The 2010 Handbook of

Application Delivery

The Emergence of the Application Delivery 2.0 Era

By Dr. Jim Metzler

Sponsored in part by:

ipanema Technologies

Produced by

Webtorials

www.webtorials.com

Table of Contents

1.0 EXECUTIVE SUMMARY 3
2.0 INTRODUCTION 12
3.0 APPLICATION DELIVERY CHALLENGES 16
4.0 VIRTUALIZATION 23
5.0 CLOUD COMPUTING 34
6.0 A REALITY CHECK 43
7.0 PLANNING 50
8.0 NETWORK AND APPLICATION OPTIMIZATION 58
9.0 MANAGED SERVICE PROVIDERS 81
10.0 MANAGEMENT 93
11.0 CONTROL 102
12.0 CONCLUSION 108
IPANEMA 111

1.0 Executive Summary

Introduction

While ensuring acceptable application delivery has always been important, it historically was not a top of mind issue for the majority of IT organizations. That changed several years ago when IT organizations began to develop a concerted focus on it. Throughout this handbook, the application delivery challenges and solutions of this era will be referred to as Application Delivery 1.0.

At the same time that many IT organizations are still in the process of implementing solutions that respond to the challenges of the Application Delivery 1.0 era, a new generation of challenges is emerging. These challenges are driven in large part by the:

- Emergence of a sophisticated mobile workforce
- Shifting emphasis and growing sophistication of cyber crime
- Adoption of varying forms of virtualization
- Adoption of cloud computing

Throughout this handbook, the emerging generation of application delivery challenges and solutions will be referred to as Application Delivery 2.0.

As we enter the Application Delivery 2.0 era, leading edge IT organizations must develop plans for implementing an application delivery function that can respond to both the existing and the emerging challenges. The development and implementation of those plans will not be easy, in part because the plans must cross myriad organizational boundaries and involve rapidly changing technologies never before developed by vendors, nor planned, designed, implemented and managed by IT organizations in a holistic fashion. The primary goal of this handbook is to help IT organizations plan for that transformation.

Application Delivery Challenges

The following challenges are associated with the Application Delivery 1.0 era:

- Limited focus on application development
- Chatty protocols and applications
- The Webification of applications
- Security vulnerabilities
- Server consolidation

- Data center consolidation and single hosting
- Distributed employees
- Distributed applications
- Complexity

All of the challenges listed above continue to impact IT organizations. In the Application Delivery 2.0 era, however, some of the challenges have morphed. For example, supporting distributed employees has morphed to where a large percentage of those employees are now mobile. In addition, the security vulnerabilities that were a challenge in the Application Delivery 1.0 era, are now even more of a challenge in the Application Delivery 2.0 era given the shifting emphasis and growing sophistication of cyber crime. Some of the other challenges that are associated with the Application Delivery 2.0 era include:

- Service Oriented Architectures (SOA) with Web Services
- Web 2.0 and Rich Internet Applications
- The increased focus on services
- Internal Service Level Agreements (SLAs)

Virtualization

Relative to the Application Delivery 2.0 era, virtualization is a double-edged sword as it both presents challenges and solutions. Of all of the myriad forms of virtualization, server virtualization receives the most attention. Some of the challenges associated with server virtualization include:

- Contentious management of the vSwitch
- Breakdown of network design and management tools
- Limited visibility into VM-to-VM traffic
- Poor management scalability
- Multiple hypervisors
- Inconsistent network policy enforcement
- Manual network reconfiguration to support VM migration
- Over subscription of server resources
- Layer 2 network support for VM migration
- Storage support for virtual servers and VM migration
- Complex troubleshooting on a per-VM basis

At the present time, there is no overarching solution for the comprehensive management of a computing environment composed of virtualized servers, storage, and networks. Listed below are some the key developments that can help IT departments meet the challenges of virtualization.

- Dynamic infrastructure management
- Virtualized performance and fault management
- Distributed virtual switching
- Edge virtual bridges
- Orchestration and provisioning

While it does not receive as much attention as server virtualization does, there is a growing interest in the part of IT organizations to deploy desktop virtualization. The two fundamental forms of desktop virtualization are:

- Server-side application/desktop virtualization
- Client-side application/desktop virtualization

With server-side virtualization, the client device plays the familiar role of a terminal accessing an application or desktop hosted on a central presentation server. There are two primary approaches to server-side application/desktop virtualization. They are:

- Server Based Computing (SBC)
- Virtual Desktop Infrastructure (VDI)

Client-side application virtualization is based on a model in which applications are streamed on-demand from central servers to client devices. On the client-side, streamed applications are isolated from the rest of the client system by an abstraction layer inserted between the application and the local operating system.

One of the primary challenges that are associated with implementing desktop virtualization is achieving an acceptable user experience for client-to-server connections over a WAN. For example, VDI requires at least 200 Kbps of bandwidth per simultaneous user and the minimum peak bandwidth required for a PCoIP connection is one Mbps. In most cases, the successful deployment of desktop virtualization requires the broad deployment of WAN optimization techniques that focus on the particular characteristics of the various traffic streams that are associated with desktop virtualization.

While not widely deployed currently, IT organizations are showing a significant interest in implementing virtualized appliances. A *Virtual Appliance* is based on the appropriate software running in a VM. Virtual appliances can include WOCs, ADCs, firewalls, and performance monitoring solutions among others. An important set of synergies exist between virtual servers, virtual desktops and virtual appliances such as a WOC or a performance monitoring solution. Perhaps the most important synergy is that virtual appliances are of particular interest to IT organizations in those instances in which server virtualization technology has already been implemented in both branch offices and in the data center.

A critical factor that must be considered when evaluating the deployment of virtual appliances in a dynamic, on-demand fashion is the degree of integration that the virtual appliance has with the virtual server management system. Ideally this management system would recognize the virtual appliances as another type of VM and understand the associations between the appliance VM and the application VMs in order to allow a coordinated migration whenever this is desirable.

Cloud Computing

As was the case with virtualization, relative to the Application Delivery 2.0 era, cloud computing is a double-edged sword as it both presents challenges and solutions. One, almost comical aspect of cloud computing is that there is a high degree of disagreement in the IT industry around the precise definition of what is meant by cloud computing. In spite of that confusion, the goal of cloud computing is quite clear. That goal is to enable IT organizations to achieve an order of magnitude improvement in the cost effective, elastic provisioning of IT services that are good enough. The phrase ***good enough*** refers in part to the fact that the SLAs that are associated with public cloud computing services such as Salerforce.com or Amazon's Simple Storage System are generally very weak.

Some of the primary characteristics of a cloud computing solution are:

- The centralization of applications, servers and storage resources
- The extensive virtualization of every component of IT
- The standardization of the IT infrastructure.
- The simplification of the applications and services provided by IT
- Technology convergence
- Service orchestration
- Automation
- Self-service
- Usage sensitive chargeback
- The dynamic movement of resources

While there is some disagreement, the general consensus is that there are three primary classes of cloud computing solutions. They are:

- Public
- Private
- Hybrid

Cloud computing is based on a number of familiar concepts including time-sharing, automation, virtualization and the rental of applications. What is new about cloud computing is the synthesis of these concepts combined with the dynamic creation and movement of IT resources.

As is true with any new technology or way to deliver technology based services, there are risks associated with the adoption of all three classes of cloud computing. While the security risks get the most attention, cloud computing also presents significant management and performance challenges. However, the biggest risk accrues to those companies that don't implement any form of cloud computing.

A Reality Check

The most important optimization tasks facing IT organizations are:

- Relating the performance of applications to their impact on the business
- Ensuring acceptable performance for VoIP traffic
- Improving the performance of applications used by mobile workers
- Ensuring acceptable performance for video or telepresence traffic
- Ensuring acceptable performance for the applications that you acquire from a Software-as-a-Service provider

The tasks listed above represent a mixture of challenges from both the Application Delivery 1.0 and 2.0 eras.

Roughly one third of IT organizations have broadly implemented optimization functionality throughout their company. The penetration of optimization functionality will increase over the next year, in part based on the plans that IT organizations have to implement virtual appliances.

The most important management tasks facing IT organizations are:

- Rapidly identify the root cause of degraded application performance
- Identify malicious traffic and eliminate it
- Effectively manage QoS
- Prevent large-scale DDoS attacks
- Identify the components of the IT infrastructure that support the company's critical business applications

Given the added complexity caused by the adoption of virtualization and cloud computing, it is not surprising that the most important management task is to rapidly identify the root cause of degraded application performance. IT organizations have historically focused their management efforts on individual technology domains, such as LAN, WAN and servers. While that is still the most common approach, almost half of IT organizations also have a focus on applications and over a third of IT organizations also have a focus on managing services, where services are comprised of multiple, inter-related applications.

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.

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 the goal of which is to optimize the performance of networks and applications as part of assuring acceptable application performance. The primary role that these techniques play is to:

- Reduce the amount of data 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;
- Overcome the packet delivery issues that are common in shared (i.e., over-subscribed) networks;
- Mitigate the inefficiencies of certain protocols and/or applications;
- Offload computationally intensive tasks from client systems and servers

It is possible for an IT organization to acquire network and application optimization as a service from a managed service provider as described below. It is also possible for IT organizations to acquire network and application optimization products. These two approaches are complimentary.

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.

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 at the branch office or end user device. WOCs implement a wide variety of technologies, including caching, compression, congestion control, forward error correction, protocol acceleration, as well as request prediction and spoofing.

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. ADCs are often referred to as ***asymmetric solutions*** because they only require an appliance in the data center.

Managed Service Providers (MSP)

The last few years has seen the development of a new class of MSP – the Application Delivery MSP (ADMSP). 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, virtual ADC, WOC, virtual 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 WAN services such as ATM or MPLS. 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 service 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

One of the key challenges associated with ensuring acceptable application delivery is that currently in the vast majority of instances it is the end user, and not the IT organization, that 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.

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

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.

Application performance management (APM) is a relatively new management discipline. Part of the growing interest in APM results from the fact that as previously mentioned, a number of IT organizations have begun to offer SLAs to their company's business and functional managers for at least a handful of business critical applications. 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.

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.

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.
- 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.
- 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.
- Web Application Firewall (WAF) or WAF Service

In order to overcome challenges such as the previously mentioned Port 80 Black Hole, IT organizations need to implement WAFs and/or acquire a WAF service from a service provider.

2.0 Introduction

Background and Goals

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

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

While ensuring acceptable application delivery has always been important, it historically was not a top of mind issue for the majority of IT organizations. That changed several years ago when IT organizations began to develop a concerted focus on it. As part of this focus, many IT organizations began to deploy a first generation of solutions that were intended to protect against a growing number of security attacks, to mitigate the impact of chatty protocols[1] such as CIFS (Common Internet File System), to offload computationally intensive processing (e.g., TCP termination and multiplexing) from servers and to manage application performance on an end-to-end basis. Throughout this handbook, the application delivery challenges and solutions of this era will be referred to as Application Delivery 1.0[2].

At the same time that many IT organizations are still in the process of implementing solutions that respond to the challenges of the Application Delivery 1.0 era a new generation of challenges is emerging. These challenges are driven in large part by the:

- Emergence of a sophisticated mobile workforce
- Shifting emphasis and growing sophistication of cyber crime
- Adoption of varying forms of virtualization
- Adoption of cloud computing

As will be discussed in this handbook, the adoption of virtualization and cloud computing will both increase the probability that the performance of an application will degrade and it will make it dramatically more difficult for IT organizations to troubleshoot the root cause

[1] As is explained below, chatty protocols require hundreds of round trips to complete a single transaction.
[2] The first in the series of application delivery handbooks was published in January 2007 with the goal of helping IT organizations evaluate this first generation of application delivery challenges and solutions

of the degradation. Throughout this handbook, this emerging generation of application delivery challenges and solutions will be referred to as Application Delivery 2.0.

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.

Successful application delivery requires the integration of:

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

This handbook details many of the Application Delivery 1.0 factors that complicate application delivery. This includes the centralization of IT resources, the decentralization of employees and the complexity associated with n-tier applications. As will be shown, each of these factors continues to present challenges in the Application Delivery 2.0 era. As will also be shown, in many cases the challenges that are associated with the Application Delivery 2.0 era are the logical extension of the challenges that were associated with the Application Delivery 1.0 era. For example, ensuring acceptable performance for VoIP traffic was a major challenge during the Application Delivery 1.0 era. While that remains a challenge in the Application Delivery 2.0 era, so is the logical extension of that challenge – ensuring the acceptable performance of video and telepresence.

In some cases, however, the challenges of the Application Delivery 2.0 era are brand new. For example, one of the advantages of server virtualization is that virtual machines (VMs) can be moved between physical servers. The task of moving a VM is a relatively simple function of the virtual server management system. There can, however, be significant challenges in assuring that the VM's network configuration state (including QoS settings, ACLs, and firewall settings) is also transferred to the new location.

As we enter the Application Delivery 2.0 era, leading edge IT organizations will develop plans for implementing an application delivery function that can respond to both the existing and the emerging challenges. This transformation will not be easy, in part because it crosses myriad organizational boundaries and it 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.

Taxonomy of Applications

The typical enterprise has 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 video and audio conferencing, as well as applications that are less delay sensitive such as email and instant messaging.

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. The most recent addition to this class of applications are social media applications such as Facebook.

6. Malicious (a.k.a., Malware)

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

IT organizations need to identify malicious applications and eliminate them. Depending on the corporate policy, they may need to eliminate recreational applications or they may need to just control their usage. While it may well make sense to optimize an application that is based on a chatty protocol, it makes no sense to optimize malware or to accelerate real time traffic such as VoIP. What does make sense is to ensure that real time traffic is not interfered with by other traffic such as bulk file transfers.

Application delivery is more complex than merely accelerating 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[3] 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.

Foreword to the 2010 Edition

As stated above, we have entered a second generation of application delivery challenges and solutions. So, while this year's version of the application delivery handbook builds on the 2009 edition of the handbook, it is also significantly different than last year's version. For example, all of the case studies and market research that were in the 2009 edition of the handbook have been removed. The only market research that is in this year's edition is contained in section 6 (A Reality Check) and reflects the results of surveys that were conducted in early 2010.

On the assumption that the majority of the challenges and solutions of the Application Delivery 1.0 era are by now relatively well understood, the description of those challenges and solutions was made more succinct. To compensate for that, any reader who would like a more detailed description of those challenges and solutions will continue to be able to access the 2009 edition of the handbook at Webtorials. In addition to making the description of the challenges and solutions of the Application Delivery 1.0 ear more succinct, three entire sections were removed from the 2009 edition. Those sections were entitled *The Role of the CIO, The Changing Network Management Function,* and *Pulling it Together.*

All of that information was removed in order to enable the 2010 edition of the handbook to focus on the Application Delivery 2.0 era. With that in mind, section 3 (Application Delivery Challenges) was significantly re-written to include a focus on describing some of the challenges of the Application Delivery 2.0 era. In addition, two new sections were

[3] Citrix XenApp was formerly Citrix Presentation Server

added – one on virtualization and one on cloud computing. In addition, the concepts (e.g., planning) described in remaining chapters were modified, where appropriate, to reflect how those concepts are impacted by the emergence of the Application Delivery 2.0 era.

3.0 Application Delivery Challenges

This section of the handbook discusses some of the aspects of the application delivery environment that make ensuring application delivery difficult. It is unlikely any IT organization will exhibit all of the dynamics described in this section. However, it is also unlikely that an IT organization will not exhibit at least some of these dynamics.

The Application Delivery 1.0 Era

The challenges listed below were initially associated with the Application Delivery 1.0 era but which continue to be challenges during the Application Delivery 2.0 era.

Limited Focus of Application Development

In most situations application development focuses on ensuring that applications are developed on time, on budget, and with relatively few security vulnerabilities. Such a 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.

Chatty Protocols and Applications

The lack of emphasis on an application's performance over the WAN often results in the deployment of chatty applications[4] as illustrated in **Figure 3.1**.

Figure 3.1: Chatty Application

To exemplify the impact of a chatty protocol or application, let's assume that a given transaction requires 200 application turns. Further assume that the latency on the LAN

[4] Similar to a chatty protocol, a chatty application requires hundreds of round trips to complete a transaction.

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.

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.

Webification of Applications

The phrase ***Webification of Applications*** refers to the growing movement to implement Web-based user interfaces and to utilize Web-specific protocols such as HTTP. Unlike CIFS, HTTP is not a chatty protocol. However, HTTP is used to download web pages. It is common for a web page to have fifty or more objects, each of which requires multiple round trips in order to be transferred. Hence, although HTTP is not chatty, downloading a web page may require hundreds of round trips.

Security Vulnerabilities

As previously noted, an integral part of ensuring application delivery is to ensure that the application can be delivered securely. IT security has been an issue of growing importance for decades. For example, in 2000, the Computer Emergency Response Team (CERT) at Carnegie Mellon University catalogued 1,090 security vulnerabilities[5]. In 2008, they catalogued 7,236 security vulnerabilities[6]. Up until a few years ago, the primary motivation of a hacker was notoriety. While notoriety continues to motivate some hackers, as will be discussed below, cyber crime has become more sophisticated and more pervasive.

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.

[5] http://www.cert.org/stats/
[6] Ibid.

Server consolidation typically results in a chatty protocol such as Common Internet File System (CIFS), which was designed to run over the LAN, running over the WAN.

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.***

Distributed Employees

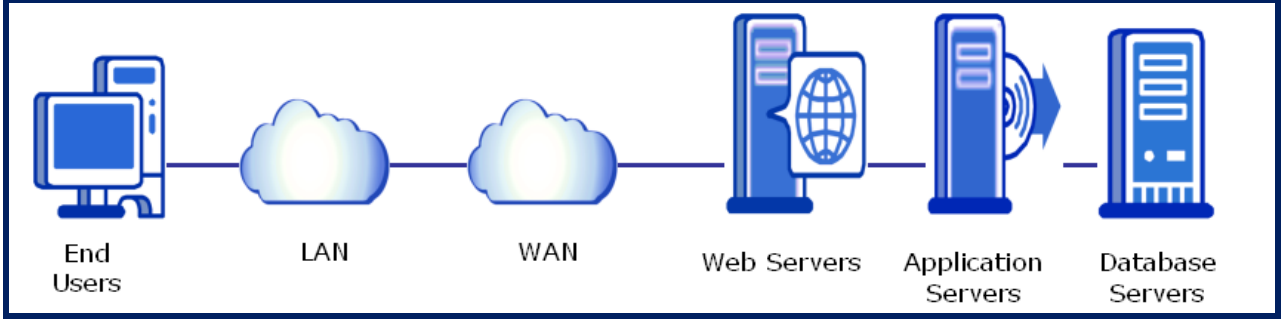
The 80/20 rule in place until a few years ago stated that 80% of a company's employees were in a headquarters facility and accessed an application over a high-speed, low latency LAN. The new 80/20 rule states that 80% of a company's employees access applications over a relatively low-speed, high latency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

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

Distributed Applications

Most IT organizations have deployed a form of distributed computing often referred to as *n-tier applications*. The browser on the user's device is typically one component of an n-tier application. The typical 4-tier application (**Figure 3.2**) is also comprised of a Web tier, an application tier and a data base tier which are implemented on a Web server(s), an application server(s) and a database server(s). In the Application Delivery 1.0 era few, if any, of the servers were virtualized.

Figure 3.2: A 4-Tier Application



Distributed applications increase the management complexity in part because each tier of the application is implemented on a separate system from which management data must be gathered. The added complexity also comes from the fact that the networks that support these applications are comprised of a variety of switches, routers, access points, WAN optimization controllers, application delivery controllers, firewalls, intrusion detection systems and intrusion protection systems from which management data must also be gathered.

Complexity

As noted in the preceding paragraph, the traditional IT distributed application environment is complex in part because there are so many components in the end-to-end flow of a transaction. If any of the components are not available, or are not performing well, the performance of the overall application or 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. Some of the implications of this complexity on performance management are that:

> ***In the vast majority of instances when the performance of a key business application is degrading, the end user, not the IT organization, first notices the degradation.***
>
> ***As the complexity of the environment increases, the number of sources of delay increases and the probability of application degradation increases in a non-linear fashion.***
>
> ***As the complexity increases the amount of time it takes to find the root cause of degraded application performance increases.***

In addition, as complexity increases so does the probability of a security intrusion. That follows because as a system becomes more complex there are more components that need to be secured. There are also more components that can be attacked.

The Application Delivery 2.0 Era

The challenges listed below are uniquely associated with the Application Delivery 2.0 era. However, because of their importance, the two factors that are most closely associated with the Application Delivery 2.0 era, virtualization and cloud computing, will be discussed in the next two sections of the handbook.

Services Oriented Architectures (SOA) with Web Services

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. Part of the appeal of an SOA is that:

- Functions are defined as reusable services where a function can be a complex business transaction such as 'Create a mortgage application' or 'Schedule Delivery'. A function can also be a simple capability such as 'Check credit rating' or 'Verify employment'.
- Services neither know nor care about the platform that other services use to perform their function
- Services are dynamically located and invoked and it is irrelevant whether the services are local or remote to the consumer of the service.

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 a traditional *n*-tier application.

Web 2.0 and Rich Internet Applications

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. One of the concepts that is typically associated with Web 2.0 is the concept of an application that is the result of aggregating other applications. This concept 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. As is the case of a Web-services based application, the WAN impacts multiple traffic flows of a mashup.

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.

As was previously mentioned, the introduction of new technologies further complicates the IT environment and leads to move security vulnerabilities. AJAX is a good example of that. AJAX is actually a group of interrelated web development techniques used on the

[7] http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid)
[8] Wikipedia on Rich Internet Applications:http://en.wikipedia.org/wiki/Rich_Internet_Application

client-side to create interactive web applications. While the interactive nature of AJAX adds significant value, it also creates some major security vulnerabilities. For example, if they are not properly validated, user inputs and user-generated content in an application can be leveraged to access sensitive data or inject malicious code into a site. According to the AJAX Resource Center[9] the growth in AJAX applications has been accompanied by a significant growth in security flaws and that this growth in security flaws "has the potential to turn AJAX-enabled sites into a time bomb."

Sophisticated Mobile workers

As previously noted, one of the challenges that was associated with the Application 1.0 era is the fact that many employees who had at one time worked in a headquarters facility now work someplace other than a headquarters facility; i.e., a regional, branch or home office. In the Application Delivery 2.0 era the challenge of supporting decentralized employees has evolved because many of these employees are now mobile[10]. What has also evolved is the type of applications that these workers access. At one time, mobile workers tended to primarily access either recreational applications or applications that are not delay sensitive; e.g., email. However, in the current environment mobile workers also need to access a wide range of business critical applications, many of which are delay sensitive. One of the issues associated with supporting mobile workers' access to delay sensitive, business critical applications is that because of the way that TCP functions, even the small amount of packet loss that is often associated with wireless networks results in a dramatic reduction in throughput.

Shifting Emphasis and Growing Sophistication of Cyber Crime

As previously noted, preventing security breaches was a key component of the Application Delivery 1.0 era. One key thing that has evolved over the last couple of years, however, is the level of sophistication of cyber crime. For example, McAfee recently published a report[11] that stated that, "Credit card fraud and identity theft have moved into the so-called 'cash cow' phase of criminal strategy. In other words, it's a source of revenue, but there's not much room for growth, so criminals are looking for the new stars of their portfolios. And intellectual property has emerged as a favorite." Also included in the report was the observation that many malware writers now have R&D and test departments.

The Center for Strategic and International Studies (CSIS) recently released a report[12] entitled "In the Crossfire – Critical Infrastructure in the Age of Cyber-War" that provided further evidence that cyber crime is becoming more sophisticated and costlier. That report included the results of a survey that was completed by six hundred IT and security

[9] http://www.ajaxtopics.com/security.html
[10] One analyst firm has predicted that there will be one billion mobile workers worldwide by 2011 - http://findarticles.com/p/articles/mi_m0EIN/is_2008_Jan_15/ai_n24230213/
[11] http://www.mcafee.com/us/about/press/corporate/2009/20090129_063500_j.html
[12] http://csis.org/event/crossfire-critical-infrastructure-age-cyber-war

executives in 14 countries. According to the report, “More than half of the executives surveyed (54 percent) said they had experienced ‘Large-scale denial of service attacks by high level adversary like organized crime, terrorists or nation-state (e.g. like in Estonia and Georgia).’ The same proportion said they had been subject to ‘stealthy infiltration” of their network by such a high-level adversary “e.g. like GhostNet’—a large-scale spy ring featuring individualized malware attacks that enabled hackers to infiltrate, control, and download large amounts of data from computer networks belonging to non-profits, government departments and international organizations in dozens of countries.”

Another aspect of the changing nature of cyber crime is the sheer scale of the attacks. In January 2010 Arbor networks published their Fifth Annual Infrastructure Security Report[13]. According to that report, the peak rate of DDoS attacks has grown from 400 Mbps in 2001 to 49 Gbps in 2009. The report also stated that, "We expect DDoS attack rates to continue to grow, but given that most enterprises are still connected to the Internet at speeds of one gigabit per second (Gbps) or less, any attack over one Gbps will be typically effective, and often trigger collateral damage to adjacent network or customer service elements as well."

The Increased Focus on Services

Just as IT organizations are getting somewhat comfortable with managing the performance of applications, they are being tasked with managing the performance of services. IT professionals use the term *service* in a variety of ways. Throughout this handbook, the definition of the term *service* will include the key characteristics of the ITIL definition of service[14]. Those characteristics include that a service:

- Is based on the use of Information Technology.
- Supports one or more of the customer's business processes.
- Is comprised of a combination of people, processes and technology.
- Should be defined in a Service Level Agreement (SLA).

The definition of service used in this handbook will also include three other key characteristics. Those characteristics are that:

- A service is often comprised of multiple inter-related applications.
- On a going forward basis, a service will increasingly be supported by an infrastructure that is virtual.
- On a going forward basis, a service will increasingly be *dynamic*. By *dynamic* is meant that the service can be provisioned or moved in a matter of seconds or minutes.

[13] http://www.marketwire.com/press-release/Arbor-Networks-Fifth-Annual-Infrastructure-Security-Report-Finds-Service-Application-1103590.htm
[14] http://www.knowledgetransfer.net/dictionary/ITIL/en/IT_Service.htm

Internal Service Level Agreements (SLAs)

IT organizations have historically insisted on receiving an SLA for services such as MPLS that they acquire from a service provider. However, IT organizations have been reluctant to offer an SLA internally to their organization's business and functional managers. That situation has changed and today roughly half of IT organizations provide internal SLAs and that percentage is expected to grow. In the current environment, IT organizations are more likely to offer an SLA for:

- Availability than for performance
- Networks than for applications
- A selected set of WAN links or applications rather than for all of the WAN or all applications

Most IT organizations, however, report that the SLAs that they currently offer internally are relatively weak and that they don't really have the tools and processes to effectively manage them. Section 7 (Planning) will outline a process that IT organizations can use to effectively manage internal SLAs.

4.0 Virtualization

Introduction

In the current environment, almost every component of IT can be virtualized. This section of the handbook will focus primarily on three forms of virtualization: server virtualization, desktop virtualization and virtualized appliances. The benefits of server and desktop virtualization have been discussed in length in various trade publications. As a result, this section will not dwell on those topics, but will instead focus on defining the challenges associated with server and desktop virtualization as well as on the technologies, both existing and emerging, that enable IT organizations to respond to those challenges. Because the benefits of virtual appliances have not been discussed in length in the trade publications, this section will discuss those benefits. This section will also discuss the challenges associated with virtual appliances as well as the technologies, both existing and emerging, that enable IT organizations to respond to those challenges. Additional information on this topic can be found in the 2010 report entitled *Virtualization: Benefits, Challenges and Solutions*[15].

This section will only briefly mention the impact that virtualization has on networking. That topic will be covered in detail in a report to be published on or about October 1, 2010. That report will be entitled *Cloud Networking*.

[15] http://www.webtorials.com/content/2010/06/virtualization.html

Server Virtualization

Challenges of Server Virtualization

Some of the specific challenges that server virtualization poses for the network infrastructure and network management include:

Contentious Management of the vSwitch

Each virtualized server includes at least one software-based virtual switch (vSwitch). This adds yet another layer to the existing data center LAN architecture. It also creates organizational stress and leads to inconsistent policy implementation.

Breakdown of Network Design and Management Tools

The workload for the operational staff can spiral out of control due to the constant stream of configuration changes that must be made to the static date center network devices in order to support the dynamic provisioning and movement of VMs.

Limited VM-to-VM Traffic Visibility

The first generation of vSwitches doesn't have the same traffic monitoring features as does physical access switches. This limits the IT organization's ability to do security filtering, performance monitoring and troubleshooting within virtualized server domains.

Poor Management Scalability

The ease with which new VMs can be deployed has led to VM sprawl. The normal best practices for virtual server configuration call for creating separate VLANs for the different types of traffic to and from the VMs within the data center. The combination of these factors strains the manual processes traditionally used to manage the IT infrastructure.

Multiple Hypervisors

It is becoming increasingly common to find IT organizations using multiple hypervisors, each with their own management system and with varying degrees of integration with other management systems. This creates islands of management within the data center.

Inconsistent Network Policy Enforcement

Traditional vSwitches lack some of the advanced features that are required to provide a high degree of traffic control and isolation. Even when vSwitches support some of these features, they may not be fully compatible with similar features offered by physical access switches. This situation leads to implementing inconsistent end-to-end network policies.

Manual Network Reconfiguration to Support VM Migration

VMs can be migrated dynamically between physical servers. However, assuring that the VM's network configuration state (including QoS settings, ACLs, and firewall settings) is also transferred to the new location is typically a time consuming manual process.

Over-subscription of Server Resources

With a desire to cut cost, there is the tendency for IT organizations to combine too many VMs onto a single physical server. The over subscription of VMs onto a physical server can result in performance problems due to factors such as limited CPU cycles or I/O bottlenecks.

Layer 2 Network Support for VM Migration

When VMs are migrated, the network has to accommodate the constraints imposed by the VM migration utility. Typically the source and destination servers have to be on the same VM migration VLAN, the same VM management VLAN, and the same data VLAN.

Storage Support for Virtual Servers and VM Migration

The data storage location, including the boot device used by the VM, must be accessible by both the source and destination physical servers at all times. If the servers are at two distinct locations and the data is replicated at the second site, then the two data sets must be identical.

Complex Troubleshooting on a per-VM Basis

Many of the same management tasks that must be performed in the traditional server environment need to be extended into the virtualized environment. An example of this is that IT organizations must be able to troubleshoot on a per-VM basis.

To put the challenge of troubleshooting on a per-VM basis into perspective, consider a hypothetical 4-tier application, similar to what was depicted in Figure 3.2, that will be referred to as TheApp. For the sake of this example, assume that TheApp is implemented in a manner such that the web server, the application server and the database server are each running on VMs on separate servers, each of which have been virtualized using different hypervisors. It is notably more difficult to troubleshoot TheApp than it is to troubleshoot the traditional 4-tier application in part because each server has a different hypervisor management system and in part because of the lack of visibility into the inter-VM traffic on a given physical server.

> ***Troubleshooting in a virtualized environment is notably more difficult than troubleshooting in a traditional environment.***

Meeting the Challenges of Server Virtualization

At the present time, there is no overarching solution for the comprehensive management of a computing environment composed of virtualized servers, storage, and networks. Listed below are some the key developments that can help IT departments meet the challenges of virtualization.

Dynamic Infrastructure Management

A dynamic virtualized environment can benefit greatly from a highly scalable and integrated DNS/DHCP/IPAM solution. Where DNS, DHCP and IPAM share an integrated database, this obviates the need to manually coordinate records in different locations.

Virtualized Performance and Fault Management

Virtual switches currently being introduced into the market can export traffic flow data to external collectors. Another approach to monitoring and troubleshooting intra-VM traffic is to deploy a virtual performance management appliance or probe[16] within the virtualized server. A third approach is to access the data in the virtual server management system.

Distributed Virtual Switching (DVS)

Most vSwitches include an integrated control and data plane. With DVS, the control and data planes are decoupled. This makes it easier to integrate the vSwitch's control plane with the control planes of other switches and with the virtual server management system.

Edge Virtual Bridges (EVBs)

With EVB, the hypervisor is relieved from all switching functions, which are now performed by the physical access and aggregation network. Using Virtual Ethernet Port Aggregator (VEPA), all traffic from VMs is forwarded to the adjacent physical access switch and directed back to the same physical server if the destination VM is co-resident on the same server.

Orchestration and Provisioning

Service orchestration is an operational technique that helps IT organizations to automate many of the manual tasks that are involved in provisioning and controlling the capacity of dynamic virtualized services.

Desktop[17] Virtualization

The two fundamental forms of desktop virtualization are:

- Server-side application/desktop virtualization
- Client-side application/desktop virtualization

With server-side virtualization, the client device plays the familiar role of a terminal accessing an application or desktop hosted on a central presentation server and only screen displays, keyboard entries, and mouse movements are transmitted across the network. This approach to virtualization is based on display protocols such as Citrix's Independent Computing Architecture (ICA) and Microsoft's Remote Desktop Protocol (RDP).

[16] This will be discussed in the subsequent analysis of virtual appliances.
[17] In this context, the term 'desktop' refers to the tradition desktop as well as to various mobile devices including laptops and smartphones.

There are two primary approaches to server-side application/desktop virtualization. They are:

- Server Based Computing (SBC)
- Virtual Desktop Infrastructure (VDI)

IT organizations have been using the SBC approach to virtualization for a long time and often refer to it as Terminal Services. Virtual Desktop Infrastructure (VDI) is a relatively new form of server-side form of virtualization in which a VM on a central server is dedicated to host a single virtualized desktop.

Client-side application virtualization is based on a model in which applications are streamed on-demand from central servers to client devices over a LAN or a WAN. On the client-side, streamed applications are isolated from the rest of the client system by an abstraction layer inserted between the application and the local operating system. In some cases, this abstraction layer could function as a client hypervisor isolating streamed applications from local applications on the same platform. Application streaming is selective in the sense that only the required application libraries are streamed to the user's device. The streamed application's code is isolated and not actually installed on the client system. The user can also have the option to cache the virtual application's code on the client system.

Challenges of Desktop Virtualization

IT organizations are showing a growing interest in desktop virtualization. However:

> ***From a networking perspective, the primary challenge in implementing desktop virtualization is achieving adequate performance and an acceptable user experience for client-to-server connections over a WAN.***

Ensuring acceptable performance for VDI presents some significant challenges. One such challenge is that, as is the case in with any TCP based application, packet loss causes the network to retransmit packets. This can dramatically increase the time it takes to refresh a user's screen. While this is a problem in any deployment, it is particularly troublesome in those situations (i.e., global deployments) in which there is a significant amount of packet loss. In addition, VDI deployments are characterized by requiring a large number of simultaneous flows that cannot be supported by many of the WOCs currently on the market (see section 8).

The ICA and RDP protocols employed by many hosted application virtualization solutions are somewhat efficient in their use of the WAN because they incorporate a number of compression techniques including bitmap image compression, screen refresh compression and general data compression. While these protocols can often provide

adequate performance for traditional data applications, they have limitations with graphics-intensive applications, 3D applications, and applications that require audio-video synchronization.

Before implementing desktop virtualization, IT organizations need to understand the network implications of that implementation. One of those implications is that other WAN traffic such as large file transfers, can negatively impact the user's experience with desktop virtualization. Another implication is that a large amount of WAN bandwidth may be required. For example, the first two columns of **Table 4.1** show estimates for the amount of WAN bandwidth required by XenDesktop as documented in a recent entry in The Citrix Blog[18].

Table 4.1: Bandwidth Requirements from a Representative Branch Office

Activity	XenDesktop Bandwidth	Number of Simultaneous Users	WAN Bandwidth Required
Office	43 Kbps	10	430 Kbps
Internet	85 Kbps	15	1,275 Kbps
Printing	573 Kbps	15	8,595 Kbps
Flash Video	174 Kbps	6	1,044 Kbps
Standard WMV Video	464 Kbps	2	928 Kbps
High Definition WMV Video	1,812 Kbps	2	3,624 Kbps
Total WAN Bandwidth			15,896 Kbps

The two rightmost columns in Table 4.1 depicts one possible scenario of what fifty branch office users are doing and identifies that the total WAN bandwidth that is required by this scenario is just less than 16 Mbps.

Compared with hosted applications, streamed applications are far less efficient as they typically use the same inefficient protocols (e.g., CIFS) that are native to the application. Furthermore, streamed applications create additional bandwidth challenges for IT organizations because of the much larger amount of data that must be transmitted across the WAN when the application is initially delivered to the branch.

Meeting the Challenges of Desktop Virtualization

As mentioned, protocols such as ICA and RDP have limitations with graphics-intensive applications, 3D applications, and applications that require audio-video synchronization. To respond to the challenges created by these types of applications, Teradici recently introduced the PC-over-IP (PCoIP) protocol. PCoIP is a proprietary protocol that renders the graphics images on the host computer and transfers compressed pixel level data to

[18] http://community.citrix.com/display/ocb/2010/05/20/How+Much+Bandwidth+Do+I+Need+for+My+Virtual+Desktop

the client device. PCoIP is the display protocol used by the recently introduced VMware View 4 VDI product, which also supports RDP.

While PCoIP resolves some challenges, it also creates others. For example, a recently published document[19] stated that, "To support the lower bandwidth typically available over a WAN, the minimum peak bandwidth required for a PCoIP connection has been reduced to 1 Mbps." While the 1 Mbps required by PCoIP to support a single user represents a worst-case situation, it does underscore the fact that a significant amount of WAN bandwidth can be required to support desktop virtualization. Another challenge associated with PCoIP is that Teradici cannot turn off encryption which makes it difficult, if not impossible, to optimize PCoIP traffic.

As mentioned, packet loss can have a very negative impact on the performance of desktop virtualization solutions. Two techniques that can be used to mitigate the impact of packet loss are Forward Error Correction (FEC) and real time Packet Order Correction (POC). Unfortunately, these techniques are not widely supported by the current generation of WOCs. As was also mentioned, one of the implications of implementing desktop virtualization is that other WAN traffic such as large file transfers, can negatively impact the user's experience with desktop virtualization. To avoid this situation, QoS needs to be implemented throughout the WAN. Given the need for QoS as well as the need to support large file transfers and to support the optimization of protocols such as CIFS and ICA:

> ***IT organizations that are implementing virtualized desktops should analyze the viability of implementing WAN and application optimization solutions***.

There are three general classes of network and application optimization solutions. Two of these classes (WAN Optimization Controllers and Application Delivery Controllers) are described in section 8 of this handbook. The third class (Application Delivery Services) is described in section 9 of this handbook. Some of the relevant optimization technique include:

- Compression
- Caching and de-duplication
- TCP Protocol optimization
- Application and protocol (e.g., CIFS, HTTP, MAPI) optimization
- Protocol (e.g., ICA, RDP, PCoIP) optimization
- QoS and traffic shaping

Although virtually all WAN Optimization Controllers (WOCs) on the market support the functions listed above, there are some significant differences in terms of how the functionality is implemented and how well it performs. For example, the ICA and RDP

[19] http://www.teradici.com/media/resources/PCoIP-WAN-brief.pdf

protocols can be difficult to optimize for a number of reasons. One of those reasons is that these protocols only send small request-reply packets. This form of communications is best optimized by byte-level caching that is not supported by many WOC vendors. In addition, before implementing any of the techniques listed above, an IT organization must determine which acceleration techniques are compatible with the relevant display protocols. For example, in order to be able to compress ICA traffic, a WOC must be able to decrypt the ICA workload, apply the optimization technique, and then re-encrypt the data stream.

In order to enable the growing population of mobile workers to access enterprise applications, the communications between the mobile worker and the data center has to be optimized. One way to optimize this communications is to deploy client software on the user's mobile device (e.g., laptop, smartphone) that provides WOC functionality. Until recently, the typical device that mobile workers used to access enterprise applications was a laptop. While that is still the most common scenario, today many mobile workers use their smartphones to access enterprise applications. Therefore, over the next few years it is reasonable to expect that many IT organizations will support the use of smartphones as an access device by implementing server-side application virtualization for those devices. This means that in a manner somewhat similar to remote workers, mobile workers will access corporate applications by running protocols such as ICA and RDP over a WAN.

Just as was the case with workers who access applications from a fixed location, in order for mobile workers to be able to experience acceptable application performance, network and application optimization is required. In many cases the mobile worker will use some form of wireless access. Since wireless access tends to exhibit more packet loss than does wired access, the WOC software that gets deployed to support mobile workers needs functionality such as forward error correction that can overcome the impact of packet loss. In addition, as workers move in and out of a branch office, it will be necessary for a seamless handoff between the mobile client and the branch office WOC.

As previously noted, application streaming creates some significant WAN performance problems that require the deployment of a WOC in part because the code for streamed applications is typically transferred via a distributed file system protocol, such as CIFS, which is well known to be a chatty protocol. Hence, in order effectively support application streaming, IT organizations need to be able to optimize the performance of protocols such as CIFS, MAPI, HTTP, and TCP. In addition, IT organizations need to implement other techniques that reduce the bandwidth requirements of application streaming. For example, by using a WOC, it is possible to cache the virtual application code at the client's site. Caching greatly reduces the volume of traffic for client-side virtualized applications and it also allows applications to be run locally in the event of network outages. Staging is a technique that is similar to caching but is based on pre-positioning and storing streamed applications at the branch office on the WOC or on a branch server. With staging, the application is already locally available at the branch when users arrive for work and begin to access their virtualized applications.

One of the challenges associated with deploying WOC functionality to support desktop virtualization is:

> ***Supporting desktop virtualization will require that IT organizations are able to apply the right mix of optimization technologies for each situation.***

For example, pre-staging and storing large virtual desktop images on the WOC at the branch office must be done in an orchestrated fashion with the corresponding resources in the data center. Another example of the importance of orchestration is the growing requirement to automatically apply the right mix of optimization technologies. For example, as noted protocols such as ICA and RDP already incorporate a number of compression techniques. As a result, any compression performed by a WAN optimization appliance must adaptively orchestrate with the hosted virtualization infrastructure to prevent compressing the traffic twice - a condition that can actually increase the size of the compressed payload.

Virtual Appliances

A *Virtual Appliance* is based on software that provides the appropriate functionality, together with its operating system, running in a VM on top of the hypervisor in a virtualized server. Virtual appliances can include WOCs, ADCs, firewalls, and performance monitoring solutions among others[20].

> ***An important set of synergies exist between virtual servers, virtual appliances such as a WOC or a performance monitoring solution and virtual desktops.***

Throughout the rest of this section, those synergies will be referred to as the Virtuous Synergies of Virtualization (VSV). The key components of the VSV are:

- The fact that IT organizations have already deployed server virtualization means that it is easier and less costly to implement virtualized appliances.
- Because it is easier and less expensive to deploy a software-based appliance than it is to deploy a hardware-based appliance, they are more likely to be broadly deployed.
- Because software-based WOCs can be broadly deployed, they can enable the deployment of virtual desktops.
- Because vProbes can be broadly deployed, they enable IT organizations to manage the performance of applications that run on virtualized servers.

[20] The argument could be made that a virtual router is a virtual appliance. Virtual routers will be discussed in *Cloud Networking.*

- Because virtual firewalls can be broadly deployed, they enable IT organizations to meet regulatory and compliance requirements for applications that run on virtualized servers.
- As part of moving a VM, virtual appliances can be easily migrated along with the VM in order to replicate the VMs's networking environment in its new location.
- Because vProbes can be broadly deployed, they eliminate some of the challenges associated with other forms of virtual appliances; i.e., WOCs, ADCs and firewalls.

A cornerstone of the VSV is that virtual appliances are of particular interest to IT organizations in those instances in which server virtualization technology has already been disseminated to branch offices and has also been implemented in the data center. When server virtualization pervades the enterprise, a wide variety of networking functions can be deployed wherever needed easily and cost effectively with virtual appliances, without the installation of additional hardware.

In the branch office, a suitably placed virtualized server could potentially host a virtual WOC appliance as well as other virtual appliances. Alternatively, a router or a WOC that supports VMs could also serve as the infrastructure foundation of the branch office. Virtual appliances can therefore support branch office server consolidation strategies by enabling a single device (i.e., server, router, WOC) to perform multiple functions typically performed by multiple physical devices. These physical devices include a WOC, router, firewall, IDS/IPS, DHCP/DNS server, client-side application virtualization staging server, local application server, etc.

One of the compelling advantages of a virtualized appliance is that the acquisition cost of a software-based appliance can be notably less than the cost of a hardware-based appliance with same functionality[21]. In many cases the cost of a software-based appliance can be a third less than the cost of a hardware-based appliance. In addition, a software-based client can potentially leverage the functionality provided by the hypervisor management system to provide a highly available system without having to pay for a second appliance[22]. As a result of these cost savings, IT organizations will be able to afford to deploy virtualized appliances more broadly than they would be able to deploy hardware-based appliances.

As discussed in the preceding section of this section, WOCs that implement the appropriate techniques can make virtual desktops a viable solution for provisioning and managing branch office desktop environments. While these advantages occur whether the WOC is hardware based or software based, the fact that a virtual WOC is so much more cost effective than a hardware based WOC means that they are more likely to be

[21] The actual price difference between a hardware-based appliance and a software-based appliance will differ by vendor.
[22] This statement makes a number of assumptions, including the assumption that the vendor does not charge for the backup software-based appliance.

deployed. Because virtual WOCs can be deployed broadly in a cost effective manner, IT organizations are more likely to be successful with desktop virtualization.

Another advantage of a virtual appliance is that as previously mentioned, in many instances the benefits of the dynamic movement of a VM from one server to another are maximized if the supporting infrastructure can also be dynamically moved. If virtualized appliances have been deployed, then it is notably easier than it is in a more traditional environment for various networking functions (WOC, ADC, firewall, etc.) to be migrated along with VMs in order to replicate the VMs's networking environment in its new location.

As previously noted, one approach to monitoring and troubleshooting inter-VM traffic is to deploy a virtual performance management appliance or probe (vProbe). However, one of the characteristics of a virtualized server is that each virtual machine 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, in order to be effective, a vProbe must not consume significant resources. The way that a vProbe works is similar to how many IT organizations monitor a physical switch. In particular, the vSwitch has one of its ports provisioned to be in promiscuous mode and hence forwards all inter-VM traffic to the vProbe. As a result, the use of a vProbe gives the IT organization the necessary visibility into the inter-VM traffic.

A virtual firewall appliance can help IT organizations meet some of the challenges associated with server virtualization. That follows because virtual firewall appliances can be leveraged to provide isolation between VMs on separate physical servers as well as between VMs running on the same physical server. Ideally, the firewall virtual appliance would use the same software as the physical firewalls already in use in the data center. The security appliance can potentially provide highly integrated functionality to help secure virtual machines, applications, and traffic. This includes firewall, VPN, anti-malware, IDS/IPS, integrity monitoring (e.g., registry changes), and log inspection functionality.

A potential issue to keep in mind is that each virtual appliance is running in a VM, so all of the challenges of managing a virtual server environment described earlier are also applicable to virtual appliances. In particular, visibility of VM-to-VM traffic would be more critical in order to troubleshoot a virtual server environment where traffic traverses several virtual appliances on its way to a destination VM located in the same physical server. In this instance, the deployment of one form of virtual appliance, a vProbe, eliminates some of the challenges associated with implementing other forms of virtual appliances; i.e., WOCs, ADCs and firewalls.

One of the potential downsides of a virtual appliance is performance. The conventional wisdom in our industry is that a solution based on dedicated, purpose-built hardware performs better than a solution in which software is ported to a generic piece of hardware, particularly if that hardware is supporting multiple applications. For example,

no virtual WAN optimization controller (see section 8) currently on the market supports a WAN link greater than 45 Mbps.

However, conventional wisdom is often wrong. Some of the factors that enable a virtualized appliance to provide high performance include:

- Moore's law that states that the price performance of off the shelf computing devices doubles every 18 months.
- The deployment of multiple core processors to further increase the performance of off the shelf computing devices.
- The optimization of the software on which the virtual appliance is based.

Because of the factors listed above and because of the advantages that they provide, IT organizations should evaluate the performance of a virtual appliance to determine if a virtual appliance is an appropriate solution.

Another critical factor when evaluating the deployment of virtual appliances in a dynamic, on-demand fashion is the degree of integration of the virtual appliance with the virtual server management system. Ideally this management system would recognize the virtual appliances as another type of VM and understand associations between appliance VM and application VMs to allow a coordinated migration whenever this is desirable. In addition to VM migration, integration with the virtual server management system should support other management features, such as:

- Provisioning of Virtual Appliances
- Resource Scheduling and Load Balancing
- High Availability
- Business Continuance/Disaster Recovery

5.0 Cloud Computing

Within the IT industry there is considerable confusion and disagreement relative to exactly what is meant by the phrase ***cloud computing***. An example of that confusion is the fact that the January 2009 edition of The Cloud Computing Journal published an article[23] that had twenty one definitions of cloud computing. The position taken in this handbook is that creating yet one more definition of cloud computing would only add to the confusion. This handbook also takes the position that it is notably less important to define exactly what is meant by the phrase ***cloud computing*** than it is to identify the goal of cloud computing.

[23] Twenty-One Experts Define Cloud Computing, http://cloudcomputing.sys-con.com/node/612375

The goal of cloud computing is to enable IT organizations to achieve an order of magnitude improvement in the cost effective, elastic provisioning of IT services that are good enough.

The phrase ***good enough*** refers in part to the fact that the SLAs that are associated with public cloud computing services such as Salerforce.com or Amazon's Simple Storage System are generally weak. For example, most of the SLAs don't contain a goal for the performance of the service. In addition, it is common to access these services over the Internet and nobody provides an SLA for the availability of performance of the Internet. As such, organizations that use these services do so with the implicit understanding that if the level of service they experience is not sufficient, their only recourse is to change providers.

There are several proof points that indicate that the goal of cloud computing as stated above is achievable. For example, an article in Network World identified some of the potential cost savings that are associated with cloud computing[24]. In that article, Geir Ramleth the CIO of Bechtel stated that he benchmarked his organization against some Internet-based companies. As a result of that activity, Ramleth determined that the price that Amazon charges for storage is roughly one fortieth of his internal cost for storage. Ramleth also estimated that YouTube spends between $10 and $15 per megabit/second for WAN bandwidth, while Bechtel is spending $500 per megabit/second for its Internet-based VPN.

Relative to the provisioning of IT services, historically it has taken IT organizations several weeks or months from the time when someone first makes a request for a new server to the time when that server is in production. In the last few years many IT organizations have somewhat streamlined the process of deploying new resources. However, in the traditional IT environment in which IT resources have not been virtualized, the time to deploy new resources is still measured in weeks if not longer. This is in sharp contrast to a public cloud computing environment where the time it takes to acquire new IT resources from a cloud computing service provider is measured in seconds or minutes.

Additional information on the topic of cloud computing can be found in two recent report: *A Guide for Understanding Cloud Computing*[25] and *Cloud Computing: A Reality Check and Guide to Risk Mitigation*[26].

The Primary Characteristics of Cloud Computing

In spite of the confusion as to the exact definition of cloud computing, the following set of characteristics are typically associated with cloud computing.

[24] The Google-ization of Bechtel, Carolyn Duffy Marsan, Network World, October 28, 2008
[25] http://www.webtorials.com/content/2009/11/a-guide-for-understanding-cloud-computing.html
[26] http://www.webtorials.com/content/2009/12/cloud-computing-a-reality-check-guide-to-risk-mitigation.html

- **Centralization** of applications, servers and storage resources. Many companies either already have, or are currently in the process of consolidating applications, servers and storage out of branch offices and into centralized data centers. This consolidation reduces cost and enables IT organizations to have better control over the company's data.
- Extensive **virtualization** of every component of IT. This includes servers, desktops, applications, storage, networks and appliances such as WAN optimization controllers, application delivery controllers and firewalls. The reason that virtualization is so often associated with cloud computing is that virtualization tends to reduce cost and increase the elasticity of service provisioning.
- **Standardization** of the IT infrastructure. Complexity drives up cost and reduces agility and elasticity. As such, complexity is the antithesis of cloud computing. One source of complexity is having multiple suppliers of equipment such as switches and routers, as well as having multiple operating systems (i.e., Linux, Windows, Solaris), or even multiple versions of the same network operating system such as IOS.
- **Simplification** of the applications and services provided by IT. In a simplified IT environment, the IT organization rarely develops a custom application or customizes a third party application, has just one system for functions such as ERP and SCM, and only supports one version of a given application.
- **Technology convergence**. Cisco recently announced its Unified Computing System[27] (UCS). UCS is intended to enable the convergence of technologies such as servers, networks, storage and virtualization. Cisco's stated rational for technology convergence is to lower the cost and improve the elasticity of the data center infrastructure.
- **Service orchestration** is an operational technique that helps IT organizations to automate many of the manual tasks that are involved in provisioning and controlling the capacity of dynamic virtualized services. This enables IT to streamline operational workloads and overcome technology and organizational silos and boundaries.
- **Automation** of as many tasks as possible; e.g., provisioning, troubleshooting, change and configuration management. Automation can enable IT organizations to reduce cost, improve quality and reduce the time associated with management processes.
- **Self-service** allows end users to select and modify their use of IT resources without the IT organization being an intermediary. This concept is often linked with usage sensitive chargeback (see below) as well as the concept of providing IT services on-demand.
- **Usage sensitive chargeback** is often referred to as pay-as-you-go. One part of the rational for implementing usage sensitive chargeback is that it gives the users greater control over their IT spend because they determine how much of the IT services they consume. Another part of that the rationale is that it enables the IT

[27] http://newsroom.cisco.com/dlls/2009/prod_031609.html

organization providing the services to focus on what they can control - the unit cost of the services.

- The **dynamic movement of resources** such as virtual machines and the associated storage. This capability also helps to streamline the provisioning of new applications, improve backup and restoration operations and enable zero-downtime maintenance.

Classes of Cloud Computing

Cloud Computing Service Providers (CCSPs) that provide their services either over the public Internet or over other WAN services are offering a class of solution that is often referred to as the *public cloud* or *public cloud computing.*

The primary types of services provided by CCSPs are:

- Software-as-a-Service (SaaS)
  In the typical SaaS offering, an independent software vendor (ISV) offers access to their applications on a subscription basis. An example of this type of service offering is SalesForce.com.
- Infrastructure-as-a-Service (IaaS)
  Infrastructure services are comprised of the basic compute, storage, and network services required to run applications. An example of this type of service is Amazon's Simple Storage Service (S3).
- Platform-as-a-Service (PaaS)
  Platform services provide software development environments, including application programming interfaces (APIs) and middleware that abstract the underlying IaaS infrastructure in order to support rapid application development and deployment. An example of this type of service is Force.com.

Referring back to Geir Ramleth the CIO of Bechtel, the decision that he reached was not that he was going to rely on third parties to supply all of his IT requirements. Rather, he decided that Bechtel would adopt the characteristics of cloud computing (e.g., virtualization, automation) within Bechtel's internal IT environment. In many, but not all instances, the approach that Ramleth is taking is referred to as *Private Cloud* or *Private Cloud Computing*. Private Clouds have the advantages of not being burdened by many of the potential security vulnerabilities, data confidentiality and control issues that are associated with public clouds and that are discussed in a subsequent section of this paper. A number of industry pundits, however, believe that the concept of a private cloud is an oxymoron. They believe that the only form of cloud computing is public cloud computing.

In those instances in which an enterprise IT organization uses a mixture of public and private cloud services, the result is often referred to as a *Hybrid Cloud*. The hybrid cloud approach can offer the scalability of the public cloud coupled with the higher degree of

control offered by the private cloud. Hybrid clouds, however, do present significant management challenges. For example, the preceding section of the handbook discussed a hypothetical 4-tier application that was referred to as TheApp. As that section pointed out, it is notably more difficult to troubleshoot TheApp in a virtualized environment than it would be to troubleshoot the same application in a traditional environment. Now assume that TheApp is deployed in such a way that the web tier is supported by a CCSP and the application and database tiers are provided by the IT organization. This increases the difficulty of management yet again because all of the management challenges that were discussed previously still exist and added to them are the challenges associated with having multiple organizations involved in managing the application.

Troubleshooting in a hybrid cloud environment will be an order of magnitude more difficult than troubleshooting in a traditional environment.

As noted, the preceding paragraphs define what is commonly meant by public cloud computing, private cloud computing and hybrid cloud computing. As shown in **Table 5.1**, there is an emerging view of cloud computing that is being advocated by a number of vendors including Cisco[28]. The first column of Table 1 reflects the dominant current view that as described, there are three general classes of cloud computing: private, public and hybrid. The second column of Table 5.1 reflects the emerging view of cloud computing. The primary conceptual difference between the current and the emerging views is how a private cloud is defined. In the current view, private cloud refers to IT resources provided

Table 5.1: Classes of Cloud Computing

Current View	Emerging View	Definition
Private Cloud	Internal Cloud	Apply the characteristics of cloud computing solutions to the resources wholly owned and managed by the organization consuming the services.
Public Cloud	External Cloud; Public Cloud	Obtain IT services from a CCSP.
Hybrid Cloud	Private Cloud	Provide IT services from a combination of external and internal resources but both sets of services are under the control of the IT organization that is acquiring the services.
Hybrid Cloud	Hybrid Cloud	Uses a combination of external and internal resources. The external resources are under the control of the CCSP and the internal resources are under the control of the IT organization that is acquiring the services.

[28] Private Cloud Computing for Enterprises: Meet the Demands of High Utilization and Rapid Change, http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns836/ns976/white_paper_c11-543729.html

entirely by an IT organization such as Bechtel. In the emerging view of a private cloud, there is management functionality that allows the enterprise IT organization to control not just the resources that it provides, but also the resources of one or more CCSPs.

The capability for IT organizations to control both their own resources as well as the resources of one or more CCSPs is extremely powerful. This capability would require, however, a tight integration between the management functionality that the IT organization has and the management functionality that the CCSPs have.

> ***The emerging view of private cloud computing requires the ability to move massive files between the IT organization and the CCSPs in a small amount of time.***

Because the enabling functionality either does not exist, or has not been widely deployed, the rest of this report will use the current view of the classes of cloud computing.

Cloud Computing: A Lot of Things Old, A Few Things New

Most of the concepts and characteristics of cloud computing are not new. Acquiring computing cycles from a third party (currently referred to as Infrastructure as a Service) was first done in the 1960s and was called time-sharing. It is hard to argue that virtualized servers are a new concept since IBM was shipping virtualized mainframes thirty-five years ago. The dot.com era spawned an early attempt at outsourcing applications from a class of provider referred to at the time as Application Service Providers (ASPs). It is certainly possible to say that from a conceptual perspective that ASPs and the current generation of Software-as-a-Service (SaaS) providers are very similar and that the only real difference between them is how they implement their services. Those implementation differences, however, can make a huge difference in terms of the likely success or failure of each type of vendor. For example, the IT infrastructure that the ASP model of the 1990s was built on did not differ significantly from the typical enterprise IT infrastructure of that era. As such, the traditional ASP did not provide a significant cost savings vs. what an enterprise IT organization could provide on its own. The following subsection outlines the factors that indicate whether or not a CCSP is notably more cost effective than an ASP or an IT organization.

There are, however, some new concepts that are associated with cloud computing. One of the new concepts associated with cloud computing is the on-demand provisioning of IT resources. As noted, in a traditional IT environment it can take weeks or even months to provision a new server. As previously noted, in a public cloud computing environment, a new virtual machine can be implemented in seconds or minutes. Another one of the new concepts associated with cloud computing is the dynamic movement of IT resources. One way that his concept is manifested is through functionality that VMware refers to as VMotion. VMotion enables the live migration of a virtual machine from one physical server to another. This functionality is intended to:

- Optimize IT resources for maximum utilization, flexibility and availability
- Allow IT organizations to perform hardware maintenance without scheduled downtime
- Improve performance by moving virtual machines away from failing or under performing servers

The CCSP Economic Value Proposition

One of the underpinnings of the theoretical economic value proposition of CCSPs such as Amazon, Google or Salesforce.com is that they are large and as such enjoy an economy of scale that most IT organizations don't. Another one of the underpinnings is that some CCSPs have invested to develop functionality that is not commercially available. A third underpinning is that CCSPs have broadly implemented the majority of the characteristics of cloud computing that were discussed in the preceding section and as a result are notably more efficient than are the vast majority of IT organizations.

There is some anecdotal evidence that indicates that at least in some instances that the theoretical value proposition of CCSPs is indeed accurate. One piece of that anecdotal evidence is the previously described work that Bechtel did to benchmark themselves vs. Internet startups. However, there is nothing that says that all CCSPs are so notably cost effective that even when they layer their cost of sales and their profit margin on top of the direct costs of providing the services that they offer, that they are notably more cost effective than are the vast majority of IT organizations. In addition, while most IT organizations are not as large as Amazon of Google, many are large enough to enjoy a significant economy of scale. If these large IT organizations also successfully implement most of the key characteristics of cloud computing, based on both the currently available technologies as well as those that will be deployed in the near term, then they can potentially significantly reduce, or even eliminate, the potential CCSP cost advantage.

The Risks of Public Cloud Computing

One of the risks associated with public cloud computing is performance. One of the primary causes of those performance problems is that most cloud computing platforms (i.e., Amazon's EC2) are built on a small number of large data-centers that users access over the Internet. As a result of this design, the majority of users of the platform are a considerable distance removed from the data-center. As is always the case, the user's experience tends to degrade as the user gets further removed from the data-center. Even users in the same country as the data-center can be subject to unacceptable performance as a result of sub-optimal routing and the inefficient protocols used within the Internet. The limitations of the Internet are discussed in section 9. That section also discusses how Internet-based Application Delivery Services can be used to mitigate those limitations.

However, as is true with any new technology or way to deliver technology based services, there are risks associated with the adoption of all three classes of cloud computing. However:

The biggest risk accrues to those companies that don't implement any form of cloud computing.

IT organizations that don't implement any form of cloud computing guarantee that their company will not realize the order of magnitude improvement in the cost effective elastic provisioning of cloud computing that is the goal of cloud computing. Partially because of that, IT organizations that don't implement any form of cloud computing run the risk of being bypassed by business and functional managers that are demanding solutions that have a level of cost and agility that the IT organization cannot provide.

Private cloud computing has the advantage of not being burdened by many of the potential security vulnerabilities, data confidentiality and control issues that are associated with public cloud computing. Because of that fact, this section will focus on three categories of risk that are associated with public cloud computing and that IT organizations need to evaluate prior to using public cloud computing services. Based on the definition of the phrase ***ensuring acceptable application delivery*** that was presented in the introduction to this handbook, those categories are security, management and performance.

In particular, as part of performing due diligence prior to acquiring public cloud computing serves, IT organizations need to do a thorough assessment of a CCSP's capabilities in the following three areas:

Security

- Can the CCSP pass the same security audits (e.g., PCI, HIPAA) to which the IT organization is subject?
- Does the CCSP undergoes regular third party risk assessment audits and will the CCSP make the results of those audits available to both existing and potential customers?
- What are the encryption capabilities that the CCSP provides.
- To what degree does the CCSP follow well-established guidelines such as the Federal Information Security Management Act (FISMA) or National Institute of Science and Technology (NIST) guidelines?
- Has the CCSP achieved SAS 70 Type II security certification?
- Is it possible for the IT organization to dictate in which countries their data will be stored?
- What tools and processes has the CCSP implemented to avoid unauthorized access to confidential data?
- Will the CCSP inform the IT organization when someone accesses their data?

- Does the CCSP have the right and/or intention to make use of the data provided to it by the IT organization; e.g., analyzing it to target potential customers or to identify market trends?
- What are the CCSP's policies and procedures relative to data recovery?
- What procedures does the CCSP have in place to avoid issues such as virus attacks, Cross-site scripting (XSS) and man in the middle intercepts?
- How well trained and certified is the CCSP's staff in security matters?

Management

- What is the ability of the CCSP to manage the challenges associated with virtualization that were discussed in the preceding section of this handbook?
- What management data will the CCSP make available to the IT organization?
- What is the ability of the CCSP to troubleshoot performance or availability issues?
- What are the CCSP's management methodologies for key tasks such as troubleshooting?
- Does the CCSP provide tools such as dashboards to allow the IT organization to understand how well the service they are acquiring is performing?
- Does the CCSP provide detailed information that enables the IT organization to report on their compliance with myriad regulations?
- What are the primary management tools that the CCSP utilizes?
- What is the level of training and certification of the CCSP's management personnel?
- What are the CCSP's backup and disaster recovery capabilities?
- What approach does the CCSP take to patch management?
- What are the specific mechanisms that the IT organization can use to retrieve its data back in general and in particular if there is a dispute, the contract has expired or the CCSP goes out of business?
- Will the CCSP allow the IT organization to test the data retrieval mechanisms on a regular basis?
- What is the escalation process to be followed when there are issues to be resolved?
- How can the service provided by the CCSP be integrated from a management perspective with other services provided by either another CCSP and/or by the IT organization?
- How can the management processes performed by the CCSP be integrated into the end-to-end management processes performed by the IT organization?

Performance

- What optimization techniques has the CCSP implemented?
- What is the ability of the CCSP to identify and eliminate performance issues?
- What are the procedures by which the IT organization and the CCSPs will work together to identify and resolve performance problems?
- What is the actual performance of the service and how does that vary by time of day, day of week and week of the quarter?

- Does the IT organization have any control over the performance of the service?
- What technologies does the CCSP have in place to ensure acceptable performance for the services it provides?
- Does the CCSP provide a meaningful SLA? Does that SLA have a goal for availability? Performance? Is there a significant penalty if these goals are not met? Is there a significant penalty if there is a data breach?
- To what degree is it possible to customize an SLA?
- What is the ability of the CCSP to support peak usage?
- Can the CCSP meet state and federal compliance regulations for data availability to which the IT organization is subject?

6.0 A Reality Check

This section of the report will focus on the current status of the optimization and management components of application delivery. In particular, this section will identify both the optimization and management challenges that IT organizations find to be the most important as well as what they intend to do in the next year to respond to those challenges. This section is based in large part on two surveys that were administered in the first quarter of calendar year 2010 to over three hundred subscribers of Webtorials[29]. Throughout this report, the IT professionals who completed those surveys will be referred to as The Survey Respondents.

Optimization

Optimization Challenges

The Survey Respondents were asked to indicate how important it is to their organization to get better at seventeen optimization tasks over the next year. They were given the following five-point scale:

1. Not at all important
2. Slightly important
3. Moderately important
4. Very important
5. Extremely important

The Survey Respondents were also asked to indicate how difficult it would be for their organization to get better at each of the seventeen optimization tasks over the next year. They were given the following five-point scale:

1. Not difficult
2. Slightly difficult

[29] Additional survey results can be found at http://www.webtorials.com/content/2010/04/2010-app-del.html

3. Moderately difficult
4. Significantly difficult
5. Very significantly difficult

Table 6.1 shows the ten application delivery related optimization tasks that are the most important for IT organizations to improve on in the next year. Included in Table 6.1 are the task and the percentage of The Survey Respondents who indicated that the task was either very or extremely important to get better at over the next year. Also included is the percentage of The Survey Respondents who indicated that it would either be significantly or very significantly difficult for their IT organization to get better at the task in the next year.

Table 6.1: Top Ten Optimization Tasks

Optimization Task	Importance: Very or Extremely	Difficulty: Significant or Very Significant
Relating the performance of applications to their impact on the business	70%	39%
Ensuring acceptable performance for VoIP traffic	68%	29%
Improving the performance of applications used by mobile workers	60%	38%
Ensuring acceptable performance for video or telepresence traffic	57%	31%
Ensuring acceptable performance for the applications that you acquire from a Software-as-a-Service provider	56%	32%
Optimizing the performance of TCP	54%	19%
Controlling the cost of the WAN by reducing the amount of traffic that transits the WAN	50%	31%
Optimizing the Web tier of a multi-tiered application for peak utilization	50%	23%
Optimizing the performance of specific applications such as SharePoint	49%	19%
Optimizing the performance of protocols other than TCP; e.g., HTTP and MAPI	49%	24%

It is interesting to note that ensuring acceptable performance for VoIP traffic is the second most important optimization task and that ensuring acceptable performance for video or telepresence traffic is the fourth most important optimization task.

Ensuring acceptable performance of real time traffic is critically important to IT organizations.

An analysis of the right hand column in Table 6.1 (the column that indicates the difficulty of getting better at the task) shows that the relative difficulty of the tasks as indicated by

The Survey Respondents seems reasonable. For example, The Survey Respondents indicated that of the ten tasks listed in Table 6.1, "Relating the performance of applications to their impact on the business" was the most difficult and "Optimizing the performance of TCP" was the least difficult. This makes sense as the former task involves both implementing sophisticated tools and changing organizational culture while the subsequent task involves implementing well understood, somewhat narrowly focused technology.

One obvious conclusion that can be drawn from the data in Table 6.1 is that there are a large number of optimization tasks that are important for IT organizations to get better at in the next year. The data also indicates that IT organizations don't think that getting better at those tasks is very difficult. One possible reason for that is the *can do* attitude of most IT professionals. However, even if getting better at a single task, such as optimizing the performance of TCP, is not that difficult unto itself, marshalling the resources to get better at multiple tasks will be very difficult for most IT organizations.

The data in Table 6.1 does raise some interesting questions. One such question is that given that VoIP has been widely deployed for many years, why is it still so important to get better at ensuring acceptable performance for VoIP traffic if the task is not that difficult? Also, as previously mentioned, most Software-as-a-Service (SaaS) providers do not provide any form of an SLA relative to the performance of the applications that they provide. In addition, in the majority of instances users access applications from a SaaS provider by using the Internet, and there is no SLA associated with the use of the Internet. Because of factors such as these, it is reasonable to suggest that The Survey Respondents underestimated how difficult it will be to get better at certain tasks such as ensuring acceptable performance for the applications that they acquire from a SaaS provider.

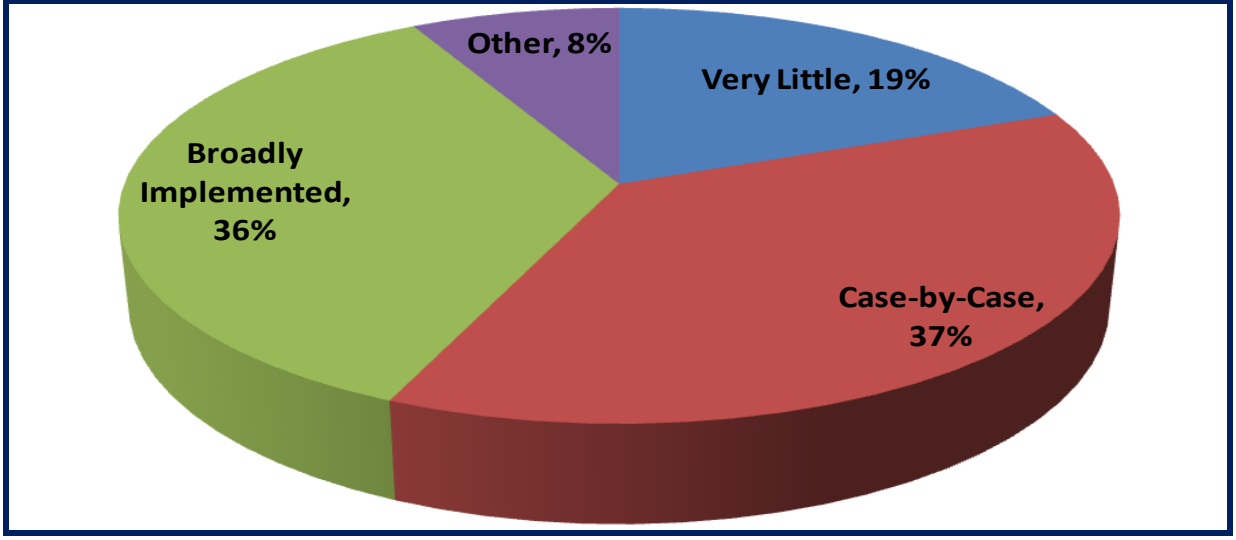
Optimization Solutions

The Survey Respondents were asked to indicate their organization's approach to optimizing network and application performance. They were given four choices:

- We implement very little if any functionality specifically to optimize network and application performance.
- We implement optimization functionality primarily on a case-by-case basis in response to high visibility problems.
- We have broadly implemented optimization functionality throughout our organization.
- Other

Their responses are shown in **Figure 6.1**.

Figure 6.1: Current Deployment of Optimization Functionality



The data in Figure 6.1 is a classic good news – bad news situation. The good news is:

Just over a third of IT organizations have broadly implemented optimization functionality throughout their organization.

The bad news is that two thirds of IT organizations have not.

In order to quantify the current and planned deployment of WAN Optimization Controllers (WOCs) The Survey Respondents were asked to indicate the percentage of locations (e.g., branch offices and end users' computers) to which their organization has either already deployed WOC functionality or will deploy it within the next year. The Survey Respondents were given three possible form factors:

- WOCs that are appliance based
- A soft or virtualized WOC
- A software client running on a user's desktop

Roughly two thirds of IT organizations have already deployed WOCs that are appliance based, although only slightly more than a quarter of IT organizations have deployed WOCs to the majority of their locations. The Survey Respondents indicated that they would make only a modest increase in the deployment of appliance based WOCs over the next year.

Less than half of The Survey Respondents indicated that their organization had either already deployed a soft WOC or a software client running on users' desktops. However, The Survey Respondents indicated that their IT organization would increase their deployment of both forms of optimization tools in the next year.

In order to quantify the current and planned deployment of Application Delivery Controllers (ADCs) The Survey Respondents were asked to indicate the percentage of their application traffic that is front-ended by an ADC either currently or within the next year. The Survey Respondents were given two possible form factors:

- Appliance based ADCs
- Software based ADCs

Over half of IT organizations already front-end at least some of their application traffic with an appliance based ADC, and roughly a quarter of IT organizations front-end the majority of their application traffic with an appliance based ADC. Both of those percentages will increase slightly over the next year. A much smaller percentage of IT organizations already front-end at least some of their application traffic with an software based ADC, and that percentage will increase slightly over the next year.

Management

Management Challenges

In addition to evaluating seventeen optimization tasks, The Survey Respondents were also asked to evaluate twenty management tasks. **Table 6.2** shows the ten management tasks that are the most important for IT organizations to improve on in the next year.

Table 6.2: Top Ten Management Tasks

Management Task	Importance: Very or Extremely	Difficulty: Significant or Very Significant
Rapidly identify the root cause of degraded application performance	76%	47%
Identify malicious traffic and eliminate it	71%	31%
Effectively manage QoS	67%	29%
Prevent large scale DDOS attacks	66%	32%
Identify the components of the IT infrastructure that support the company's critical business applications	66%	20%
Obtain performance indicator metrics and granular data that can be used to detect and eliminate impending problems	64%	38%
Effectively mange services, where services are comprised of multiple, inter-related applications	61%	39%
Effectively manage SLAs for one or more business critical applications	61%	35%
Obtain real-time, or nearly real-time, insight into how specific applications and end user sessions are performing	59%	43%
Track end user experience and relate it to factors such as Web response time	51%	40%

Similar to the format in the preceding section, included in Table 6.2 are the tasks and the percentage of The Survey Respondents who indicated that each task was either very or extremely important for their organization to get better at over the next year. Also included is the percentage of The Survey Respondents who indicated that it would either be significantly or very significantly difficult for their IT organization to get better at the task in the next year.

As was the case with optimization, the data in Table 6.2 indicates that there is a large number of management tasks that are important for IT organizations to get better at in the next year. Comparing the data in Table 6.1 with the data in Table 6.2 indicates that The Survey Respondents believe that getting better at the top ten management tasks will be more difficult than getting better at the top ten optimization tasks.

Given the previously discussed shifting emphasis and growing sophistication of cyber crime, it was not surprising to see that two of the management tasks that are most important to The Survey Respondents are identifying malicious traffic and eliminating it, and preventing large scale DDoS attacks. It was also not surprising that:

> ***In 2010 the most important management task for IT organizations to get better at is the rapid identification of the root cause of degraded application performance.***

That follows because as previously noted, in the current environment the end user typically notices application degradation before the IT organization does and when the IT organization is made aware of the fact that the performance of an application is degrading, it typically is unaware of the cause of the degradation. The fact that 47% of The Survey Respondents stated that getting better at identifying the root cause of degraded application performance would be either significantly difficult or very significantly difficult is in line with the fact that as previously discussed:

> ***The movement to adopt virtualization and cloud computing will make troubleshooting an order of magnitude more difficult than it is currently.***

One of the interesting results of the survey is that effectively managing services, where services are comprised of multiple, inter-related applications is very important to IT organizations. That is interesting in large part because over the last few years that has been considerable discussion about the fact that IT organizations need to move away from focusing on managing individual technology domains and need to focus more on managing the end-to-end performance of applications. There hasn't been anywhere near as much discussion about the need to focus on managing services where a service is comprised of multiple, inter-related applications.

The Survey Respondents were asked to identify the difficulty of several tasks that are associated with server virtualization. Their responses are shown in **Table 6.3**.

Table 6.3: Importance of Server Virtualization Tasks		
Server Virtualization Management Task	**Importance: Very or Extremely**	**Difficulty: Significant or Very Significant**
Perform traditional management tasks such as troubleshooting and performance management, on a per VM basis	49%	20%
Keep track of VMs as they are moved between physical servers	38%	17%
Dynamically move VMs, and all of the supporting management functionality, between physical servers	37%	20%
Discover VMs	33%	12%
Manage the traffic that goes between virtual machines (VMs) on a single physical server	31%	23%

As shown in Table 6.3, The Survey Respondents indicated that getting better at many of the individual challenges associated with server virtualization is important to their organization. In addition, it is reasonable to look at the five challenges contained in Table 6.3 as being a single challenge - managing server virtualization. When looked at that way, getting better at server virtualization is extremely important to The Survey Respondents. However, The Survey Respondents don't find any of the individual challenges associated with server virtualization as being very difficult. The reality is that some tasks such as dynamically moving VMs, and all of the supporting management functionality between physical servers, is extremely difficult.

Management Solutions

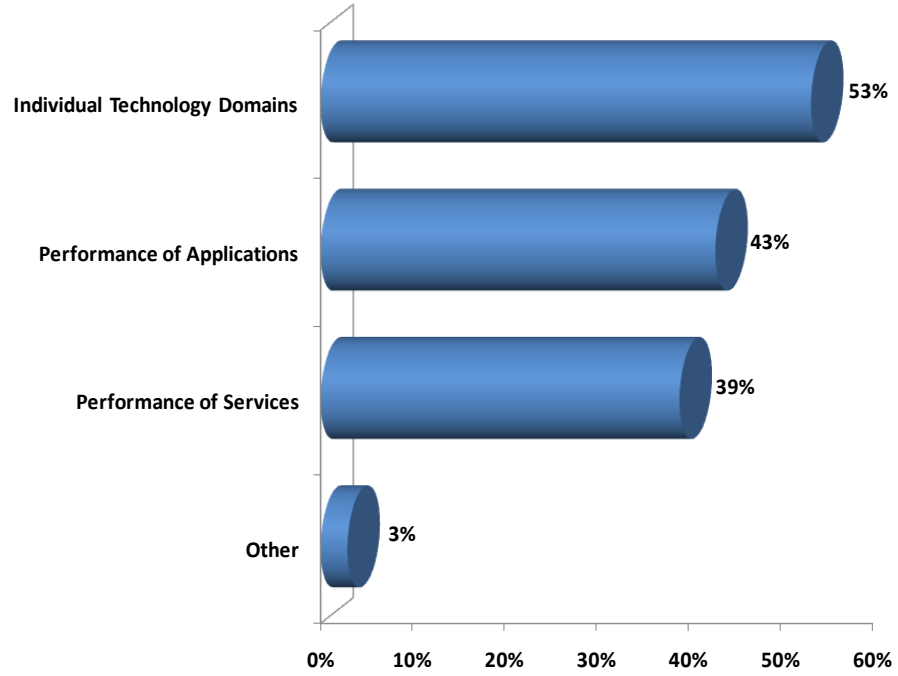
The Survey Respondents were asked to indicate the approach their organization takes to management. They were given the following choices and allowed to choose all that applied to their environment.

- We have a focus primarily on individual technology domains such as LAN, WAN and servers
- We have a focus on managing the performance of applications as seen by the end user
- We have a focus on managing the performance of services as seen by the end user, where service refers to multiple, inter-related applications
- Other

Their responses are shown in **Figure 6.2.**

Roughly half of The Survey Respondents indicated that their organization has a well-understood, effective strategy that describes how they will manage each component of IT. Roughly half of The Survey Respondents also indicated that the effective management of IT is a very important topic for their senior IT managers. It was somewhat of a surprise to see that whether or not an organization's senior IT managers regard effective management as being very important didn't significantly impact whether or not that IT organization had a well-understood, effective management strategy.

Figure 6.2: Focus of Management



7.0 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."

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. As previously mentioned, roughly fifty percent of IT organizations currently perform one such function: the establishment of SLAs for at least a core set of business applications. Unfortunately, as was also previously discussed, the internal SLAs offered by most IT organizations are weak and IT organizations don't have the tools and processes to effectively manage them. The steps below are intended to be a framework for a process that IT

organizations can use to better manage internal SLAs for a select set of business critical applications (The Applications) as well as for new business critical applications.

- Establish SLAs for The Applications.
- Identify the key elements (e.g., specific switches and routers) for each component of the IT infrastructure (e.g., servers, databases, networks) that support each of The Applications.
- Baseline the performance of each of The Applications.
- Baseline the performance of the key elements for each component of the IT infrastructure that support each of The Applications.
- Establish SLAs for the key elements for each component of the IT infrastructure that support each of The Applications. As explained in the next subsection, do this in such a way that if the SLAs for each key element are met, then the SLA for each of The Applications is met.
- Establish application design guidelines that ensure that each of The Applications will perform optimally when run over a WAN.
- Profile an application prior to deploying it, including running it in conjunction with a WAN emulator, to quantify the performance that can be expected.
- Perform an assessment of the IT infrastructure to support a new application prior to deploying it.
- Identify in advance the impact of a change to each of The Applications.

WAN Emulation

Chapter 3 (Application Delivery Challenges) 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 a WAN emulator. Used during application development and quality assurance (QA), these tools simulate 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 an 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 7.1** depicts the results of a lab test done using a WAN emulator. The emulator was used to perform one of the steps (e.g., Baseline the performance of the key elements for each component of the IT infrastructure that support each of The Applications) in the preceding framework for how to better manage internal SLAs. In

particular, the emulator was used to quantify the affect that WAN latency would have on an inquiry-response application that has a target response time of 5 seconds.

Table 7.1: Impact of Latency on Application Performance

Network Roundtrip Latency	Measured Response Time
0 ms	2 seconds
25 ms	2 seconds
50 ms	2 seconds
75 ms	3 seconds
100 ms	7 seconds
125 ms	10 seconds
150 ms	18 seconds

As Table 7.1 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. However, if the network latency goes above 75 ms, the response time of the application degrades rapidly. Referring to the preceding framework for how to better manage internal SLAs, any IT organization that is implementing this application needs to establish an SLA for the WAN that calls for the roundtrip delay to not exceed 75 ms.

Using a WAN emulator 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 profiles 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.

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, which enables the IT organization to identify abnormal behavior. Abnormal behavior could be the result of an operational issue, a security violation, or both.

The Key Steps

Four principal steps comprise baselining:

I. Identify the Key IT Resources

The key IT resources are the resources that support the company's critical business applications; e.g., The Applications.

II. Quantify the Utilization of the IT resources over a Sufficient Period of Time

Organizations must compute the baseline over a normal business cycle. In most cases, baselining focuses on measuring the utilization of resources. Utilization is a surrogate for what is actually needed:

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

III. Determine how the Organization Uses its 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?

IV. Effective Use of the Information

The information gained from baselining has many uses, including capacity planning, budget planning and chargeback/showback. 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.

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.

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. 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. It is also important to categorize those applications using an approach similar to that described in Chapter 2 (Taxonomy of Applications). Part of the value of this activity is to identify unauthorized use of the network; i.e., on-line gaming and streaming radio or video. Blocking recreational use can free up additional WAN bandwidth. 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. As noted, companies should modify how they think about baselining to focus not on utilization, but on delay. In some instances, however, measuring delay is not enough. If, for example, a company is about to deploy VoIP then the pre-assessment baseline must also measure the current levels of jitter and packet loss, because 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. 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. 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.

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.

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:

- 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.
- 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 the 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, complex 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 11, 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.

Route Analytics

Another class of management tool that can facilitate the integration of planning and operations is typified by an IP route analytics solution[30]. The goal of route analytics is to

[30] More information on this topic can be found at: http://www.webtorials.com/content/2008/12/the-mandate-to-better-integrate-network-planning-and-network-operations.html

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. 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.

Planning for Cloud Computing

Most IT organizations that have already implemented either public or private cloud computing have not done so in a highly systematic fashion. In some cases, they used a trial and error approach to choosing a SaaS provider, while in other cases they evaluated one aspect of private cloud computing (e.g., server virtualization) without considering other aspects of private cloud computing and did not plan for the impact that server virtualization would have on other components of IT, such as management or the design of the data center LAN.

In order to maximize the benefit of cloud computing, IT organizations need to develop a plan (The Cloud Computing Plan) that they update on a regular basis. The Cloud Computing Plan should identify the opportunities and risks associated with both public and private cloud computing. The Cloud Computing Plan must identify a roadmap of what steps the IT organization will take on a quarter-by-quarter basis for the next two to three years and ensure that the steps are in line with the corporate culture. This includes identifying:

- What functionality (e.g., applications, storage) needs to remain under the tight control of the IT organization and what functionality is appropriate to hand over to a CCSP.
- What levels of service are *good enough* for each class of application and for the myriad storage requirements.
- How the IT organization will evolve over time the twelve characteristics of a cloud computing solutions; e.g., virtualization, automation, simplification.

- How the IT organizations will evolve its data center LAN architecture to support cloud computing.
- How the IT organizations will evolve its use of WAN services to support cloud computing.
- How the IT organization will minimize the security and confidentiality risks associated with public cloud computing services.
- What management functionality must be present in the management domain controlled by the IT organization as well as provided by the relevant network service providers and CCSP(s).
- How the IT organization will overcome potential performance bottlenecks.

The Cloud Computing Plan should look systematically across multiple technologies because of the interconnected nature of the technologies. As part of creating this plan, IT organizations need to understand the cloud computing strategy of their existing and potential suppliers, including the partnerships that the suppliers are establishing between and amongst themselves.

8.0 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;
- Overcome the packet delivery issues that are common in shared (i.e., over-subscribed) networks;
- Mitigate the inefficiencies of protocols;
- Offload computationally intensive tasks from client systems and servers

As is explained in the following section of the handbook, some managed service providers offer network and application optimization as a service. It is also possible for IT organizations to acquire network and application optimization products. As is also explained in the following section, these two approaches are complimentary.

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). WOCs are often referred to as *symmetric solutions* because they typically require functionality in both the data center as well as the branch office.

The second category of products is often referred to as an 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.

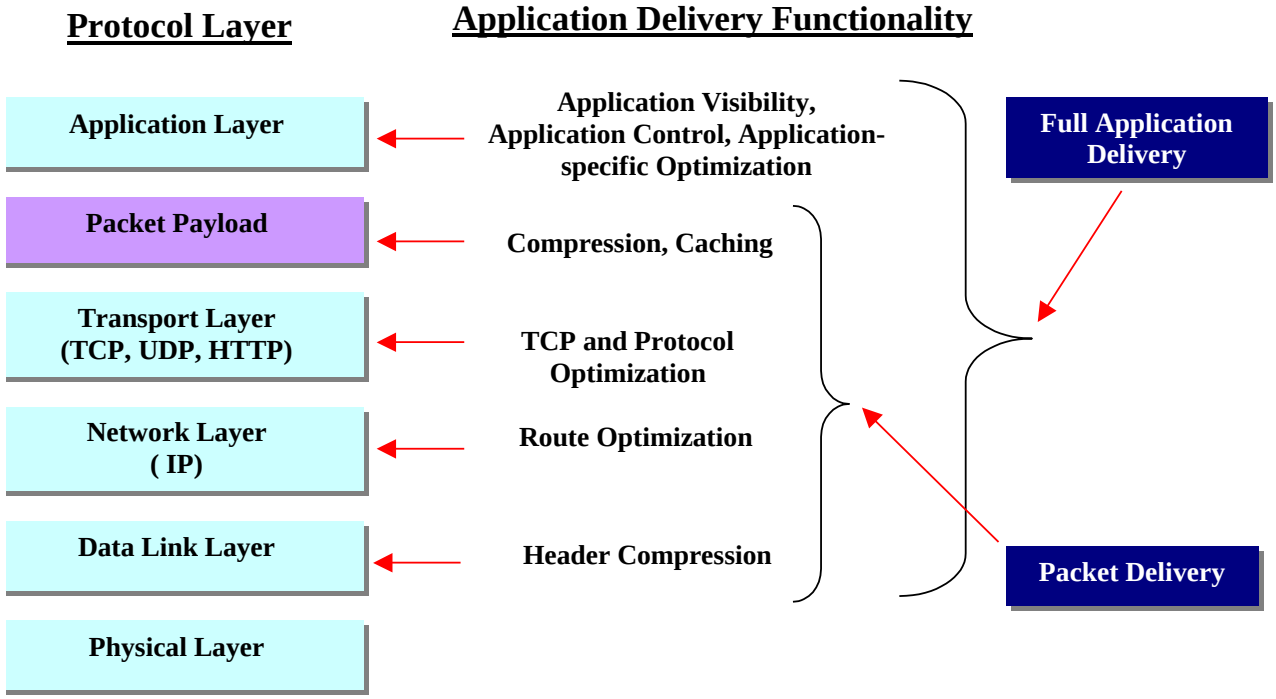
Application Delivery Network Defined

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
- Increase the ability of the IT organization to more large volumes of data across the WAN such as when performing backups

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 8.1**.

Figure 8.1: 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.

In order to choose the most appropriate optimization solution, IT organizations need to understand their environment. For example, consider the types of advanced compression solutions that are available. The effectiveness of these solutions depends on two factors. One is the quality of the compression techniques implemented in a solution. However, 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 these solutions is the amount of redundancy in 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.

Quantifying Application Response Time

A model is helpful to illustrate the potential performance bottlenecks in the performance of an application. The following model (**Figure 8.2**) is a variation of the application response time model created by Sevcik and Wetzel[31]. 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).

Figure 8.2: Application Response Time Model

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

The WOCs and ADCs that are described below are intended to mitigate the impact of the factors in the preceding equation.

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. **Table 8.1** lists some of WAN characteristics that impact application delivery and identifies WAN optimization techniques that can be used to mitigate the impact of the WAN.

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

Table 8.1: Techniques to Improve Application Performance

WAN Characteristics	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) Packet Reordering
Network Contention	Quality of Service (QoS)

Below is a description of 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.

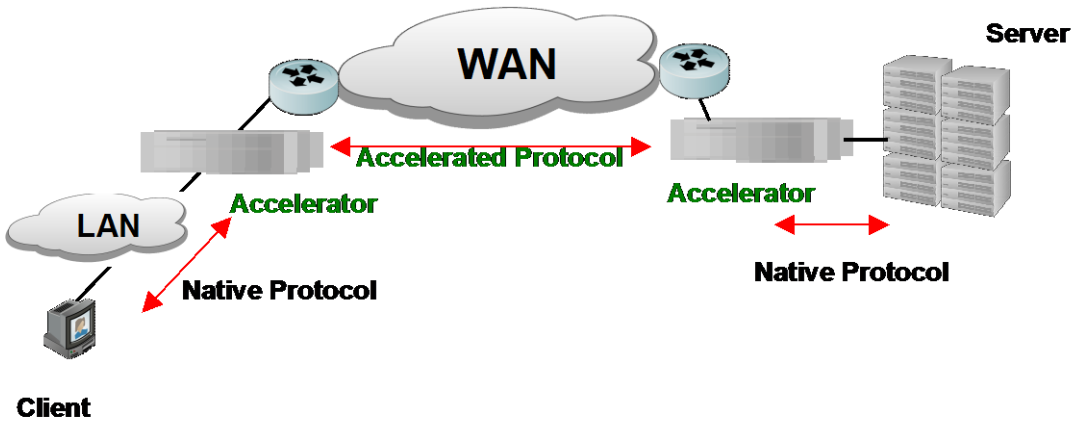
Forward Error Correction (FEC)

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. A subsequent subsection 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 8.3**. 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 8.3: 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.

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.

WOC Selection Criteria

The recommended criteria for evaluating WAN Optimization Controllers are listed in **Table 8.2**. 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 8.2 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 (Σ 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.

Table 8.2: Criteria for WAN Optimization Solutions

Criterion	Weight Wi	Score for Solution "A" Ai	Score for Solution "B" 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			
Bulk Data Transfers			
Support for Meshed Traffic			
Support for Real Time Traffic			
Individual and/or Mobile Clients			
Branch Office Consolidation			
Total Score		Σ WiAi	Σ WiBi

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 and so might be negatively impacted by changes in the application.

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.

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 WOCs 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, or 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. Also, as discussed in section 4, the adoption of VDI means that protocols such as ICA, RDP and PCoIP need to be supported. As a result, if VDI is being deployed, WOC performance for remote display protocols should be a significant evaluation criterion.

In addition to evaluation how a WOC improves the performance of a protocol, it is also important to determine if the WOC 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 example, 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.

Bulk Data Transfers

An additional benefit of server virtualization is the efficiency it lends to disaster recovery and backup operations. Virtual images of mission critical applications can be maintained at backup data centers or the data centers of providers of public cloud-based backup/recovery services or cloud-based virtual data centers. Compared to conventional files, the virtual image has the added benefit of being readily restarted on another virtualized server to quickly resume business operations. Client-side application virtualization also involves high volume data

transfers from the data center to the remote site. WOCs optimized for backup operations and other bulk transfers among sites may have special features, such as FEC (discussed in the next section), recovery from out of order packets, enhanced data de-duplication capabilities, support for specific backup applications, and support for specialized transfer protocols.

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. Most real-time applications use UDP, not TCP, as a transport protocol. As a result, they are not significantly addressed by TCP-only acceleration solutions. In addition, the payloads of VoIP and live video packets can't be compressed by the WOC because of the delay sensitive nature of the traffic and the fact that these streams are typically already highly compressed. WOC support for UDP real-time traffic is therefore generally provided in the form of header compression, QoS, and forward error correction. As the WOC performs these functions, it must be able to do so without adding a significant amount of latency.

Individual and/or Mobile Clients

As the enterprise workforce continues to become more mobile and more de-centralized, accessing enterprise applications from mobile devices or home offices is becoming a more common requirement. As discussed in section 4, accelerating application delivery to these remote users involves a soft WOC or WOC client that is compatible with a range of remote devices, including laptops, PDAs, and smart phones. The WOC client must also be compatible with at least a subset of the functionality offered by the data center WOC. Another issue with WOC clients is whether the software can be integrated with other client software that the enterprise requires to be installed on the remote device. Installation and maintenance of numerous separate pieces of client software on remote devices can become a significant burden for the IT support staff.

Branch Office Platform

As previously noted, many enterprises are consolidating servers into a small number of central sites in order to cut costs and to improve the manageability of the branch office IT resources. Another aspect of branch office consolidation is minimizing the number of standalone network devices and hardware appliances in the branch office network. As noted in section 4, one approach to branch office consolidation is to install a virtualized server at the branch office that provides local

services and also supports virtual appliances for various network functions. A variation on this consolidation strategy involves using the WOC as an integrated (or virtualized) platform that supports a local branch office server and possibly other networking functions, such as DNS and/or DHCP. Another variation is to have WOC functionality integrated into the router in the branch office.

The Data Replication Bottleneck

While packet loss and out of order packets are merely a nuisance for a network that supports typical data applications[32] such as file transfer and email, it is a very serious problem when performing bulk transfers for 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 re-transmitting 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. As discussed in section 5 (Cloud Computing):

> ***Efficient bulk transfers and data replication are critical requirements to gain many of the potential benefits of both private and public cloud computing.***

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[33]. 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:

[32] 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.

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

Figure 8.4: Factors that Impact Throughput

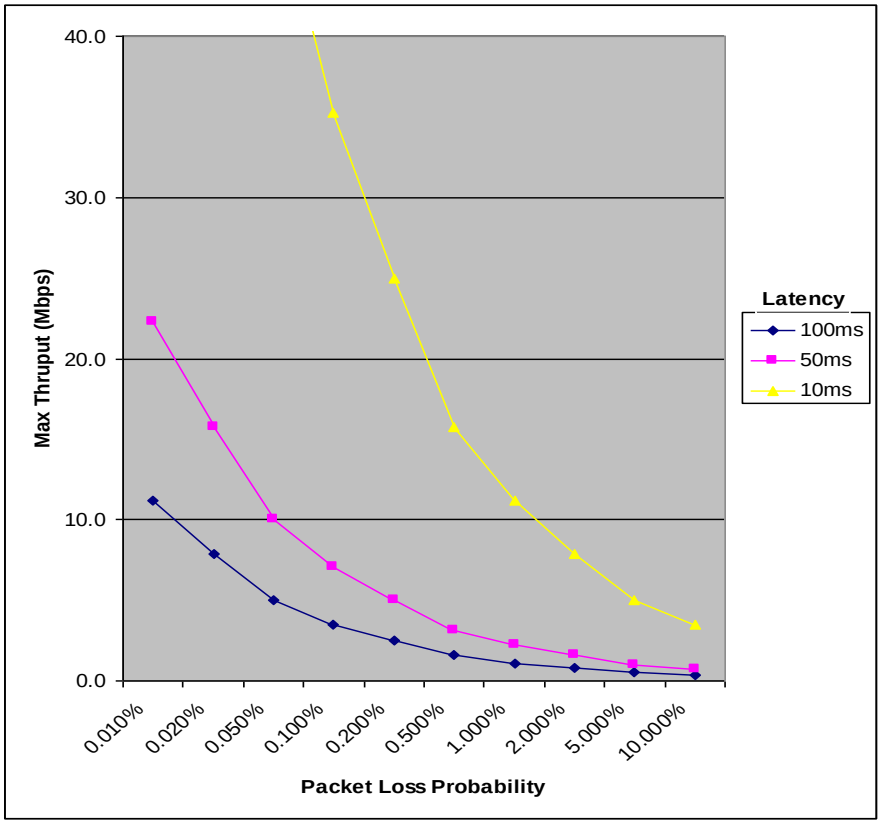
$$Throughput <= (MSS/RTT)*(1/sqrt\{p\})$$

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 8.5** 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.

Figure 8.5: Impact of Packet Loss on Throughput



One conclusion we can draw from Figure 8.5 is:

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

For example, on a WAN link 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. Section 9 expands on the impact of packet loss and out of order packets on throughput.

Techniques for Coping with Loss and Out of Order Packets

The data in Figure 8.5 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)[34], 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 dynamically 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 was mentioned earlier in this section, 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

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

efficiency and provide asymmetrical functionality to accelerate the delivery of applications from the data center to individual remote users. In particular, the ADC can allow a number of compute-intensive functions, such as SSL processing and TCP session processing, to be offloaded from the server. Server offload can increase the transaction capacity of each server, reducing the number of servers required for a given level of business activity. Where the ADC uses purpose-built hardware to execute the compute-intensive functions, each ADC can support more transactions and therefore more servers, further increasing the value of the ADC deployment.

An ADC provides more sophisticated functionality than a SLB does.

The deployment of an SLB enables an IT organization to get a *linear benefit* out of its servers. That means that if an IT organization that has implemented an SLB doubles the number of servers supported by that SLB that it should be able to roughly double the number of transactions that it supports. The traffic at most Web sites, however, is not growing at a linear rate, but at an exponential rate. To exemplify the type of problem this creates, assume that the traffic at a hypothetical company's (Acme) Web site doubles every year. If Acme's IT organization has deployed a linear solution, such as an SLB, after three years it will have to deploy eight times as many servers as it originally had in order to support the increased traffic. However, if Acme's IT organization were to deploy an effective ADC then after three years it would still have to increase the number of servers it supports, but only by a factor of two or three – not a factor of eight. The phrase ***effective ADC*** refers to the ability of an ADC to have all features turned on and still support the peak traffic load.

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.

- **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**

XML is a verbose protocol that is CPU-intensive. Hence, another function that can be provided by the ADC is to offload XML processing from the servers by serving as an XML gateway.

- **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. 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.

- **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 a cache in the ADC 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 offloads the server TCP session processing by terminating the client-side TCP sessions and multiplexing numerous short-lived network sessions initiated as client-side object requests into a single longer-lived session between the ADC and the Web servers. Within a virtualized server environment the importance of TCP offload is amplified significantly because of the higher levels of physical server utilization that virtualization enables. Physical servers with high levels of utilization will typically support significantly more TCP sessions and therefore more TCP processing overhead.

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.

- **Support for Server Virtualization**

With server virtualization, there are two primary tasks associated with the dynamic creation of a new VM. The first task is the spawning of the new VM and the second task is ensuring that the network switches, firewalls, and ADCs are properly configured to direct and control traffic destined for that VM. For the ADC (and other devices) the required configuration changes are typically communicated from an external agent via one of the control APIs that the device supports. These APIs are usually based on SOAP, a CLI script, or direct reconfiguration. The exteranl agent could be a start-up script inside of the VM or it could be the provisioning/management agent that initiated the provisioning of the VM. The provisioning/management agent could be part of an external workflow orchestration system or it could be part of the orchestration function within the hypervisor management system. Preferably the process of configuring network elements, including ADCs, to support new VMs and the movement of VMs within the data center can readily be automated and integrated within the enterprise's overall architecture for managing the virtualized server environment.

ADC Selection Criteria

The ADC evaluation criteria are listed in **Table 8.3**. 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.

Table 8.3: Criteria for Evaluating ADCs

Criterion	Weight Wi	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		Σ WiAi	Σ WiBi

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. When choosing an ADC, IT organizations obviously need to understand the features that it supports. However, as this class of product continues to mature, the distinction between the features provided by competing products is lessening. This means that when choosing an ADC, IT organizations should pay the most attention to the ability of the ADC to have all features turned on and still support the peak traffic load, because this is what determines the ability of the ADC to cost effectively support exponential traffic growth. In order to support the peak traffic load with all features turned on, the ADC should perform the most computationally intensive tasks, such as TCP offload, in hardware that is itself purpose-built for the task.

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.

As is the case with WOCs, 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 production application environment where it will be installed. As noted above, an important part of these trails is to identify any performance degradation that may occur as the full suite of desired features and functions are activated or as changes are 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. That said, it is still very important to be able to deploy an ADC 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 should 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 has become a key technology for realizing data center consolidation and its related benefits. The degree of integration of an ADC's configuration management capabilities with the rest of the solution for managing the virtualized environment may be an important selection criterion. For example, it is important to know how the ADC interfaces with the management system of whatever hypervisors that the IT organization currently supports, or expects to support in the near term. It is also important to know how the ADC supports the creation and movement of VMs within a dynamic production environment.

Section 4 (Virtualization) described one way of virtualizing an ADC. 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.

Security

The ADC 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, 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.

9.0 Managed Service Providers

Driven by factors such as virtualization and cloud computing, IT 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 in part 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.

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.

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:

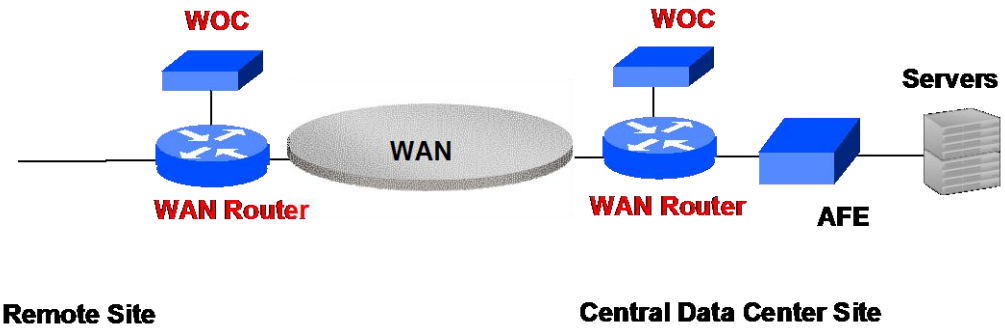
1. Site-based services comprised of managed WOCs and/or ADCs installed at participating enterprise sites

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 9.1**. 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 8.

Figure 9.1: 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, 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[35].

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.

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[36], 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 tends to be availability and performance bottlenecks at the peering points. Another affect is that since there is not a single, end-to-end provider,

[35] Private WAN services refers to services such as private lines, Frame Relay, ATM and MPLS.
[36] http://en.wikipedia.org/wiki/Internet

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 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 section 8. 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:

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

where:

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

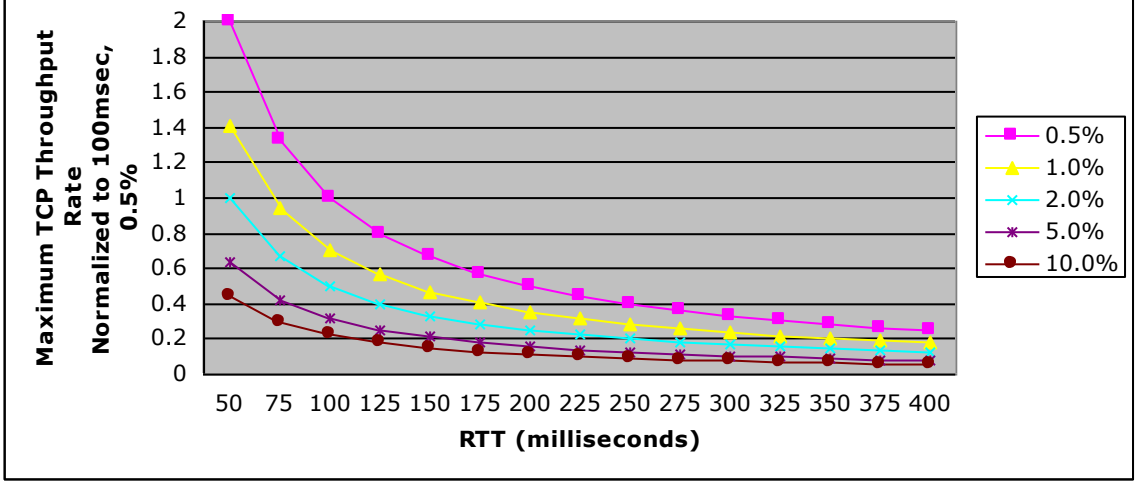
TCP throughput on a single session decreases as either the round trip time or the packet loss increases.

Figure 9.2 demonstrates the impact of delay and packet loss on TCP throughput given an MSS of 1,460 bytes. The data in Figure 9.2 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 9.2 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.

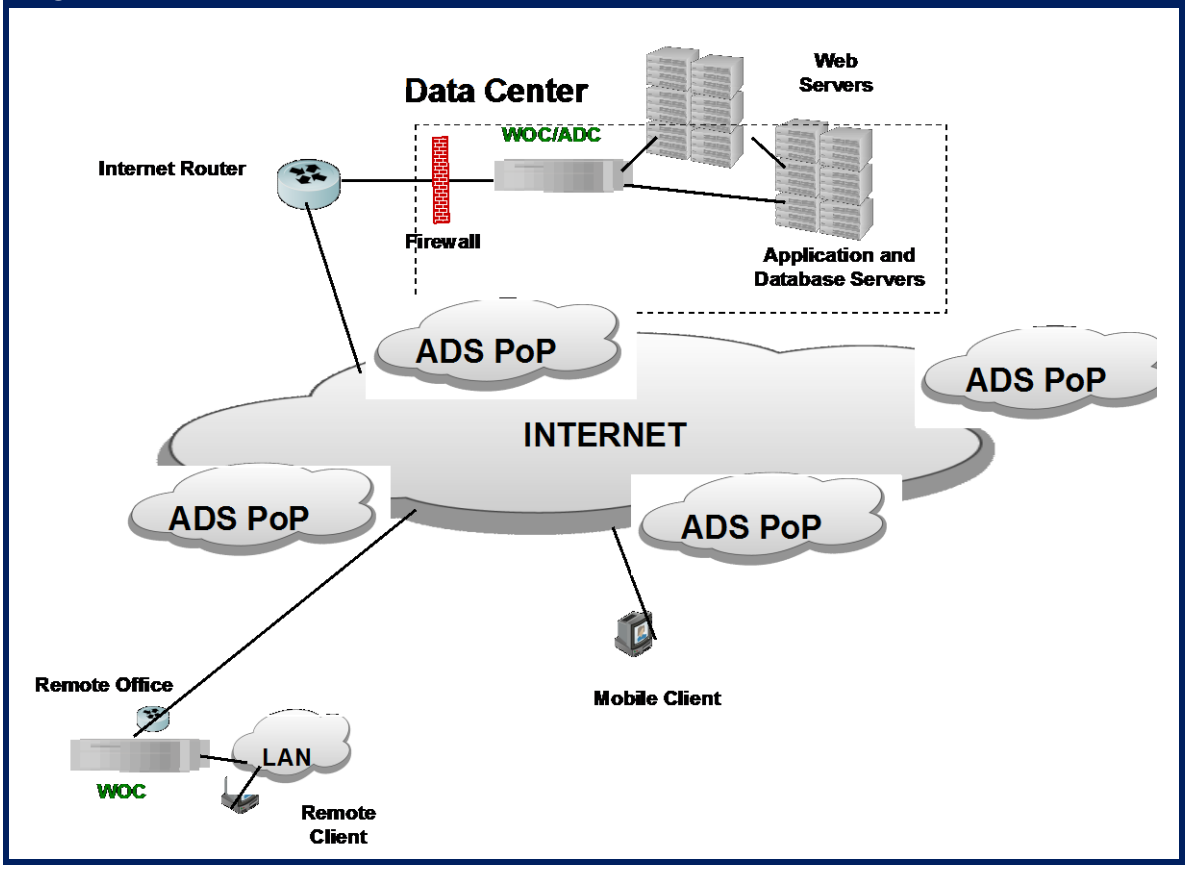
Figure 9.2: Impact of Delay and Packet Loss on TCP Throughput



Internet-Based Application Delivery Optimization

The traditional classes of application delivery solutions (e.g., ADC, WOC) that were described in Chapter 8 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 9.3**, 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 9.3: 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.

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.

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.

Web Application Firewall Services

As described in section 3, one of the characteristics of the Application Delivery 2.0 era is the shifting emphasis and growing sophistication of cyber crime.

Role of a Traditional Firewall: Protect the Perimeter

Roughly twenty years ago IT organizations began to implement the first generation of network firewalls, which were referred to as packet filters. These devices were placed at the perimeter of the organization with the hope that they would prevent malicious activities from causing harm to the organization.

Today most network firewalls are based on stateful inspection. A stateful firewall holds in memory attributes of each connection. These attributes include such details as the IP addresses and ports involved in the connection and the sequence numbers of the packets traversing the connection. One of the weaknesses associated with network firewalls is that they are typically configured to open up ports 80 and 443 in order to allow passage of all HTTP and SSL traffic. Given that ports 80 and 443 are generally configured to be open, this form of perimeter defense is porous at best.

Whereas network firewalls are focused on parameters such as IP address and port numbers, a more recent class of firewall, referred to as a Web application firewall, analyzes messages at layer 7 of the OSI model. Web application firewalls are typically deployed as a hardware appliance and they sit behind the network firewall and in front of the Web servers. They look for violations in the organization's established security policy. For example, the firewall may look for abnormal behavior, or signs of a known attack. It may also be configured to block specified content, such as certain websites or attempts to exploit known security vulnerabilities. Because of their ability to perform deep packet inspection at layer 7 of the OSI model, a Web application firewall provides a level of security that cannot be provided by a network firewall. Web application firewalls will be discussed in detail in section 11.

Defense in Depth: The Role of a Web Application Firewall Service

There are fundamental flaws with an approach to security that focuses only on the perimeter of the organization. To overcome these flaws, most IT organizations have moved to an approach to security that is typically referred to as *defense in depth*. The concept of defense in depth is not new. This approach was widely used during the Application Delivery 1.0 era as IT organizations often deployed multiple layers of security functionality including virus scanning, authentication, firewalls, intrusion detection systems and intrusion protection systems. In the Application Delivery 1.0 era, however, all of the layers of security functionality were typically deployed onsite. What is new in the Application Delivery 2.0 era is the deployment of a layer of security, such as a Web application firewall service, that is distributed throughout the Internet so that it is close to the source of security attacks and hence can prevent many security attacks from reaching the organization. The distribution of security functionality on the part of a Web application firewall service is analogous to the distribution of optimization functionality on the part of an application delivery service that was discussed in the preceding subsection.

As described in section 3, in the current environment DDoS attacks can generate 50 Gbps of traffic. Attacks of this magnitude cannot be prevented by onsite solutions. They can, however, be prevented by implementing security functionality such as a Web application firewall service that identifies and mitigates the DDoS-related traffic close to attack traffic origin.

There is a wide range of ways that a DDoS attack can cause harm to an organization in a number of ways, including the:

- Consumption of computational resources, such as bandwidth, disk space, or processor time.
- Disruption of configuration information, such as routing information.
- Disruption of state information, such as the unsolicited resetting of TCP sessions.
- Disruption of physical network components.
- Obstructing the communication media between the intended users and the victim so that they can no longer communicate adequately.

Because there are a variety of possible DDoS attacks, IT organizations need to implement a variety of defense in depth techniques. This includes:

- Minimizing the points of vulnerability
  If an organization has most or all of its important assets in a small number of locations, this makes the organization more vulnerable to successfully being attacked as the attacker has fewer sites on which to concentrate their attack.

- Protecting DNS
  Many IT organizations implement just two or three DNS servers. As such, DNS is an example of what was discussed in the preceding bullet – how IT organization are vulnerable because their key assets are located in a small number of locations.

- Implementing robust, multi-tiered failover
  Many IT organizations have implemented disaster recovery plans that call for there to be a stand-by data center that can support at least some of the organization's key applications if the primary data center fails. Distributing this functionality around a global network increases overall availability in general, and dramatically reduces the chance of an outage due to a DDoS attack in particular.

  In order to be effective, a Web application firewall service needs to be deployed as broadly as possible, preferably in tens of thousands of locations. When responding to an attack, a Web application firewall service must also be able to:

- Block or redirect requests based on characteristics such as the originating geographic location and whether or not the originating IP addresses are on either a whitelist or a blacklist.
- Direct traffic away from specific servers or regions under attack.
- Issue slow responses to the machines conducting the attack. The goal of this technique, known as tarpits[37], is to shut down the attacking machines while minimizing the impact on legitimate users.
- Direct the attack traffic back to the requesting machine at the DNS or HTTP level.

An ADC that supports Web application firewall functionality is complimentary to an ADS that supports a Web application firewall service. That follows because while a Web application firewall service can perform many security functions that cannot be performed by a Web application firewall, there are some security functions that are best performed by a Web application firewall. An example of that is protecting an organization against information leakage by having an onsite Web application firewall perform deep packet inspection to detect if sensitive data such as a social security number or a credit card number is leaving the site. If sensitive data is leaving the site, the onsite Web application firewall, in conjunction with other security devices, can determine if that is authorized and if it is not, prevent the data from leaving the site.

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

[37] http://en.wikipedia.org/wiki/Tarpit_(networking)

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? Are they acceptable?
- 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?

10.0 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:

- (As previously discussed) The 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.

Below is a discussion of some of the factors that complicate and/or enable IT organizations to perform the tasks listed above.

The Port 80 Black Hole

As noted above, 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 (see below).

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. As was previously noted, 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 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 standard SIP port or a standard 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

The IT 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.

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 previously discussed, when implementing techniques to gain end-to-end visibility IT organizations 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.

Relative to choosing an end-to-end visibility solution, 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 have attempted to manage performance was by setting static threshold performance-based alarms. The vast majority of IT organizations that set thresholds set them against WAN traffic utilization or some other network parameter. Relatively few IT organizations have traditionally set thresholds against application-response time.

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

- If the network is heavily utilized, the applications are performing poorly.
- 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.

Application management should focus directly on the characteristics of the business critical application and not just on factors that have the potential to influence generic application performance.

The vast majority of IT organizations that set thresholds set them at a high water mark which has the affect of minimizing the number of alarms that the IT organization receives. While this approach makes sense operationally, it leads to an obvious conclusion:

Most IT organizations ignore the majority of the performance alarms.

Proactive Alarms

Because the majority of IT organizations that implement static performance alarms set thresholds at the high water mark, this means that the use of static performance alarms is reactive. In particular, 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 previously defined concept of baselining and applies this concept 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

Section 7 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. In simple networks logical factors are typically not a significant source of application degradation. However, in large complex networks:

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 subsection 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.

Another 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. One more 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.

11.0 Control

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 previously discussed in the context 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:

- Measurement
  Measure the performance (i.e., availability, delay, packet loss, and jitter) of each path through the network.
- Analysis and Decision Making
  Use the performance measurements to determine the best path. This analysis must occur in real time.
- Automatic Route Updates
  Once the decision has been made to change paths, update the routers to reflect the change.
- Reporting
  Report on the performance of each path as well as the overall route optimization process.

SSL VPN Gateways

The SSL protocol[38] 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 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 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:

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

- 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.

> ***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 previously discussed 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.

Web Application Firewall

Section 9 (Managed Service Providers) discussed the value of a Web application firewall service as a compliment to an onsite Web application firewall. This subsection will discuss the role of a Web application firewall in detail.

Traditional 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.

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

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.

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 previously mentioned, some applications such as Skype 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[39].

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.

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

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

that is deemed to be *bad*. This model had some 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.

12.0 Conclusion

Ensuring acceptable application delivery was never easy. However the ongoing emergence of a new generation of challenges (e.g., virtualization, cloud computing, sophisticated mobile workers and the shifting emphasis and growing sophistication of cyber crime) will dramatically increase the difficulty of ensuring acceptable application delivery. In order to be successful in this challenging environment, IT organizations need to develop a systematic approach to applications delivery. To help IT organizations with this task, this handbook identified a number of conclusions that an IT organization can use when formulating their approach to ensuring acceptable application delivery. Those conclusions are:

- Successful application delivery requires the integration of:
  - Planning
  - Network and application optimization
  - Management
  - Control.
- Application delivery is more complex than merely accelerating the performance of all applications
- 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.

- A relatively small increase in network delay can result a significant increase in application delay.
- While server consolidation produces many benefits, it can also produce some significant performance issues.
- One of the effects of data center consolidation and single hosting is that it results in additional WAN latency for remote users.
- In the vast majority of situations, when people access an application they are accessing it over the WAN instead of the LAN.
- In the vast majority of instances when the performance of a key business application is degrading, the end user, not the IT organization, first notices the degradation.
- As the complexity of the environment increases, the number of sources of delay increases and the probability of application degradation increases in a non-linear fashion.
- As the complexity increases the amount of time it takes to find the root cause of degraded application performance increases.
- Troubleshooting in a virtualized environment is notably more difficult than troubleshooting in a traditional environment.
- From a networking perspective, the primary challenge in implementing desktop virtualization is achieving adequate performance and an acceptable user experience for client-to-server connections over a WAN.
- IT organizations that are implementing virtualized desktops should analyze the viability of implementing WAN and application optimization solutions.
- Supporting desktop virtualization will require that IT organizations are able to apply the right mix of optimization technologies for each situation.
- An important set of synergies exist between virtual servers, virtual desktops and virtual appliances such as a WOC or a performance monitoring solution.
- The goal of cloud computing is to enable IT organizations to achieve an order of magnitude improvement in the cost effective, elastic provisioning of IT services that are good enough.
- Troubleshooting in a hybrid cloud environment will be an order of magnitude more difficult than troubleshooting in a traditional environment.
- The emerging view of private cloud computing requires the ability to move massive files between the IT organization and the CCSPs in a small amount of time.
- The biggest risk accrues to those companies that don't implement any form of cloud computing.
- Ensuring acceptable performance of real time traffic is critically important to IT organizations.
- Just over a third of IT organizations have broadly implemented optimization functionality throughout their organization.
- In 2010 the most important management task for IT organizations to get better at is the rapid identification of the root cause of degraded application performance.

- The movement to adopt virtualization and cloud computing will make troubleshooting an order of magnitude more difficult than it is currently.
- Hope is not a strategy. Successful application delivery requires careful planning, coupled with extensive measurements and effective proactive and reactive processes.
- The application delivery function needs to be involved early in the application development cycle.
- IT organizations need to modify their baselining activities to focus directly on delay.
- Organizations should baseline by measuring 100% of the actual traffic from the real users.
- Understanding the performance gains of any network and application optimization solution requires testing in an environment that closely reflects the live environment.
- Efficient bulk transfers and data replication are critical requirements to gain many of the potential benefits of both private and public cloud computing.
- 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.
- 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 the mouse button when they receive a response from an application.
- Application management should focus directly on the characteristics of the business critical application and not just on factors that have the potential to influence generic 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.
- 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.

Ipanema: Make a 1,000 Site Network Feel Like One.

ipanema Technologies

Ipanema enables any large enterprise to have full control of their global network. Our unique, patented, Autonomic Networking system guarantees business application performance and continuity in all enterprise environments. Ipanema simplifies network operations, reduces IT costs, and guarantees optimal performance of your critical business applications. Using Ipanema, a 1,000 site network runs as one.

- Guarantee user experience
  Far beyond the capabilities of traditional WAN technologies, the Ipanema System is self-learning, self-adapting and self-healing. It provides an unprecedented ability to guarantee end-user Quality of Experience (QoE) with critical business applications, regardless of traffic or application mix.
- Accelerate business applications
  Through state-of-the-art WAN Optimization, Ipanema unleashes the power to rapidly deploy business applications, dramatically reduce application response times for end users while automatically controlling traffic in real time.
- Unify hybrid networks
  With Ipanema's Hybrid Network Unification and Dynamic WAN Selection technology, hybrid [MPLS + Internet] networks become flexible assets that any enterprise can rely on for more effective business communications.

- Save on IT costs
  Enterprise WANs are pivotal to executing successful business strategies. Ipanema solutions improve applications performance and continuity while substantially reducing IT costs, enabling enterprises to leverage their WAN for greater competitive advantage
- Enable global WAN Governance
  Ipanema enables enterprises to institute WAN Governance so they can manage the WAN coherently, predictably and aligned with business needs - driving higher levels of enterprise performance

With Ipanema, enterprises are ready to perform today and transform for tomorrow. Rollout new applications, implement desktop virtualization, consolidate servers, deploy VoIP, videoconferencing or telepresence, embrace cloud computing and more.

With WAN Governance improve business performance

After budget cuts during the global recession, many enterprises in 2010 will operate with IT spending at 2005 budget levels. Given the dependence on wide area networks for operational performance and enterprise growth, WAN Governance, a strategic and business - driven approach to network management within IT Governance, will become even more critical for executives to achieve business goals.

Using WAN Governance, organizations have been able to:

- Roll out business applications across their enterprise without long, costly network reconfigurations - accelerating application time to value and ROI while saving millions in IT costs
- Increase their global WAN capacity by 3x to 4x with no additional network investment - reducing costs for expensive bandwidth and delaying need for network upgrades
- Improve the performance of business applications over their corporate network - eliminating up to 90% of network - related application performance incidents and reducing time - to - repair incidents by 80%
- Increase availability and response times of business applications –improving workforce productivity, customer service and relationships with partners who depend on network access
- Improve usage of software assets – optimizing hardware and license utilization while reducing maintenance expenses
- Reduce overall OPEX and CAPEX costs while achieving higher levels of business performance

"Ipanema fit perfectly to our requirements: a "all in one solution" which would provide: Monitoring at application level (with pro-active alerting), full range of optimization functions: bandwidth allocation (shaping), compression, acceleration, caching (for wide area file service functionality), traffic prioritization configurable by application and ability to re-charge network costs according to volume/priority of traffic."

David Dodds, Global Network & Security Manager,

SGS

For More Information on Ipanema's WAN Governance approach:

Link to WAN Governance Executive Primer White Paper

Link to WAN Governance website section

With Hybrid Network Unification solution, maximize performance, continuity and IT savings

This automated solution uses "sense and respond" intelligence to monitor, control, accelerate and select the best path among MPLS and Internet networks for each individual user's traffic – all in real-time.

Hybrid networking is the simultaneous usage of different networks— MPLS and Internet VPN—to interconnect an enterprise's headquarters, data centers and remote sites. Ipanema's solution for Hybrid Network Unification automatically combines the performance and quality of business grade MPLS, the high capacity and lower-cost of Internet VPN and fully integrated QoS and WAN optimization techniques. With Ipanema's Hybrid Network Unification offering, companies can eliminate the trade-offs between the performance and quality of MPLS and lower-cost Internet VPN. This dynamic mix enables enterprises to get the best of both worlds: a fast and large unified network with 99.99% reliability, in a very cost effective way.

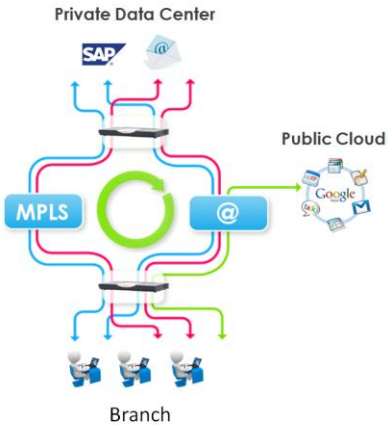
"Sense and respond," application-aware intelligence unifies traffic management automatically

Hybrid networking is not a new concept, but has been complex and difficult to implement so far. Until now, performance and continuity have been less than optimal, mitigating the expected business benefits.
Ipanema unifies hybrid network performance and optimizes application delivery through automated, second-by-second decision making, using distributed components that can exchange information quickly and accurately. This "sense and respond" intelligence dynamically monitors end-to-end network performance and load, while recognizing

each application's business criticality, performance objective and information sensitivity – and shifts and optimizes traffic accordingly, all in real time.

The only "all in one" integrated network management solution

Ipanema's Hybrid Network Unification is the newest offering in its "All in One" system, encompassing Application Visibility, Quality of Service (QoS) and Control, WAN Optimization, and now, Dynamic WAN Selection. Based on Ipanema's patented Autonomic Networking, technology, the result is adaptive, globally coordinated, automated processing of all application flows on the network.



Companies using Ipanema's all-in-one traffic management solution can:

- **Guarantee unified application performance across [MPLS + Internet] networks**: Enabling enterprises to consolidate data centers and adopt new computing platforms faster, including complete visibility and management of external or private cloud networks.
- **Improve business communication continuity**: Combining business-grade MPLS with plain Internet access provides 99.99% availability.
- **Exploit large network capacity at low cost:** Customers using Ipanema's solution have shown a net cost decrease from 50% to 80% per transferred Gbyte for a hybrid [MPLS + Internet] unified network compared to full MPLS.
- **Benefit from Internet immediacy and ubiquity:** For enterprises that need to quickly deploy or modify their infrastructure, Ipanema enables everywhere and anytime access to the Internet, even where MPLS networks are out of reach.
- **Turn back-up lines into business lines:** Allowing enterprises to use a resource they are paying for and would remains unused 99.9% of the time otherwise.

"Hybrid Network Unification by Ipanema provides a fully integrated Visibility, Control and WAN Optimization feature-set that ideally combines Internet VPN and MPLS in parallel to maximize performance and continuity over our global network, while minimizing our IT cost – all of this from an All-in-One system." Stuart Miles, Central Infrastructure Service Manager, Rhodia

Rhodia
Chemistry is our world, Responsibility is our way

For More Information on Ipanema's Hybrid Network Unification solution:

Link to Hybrid Network Unification White Paper

Link to Hybrid Network Unification website section

Link to Hybrid Network Unification video