies and facilities that are beyond the reach of the typical IT budget.

Timely Deployment of Technology

Incorporating a complex application delivery solution in the enterprise network can be quite time consuming, especially where a significant amount of training or hiring is required. In contrast, with a managed service, the learning curve is essentially eliminated, allowing the solution to be deployed in a much more timely fashion.

Better Strategic Focus

The availability of managed application delivery services can free up enterprise IT staff facilitating the strategic alignment of in-house IT resources with the enterprise business objectives. For example, in-house IT can focus on a smaller set of technologies and in-house services that are deemed to be of greater strategic value to the business.

Enhanced Flexibility

Managed application delivery services also provide a degree of flexibility that allows the enterprise to adapt rapidly to changes in the business environment resulting from competition or mergers/acquisitions. In addition, with an ADMSP, the enterprise may be able to avoid being locked in to a particular equipment vendor due to large sunk costs in expertise and equipment.

Different Types of Managed Application Delivery Services

There are two primary categories of managed application delivery service environments:

1. Site-based services comprised of managed WOCs and/or ADCs installed at participating enterprise sites
2. Internet-based services that deal with acceleration of applications (e.g, web access and SSL VPN access) that traverse the Internet

Site-based Services

These services are usually based on the deployment of managed WOCs at the central data center and at each remote site participating in the application optimization project, as illustrated in Figure 7.1. The WAN depicted in the figure is typically a private leased line network or a VPN based on Frame Relay, ATM or MPLS. The application optimization service may be offered as an optional add-on to a WAN service or as a standalone service that can run over WAN services provided by a third party. Where the application delivery service is bundled with a managed router and WAN service, both the WOC and the WAN router would be deployed and managed by the same MSP. The ADC shown in the figure is performing firewall, load balancing, and similar functions that may or may not be included in the MSP offering. Site-based services are generally based on MSP deployment of WOCs and/or ADCs that were described in detail in Chapter 6.

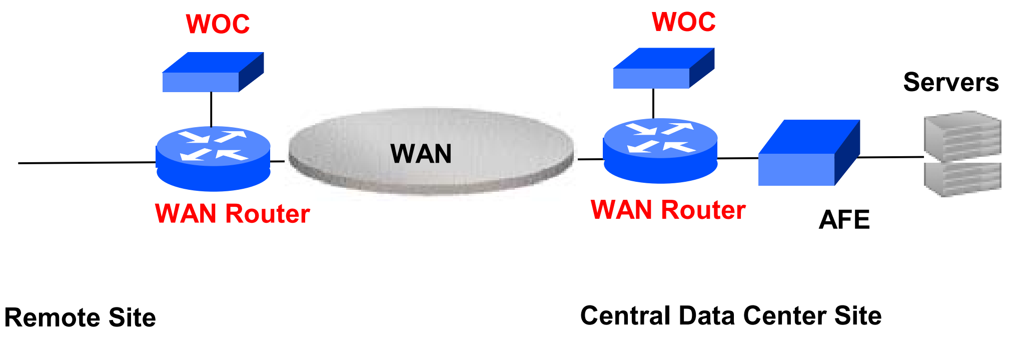


Figure 7.1: Site-Based Application Delivery Services

Internet-based Services

An increasing amount of enterprise WAN traffic is traversing the Internet. This is due to the attractiveness of the Internet as a lower cost alternative to WAN services such as Frame Relay and MPLS, and to the fact that for some of the enterprise's user constituencies (e.g., customers, suppliers, distributors) the Internet is the only viable WAN connectivity option. As the boundaries of the typical enterprise continue to be blurred due to a increasingly diverse user community, as well as the adoption of new distributed application architectures (e.g., Web-enabled applications and business processes, SOA/Web Services, SaaS, and Cloud Computing) that often traverse multiple enterprises, enterprise usage of the Internet for WAN connectivity is expected to continue to expand.

Over the last few years that IT organizations have focused on application delivery, the vast majority of that focus has been on either making some improvements within the data center or on improving the performance of applications that are delivered to branch office employees over private WAN services[16].

A comprehensive strategy for optimizing application delivery needs to address both optimization over the Internet and optimization over private WAN services.

Optimizing the delivery of applications that transit the Internet requires that flows be optimized within the Internet itself. This in turn requires subscription to an Application Delivery Service (ADS) offered by an MSP. Internet-based services are based primarily on proprietary application acceleration and WAN optimization servers located at MSP points of presence (PoPs) distributed across the Internet and do not require that remote sites accessing the services have any special hardware or software installed.

The benefits of these services include complete transparency to both the application infrastructure and the end-users. This transparency ensures the compatibility of the ADS with complementary application acceleration technologies provided by WOCs or ADCs deployed in the data center or at remote sites. ADSs are available for optimizing all IP-based applications as well as Web applications and these ADSs provide the visibility into the Internet traffic that is comparable to the visibility most IT organizations have relative to the traffic that transits private WAN services.

[16] Private WAN services refers to services such as private lines, Frame Relay, ATM and MPLS.

The Limitations of the Internet

When comparing the Internet with private WAN services, the primary advantages of the private WAN services are better control over latency and packet loss, as well as better isolation of the enterprise traffic and of the enterprise internal network from security threats. As will be discussed in this section, the limitations of the Internet result in performance problems. These performance problems impact all applications, including bulk file transfer applications as well as delay sensitive applications such as Voice over IP (VoIP), video conferencing and telepresence.

The primary reason for the limitation of the Internet is that as pointed out by Wikipedia[17], the Internet "Is a 'network of networks' that consists of millions of private and public, academic, business, and government networks of local to global scope." In the case of the Internet, the only service providers that get paid to carry Internet traffic are the providers of the first and last mile services. All of the service providers that carry traffic between the first and last mile do so without compensation. One of the affects of this business model is that there tend to be availability and performance bottlenecks at the peering points. Another affect is that since there is not a single, end-to-end provider, service level agreements (SLAs) for the availability and performance of the Internet are not available.

As noted, the primary source of packet loss within the Internet occurs at the peering points. Packet loss also occurs when router ports become congested. In either case, when a packet is dropped, TCP-based applications (including most critical enterprise data applications) behave as good network citizens, reacting to a lost packet by reducing offered load through halving the transmission window size and then following a slow start procedure of gradually increasing the window size in a linear fashion until the maximum window size is reached or another packet is dropped and the window is halved again.

With UDP-based applications, such as VoIP, Videoconferencing, and streaming video, there is no congestion control mechanism triggered by packet loss. As a result, the end systems continue to transmit at the same rate regardless of the number of lost packets. In the Internet, the enterprise subscriber has no control of the amount of UDP-based traffic flowing over links that are also carrying critical TCP application traffic. As a result, the enterprise subscriber cannot avoid circumstances where the aggregate traffic consumes excessive bandwidth which increases the latency and packet loss for TCP applications.

Another aspect of the Internet that can contribute to increased latency and packet loss is the use of the BGP routing protocol for routing traffic among Autonomous Domains (ADs). When choosing a route, BGP strives to minimize the number of hops between the origin and the destination networks. Unfortunately, BGP does not strive to choose a route with the optimal performance characteristics; i.e., the lowest delay or lowest packet loss. Given the dynamic nature of the Internet, a particular network link or peering point router can go through periods exhibiting severe delay and/or packet loss. As a result, the route that has the fewest hops is not necessarily the route that has the best performance.

Virtually all IT organizations have concerns regarding security intrusions via the Internet and hence have decided to protect enterprise private networks and data centers with firewalls and other devices that that can detect and isolate spurious traffic. At the application level, extra security is provided by securing application sessions and transactions using SSL authentication and encryption. As previously noted, processing of SSL session traffic is very compute-intensive and this has the affect of reducing the number of sessions that a given server can terminate. SSL processing can also add to the session latency even when appliances that can provide hardware-acceleration of SSL are deployed.

TCP has a number of characteristics that can cause the protocol to perform poorly when run over a lossy, high latency network. One of these characteristics is TCP's retransmission timeout. This parameter controls how long the transmitting device waits for an acknowledgement from the receiving device before assuming that the packets were lost and need to

[17] http://en.wikipedia.org/wiki/Internet

be retransmitted. If this parameter is set too high, it introduces needless delay as the transmitting device sits idle waiting for the timeout to occur. Conversely, if the parameter is set too low, it can increase the congestion that was the likely cause of the timeout occurring.

Another important TCP parameter is the previously mentioned TCP slow start algorithm. The slow start algorithm is part of the TCP congestion control strategy and it calls for the initial data transfer between two communicating devices to be severely constrained. The algorithm calls for the data transfer rate to increase linearly if there are no problems with the communications. When a packet is lost, however, the transmission rate is cut in half each time a packet loss is encountered.

The affect of packet loss on TCP was discussed in the section of Chapter 6 entitled "The Impact of Loss and Out of Order Packets". That section presented a formula to provide insight into the maximum TCP throughput on a single session when there is packet loss. That formula is:

Throughput <= (MSS/RTT)*(1 / sqrt{p})

where:

MSS: maximum segment size
RTT: round trip time
p: packet loss rate.

TCP throughput on a single session decreases as either the round trip time or the packet loss increases.

Similar to the example that was presented in the section of Chapter 6 entitled "The Impact of Loss and Out of Order Packets", assume that MSS is 1,460 bytes, RTT is 100 ms., and p is 0.5%. In this case, the maximum TCP throughput is 1.65 Mbps independent of the size of the WAN link.

Figure 7.2 demonstrates the impact of delay and packet loss on TCP throughput given an MSS of 1,460 bytes. The data in Figure 7.2 is normalized relative to an RTT of 100 ms. and packet loss of 0.5%. These values for RTT and p will be referred to as The Normalized Parameters. The statement that the data in Figure 7.2 is normalized means that the maximum TCP throughput with The Normalized Parameters is 1.0. It also means that the maximum TCP throughput for other values of RTT and p are depicted in the graph relative to the maximum TCP throughput for The Normalized Parameters.

For example, if the packet loss increases from 0.5% to 1.0%, then the normalized TCP throughput drops to approximately 0.7. This means that the maximum TCP throughput is reduced by 30%. Since the maximum TCP throughput with The Normalized Parameters is 1.65 Mbps, this results in a maximum TCP throughput of 1.15 Mbps.

If the packet loss were to increase to 2.0%, the maximum TCP throughput is reduced by approximately 50%. Analogously, if the packet loss stays fixed at 0.5%, but the RTT increases to 200 ms. then the maximum TCP throughput is also reduced by approximately 50%. In both cases, the maximum TCP throughput is roughly 0.83 Mbps independent of the size of the WAN link.

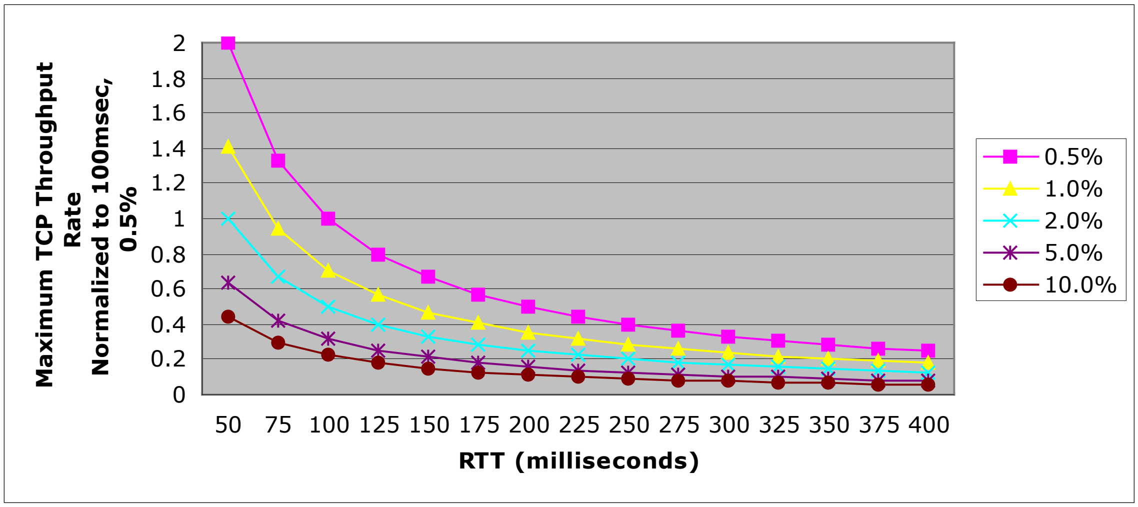


Figure 7.2: Impact of Delay and Packet Loss on TCP Throughput

Internet-Based Application Delivery Optimization

The traditional classes of application delivery solutions (ADC, WOC, soft WOC) that were described in Chapter 6 were designed to address application performance issues at both the client and server endpoints. These solutions make the assumption that performance characteristics within the WAN itself are not optimizable because they are determined by the relatively static service parameters controlled by the WAN service provider. This assumption is reasonable in the case of private WAN services. However, this assumption does not apply to enterprise application traffic that transits the Internet because there are significant opportunities to optimize performance within the Internet itself based on Application Delivery Services (ADSs). An ADS leverages service provider resources that are distributed throughout the Internet in order to optimize the performance, security, reliability, and visibility of the enterprise's Internet traffic. As shown in Figure 7.3, all client requests to the application's origin server in the data center are redirected via DNS to an ADS server in a nearby point of presence (PoP). This edge server then optimizes the traffic flow to the ADS server closest to the data center's origin server.

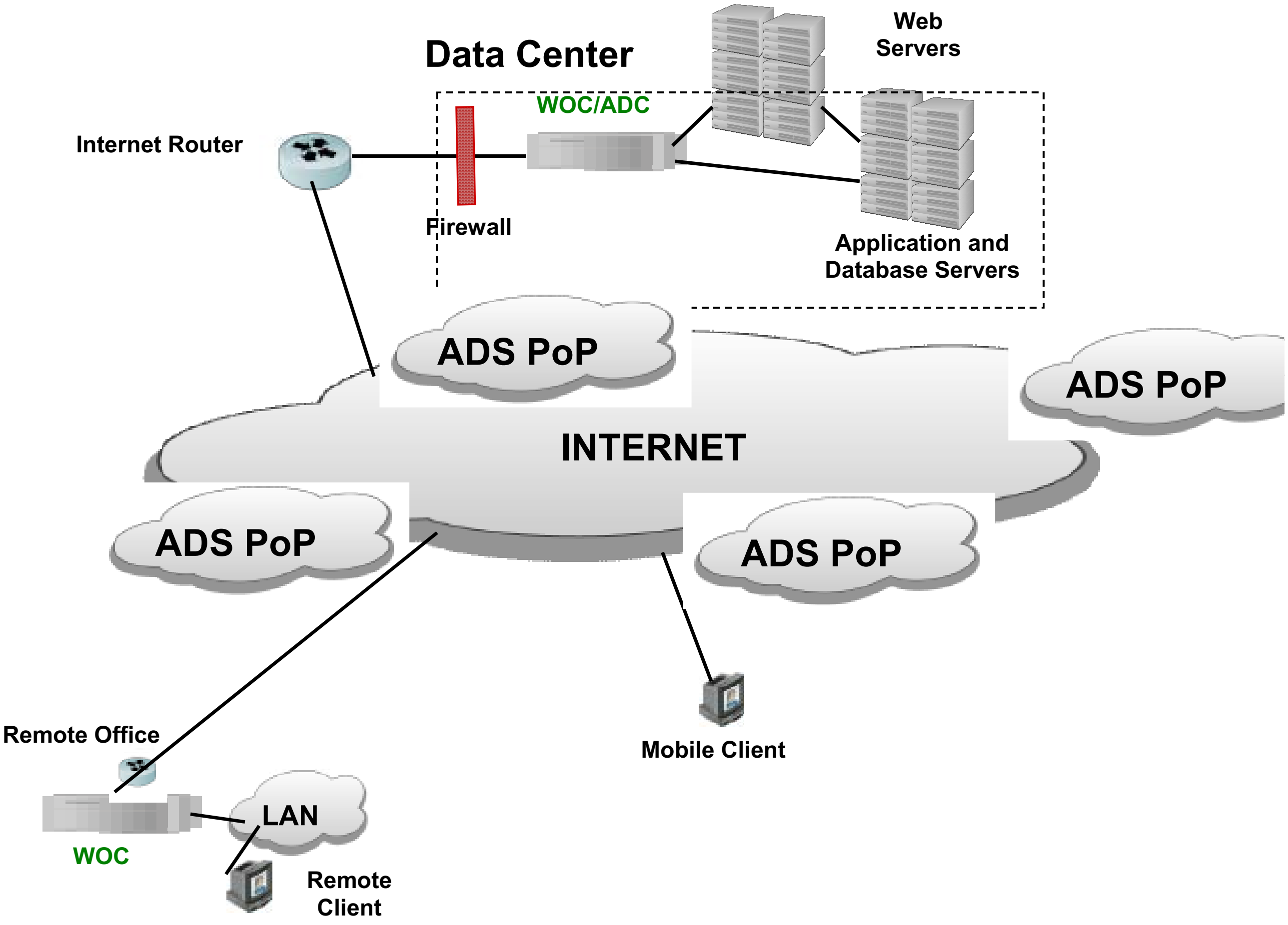


Figure 7.3: The Internet Infrastructure for an ADS

The servers at the ADS provider's PoPs perform a variety of optimization functions that generally complement the traditional application delivery solutions rather than overlap or compete with them. Some of the ADS functions include:

Route Optimization

Route optimization is a technique for circumventing the limitations of BGP by dynamically optimizing the round trip time between each end user and the application server. A route optimization solution leverages the intelligence of the ADS servers throughout the PoPs to measure the performance of multiple paths through the Internet and chooses the optimum path from origin to destination. The selected route factors in the degree of congestion, traffic load, and availability on each potential path to provide the lowest possible latency and packet loss for each user session.

As shown in Figure 7.4, the impact of route optimization can be dramatic. The data in Figure 7.4 depicts the round trip latency between Los Angeles, CA and Bangalore, India. The red graph reflects the round trip latency over the Internet and the green graph represents the round trip latency that results from using an ADS. Not only is the round trip latency significantly reduced by using an ADS, but the spikes in latency are removed.

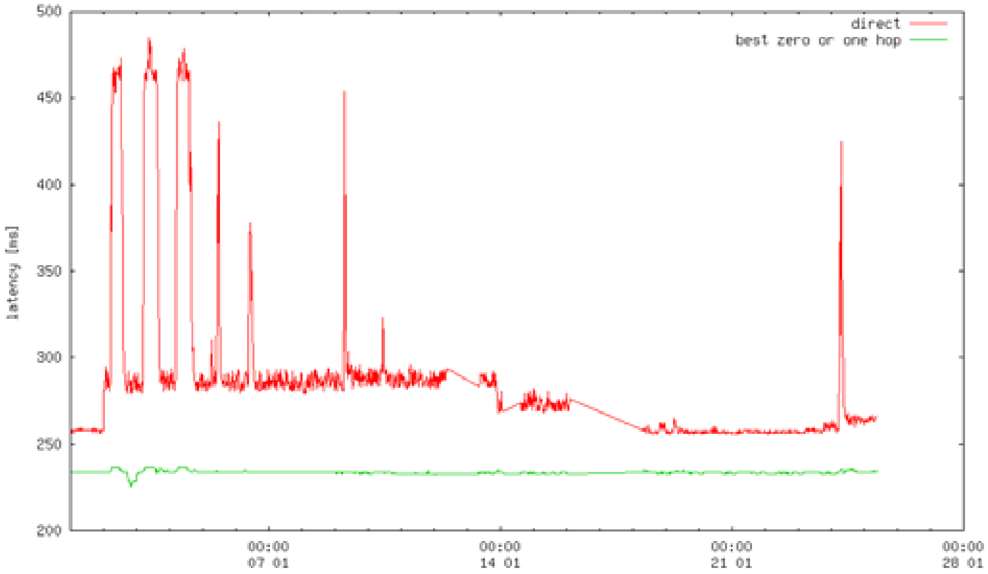


Figure 7.4: Internet Performance With and Without Route Optimization

Transport Optimization

TCP performance can be optimized by setting retransmission timeout and slow start parameters dynamically based on the characteristics of the network such as the speed of the links and the distance between the transmitting and receiving devices. TCP optimization can be implemented either asymmetrically (typically by an ADC) or symmetrically over a private WAN service between two WOCs, or within the Internet cloud by a pair of ADS servers in the ingress and egress PoPs. The edge ADS servers can also apply asymmetrical TCP optimization to the transport between the subscriber sites and the ADS PoPs. It should be noted that because of its ability to optimize based on real time network parameters, symmetrical optimization is considerably more effective than is asymmetrical optimization.

Another approach to transport optimization is to replace TCP with a higher performing transport protocol for the traffic flowing over the Internet between in the ingress and egress ADS servers. By controlling both ends of the long-haul Internet connection with symmetric ADS servers, a high performance transport protocol can eliminate most of the inefficiencies associated with TCP, including the three-way handshake for connection setup and teardown, the slow start algorithm, and re-transmission timer issues. For subscriber traffic flowing between ADS servers, additional techniques are available to reduce packet loss, including forward error correction and packet replication.

There is a strong synergy between route optimization and transport optimization because both an optimized version of TCP or a higher performance transport protocols will operate more efficiently over route-optimized paths that exhibit lower latency and packet loss. Unfortunately, as shown in Table 7.2, conventional Internet routes often exhibit high levels of latency and packet loss[18]. What is even worse for some applications is that, as shown in Table 7.2, the latency over the Internet can vary widely.

[18] The data in Table 7.2 was compiled over two weeks in March 2008.

Destination	Median RTT	Maximum RTT	Peak Packet Loss
NYC, New York	22 ms	185 ms	28%
Dallas, TX	22 ms	65 ms	8%
Paris, France	94 ms	148 ms	14%
Beijing, China	280 ms	510 ms	54%
Seoul, South Korea	275 ms	580 ms	15%
Tel Aviv, Israel	193 ms	295 ms	16%

Table 7.2: Internet Performance with Chicago as the Origin

HTTP Protocol Optimization

HTTP inefficiencies can be eliminated by techniques such as compression and caching at the edge ADS server with the cache performing intelligent pre-fetching from the origin. With pre-fetching, the ADS edge server parses HTML pages and brings dynamic content into the cache. When there is a cache hit on pre-fetched content, response time can be nearly instantaneous. With the caches located in nearby ADS PoPs, multiple users can leverage the same frequently accessed information.

Content Offload

Static content can be offloaded out of the data-center to caches in ADS servers and through persistent, replicated in-cloud storage facilities. Offloading content and storage to the Internet cloud reduces both server utilization and the bandwidth utilization of data center access links, significantly enhancing the scalability of the data center without requiring more servers, storage, and network bandwidth. ADS content offload complements ADC functionality to further enhance the scalability of the data center.

Security

The ADS servers can also be used to move the outer limits of the enterprise security perimeter from the data center into the cloud. Security services in the cloud can provide firewall-like traffic screening with Level 3-7 intelligence for access control, filtering, and validity checking that can keep malicious traffic outside of the data-center. The extra layer of security can also isolate the data center from DDoS attacks.

Availability

Dynamic route optimization technology can improve the effective availability of the Internet itself by ensuring that viable routes are found to circumvent outages, peering issues or congestion. For users accessing applications over the Internet, availability of the cloud is just as important as the availability of data center resources.

Visibility

Intelligence within the ADS servers can also be leveraged to provide extensive monitoring, configuration control and SLA monitoring of a subscriber's application with performance metrics, analysis, and alerts made visible to the subscriber via a Web portal.

MSP Selection Criterion

The beginning of this Chapter listed a number of benefits that an IT organization may gain from using an MSP for application delivery. These benefits are criteria that IT organizations can use as part of their evaluation of ADMSPs. For example, IT organizations should evaluate the degree to which using a particular ADMSP would allow them to lower their total cost of ownership or leverage that ADMSP's management processes.

Independent of whether an IT organization is evaluating a site-based service or an Internet-based service, they should consider the following criteria:

- Is the MSP offering a turnkey solution with simple pricing?
- Does the MSP provide network and application performance monitoring?
- Does the MSP provide a simple to understand management dashboard?
- What functionality does the MSP have to troubleshoot problems in both a proactive and a reactive fashion?
- What professional services (i.e., assessment, design and planning, performance analysis and optimization, implementation) are available?
- What technologies are included as part of the service?
- What is the impact of these technologies on network and application performance?
- Does the MSP offer application level SLAs?
- What is the scope of the service? Does it include application management? Server management?
- Is it possible to deploy a site-based application delivery service and not deploy WAN services from the same supplier?

8.0 Management

Some of the primary management tasks associated with application delivery are:

- Discovering the applications running over the network and identifying how they are being used.
- Gathering the appropriate management data on the performance of the applications and the infrastructure that supports them.
- Providing end-to-end visibility into the ongoing performance of the applications and the infrastructure.
- Identifying the sources of delay in the performance of the applications and the infrastructure.
- Supporting the deployment of new technologies such as virtualized servers.
- Automatically identifying performance issues and resolve them.

As Chapter 2 mentioned, Webtorials 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.***

As part of that survey, Webtorials also asked the survey respondents to indicate what component of IT was the biggest cause of application degradation. Figure 8.1 summarizes their answers. In this figure, the answer *shared equally* means that multiple components of IT are equally likely to cause application degradation.

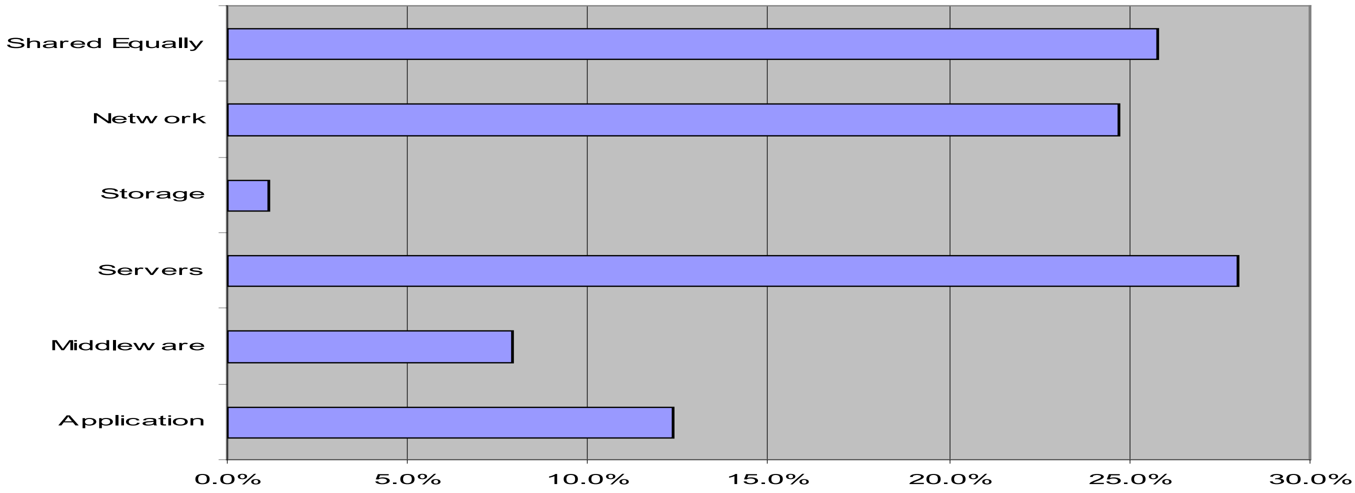


Figure 8.1: Causes of Application Degradation

The data in Figure 8.1 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 Dynamic

To understand how IT organizations respond to application degradation, Webtorials 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 8.1 contains their answers.

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%

Table 8.1: Organization Responsible for Application Performance

Webtorials recently asked over 200 IT professionals "How would you characterize the current relationship between your company's application development organization and the network organization?" Their responses are depicted in Table 8.2.

Response	Percentage of Respondents
Highly adversarial	0.0%
Moderately adversarial	7.9%
Slightly adversarial	17.2%
Neutral	33.0%
Slightly cooperative	13.3%
Moderately cooperative	24.6%
Highly cooperative	3.9%

Table 8.2: The Relationship between IT Groups

In roughly twenty-five percent of companies there is an adversarial relationship between the applications development groups and the network organization.

The data in Tables 8.1 and 8.2 speak to the organizational dynamic associated with managing application performance. Taken together with the data in Figure 8.1, managing application performance clearly is complex, both technically and organizationally.

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

In order to put the technical and organizational complexity that is associated with application delivery into context, Webtorials asked 200 IT professionals "When an application is degrading, how difficult is it for you to identify the root cause of the degradation? An answer of neutral means that identifying the root cause of application degradation is as difficult as identifying the root cause of a network outage." Their responses are contained in Table 8.3.

Extremely Easy: 1	2	3	Neutral: 4	5	6	Extremely Difficult: 7
1.6%	6.0%	12.0%	23.4%	23.9%	25.0%	8.2%

Table 8.3: The Difficulty of Identifying the Cause of Application Degradation

Identifying the root cause of application degradation is significantly more difficult than identifying the root cause of a network outage.

The Process Barriers

Webtorials asked hundreds of IT professionals if their companies have a formalized set of processes for identifying and resolving application degradation. Table 8.4 contains their answers. The data in Table 8.4 clearly indicate that the majority of IT organizations either currently have processes, or soon will, to identify and resolve application degradation.

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%

Table 8.4: Existence of Formalized Processes

Webtorials 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 8.5 shows the answers that received the highest percentage of responses.

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%

Table 8.5: Impediments to Effective Application Delivery

The data in Table 8.5 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 8.4. In particular, the data in Table 8.4 indicates that the vast majority of IT organizations either have formalized processes for identifying and resolving application degradation, or are developing these processes. However, the data in Table 8.5, also indicates that, in many cases, these processes are inadequate. In Chapter 9, we will discuss the interest on the part of IT organizations to leverage ITIL (IT Infrastructure Library) to develop more effective IT processes.

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.

Discovery

Chapter 5 of this handbook 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.

Chapter 4 mentioned one reason why 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 and/or control the applications that are running on the network and have no business relevance. To put this in context, Figure 8.2 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 multi-user gaming 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 even higher than what Figure 8.2 reflects.

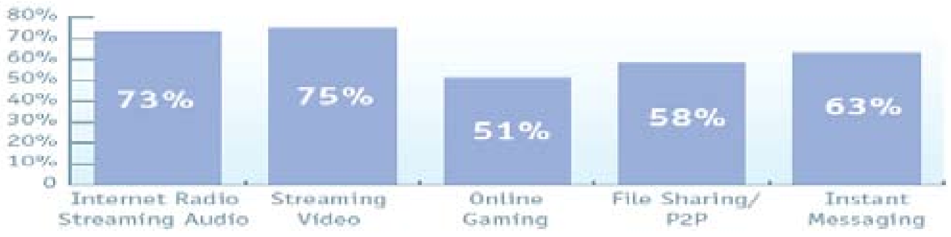


Figure 8.2: Occurrence of Recreational Applications

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

The Port 80 Black Hole

As noted, identifying the applications running on a network is a critical part of managing application performance. Unfortunately, there are many applications whose behavior makes this a difficult task; in particular, those that use *port hopping* to avoid detection.

In IP networks, TCP and UDP ports are endpoints to logical connections and provide the multiplexing mechanism to allow multiple applications to share a single connection to the IP network. Port numbers range from 0 to 65535. As described in the IANA (Internet Assigned Numbers Authority) Port Number document (www.iana.org/assignments/port-numbers), the ports that are numbered from 0 to 1023 are reserved for privileged system-level services and are designated as *well-known ports*. A well-known port serves as a contact point for a client to access a particular service over the network. For Example, port 80 is the well-known port for HTTP data exchange and port 443 is the well-known port for secure HTTP exchanges via HTTPS.

Because servers listen to port 80 expecting to receive data from Web clients, a firewall can't block port 80 without eliminating much of the traffic on which a business may depend. Taking advantage of this fact, many applications will port-hop to port 80 when their normally assigned ports are blocked by a firewall. This behavior creates what is referred to as *the port 80 black hole.*

Lack of visibility into the traffic that transits port 80 is a major vulnerability for IT organizations.

The port 80 black hole can have four primary effects on an IT organizations and the business it serves:

- Increased vulnerability to security breaches
- Increased difficulty in complying with government and industry regulations
- Increased vulnerability to charges of copyright violation
- Increased difficulty in managing the performance of key business-critical, time-sensitive applications

Port Hopping

Two applications that often use port hopping are instant messaging (IM) and peer-to-peer (P2P) applications such as Skype.

Instant Messaging

An example of a port-hopping instant messaging client is AOL's Instant Messenger (AIM). AOL has been assigned ports 5190 through 5193 for its Internet traffic, and AIM is typically configured to use these ports. If these ports are blocked, however, AIM will use port 80. As a result, network managers might well think that by blocking ports 5190 – 5193 they are blocking the use of AIM when in reality they are not.

The point of discussing AIM is not to state whether or not a company should block AIM traffic. That is a policy decision that needs to be made by the management of the company. Some of the reasons why a company might choose to block AIM include security and compliance. AIM can present a security risk because it is an increasingly popular vector for virus and worm transmission. As for compliance, a good example is the requirement by the Securities and Exchange Commission that all stock brokers keep complete records of all communications with clients. This requires that phone calls be recorded, and both email and IM archived. However, if AIM traffic is flowing through port 80 along with lots of other traffic, most network organizations will not even be aware of its existence.

Peer-to-Peer Networks and Skype

A peer-to-peer computer network leverages the connectivity between the participants in a network. Unlike a typical client-server network where communication is typically to and from a central server along fixed connections, P2P nodes are generally connected via largely ad hoc connections. Such networks are useful for many purposes, including file sharing and IP telephony.

Skype is a peer-to-peer based IP telephony and IP video service developed by Skype Technologies SA. The founders of Skype Technologies SA are the same people who developed the file sharing application Kazaa. Many IT organizations attempt to block peer-to-peer networks because they have been associated with distributing content in violation of copyright laws, and are often easy targets for security breaches.

Many peer-to-peer applications, including Skype, change the port that they use each time they start. Consequently, there is no standard 'Skype port' like there is a 'SIP port' or 'SMTP port'. In addition, Skype is particularly adept at port-hopping with the aim of traversing enterprise firewalls. Once inside the firewall, it then intentionally connects to other Skype clients. If one of those clients happens to be infected, then the machines that connect to it can be infected with no protection from the firewall. Moreover, because Skype has the ability to port-hop, it is much harder to detect anomalous behavior or configure network security devices to block the spread of the infection."

FIX-Based Applications

Another component of the port 80 black hole is the existence of applications designed to use port 80 but which require more careful management than the typical port 80 traffic. A good example of this is virtually any application based on the Financial Information eXchange ('FIX') protocol. The FIX protocol is a series of messaging specifications for the electronic communication of trade-related messages. Since its inception in 1992 as a bilateral communications framework for equity trading between Fidelity Investments and Salomon Brothers, FIX has become the de-facto messaging standard for pre-trade and trade communications globally within equity markets, and is now experiencing rapid expansion into the post-trade space, supporting Straight-Through-Processing (STP) from Indication-of-Interest (IOI) to Allocations and Confirmations.

End-to-End Visibility

Our industry uses the phrase *end-to-end visibility* in various ways. Given that one of this handbook'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 handbook 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 communications once users hit ENTER or click the mouse button 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 needs 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 so it can fix the problem quickly.
- Enables better cross-functional collaboration. As previously 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.

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

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, access points, firewalls, ADCs, WOCs, intrusion detection and intrusion prevention appliances. 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 Chapter 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 either dedicated instrumentation or software agents 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 XenApp.
- 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.
- Add minimum management traffic overhead.
- Support granular data collection.
- Capture performance data as well as events such as a fault.
- 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.
- Integrate with other management systems.
- 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-time 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 8.6 contains their answers to that question. Survey Respondents were instructed to indicate as many parameters as applied to their situation.

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%

Table 8.6: Percentage of Companies that Set Specific Thresholds

As Table 8.6 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 8.7 contains their answers.

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%

Table 8.7 : Approach to Setting Thresholds

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 we can draw from Table 8.5 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 operationally, it leads to an obvious conclusion:

Most IT organizations ignore the majority of the performance alarms.

Proactive Alarms

As the survey response illustrates, 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: problems 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 Chapter 5 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.

Route Analytics

The section of Chapter 5 entitled "Predicting the Impact of Change" discussed the importance of using the same tools for planning and for operations. It also discussed the use of route analytics for planning. This section of the handbook will expand on the use of route analytics for operations.

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. This distributed intelligence is both a strength and a weakness of IP. 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.

	Percentage of Respondents	Percentage of Respondents *not* including "don't know" respondents
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%	

Table 8.8: Impact of Logical vs. Device Specific Factors

As Table 8.8 shows, a high percentage of survey respondents answered *don't know.* To compensate for this, the far right column of Table 8.8 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.

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 mis-configurations.

The preceding section 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 traffic 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.

Criteria to evaluate a route analytics solution is the ability of the solution to:

- Listen to and participate in the routing protocol exchanges between routers as they communicate with each other.
- Compute 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.
- Map Netflow traffic data, including application composition, across all paths and links in the map.
- Monitor and display routing topology and traffic flow changes as they happen.
- Detect and alert on routing events or failures as routers announce them, and report on correlated traffic impact.
- Correlate routing events with other information, such as performance data, to identify underlying cause and effect.
- Record, analyze and report on historical routing and traffic events and trends.
- Simulate 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 misconfiguration 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.

Two hundred IT professionals were given the following question: "Sometimes logical problems such as routing issues are the source of application degradation and application outages. Which of the following describes how you resolve those types of logical issues?" Their answers are shown in Table 8.9.

Approach	Percentage of Respondents
Lots of hard work – typically by digging deeply into each device	38.7%
Employee specific tools such as route analytics	24.9%
N/A or don't know	19.9%
Waiting for it to happen again and trying to capture it in real time	13.3%
Other	3.3%

Table 8.9: Resolving Logical Issues

As table 8.9 shows, many IT organizations still rely on laborious manual processes, or simply hope to be able to catch recurrences of issues for which there are no obvious physical/device explanations. This indicates that most network management toolsets lack the ability to address the logical issues for which route analytics tools are useful. In particular, the ability of route analytics to *rewind* the entire recorded history of network-wide routing and traffic helps network engineers look into logical issues as if they were seeing them currently. This level of automation can greatly speed problem localization and root cause analysis. Since many logical problems exhibit symptoms only intermittently, getting to the root of these problems rather than hoping to solve them in the future also can help increase the overall stability of application delivery and performance.

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.

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 somewhat 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 8.10 depicts those categories, and the numerical rating associated with each.

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

Table 8.10: Mean Opinion Scores and Speech Quality

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 8.11[19] depicts the relationship between 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

Table 8.11: Comparison of R-Values and Mean Opinion Scores

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 Alliance[20] 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.

[19] Overcoming Barriers to High-Quality Voice over IP Deployments, Intel
[20] http://www.apdex.org/index.html

Application Performance Management

Application performance management (APM) is a relatively new management discipline. The newness of APM is attested to by the fact that ITIL has yet to create a framework for APM. Successful APM requires a holistic approach based on integrated management of both the application itself as well as the end-to-end IT infrastructure. This approach must focus on the experience of the end user of the application or service and must address most, if not all, of the following aspects of management:

- Adoption of a system of service level agreements (SLAs) at levels that ensure effective business processes and user satisfaction.
- End-to-end monitoring of all end user transactions. Monitoring actual user transactions in production environments provides valuable insight into the end-user experience and provide the basis for the ability to quickly identify, prioritize, triage, and resolve problems that can affect business processes.
- Automatic discovery of all the elements in the IT infrastructure that support each service. This provides the basis for the ability to create two-way mappings between the services and the supporting infrastructure components. These mappings, combined with event correlation and visualization, can facilitate root cause analysis, significantly reducing mean-time-to-repair.
- With service-infrastructure mappings, monitoring can be extended to identify when services are about to begin to degrade because of problems in the infrastructure. As part of this monitoring, predictive techniques such as heuristic-based trending of software issues and infrastructure key performance indicators can be employed to identify and alert management of problems before they impact end users.
- Outages and other incidents that generate alerts can be prioritized based on potential business impact. Prioritization can be based on a number of factors, including: the affected business process and its value to the enterprise, the identity and number of users affected, and the severity of the issue.
- Triage and root cause analysis can be applied at the application and infrastructure levels. When applied directly to applications, triage and root cause analysis can identify application issues such as the depletion of threads and pooled resources, memory leaks or internal failures within a Java server or .NET server. At the infrastructure level, root cause analysis can determine the subsystem within the component that is causing the problem.
- Automated generation of performance dashboards and historical reports allows both IT and business managers to gain insight into SLA compliance and performance trends. These insights can be applied to further enhancements in IT support for business processes, capacity planning, and the adoption of new technologies that can further improve the optimization, control, and management of service performance.

Managing Virtualized Servers

One of the primary management tasks associated with application delivery is to support the deployment of new technologies such as virtualized servers. Chapter 4 discussed some of the benefits and challenges associated with deploying virtualized servers. This section of the handbook will discuss in detail some of the management challenges that come as a result of implementing virtualized servers.

Today the majority of IT organizations that have implemented virtualized servers have used VMware®. There are, however, a number of other vendors, including Microsoft and Citrix, who have also entered this market. As such, for most IT organizations managing virtual servers will soon take on the added complications that result from a multi-vendor environment.

Managing Virtual Machines

Virtual machines that reside on a given physical server communicate with each other using a virtual switch often referred to as a vSwitch. Unfortunately, unlike the typical physical switch, a vSwitch provides limited traffic visibility. For example, while most embedded virtualization management tools can identify the total volume of traffic within the entire virtual environment, they cannot provide information on individual network services such as HTTP or FTP. A vSwitch also provides little if any inherent security visibility. As a result, it is typically not possible for an IT organization to know if unsanctioned network services have been enabled, or if worms or other forms of malware are propagating within the virtual environment.

Traditional methods of monitoring traffic on a physical LAN switch cannot be used to monitor traffic that goes between virtual machines.

One of the characteristics of a virtual machine is that it only has at its disposal a fraction of the resources (i.e., CPU, memory, storage) of the physical server on which it resides. As a result, any effective management tool must not consume significant resources. One viable option to gather detailed management data without consuming significant resources is to use NetFlow.

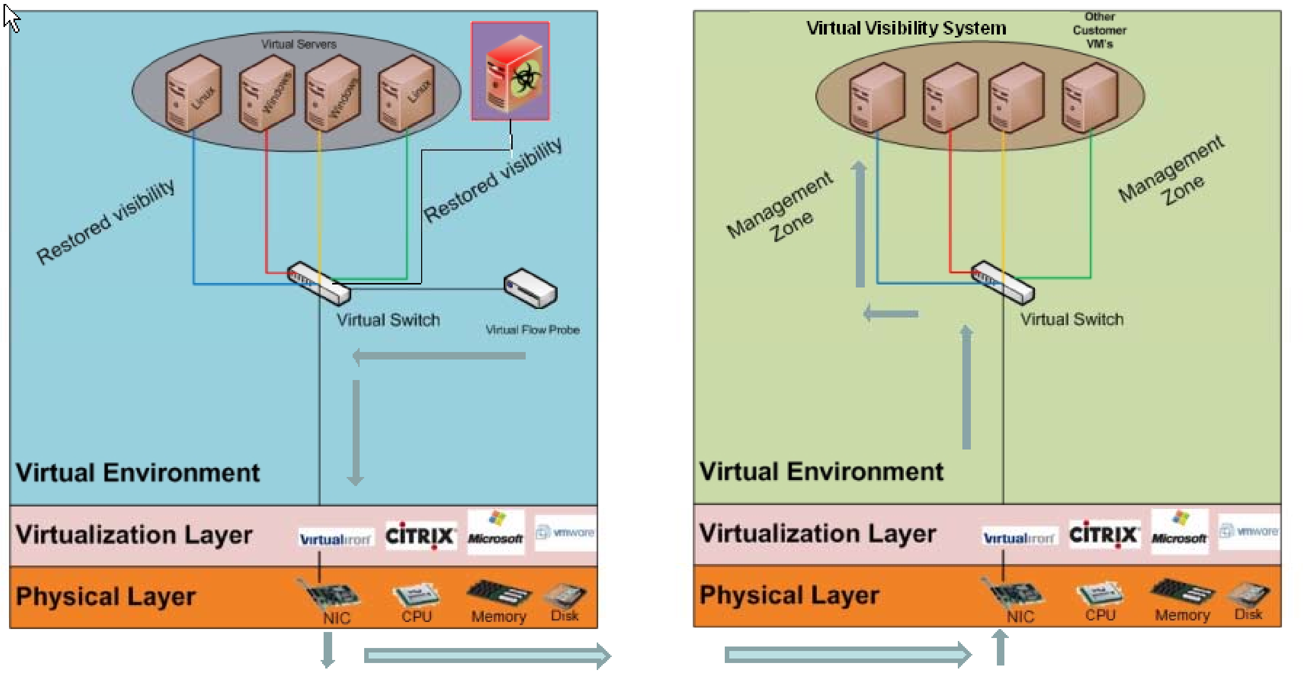


Figure 8.3: Management of Virtual Machines

As shown in Figure 8.3, virtual machines are managed by the use of a virtual probe, which is a virtual machine with software on it that does packet capture and converts the captured data to NetFlow records. The way this works is that, similar to how many IT organizations monitor a physical switch, the vSwitch has one of its ports provisioned to be in promiscuous mode and as a result, it forwards all inter-VM traffic to this probe software. The probe software then converts the data to NetFlow records.

The NetFlow management data can be used to identify which network users or applications are consuming the bandwidth and this information can be used to perform a number of key management tasks, some of which are unique to virtualized servers. For example, VMotion is functionality provided by VMware that is capable of transferring an entire running virtual server from one physical server to another. While there are some significant benefits associated with VMotion, there are also some significant management challenges, such as identifying and tracking events that may trigger VMotion to transfer a virtual server to a different physical server. The NetFlow data can be used to respond to these management challenges.

Automated Problem Identification and Resolution

As mentioned in the introduction to this chapter, automatically identifying performance issues and resolving them is an important part of managing the application delivery process. Throughout this section, automatically identifying performance issues and resolving them will be referred to as APIR. The successful implementation of APIR is built upon an integrated system that provides three key management functions: find, configure and monitor.

Find

Particularly in large IT organizations, it is difficult if not impossible to track all of the components of the IT infrastructure (i.e., networking equipment, servers, storage and applications) using manual methods, as these methods are both time consuming and error prone. As a minimum, what is needed is the ability to automatically discover all of the IT infrastructure elements.

IT infrastructure elements such as network switches and routers are important unto themselves. These elements, however, are typically used to provide resources for services such as a VLAN (Virtual LAN) or a VPN (Virtual Private Network). As a result, what is also needed is the ability to discover services and to be able to perform deep subcomponent discovery in order to understand the relationships between deployed services and the subtending infrastructure elements, down to the level of individual ports and interfaces.

Configure

IT organizations are continually modifying their management processes. Given this, a successful implementation of APIR should allow IT organizations to easily modify their processes over time but must not require that they modify them as part of the implementation of the tool. As a result, an APIR tool must allow the IT organization to configure a wide range of devices using a broad range of approaches; i.e., GUIs, CLI, etc. Configuration should include exposing capabilities as automation tasks at both device and service level.

Because services such as VLANs typically comprise multiple devices, an APIR tool must enable an IT organization to automatically configure a service and all of the subtending IT infrastructure elements.

Monitor

Monitoring can be done either passively or actively. Using a passive approach, an infrastructure element such as a network switch would inform the management tool of a problem. Using an active approach, the management tool would interrogate infrastructure elements and would determine the health of the individual components and/or the associated service. Given that both approaches have their advantages and disadvantages, an effective solution requires both. In order to enable automated remediation, it is necessary that the monitoring tool capture a level of information that is granular enough to enable root cause analysis down to the subinterface level.

Historically, such functionality would involve multiple IT disciplines using multiple tools. However, in order to support the expanding role of the network manager (see Chapter 9) and to reduce the negative impact of technological and organizational stovepipes, it is important that all of these capabilities are available on a single system that features a customizable user interface.

Another critical success factor relative to implementing APIR is the use of a Configuration Management Data Base (CMDB). A CMDB is a repository of information related to all the components of an information system. ITIL coined the term CMDB [21] and has created a detailed description of a number of important IT processes (i.e., problem management, configuration management, incident management) with a set of comprehensive checklists, tasks and procedures that can be tailored to any IT organization.

In the ITIL context, a CMDB represents the authorized configuration of the significant components of the IT environment. A key goal of a CMDB is to help an organization understand the relationships between these components and track their configuration. A CMDB also stores contextual information about IT assets, finance, and organizational structure. A CMDB is a fundamental component of the ITIL framework for an effective configuration management process. CMDB implementations often involve integration with other systems, such as Asset Management Systems.

Implementing Automation

Automation is best applied where performing the management tasks manually is repetitious, time-consuming, and prone to human error. The remainder of this section provides an overview of where automation can be best applied to improve operational efficiency and effectiveness for key management functions.

Configuration Management

Minimizing the need for manual configuration of IT infrastructure elements reduces the operational workload and helps to eliminate human errors that can affect the reliability and security of the network. An automated configuration management system discovers the infrastructure elements and stores their configuration and associated business data in a CMDB, which can be used as the basis for generating both physical and logical perspectives of the entire IT infrastructure and the services it supports. As new elements or services are brought on-line, or changes are made to the configuration of the infrastructure, configurations can be automatically downloaded from the CMDB to ensure consistency and accuracy.

As noted, IT infrastructure elements (i.e., switches, routers, access points, firewalls, IDSs, IPSs, Windows servers, Linux servers, clients, printers, etc.) are typically combined into network services such as VLANs and VPNs. As such, an effective APIR tool must be able to assist in the rapid creation and/or modification of network services such as automating the creation and/or modification of a VLAN across multiple switches and switch ports.

The CMDB can also be leveraged to enforce policies related to software updates, authorized device access, and configuration changes. Maintaining a documented audit trail of actions taken by operational personnel, including addition, removal and modification of Configuration Items (CIs) reduces errors and helps automate the processes required for regulatory compliance.

[21] http://www.itil-officialsite.com/home/home.asp

Event Management

Event management systems are required to maximize network and service availability by reducing the downtime of critical devices and subsystems. Event management systems typically provide error detection and some degree of error analysis, together with alarm generation. Automated event management systems can proactively monitor IT infrastructure elements for conditions that may lead to fault events and can use automated event correlation and root cause analysis to determine the precise nature and location of the faults that do occur. Effective root cause analysis relies heavily on the service dependencies that are discovered, as well as an accurate physical and logical topology of the network, such as one derived from a CMDB. Furthermore, in addition to raising alarms, the automated event management system can use the results of the event detection and analysis processes to trigger automated remediation actions.

Service Level Management

Management at the service level is essential to ensure user satisfaction with services such as VoIP or video conferencing, as well as meeting Service Level Agreements (SLAs) for user access to ERP, CRM, and other business-critical application services. Automated service level management involves proactive monitoring of service level metrics such as availability, as well as service and application response times. Results from service level monitoring can be used to trigger actions and to create SLA violation events that can be automatically fed into the fault management system for diagnosis and resolution. Automated integration of service level management, fault management systems, and trouble ticket systems can help minimize service downtimes or SLA violations.

Security Management

Automated security management includes continual monitoring of the security configuration and status of the various elements of the network to ensure that security policies are enforced. Security audits can be automated as regularly scheduled scans or as responses driven by various security events. Security events can also by automatically correlated with other network events to minimize the number of false positives requiring management attention. Validated security events can then trigger automatic responses, such as reconfiguration of security devices, server reconfiguration, patch deployments, or revocation of user access privileges.

9.0 The Changing Network Management Function

The preceding chapter discussed the organizational complexity and the process barriers associated with ensuring acceptable application performance. This chapter describes the changing role of one of the key players in the application delivery function – the Network Operations Center (NOC). As part of that description this chapter examines the current and emerging role of the NOC, the attempt on the part of many NOCs to improve their processes, and highlights the shift that most NOCs are taking from where they focus almost exclusively on the availability of networks to where they are beginning to also focus on the performance of networks and applications.

It is now somewhat common to have the NOC heavily involved in managing the performance of applications. This chapter also examines how the NOC must change in order to reduce the meant time to repair associated with application performance issues, and details the ways that IT organizations justify an investment in performance management.

Today's NOC

Perceptions of the NOC

The survey respondents were asked if they thought that working in the NOC is considered to be prestigious. The NOC-associated respondents [22] were evenly split on this issue. That was not the case for the Non-NOC respondents[23]. By roughly a 2-to-1 margin, these respondents indicated that they do not think that working in the NOC is prestigious.

The survey respondents were asked a series of questions regarding senior IT management's attitude towards the NOC. The results are shown in Table 9.1.

Our senior IT management believes that...	Agree/ Tend To Agree	Disagree/ Tend to Disagree
...the NOC provides value to our organization.	90.7%	9.3%
...the NOC is a strategic function of IT.	87.9%	12.1%
...the NOC is capable of resolving problems in an effective manner.	82.4%	17.6%
...the NOC will be able to meet the organization's requirements 12 months from now.	81.4%	18.6%
...the NOC works efficiently.	80.6%	19.4%
...the NOC meets the organization's current needs.	71.9%	28.1%

Table 9.1: IT Management's Perception of the NOC

[22] NOC-associated respondents will refer to survey respondent who work in the NOC
[23] Non-NOC respondents will refer to survey respondents who do not work in the NOC

Overall the data in Table 9.1 is positive. There are, however, some notable exceptions.

In over a quarter of organizations, the NOC does not meet the organization's current needs.

The Function of the NOC

When it comes to how the NOC functions, one of the most disappointing findings is that:

In the majority of cases, the NOC tends to work on a reactive basis identifying a problem only after it impacts end users.

The survey also asked the respondents about the most common type of event that causes NOC personnel to take action. The replies of the NOC-associated respondents who provided a response other than "don't know" are depicted in Figure 9.1. The data in this figure indicates that roughly half the time either someone in the NOC or an automated alert causes the NOC to take action. This data, however, does not address the issue of whether or not this occurs before the user is impacted.

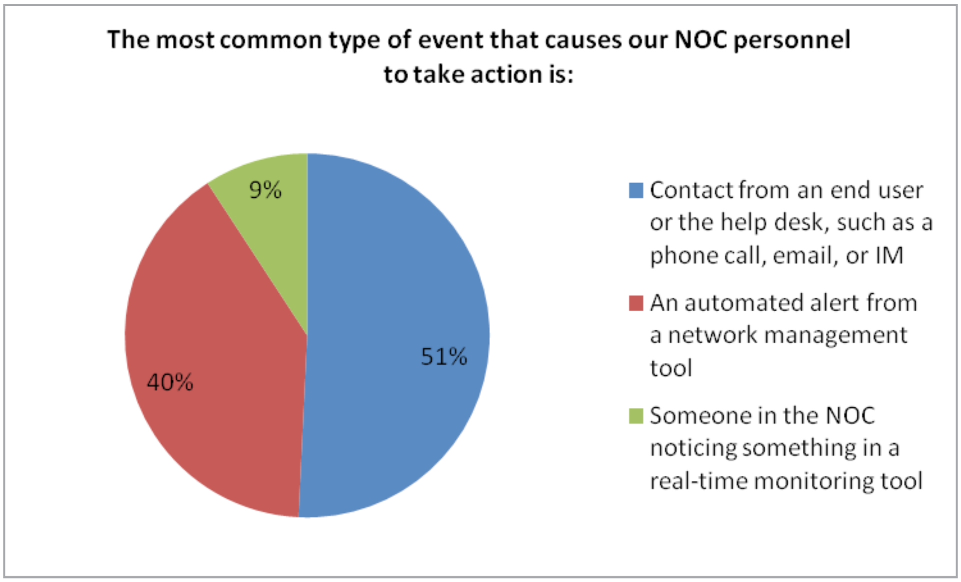


Figure 9.1: Events that Cause the NOC to Take Action

The conventional wisdom in our industry is that NOC efficiency is reduced because of the silos that exist within the NOC. In this context, silos means that the workgroups have few common goals, terminology, processes and tools. The survey respondents validated that conventional wisdom.

Just under half of NOCs are organized around functional silos.

A majority of NOCs use many management tools that are not well integrated.

Where Does the NOC Spend Most of Its Time?

To identify the areas in which NOC personnel spend most of their time, the survey contained three questions where each question contained a number of possible answers:

- During the past 12 months, our NOC personnel have spent the greatest amount of time addressing issues with…
- During the past 12 months, our NOC personnel have spent the second greatest amount of time addressing issues with…
- During the past 12 months, our NOC personnel have seen the greatest increase in time spent addressing issues with…

Table 9.2 shows the answers of the NOC-associated respondents.

	Greatest Amount of Time	Second Greatest Amount of Time	Greatest Increase in Time
Applications	39.1%	16.9%	45.0%
Servers	14.1%	21.5%	21.7%
LAN	10.9%	15.4%	5.0%
WAN	23.4%	30.8%	11.7%
Security	9.4%	6.2%	10.0%
Storage	3.1%	9.2%	6.7%

Table 9.2: Where the NOC Spends the Most Time

There are many conclusions that can be drawn from the data in Table 9.2, including:

NOC personnel spend the greatest amount of time on applications and that is a relatively new phenomenon.

NOC personnel spend an appreciable amount of their time supporting a broad range of IT functionality.

What Do NOC Personnel Monitor?

The Survey Respondents were asked four questions about what NOC personnel in their organization monitor; the results from NOC-associated respondents are shown in Figure 9.2.

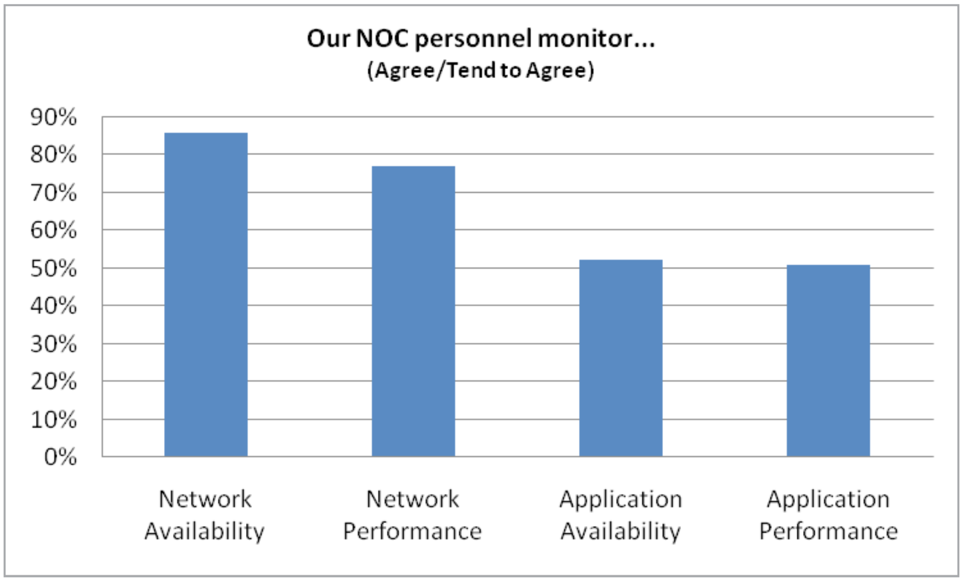


Figure 9.2: What the NOC Monitors

One of the conclusions that can be drawn from the data in Figure 9.2 is:

The NOC is almost as likely to monitor performance, as it is to monitor availability.

In addition, while there is still more of a focus in the NOC on networks, there is a significant emphasis on applications.

What Else Does the NOC Do?

We also asked the Survey Respondents about other tasks or responsibilities that NOC personnel are involved in. Figure 9.3 shows the responses for both NOC-associated and non-NOC personnel.

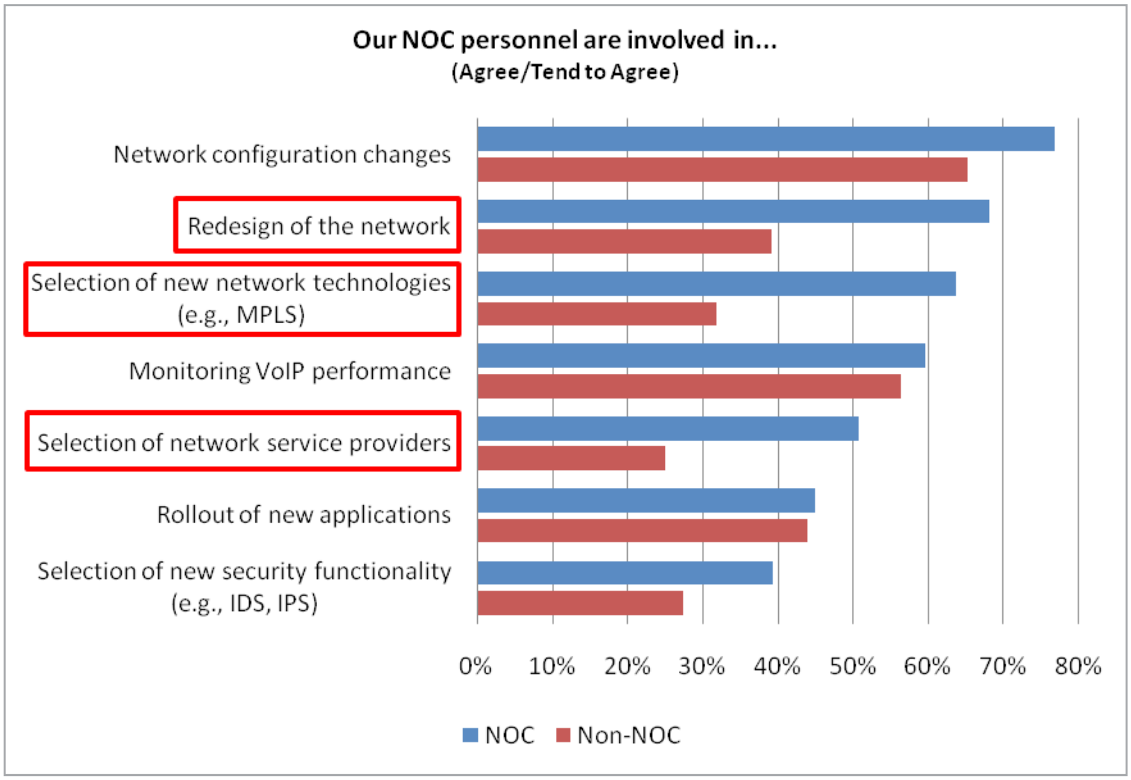


Figure 9.3: Tasks in which the NOC is involved

The most obvious conclusion that can be drawn from Figure 9.3 is that NOC personnel are involved in myriad tasks beyond simple monitoring. As shown, NOC personnel are typically involved in traditional network activities such as configuration changes, selection of new network technologies and the selection of network service providers. The NOC is less likely to be involved in application rollout and the selection of security functionality.

The responsibilities that are highlighted in Figure 9.3 are where NOC-associated and non-NOC respondents differed most in their responses. Interestingly, the areas where there were the greatest differences are all traditional networking activities.

The Use of ITIL

There has been significant discussion over the last few years about using a framework such as ITIL to improve network management practices. To probe the use of ITIL, the survey respondents were asked if their organization either now has an IT service management process such as ITIL in place, or intended to adopt such a process within the next 12 months. The majority of respondents (62%) indicated that their organization did have such a process in place. Of those respondents who did not, a similar percentage (63%) believed that their organization would put such a process in place within the next 12 months. The fact that 86% of respondents stated that their organization either had or would have within 12 months a service management process in place indicates the emphasis being placed within the NOC to improve their processes.

There is a lot of interest in ITIL, but it is too soon to determine how much impact the use of ITIL will have.

Routing Troubles

The vast majority of organizations have at least a simple escalation process in place for problem response. In particular, over 90% of Survey Respondents indicated that their organization has a help desk that assists end users, and over 80% agree that the help desk does a good job of routing issues that it cannot resolve to whatever group can best handle them. It should be noted that of the latter group of respondents (those agreeing), better than three-quarters stated that the help desk typically routes issues that it cannot resolve to the NOC. One of the reasons that the help desk routes so many calls to the NOC is the following:

In the vast majority of instances, the assumption is that the network is the source of application degradation.

The Survey Respondents were asked to indicate their degree of agreement with the statement: "Our NOC personnel not only identify problems, but are also involved in problem resolution." Their responses are depicted in Figure 9.4.

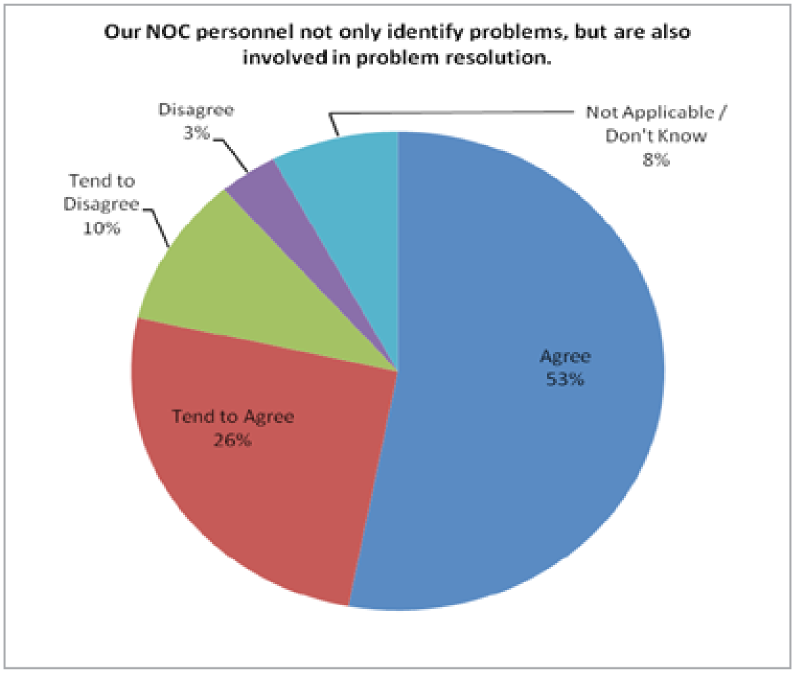


Figure 9.4: Role of the NOC in problem resolution

The data in Figure 9.4 is further evidence of the fact that NOC personnel do a lot more than just monitor networks.

In the majority of instances, the NOC gets involved in problem resolution.

Change in the NOC

Factors Driving Change

As shown in Table 9.1, over a quarter of the total base of survey respondents indicated that the NOC does not meet the organization's current needs. This level of dissatisfaction with the NOC is in line with the fact that as shown in Figure 9.5, almost two thirds of the respondents indicated that their organization would attempt to make any significant changes in their NOC processes within the next 12 months.

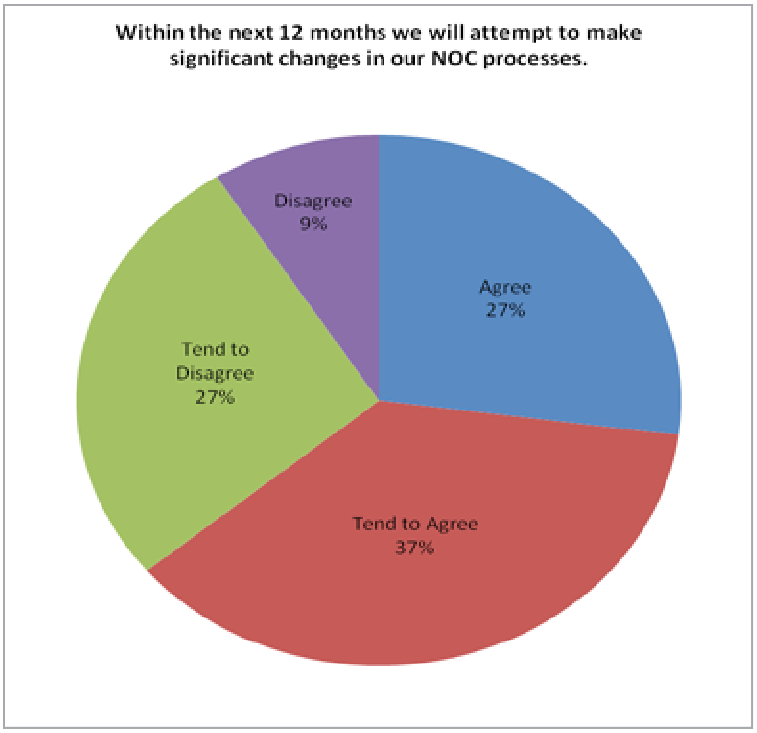


Figure 9.5: Interest in changing NOC processes

The survey respondents were asked to indicate which factors would drive their NOC to change within the next 12 months. Their responses are shown in Figure 9.6.

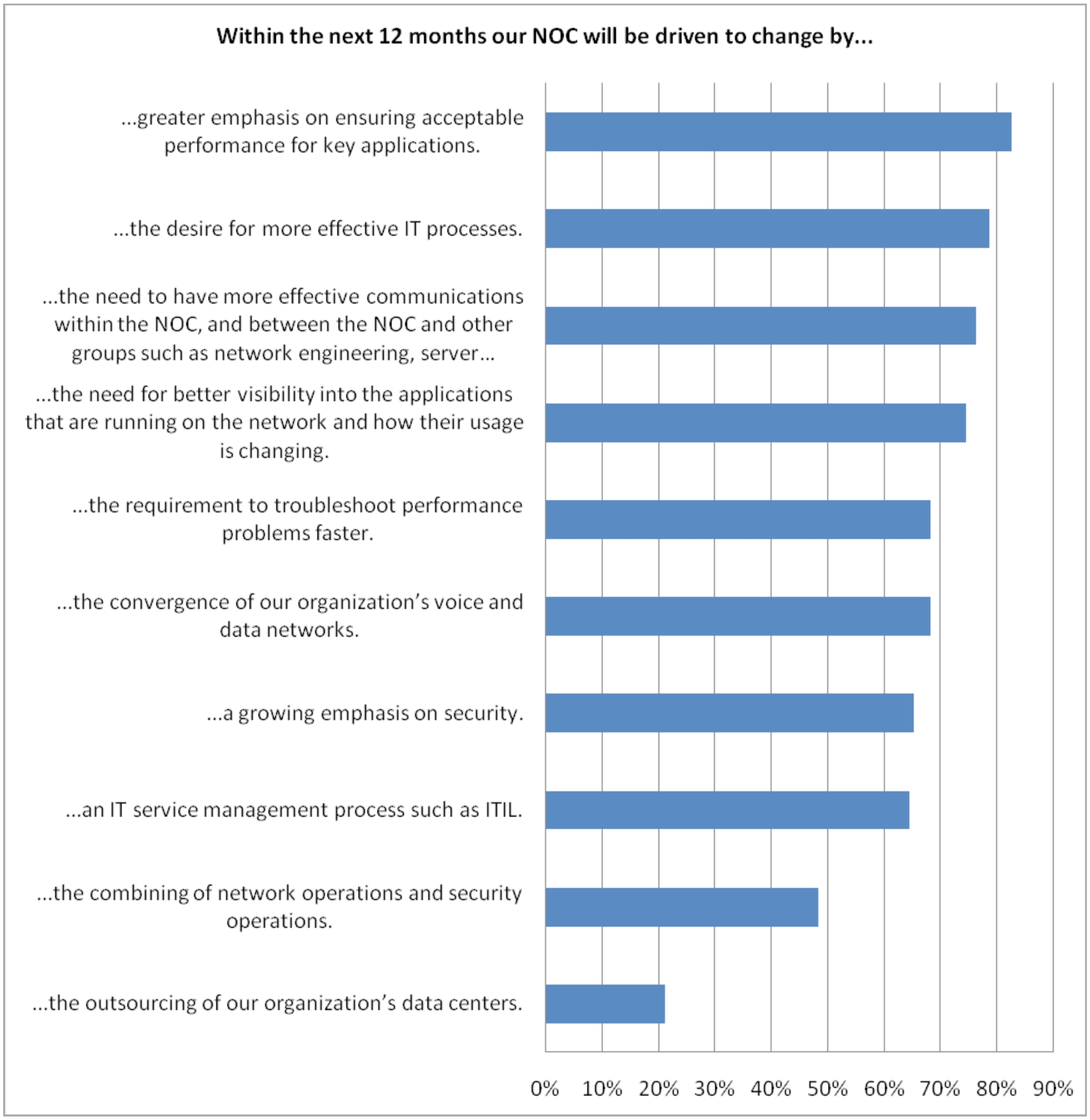


Figure 9.6: Factors driving change in the NOC

One clear conclusion that can be drawn from the data in Figure 9.6 is that a wide range of factors are driving change in the NOC. Given that NOC personnel spend the greatest amount of time on applications, it is not at all surprising that:

The top driver of change in the NOC is the requirement to place greater emphasis on ensuring acceptable performance for key applications.

And a related driver, the need for better visibility into applications, is almost as strong a factor causing change in the NOC.

As shown in Table 9.2, NOC personnel do not spend a lot of their time today on security. However, that is likely to change in the next year as roughly half of the Survey Respondents indicated that combining network and security operations would impact their NOC over the next 12 months. In addition, two thirds of the Survey Respondents also indicated that a growing emphasis on security would impact their NOC over the next 12 months.

Factors Inhibiting Change

Particularly within large organizations, change is difficult. To better understand the resistance to change, we asked the Survey Respondents to indicate what factors would inhibit their organization from improving their NOC. Their responses are shown in Figure 9.7.

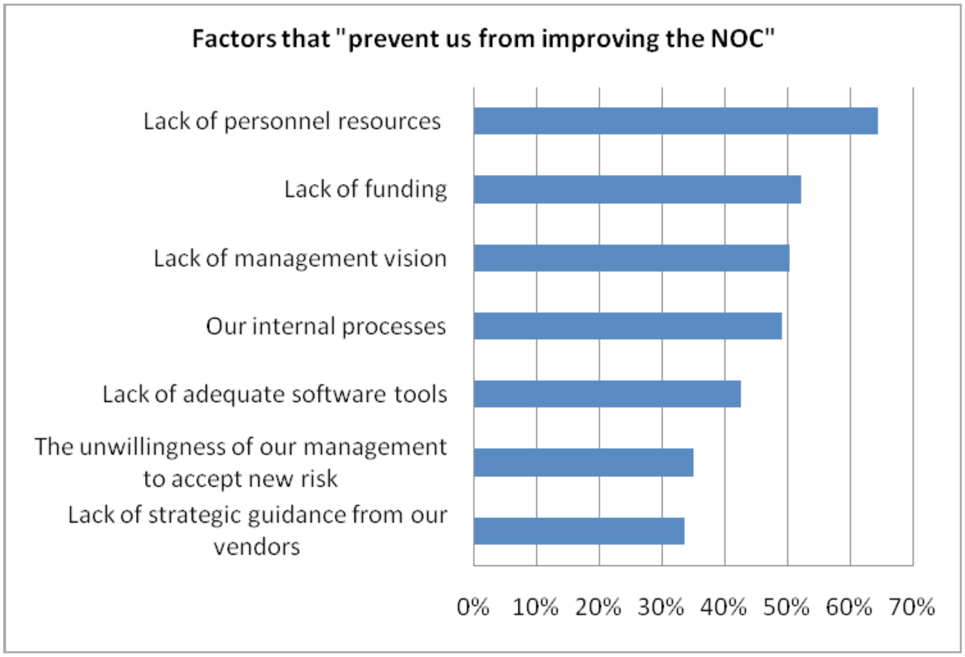


Figure 9.7: Factors inhibiting change in the NOC

It was not surprising that the two biggest factors inhibiting change are the lack of personnel resources and the lack of funding. This is in line with the general trend whereby IT budgets are increasing on average by only single digit amounts and headcount is often being held flat. It is also not surprising that internal processes are listed as a major factor inhibiting change. The siloed NOC, the interest in ITIL and the need to make significant changes to NOC processes have been constant themes throughout this chapter.

The lack of management vision and the NOC's existing processes are almost as big a barrier to change as are the lack of personnel resources and funding.

Call to Action: The Next-Generation Integrated Operations Center

The market research that was presented in this chapter demonstrates that there is considerable dissatisfaction with the role currently played by the NOC and as a result there is also widespread interest in making significant changes to the

NOC. Given the interest in making significant changes to the NOC, this section will describe the key characteristics of a truly next generation NOC – one that integrates the operations of each component of IT.

An Integrated Operations Center (IOC) would not have to be housed in a single facility, nor would it necessarily have to be provided by a single organization within the IT function. However, independent of how it is organized, the IT professionals who work in an IOC must have a common terminology [24] and common goals. Below is a listing of the other key characteristics of an IOC as well as a summary of where the bulk of IT organizations currently stand relative to each characteristic.

Efficient Processes

There is clear recognition on the part of the survey base that the NOC needs to improve its processes. There is also clear acknowledgement that the vast majority of IT organizations will use ITIL as part of their process improvement efforts.

Focus on Performance

Today's NOC is almost as likely to focus on performance as it is to focus on availability. This focus on performance will likely increase in the near term in part because placing greater emphasis on ensuring acceptable application performance for key applications is the strongest factor driving change in the NOC.

Skilled Staff

In general, the skill set of NOC personnel has been increasing and the majority of NOC personnel are now performing functions that until recently were considered to be Level 2 or Level 3 functions.

Automation & Intelligent Tools

Many NOCs have begun the shift away from having NOC personnel sitting at screens all day waiting for green lights to turn yellow or red. In addition, over a quarter of the NOC respondents indicated that their company has "eliminated or reduced the size of our NOC because we have automated monitoring, problem detection and notification." This trend, combined with the trend to increase the skill set of NOC personnel, indicates that more intelligence is being placed in the NOC, and that intelligence is a combination of people and tools.

Integrated set of Tools

Many IT organizations have stated that having management tools that are not well integrated is a fact of life. However, a common theme of the market research is that tool integration is one of the biggest issues IT organizations hope to address when they initiate a NOC redesign project.

Focus on Applications

NOCs currently have a significant focus on managing application performance and there is interest in increasing that focus. As a result, it is highly likely that within the next two years the vast majority of operations centers will be responsible for managing application performance.

Focus on Security

NOC personnel do not currently spend a lot of their time on security. However, two thirds of the survey respondents indicated that a growing emphasis on security will impact their NOC over the next 12 months. In addition, almost half of

[24] An example of having a common terminology is that everyone in the IOC has the same definition for the word service.

the Survey Respondents indicated that combining network and security operations will impact their NOC over the next 12 months.

Being Proactive

In spite of the widespread interest in being proactive, the majority of the NOCs currently work on a reactive basis, identifying a problem only after it impacts end users.

Relate IT Management Tasks with Business Processes

IT organizations need to identify their company's key business processes and then determine what IT assets (i.e., switches, routers, servers, databases, applications) support those processes. As a minimum, this allows IT to put any discussion of the performance of those assets into business terms. It also positions IT organizations to perform tasks such as prioritizing incidents based on their business impact.

The migration away from today's stove-piped, reactionary NOC to an effective IOC that exhibits the characteristics described above will not be easy. As described in chapter 3 this migration will require the active involvement of the IT organization's senior management. Part of senior management's role is to articulate a clear vision of the future role of the operations center, and to be the champion of that role, both inside of the IT organization as well as more broadly within the company. In addition, senior management must ensure the creation of a roadmap that leads to an effective IOC and must also closely manage the journey.

While it is the role of senior management to create the vision and the roadmap, a major part of the role of the rank and file members of the operations function is to ground senior management in terms of what is possible in what timeframe. The rank and file must also work with senior management to establish a program comprised of formal training, on-the-job training, and job rotations that leads to increasing and broadening the skills of the operations group. In addition, the rank and file must embrace change as their jobs five years from now will have very little in common with what their jobs were five years ago.

CASE STUDY: Service Oriented Network Management and Automation

Jerome Oglesby
CTO
Information Technology Services
Deloitte Services LP

Deloitte Profile

Deloitte is one of the world's top accounting and professional services firms providing audit, tax, consulting and financial advisory services to companies and organizations.

With headquarters in the U.S., Deloitte is part of a global network of 70 firms in 142 countries under the name Deloitte Touche Tohmatsu. In fiscal 2007, the global network had $23.1 billion in total revenues, up from $20.1 billion the previous year.

The U.S. division generated $9.85 billion in revenues in 2007. Audit and enterprise risk services made up 44% of U.S. revenue, 30% from consulting, 22% from tax and 4% from financial advisory services.

There are 40,998 employees in the U.S. with 2,758 partners, 29,725 staff and 8,515 administrative staff. Deloitte has 101 U.S. offices in 92 cities and 8,108 CPAs.

Business Challenge

Our challenges were similar to those that many organizations face as they strive to meet zero defects and zero downtime.

Our Enterprise Systems Management (ESM) landscape consisted of multiple point solutions with no integration, no holistic view of the services that we provided, and no way to give the business a view into IT services and service levels.

Our toolsets were device focused and we had no automated event correlation, relying mostly human and manual correlation of events from servers, applications, and other devices. We provided the business very few metrics around our performance and our delivery of services

We needed to transform our Enterprise Systems Management capabilities. We needed to make ESM an enabler and provide more than just monitoring of events.

We needed a service-oriented platform that provided Enterprise Level Management, transparency and could operate across distributed services. We also needed tool sets that could provide visualization of discrete services, while also providing event and business level correlations.

Technical Process

To improve our operational excellence experience, we needed to develop founding principles and implement a best practice framework for operations. We adopted the ITIL framework and sought to implement systems around that framework.

To move toward a Service Oriented Network Management, we built our platforms around Enterprise Systems Management components. But we also needed sound processes to complement our ESM platform and our operations framework.

We investigated several solutions to make sure that our tools selection matched our environment and our needs.

Out-of-the-box capability was one criterion that we prioritized. Being able to incorporate out-of-the-box capabilities offered us the best opportunities to lower our Total Cost of Ownership, get immediate use of the tools, and helped ensure that the tools we selected would follow the manufacture's technology roadmap.

However, the most significant criteria was to provide service-oriented monitoring, integrated with business activity monitoring, and ITIL based tools, such as change management, configuration management and problem management.

Solution

We choose to implement a leading Enterprise Management System that integrated with our ITIL based Management Suite. The implementation included a problem management module, change management module and configuration management database.

We integrated these solutions with a custom developed Business Activity Monitoring tool that serves as our customer-facing dashboard and shows the business impact based on our key business processes. By providing the business transparency of our core services and issues, we reduced calls to our help desk.

We provide complete transparency for over 300 applications and over 50 services that are mapped to our key business processes.

Benefits

We measured success by customer satisfaction and improved service levels for our customers. Today, we have more and more web-based transactions, and synthetic transactions that provide views from a distributed standpoint.

This drove an increase in the number of proactive resolutions by 60%. Many of these events would have gone unaddressed in the past, until they became events. Our Business impacting events have been reduced by 40% as we increased our proactive resolutions. We are solving issues before they become impacting events.

With our correlation capabilities, we are able to decrease our mean-time-to-restore by 20% with graphical drill downs on communications paths, applications services and device elements that pinpoint problems that would have traditional required manual troubleshooting.

And providing business transparency has been a tremendous advantage in reducing our help desk calls by 10% as users now have the information they need when we have service impacting events. We let them know what's wrong, and when they can expect it to be restored.

Lessons Learned

We did make mistakes, but those mistakes allowed us to refine our approach to network management. First, we needed to ensure we had a true evaluation system and process for selecting an Enterprise Systems Management platform.

Second, we needed to ensure that we did not focus only on the technology. There were many things that we needed to consider, particularly our processes and our people.

Looking back, our decision to focus first on a sound operational framework such as ITIL was one of the best things we did. This allowed us to focus our efforts around tools on sound principles and methodologies.

Network management and control, it is not all about the technology. People and processes are what make or break the implementation's success.

At the end of the day, we had to be able to measure performance based on a number of criteria and also have the ability to share this effectively with the business and our customers. Our new Service Oriented Network Management process helped build trust with our customers and IT is now viewed as a trusted advisor, instead of simply advisory.One company that has made great strides towards implementing an IOC is Deloitte Services LP. In the case study entitled *Service Oriented Network Management and Automation*, Jerome Oglesby, Deloitte's CTO talks about the process that they went through to change their approach to management to where it was more automated and proactive, used more intelligent tools, had an increased focus on both applications and effective processes, and which maps applications and services to key business processes.

Rethinking MTTR for Application Delivery

The Changing Concept of MTTR

Mean Time To Repair (MTTR)—the mean or average time that it takes the IT organization to repair a problem—is a critical metric for measuring performance. The understanding of MTTR, however, is changing because as explained in the preceding section, the responsibility of the network organization is expanding to include the ongoing management of application performance. This section describes those changes and their impact on network management.

The basic, three-step process for troubleshooting is not changing:

- Problem Identification
- Problem Diagnosis
- Solution Selection and Repair

However, how these steps apply to a traditional network management task, such as fault management, differs significantly from how they apply to managing application performance.

Problem Identification

Like every component of network management, application performance management can either be done proactively or reactively. In a proactive approach, the network management organization attempts to identify and resolve problems before they impact end users. In a reactive approach, network management organizations respond to the fault once end users have been impacted.

With fault management, it's relatively easy to identify that a fault exists, since the fault often leads to a readily-noticeable outage. It is also fairly easy to set alarms indicating the failure of a component. By contrast, identifying application degradation is much more difficult. For example, as previously noted most IT organizations do not have objectives for the performance of even their key, business-critical applications, and few monitor the end-to-end performance of their applications. As a result, the issue of whether or not an application has degraded is often highly subjective.

Problem Diagnosis

In the case of fault management, the focus of diagnosis is to determine which component of the infrastructure is not working. Part of the difficulty of diagnosing the cause of an outage is that a single fault can cause a firestorm of alarms. Although one should not understate the difficulty of filtering out extraneous alarms to find the defective component, it is easier to identify the component of the infrastructure that is not functioning than it is to identify the factor that is causing an application to perform badly.

One of the reasons that it is so difficult to diagnose the cause of application degradation is that as discussed in Chapter 8, any and every component of IT could cause the application to perform badly. This includes the network, the servers, the database and the application itself. This means that unlike fault management, which tends to focus on one technology and on one organization, diagnosing the cause of application degradation crosses multiple technology and organizational boundaries. In general, most IT organizations find it difficult to solve problems that cross multiple technology and organizational boundaries.

Solution Selection and Repair

In the case of fault management, there typically is no solution selection step. In particular, once it has been determined which component has failed, the solution is obvious: fix that component.

The situation is entirely different when managing application performance because the component of IT that is causing the application to degrade may not be the component that gets fixed or replaced. For example, sometimes the way the application was written will cause the application to perform badly. However, re-writing the application may not be an option, particularly if the application was acquired from a 3rd party. In that case, the IT organization must implement a work-around to compensate for the application's faults.

In an analogous fashion the repair component of fault management differs somewhat from the repair component of application management. In the case of fault management, once you replace the defective part you fully expect the problem to be fixed. In the case of managing application performance, once you implement the chosen solution, you are less sure that the problem will go away. As a result, in some instances the IT organization has to repeat the problem diagnosis as well as the solution selection and repair processes.

Reducing MTTR requires both credible tools and an awareness of and attention to technical and non-technical factors. In many instances it can be as much a political process as a technological one.

10.0 Control

Introduction

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

- Affect the routing of traffic through the network.
- 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.
- Identify and control the traffic that enters the IT environment over the WAN.
- Provide virtualized instances of key IT resources.

Route Optimization

Route optimization was discussed in the section of chapter 7 entitled "Internet-Based Application Delivery Optimization" in the contest of an application delivery service provided by an MSP. However, many of the same challenges that impact the performance of the Internet also impact the performance of an enterprise IP network.

As a result, a few years ago IT organizations began to deploy route optimization in enterprise IP networks. As previously noted, the goal of route optimization is to make more intelligent decisions relative to how traffic is routed through an IP network. Route optimization achieves this goal by implementing a four-step process. Those steps are:

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 must 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.

SSL VPN Gateways

The SSL protocol[25] is becoming increasingly popular as a means of providing secure Web-based communications to a variety of users including an organization's mobile employees. Unlike IPSec which functions at the network layer, SSL functions at the application layer and uses encryption and authentication as a means of enabling secure communications

[25] IPSec vs. SSL: Why Choose?, http://www.securitytechnet.com/resource/rsc-center/vendor-wp/openreach/IPSec_vs_SSL.pdf

between two devices, which typically are a web browser on the user's PC or laptop and an SSL VPN gateway that is deployed in a data center location.

SSL provides flexibility in allowing enterprises to define the level of security that best meets their needs. Configuration choices include:

- Encryption: 40-bit or 128-bit RC4 encryption
- Authentication: Username and password (such as RADIUS), username and token + pin (such as RSA SecurID), or X.509 digital certificates (such as Entrust or VeriSign)

All common browsers such as Internet Explorer include SSL support by default, but not all applications do. This necessitates either upgrading existing systems to support SSL or deploying an SSL VPN gateway in the data center. One of the purposes of an SSL VPN gateway is to communicate directly with both the user's browser and the target applications and enable communications between the two. Another purpose of the SSL VPN gateway is to control both access and actions based on the user and the endpoint device.

Among the criteria IT organizations should use when choosing an SSL VPN gateway, the gateway should be:

- Easy to deploy, administer and use
- Low cost over the lifecycle of the product
- Transparent
- Capable of supporting non-traditional devices; e.g., smartphones and PDAs
- Able to check the client's security configuration
- Able to provide access to both data and the appropriate applications
- Highly scalable
- Capable of supporting granular authorization policies
- Able to support performance enhancing functionality such as caching and compression
- Capable of providing sophisticated reporting

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 where there are one or more delay-sensitive, business-critical applications. Two examples of this type of application that have been discussed previously in this handbook are VoIP and the Sales and Distribution (SD) module of SAP.

> ***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 as discussed in chapter 8 many applications share the same port, or even hop between ports.

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 such as TCP Rate Control tell the remote servers how fast they can send content providing true bi-directional management.

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

Discovering the Application

Application discovery must occur at Layer 7. Information gathered at Layer 4 or lower allows a network manager to assign priority to their Web traffic lower than that of 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 over 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

As many applications share the 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 has identified 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 256 Kbps 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 latter approach is that it frees the IT organization from having to know how many simultaneous sessions will take place.

Many IT organizations implement QoS via queuing functionality found in their routers. Implementing QoS based on aggregate queues and class of service is often sufficient to prioritize applications. However, when those queues get oversubscribed (e.g. with voice services), degradation can occur across all connections. As a result, "access control' or "per call" QoS is sometimes required to establish acceptable quality. Another option is to implement QoS by deploying MPLS based services.

Next Generation WAN Firewall

Current Generation Firewalls

The first generation of firewalls was referred to as packet filters. These devices functioned by inspecting packets to see if the packet matched the packet filter's set of rules. Packet filters acted on each individual packet (i.e., 5-tuple consisting of the source and destination addresses, the protocol and the port numbers) and did not pay any attention to whether or not a packet was part of an existing stream or flow of traffic.

Today most firewalls are based on stateful inspection. According to Wikipedia[26], "A stateful firewall is able to hold in memory significant attributes of each connection, from start to finish. These attributes, which are collectively known as the state of the connection, may include such details as the IP addresses and ports involved in the connection and the sequence numbers of the packets traversing the connection. The most CPU intensive checking is performed at the time of setup of the connection. All packets after that (for that session) are processed rapidly because it is simple and fast to determine whether it belongs to an existing, pre-screened session. Once the session has ended, its entry in the state-table is discarded."

One reason that traditional firewalls focus on the packet header is that firewall platforms generally have limited processing capacity due to architectures based on software that runs on an industry standard CPU. A recent enhancement of the current generation firewall has been the addition of some limited forms of application level attack protection. For example, some current generation firewalls have been augmented with IPS/IDS functionality that uses deep packet inspection to screen suspicious-looking traffic for attack signatures or viruses. However, limitations in processing power of current generation firewalls prevents deep packet inspection from being applied to more than a small minority of the packets traversing the device.

The Use of Well-Known Ports, Registered Ports, and Dynamic Ports

Chapter 8 pointed out that the ports numbered from 0 to 1023 are reserved for privileged system-level services and are designated as *well-known ports*. As a reminder, a well-known port serves as a contact point for a client to access a particular service over the network. For example, port 80 is the well-known port for HTTP data exchange and port 443 is the well-known port for secure HTTP exchanges via HTTPS.

Port numbers in the range 1024 to 49151 are reserved for Registered Ports that are statically assigned to user-level applications and processes. For example, SIP uses ports 5059-5061. A number of applications do not use static port assignments, but select a port dynamically as part of the session initiation process. Port numbers between 49152 and 65535 are reserved for Dynamic Ports, which are sometimes referred to as Private Ports. One of the primary reasons that stateful inspection was added to traditional firewalls was to track the sessions of whitelist applications that use dynamic ports. The firewall observes the dynamically selected port number, opens the required port at the beginning of the session, and then closes the port at the end of the session.

Most current generation firewalls make two fundamental assumptions, both of which are flawed. The first assumption is that the information contained in the first packet in a connection is sufficient to identify the application and the functions being performed by the application. In many cases, it takes a number of packets to make this identification because the application end points can negotiate a change in port number or perform a range of functions over a single connection.

[26] http://en.wikipedia.org/wiki/Stateful_firewall

The second assumption is that the TCP and UDP well-known and registered port numbers are always used as specified by IANA. Unfortunately, while that may well have been the case twenty years ago it is often not the case today. As pointed out in chapter 8, some applications have been designed with the ability to hop between ports.

Another blind spot of current generation firewalls is for HTTP traffic secured with SSL (HTTPS). HTTPS is normally assigned to well-known TCP port 443. Because the payload of these packets is encrypted with SSL, the traditional firewall cannot use deep packet inspection to determine if the traffic either poses a threat or violates enterprise policies for network usage. These two blind spots are growing in importance because they are being exploited with increasing frequency by application-based intrusions and policy violations.

A Next Generation Firewall

Firewalls are typically placed at a point where all WAN access for a given site coalesces. This is the logical place for a policy and security control point for the WAN. Unfortunately due to performance limitations, IT organizations have resorted to implementing myriad firewall helpers[27].

It is understandable that IT organizations have deployed workarounds to attempt to compensate for the limitations of traditional firewalls. This approach, however, has serious limitations including the fact that the firewall helpers often do not see all of the traffic, and that deployment of multiple security appliances significantly drives up the operational costs and complexity.

In order for the firewall to avoid these limitations and reestablish itself as the logical policy and security control point for the WAN, we now need a next generation firewall with the following attributes:

Application Identification

The firewall must be able use deep packet inspection to look beyond the IP header 5-tuple into the payload of the packet to find application identifiers. Since there is no standard way of identifying applications, there needs to be an extensive library of application signatures developed that includes identifiers for all commonly used enterprise applications, recreational applications, and Internet applications. The library needs to be easily extensible to include signatures of new applications and custom applications. Application identification will eliminate the port 80 blind spot and allow the tracking of port-hopping applications.

Extended Stateful Inspection

By tracking application sessions beyond the point where dynamic ports are selected, the firewall will have the ability to support the detection of application-level anomalies that signify intrusions or policy violations.

SSL Decryption/Re-encryption

The firewall will need the ability to decrypt SSL-encrypted payloads to look for application identifiers/signatures. Once this inspection is performed and policies applied, allowed traffic would be re-encrypted before being forwarded to its destination. SSL proxy functionality, together with application identification, will eliminate the port 443 blind spot.

[27] Now Might Be a Good Time to Fire Your Firewall, http://ziffdavisitlink.leveragesoftware.com/blog_post_view.aspx?BlogPostID=603398f2b87548ef9d51d35744dcdda4

Control

Traditional firewalls work on a simple deny/allow model. In this model, everyone can access an application that is deemed to be *good*, and nobody can access an application that is deemed to be *bad*. This model had more validity at a time when applications were monolithic in design and before the Internet made a wide variety of applications available. Today's reality is that an application labeled *bad* for one organization might well be *good* for another. On an even more granular level, an application that might be *bad* for one part of an organization might be *good* for other parts of the organization. Going even further, given today's complex applications, a component of an application might be *bad* for one part of an organization but that same component might well be *good* for other parts of the organization.

What is needed therefore is not a simple deny/allow model, but a model that allows IT organizations to set granular levels of control to allow the good aspects of an application to be accessed by the appropriate employees while blocking all access to the bad aspects of an application.

Multi-gigabit Throughput

In order to be deployed in-line as an internal firewall on the LAN or as an Internet firewall for high speed access lines, the next generation firewall will need to perform the above functions at multi-gigabit speeds. Application Identification and SSL processing at these speeds requires a firewall architecture that is based on special-purpose programmable hardware rather on than industry standard general-purpose processors. Firewall programmability continues to grow in importance with the number of new vulnerabilities cataloged by CERT hovering in the vicinity of 8,000/year.

11.0 Pulling it Together

The goal of this chapter is to apply the application delivery framework to the concept of an application delivery network as well as to a particular application – VoIP.

Application Delivery Network Characterized

As described in the section of Chapter 6 entitled "Application Delivery Network Defined", packet delivery and the corresponding optimization techniques correspond to functionality focused on the packet payload and the lower four layers of the Internet protocol suite. The packet delivery network is quite effective in terms of providing the basic benefits of WAN Optimization as described in the introduction. However, just focusing at the packet layer is limited. In particular, the packet delivery network has limited knowledge of users and content and can not identify malicious traffic. In addition, the packet delivery network cannot leverage the application headers that contain a wealth of valuable information that can be leveraged to control the performance and security of applications in order to meet the evolving business challenges.

Application Delivery Network (ADN) is an emerging industry phrase that refers to implementing application delivery technologies that reside at or above layer 4 in the OSI stack. For example, the ADN employs deep packet inspection (DPI) to parse application headers and content and uses this information to further optimize performance monitoring, application acceleration, the management of application security and WAN access, and control of how application utilize WAN resources. Therefore, in the context of the Application Delivery Framework, the ADN provides enhanced functionality in all three areas of implementation: Optimization, Management, and Control.

Optimization

DPI enables application delivery solutions to recognize applications based on signatures in the application headers. Application recognition enables application-specific optimization techniques that can significantly minimize bandwidth consumption and mitigate the effects of WAN latency. DPI also makes it possible to distinguish between business critical Web-based applications (e.g., webified enterprise applications, as well as specific SOA and Web 2.0 applications) and other traffic that relies on HTTP. In addition, DPI makes it possible to sub-classify the network flows generated by complex enterprise applications, such as SAP and Oracle, allowing the critical operations and transactions to be afforded the highest priority access to WAN bandwidth.

Management

ADN functionality as described above also provides the visibility that allows the performance of each application and of each application user to be monitored in a highly granular fashion. This functionality provides IT organizations with the capability to identify performance issues before they impact end users. However, while identifying performance issues before they impact end users is highly desirous, that capability alone is not sufficient to ensure acceptable application performance. In particular, the ADN must also be able to control the applications that are contributing to the performance issues.

Control

The control component of application delivery focuses on performance and security. In particular, ADN functionality allows the IT organization to implement highly granular policies governing QoS and bandwidth allocation and to enforce policies governing authorized user access to specific applications. ADN DPI technology that can scan application headers and the packet payloads for application signatures provides another layer of security to the network. DPI can also be used to scan for viruses and other malware that may be contained in Web content or may be attempting to piggyback over enterprise application flows. By maintaining logs of user access to applications and by logging the results of security scans, the ADN provides another source of audit information that be used to document compliance with various privacy/ integrity regulations, such as HIPAA and PCI.

Companies of virtually all sizes and industries are under increasing pressure to demonstrate compliance with government regulations and industry standards to ensure privacy and data integrity. In addition to that pressure, in difficult economic times the occurrence of cyber hacking increases dramatically. For example, a number of recent articles have commented on the great increase in the amount of malware[28 29]. As a result, an effective ADN must support security functionality beyond what was described in the preceding paragraph. An example of the requisite additional security functionality is the ability to protect naïve users from clicking on what they believe is a legitimate URL only to introduce some form of malware into the company's IT environment. Providing this protection is complex in part because so many users access Internet based content remotely and hence are not protected by a powerful enterprise firewall. To respond to these challenges, an effective ADN must be able to use a cloud computing approach to check for and evaluate the validity of a URL before establishing a connection to the site.

Another example of the requisite additional security functionality is that an effective ADN must support content filtering to prevent data leaks[30]. An effective ADN must also support intrusion detection and intrusion protection functionality. An intrusion detection system (IDS) passively watches packets transiting the network and sets off an alarm if it finds anything suspicious. A typical intrusion protection system (IPS) has all of the features of an IDS, and in addition it can stop malicious traffic from entering the network.

Voice over IP (VoIP)

VoIP poses particular challenges for two primary reasons: the new and different protocols that VoIP requires, and its stringent availability and performance requirements. For instance, there are many different coding algorithms (codecs or codices) available to handle the task of converting a conversation from analog to digital and back to analog again, and both sides of the call must use the same codec. The negotiation to ensure this is handled by another set of protocols involved with call setup, such as H.323, the Media Gateway Control Protocol (MGCP), Cisco's Skinny Client Control Protocol (SCCP), and increasingly, the Session Initiation Protocol (SIP).

A critical concern is that because of its real-time nature, VoIP almost universally relies on UDP rather than TCP. This poses particular problems for voice management, because unlike TCP, UDP does not offer any feedback information about whether or not packets that have been sent have been received. In addition, UDP does not have any flow control mechanisms to limit transmissions in the presence of congestion.

[28] IM Malware Attacks Increase, http://www.scmagazineus.com/IM-malware-attacks-increase-report/article/109663/

[29] New report predicts massive increase in malware and phishing in 2009, http://www.chutneytech.com/new-report-predicts-massive-increase-in-malware-and-phishing-in-2009/

[30] Improve Data Protection Processes with Content Discovery, Monitoring and Filtering, http://adventuresinsecurity.com/Papers/CMF.pdf

In addition to these protocol challenges, VoIP is extremely sensitive to a number of network parameters that have far less affect on transactional applications. For example, users expect 100% availability and immediate dial tone. In addition, fairly low levels of packet loss can severely impact voice quality. End-to-end delay is also critical. At about 150 ms, voice quality will likely begin to degrade, and beyond 250 ms it will almost certainly be unusable. These are levels of latency that are barely noticeable on most transactional applications. Jitter can also negatively impact voice quality. Many network management solutions don't measure jitter because while this is a key parameter for VoIP, it has virtually no impact on the typical data application.

Planning

If VoIP deployment is not adequately planned, serious disruptions to voice and data communications are possible, together with a prolonged period of transition from the legacy system to VoIP. Planning for VoIP involves the following activities:

VoIP Characterization

Working with the chosen VoIP vendor(s), the planning team should characterize the VoIP traffic expected to flow over the LAN and WAN. The characterization should include setting target thresholds for delay, jitter and packet loss, as well as establishing estimates of the traffic loading on the LAN and WAN. Impairment thresholds will be somewhat dependent on the functionality and configuration of VoIP endpoints, including codecs, de-jitter buffering design, QoS classification setting(s), and packet loss concealment (PLC) capability. Establishing a target for inter-site and intra-site availability and latency is another important aspect of this initial planning.

Network Assessment

A careful analysis of the readiness of the existing network for voice communications is required. The VoIP characterization and the chosen model for QoS implementation should be used to provide detailed guidelines for this analysis. In most cases, ensuring the desired level of availability and performances requires some degree of network redesign and modification. An adequate network assessment involves a good understanding of the baseline performance and capacity of the existing network, as well as some form of network modeling for estimating the performance improvement expected from planned enhancements or additional capacity to the network.

VoIP Impact Analysis

VoIP deployment can potentially impact mission critical interactive applications, especially in the WAN where voice and data traffic may need to share narrow band links. Modeling the impact of VoIP traffic on the performance of mission critical applications may require further adjustments to the network design and/or the QoS implementation.

Deploy Network Enhancements

The modifications of the network identified in the network assessment are implemented and measurements are made to verify the expected improvements in network performance.

Simulate VoIP Performance

At this point it may be advisable to employ test applications or equipment that emulates voice traffic to verify network performance under simulated load conditions. This sort of simulated traffic testing can also verify QoS functionality and the effectiveness of the existing tools for monitoring and troubleshooting VoIP traffic.

Pilot Deployment

In general, the transition to VoIP is a significant change to the network environment that justifies a pilot deployment followed by a phased production deployment, possibly on a department-by-department or site-by-site basis.

Management Readiness

Ensuring that management processes will be ready for VoIP deployment is another aspect of the planning process. In additional to addressing the organization issues of data communications vis-à-vis telecommunications, an assessment of management readiness should identify any requirement for additional staff members, technical training, or management tools that are compatible with the chosen VoIP solution(s). Planning for the management of VoIP may also include consideration of outsourcing some tasks to an MSP.

Optimization

Chapter 6 describes a number of techniques that can be used to accelerate or optimize application performance; however, given that voice is a real-time application, most of those techniques do not apply. One optimization technique that does apply is IP header compression, the goal of which is to reduce overhead on voice payloads. Knowing that packet loss can have a profound negative impact on the performance of VoIP, forward error correction is another optimization technique that can improve the quality of VoIP.

Management

The combination of the user expectation of 100% uptime in voice, its sensitivity to network conditions, and its cross-domain organizational demands make VoIP a great example of the need for an integrated approach to network management, both organizationally and technically. IT organizations should avoid the all-too-common fragmented approach to network management, which generally results from the incremental adoption of point solutions to address new problems on an ad hoc basis.

Instead, IT organizations should look for an integrated solution that relates voice-specific metrics such as MOS values to the underlying network behavior that influences them, and vice-versa. To deliver this integration, a voice management solution must, at a minimum, deliver information from three sources: call signaling, NetFlow and SNMP, and -- even more important -- relate them one to another.

Call signaling (or call setup), is handled by one of a number of different protocols: either one of those standards noted above (H.323, MGCP, SCCP, or SIP), or a proprietary protocol. The ability of a solution to monitor call setup is critical. During call setup, the two ends of the conversation negotiate a common codec, establish the channels that will be used for transmitting and receiving, and generate a number of status codes. This information can be used to derive important measurements like delay to dial tone. Without this data, IT organizations won't know what went wrong if users, for example, start complaining that they can't get a dial tone.

Being able to monitor call signaling also implies being able to receive and integrate data from an IP PBX. This is particularly important for monitoring voice-specific metrics such as MOS. In order to relate this VoIP-specific data to network conditions requires network-specific data. Both NetFlow and RMON-2 data can give IT organizations insight into the protocol and class-of-service composition of the traffic. And, given the increasingly meshed nature of VoIP systems, the ability of NetFlow or RMON-2 to deliver data from many points in the network can be critical in managing VoIP.

SNMP is also a requirement. Not only does it deliver data on the health of the devices, but it can also be used to access data from other sources. For example, in a Cisco-based network, SNMP can give information from both the Cisco IP SLA and the Cisco Class-Based QoS (CBQoS) MIB. Cisco IP SLA generates synthetic transactions that can be used to emulate voice traffic across important links and derive metrics critical to understanding voice quality. The CBQoS MIB provides information about the class-based queuing mechanism in a Cisco router, enabling IT organizations to ensure that their critical traffic is being treated appropriately when bandwidth is in short supply. However, in order for IT organizations to get real time scores for calls in progress, advanced monitoring tools that measure the delay, jitter, loss and MOS for actual voice calls is required.

Once all of this data has been collected IT organizations must have a way of integrating it all into a useful overview of voice and network performance. From the management console to the reports the solution generates, what is required is a holistic overview that can relate voice quality to network performance, and vice-versa. For example, the IT organization should be able to detect that MOS values are dropping on the link between HQ and the Los Angeles office and bring up management data from the appropriate devices to check the traffic composition on the link. The IT organization should be able to use this data to answer questions such as whether the link is being flooded by packets from a scheduled backup or rogue traffic from an illicit application, and what other critical applications are being affected.

Control

As previously noted, one aspect of controlling the simultaneous delivery of VoIP and data applications involves managing the different classes of VoIP and data traffic using rate-limiting QoS features and call admission levels for voice calls. Because of the dynamic nature of the enterprise application environment, frequent adjustments may be necessary to ensure that the goals for application performance continue to be met.

Another aspect of controlling the delivery of VoIP and data applications involves extending the security model to cover the converged network as well as the continual monitoring of the network for changes in traffic patterns that may have an impact on either VoIP or mission critical data applications. The basic goal in securing the converged network is to avoid the possibility of losing both data and voice communications due to a security event by preventing intrusions from spreading from the data environment to the voice environment and vice versa. This involves the logical isolation of VoIP and data traffic using separate virtual LANs (VLANs), plus deploying internal firewalls to sADCguard IP PBXs and voice servers. Authenticating both VoIP endpoints and users to prevent intruders from using rogue devices to gain network access can also enhance the security of the voice environment.

12.0 Conclusion

For the foreseeable future, the importance of application delivery is much more likely to increase than it is to decrease. This also means that for the foreseeable future the impact of the factors making application delivery difficult 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 handbook identified a number of conclusions that IT organizations can use when formulating their approaches to ensuring acceptable application delivery. Those conclusions are:

- In the vast majority of instances when a key business application is degrading, the end user, not the IT organization, first notices the degradation.
- In situations in which the end user is typically the first to notice application degradation, the reputation of the IT organization is tarnished.
- Application delivery must have a top-down approach, with a focus on application performance as seen by the user of the application.

Successful application delivery requires the integration of:

- Planning
- Network and application optimization
- Management and
- Control.

- The complexity associated with application delivery will increase over the next few years.
- If you work in IT, you either develop applications or you deliver applications.
- Senior IT management needs to ensure that their organization evolves to where it looks at application delivery holistically and not just as an increasing number of stove-piped functions.
- The application delivery solutions that IT organizations deploy must be able to scale to support clearly discernable emerging requirements.
- CIOs must drive the vision of an Integrated Operations Centre (IOC).
- The organizational model for the IT function needs to be similar to the organizational model for the enterprise.
- In many instances, there is little overlap between a CIO's priorities and what the IT organization believes those priorities should be.
- 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.
- A relatively small increase in network delay can result a significant increase in application delay.
- Application delivery is more complex than merely optimizing the performance of all applications.
- Successful application delivery requires that IT organizations identify the applications running on the network and ensure the acceptable performance of the applications relevant to the business, while controlling or eliminating irrelevant applications.
- The "webification" of application introduces chatty protocols into the network. In addition, some or 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 accesses an application they are accessing it over the WAN instead of the LAN.
- 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.
- The CYA approach to application delivery focuses on deflecting fault when the application performs badly. The goal of the CIO approach is to rapidly identify and fix the problem without assigning blame.
- As the complexity of the environment increases, the number or sources of delay increases and the probability of application degradation increases in a non-linear way.
- Just as WAN performance impacts *n*-tier applications more than monolithic applications, WAN performance impacts Web services-based applications significantly more than WAN performance impacts *n*-tier applications.
- Many IT professionals view the phrase Web 2.0 as either just marketing hype devoid of any meaning or they associate it exclusively with social networking sites such as MySpace.
- Emerging application architectures (SOA, RIA, Web 2.0) have already begun to impact IT organizations and this impact will increase over the next year.
- In addition to a services focus, Web 2.0 characteristics include featuring content that is dynamic, rich and in many cases, user created.
- The existing generation of network and application optimization solutions does not deal with a key requirement of Web 2.0 applications: the need to massively scale server performance.
- Server and storage virtualization have crossed the chasm and are now mainstream technologies.
- The deployment of desktop virtualization lags behind that of server and storage virtualization.

- Deployment of virtualized servers can result in significant cost savings.
- Deployment of virtualized servers can result in significant management, performance and security issues.
- It is extremely difficult to make effective network and application-design decisions if the IT organization does not have well-understood and enforced 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 vast majority of IT organizations see significant value from a tool that can be used to test application performance throughout the application lifecycle.
- In the vast majority of cases, a development, testing or monitoring tool that is unduly complex is of no use to an IT organization.
- The application-delivery function needs to be involved early in the applications development cycle.
- 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 they are trying to solve.
- In order 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 it will be deployed.
- Small amounts of packet loss can significantly reduce the maximum throughput of a single TCP session.
- With a 1% packet loss and a round trip time of 50 ms. or greater, the maximum throughput is roughly 3 megabits per second no matter how large the WAN link is.
- An ADC provides more sophisticated functionality than a SLB does.
- A comprehensive strategy for optimizing application delivery needs to address both optimization over the Internet and optimization over private WAN services.
- TCP throughput on a single session decreases as either the round trip time or the packet loss increases.
- 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.
- In roughly twenty-five percent of companies, there is an adversarial relationship between the application delivery organization and the network organization.
- 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.
- Identifying the root cause of application degradation is significantly more difficult than identifying the root cause of a network outage.

- 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.
- Lack of visibility into the traffic that transits port 80 is a major vulnerability for IT organizations.
- End-to-end visibility refers to the ability of the IT organization to examine every component of IT that impacts communications once users hit ENTER or click a mouse button when they receive responses from an application.
- Application management should focus directly on the application and not just on factors that have the potential to influence application performance.
- Most IT organizations 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.
- Traditional methods of monitoring traffic on a physical LAN switch cannot be used to monitor the traffic that goes between virtual machines.
- In over a quarter of organizations, the NOC does not meet the organization's current needs.
- In the majority of cases, the NOC tends to work on a reactive – rather than proactive – basis, identifying a problem only after it impacts end users.
- Just under half of NOCs are organized around functional silos.
- A majority of NOCs use many management tools that are not well integrated.
- NOC personnel spend the greatest amount of time on applications, and that is a relatively new phenomenon.
- NOC personnel spend an appreciable amount of their time supporting a broad range of IT functionality.
- The NOC is almost as likely to monitor performance as it is to monitor availability.
- There is a lot of interest in ITIL, but it is too soon to determine how much impact its use will have.
- In the vast majority of instances, the end-user assumes that the network is the source of application degradation.
- In the majority of instances, the NOC gets involved in problem resolution.
- The top driver of change in the NOC is the requirement to place greater emphasis on ensuring acceptable performance for key applications.
- The lack of management vision and the NOC's existing processes are almost as much of a barrier to change as are the lack of personnel resources and funding.
- Reducing MTTR requires both credible tools and an awareness of and attention to technical and non-technical factors. In many instances it can be as much a political process as a technological one.
- 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.

Unified Application Delivery

An increasingly mobile workforce, virtualization, cloud-computing—it's clear that "static" is no longer a viable solution to delivering your business-critical applications.

F5® provides the next generation infrastructure that is essential and effective in delivering business-critical applications. It is a framework that offers simple and efficient mechanisms to integrate new technology into your business without having to completely re-architect existing systems. This new infrastructure is called unified application delivery and it offers customers a solution that is adaptable, intelligent, and consolidated.

Adaptable

Unified application delivery must be adaptable and agile to ensure the success of existing and future infrastructure investments. It is not a pre-defined infrastructure, but is a framework on which to build your unique solution—a common method for applying the technology and services you need to run your business. Unified application delivery enables you to define a repeatable, relevant application delivery process in a manner that is modular and provides granular control over what services are provided for what applications—limiting the risk to existing business applications. This enables you to determine when you want this new service to be used, as well as the order in which services are applied. This is partly due to adaptability and partly due to application intelligence.

Intelligent

Unlike the static plumbing that characterized common business infrastructures for the last several decades, unified application delivery is intimately aware of the applications and data you use as well as the networks, devices, and people accessing them—from any device and any location. This information is used to dynamically and intelligently adjust the delivery of those applications and data over the network.

Intelligence within the solution not only understands the applications and data, but also helps you meet your business objectives. In the event of a disaster or simply an exhaustion of local resources, an intelligent solution can dynamically redirect users to backup data centers or trigger dynamic resource creation via virtual service managers or cloud computing providers, constantly monitoring usage to ensure the most cost-effective balance. Knowing when and in what order to apply the services available is critical to successful delivery.

Consolidated

Consolidation is the final step when integrating new services; it makes services part of your standard solution set, available for any and all applications that need them. Consolidation provides efficiency in both the delivery of those applications and the management of the services.

Consolidation results in a single point of implementation, at which the application needs to be mediated between the client and the server, where all necessary services can be provided efficiently and simultaneously. The alternative is multiple points of interaction that add latency and overhead that increase with each new service added to the infrastructure. While point solutions can easily interoperate with the solution, the continued consolidation of services on integrated platforms over time helps reduce the impact of service implementation, providing a single point of control.

F5 Unified Application Delivery Networking

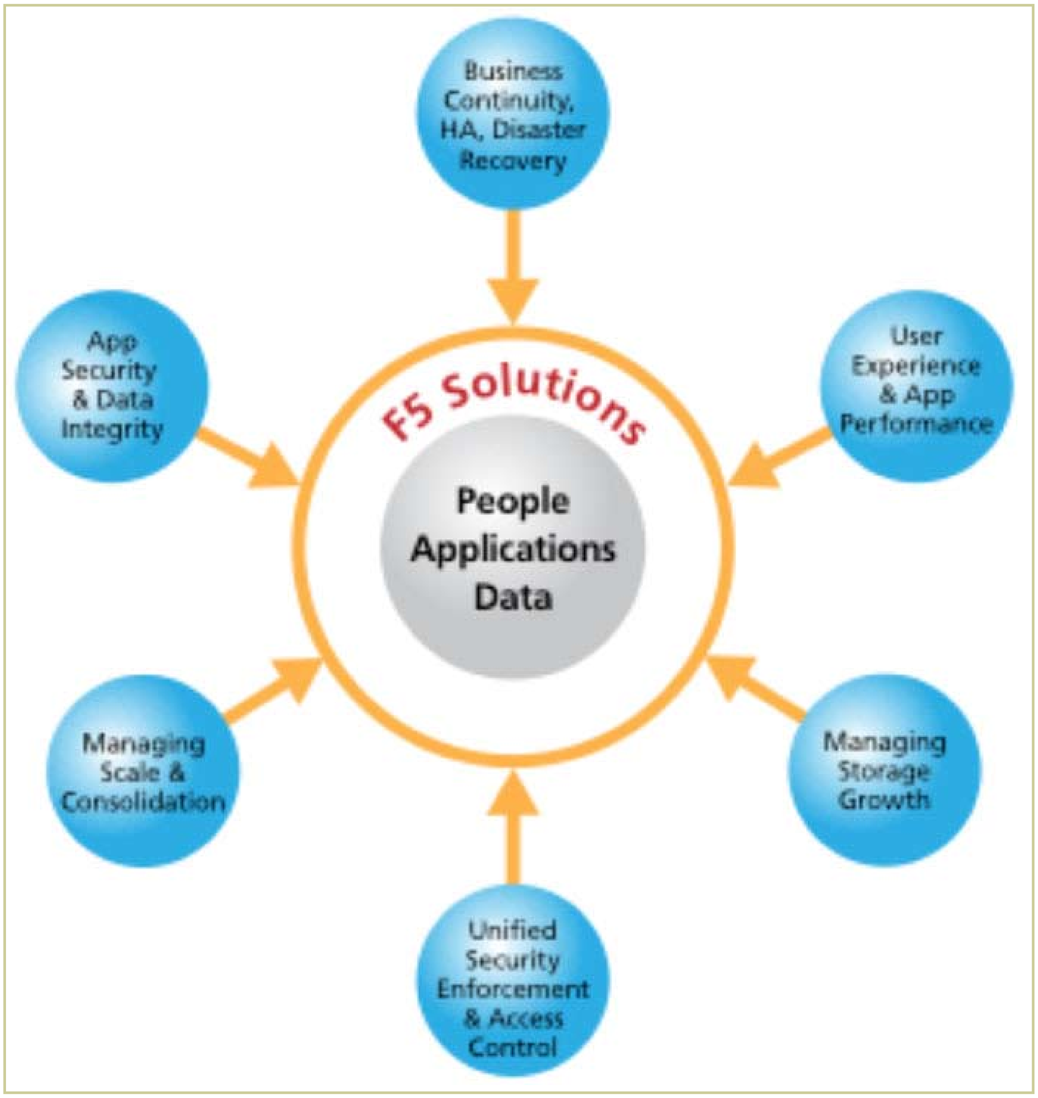
F5 has built an entire suite of Application Delivery Networking products and unified them into an adaptable, intelligent, and consolidated services solution that provides customers with the tools to build the next generation infrastructure.

Starting with the award-winning F5 BIG-IP™ Local Traffic Manager™ system to the industry-leading F5 ARX® file virtualization solution, F5 products provide an end-to-end framework of application services to ensure that users can access applications and data anytime, from anywhere, with any device, and that those applications and data will always be fast, secure, and available. With built-in, consolidated services like global distribution, link determination, load-balancing, SSL termination, caching, compression, secure access, and application security, F5 provides a comprehensive, integrated toolset for the implementation of an enterprise-wide services framework.

The modularity of the F5 TMOS™ unified software platform and the F5 VIPRION™ hardware chassis provide the ability to easily add features and functionality as well as raw performance without impacting existing services, enabling the solution to grow and change whenever your business needs change. This ensures that the solution you deploy today can be relied on to adapt to the challenges of tomorrow.

F5 iRules™ and iControl®, a network-side scripting capability and open API respectively, provide F5's Unified Application Delivery Network the intelligence to dynamically interact and adapt to application demands as well as integrate directly with applications or non-consolidated services within your infrastructure. Together these innovative tools provide the flexibility required to enable IT agility today and in the future.

F5 delivers the benefits of a decade of experience and knowledge in delivering applications over the network. F5 was the first application delivery provider to work with major software vendors (like Adobe, BEA, IBM, Microsoft, Oracle, SAP, and VMware) to build vendor-validated deployment guides and solutions for integrating F5 technology with your business-critical applications. F5 was also the first to take those deployment guides and turn them


Business Continuity, HA, Disaster Recovery
App Security & Data Integrity
User Experience & App Performance
F5 Solutions
People
Applications
Data
Managing Scale & Consolidation
Managing Storage Growth
Unified Security Enforcement & Access Control

into pre-configured application templates that can simply be "turned-on" —further reducing the costs of implementation, management, and complexity.

Because the F5 solutions are adaptable, intelligent, and consolidated, they provide a complete understanding of the applications and data they deliver and have the services and tools to dynamically optimize their delivery in any context—whether the network, user, device, or application changes. F5 led the way in the innovation from load balancers to Advanced Application Delivery Controllers and continues to innovate and adapt its technology solutions to the new applications and architecture models you need to run your business, ensuring that you can be as agile and fluid as the technology you use.

Conclusion

As you look to the technology challenges ahead of you—mobile workforce, virtualization, cloud-computing, and beyond—it is clear that "static" is no longer a viable solution for delivering your business-critical applications. Users are constantly moving between devices and networks—often public networks outside of your control. Increasingly, applications and data are now just as mobile and reside in or on networks you cannot completely control. Your infrastructure must be as dynamic, adaptable, and intelligent as the users of the applications and data—and the applications and data themselves. At the same time, you don't have an unlimited budget to constantly and manually re-architect and manage a collection of point solutions.

The solution is a Unified Application Delivery Network that provides a solid foundation on which to build your application solutions. The solution exists. Explore it. Deploy it. And run your business with it.

F5 Networks, Inc.
http://www.f5.com/