The 2011 Application & Service Delivery Handbook

By Dr. Jim Metzler, Ashton Metzler & Associates
Distinguished Research Fellow and Co-Founder
Webtorials Analyst Division

Sponsored in part by:

CISCO

Produced by:

Webtorials

Table of Contents

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 the ***2011 Application and Service Delivery Handbook***, this era that began several years ago will be referred to as the Application Delivery 1.0 era.

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
- Adoption of varying forms of virtualization
- Adoption of varying forms of cloud computing
- Shifting emphasis and growing sophistication of cyber crime

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

As we enter the Application Delivery 2.0 era, leading edge IT organizations must focus on improving their ability to ensure acceptable application and service delivery. Throughout the 2011 Application and Service Delivery Handbook, the phrase ensuring acceptable application and service delivery will refer to ensuring that the applications and services that an enterprise uses:

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

The goal of the 2011 Application and Service Delivery Handbook is to help IT organizations ensure acceptable application delivery when faced with both the first and the second generation of application delivery challenges.

Application and Service Delivery Challenges

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

- Limited focus on application performance during 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 these challenges have either become notably more difficult or have morphed to become a different challenge. An example of a challenge that has become notably more difficult is the webification of applications. Current research shows that the number of hosts for a given web based user transaction varies around the world, but it typically is in the range of six to ten. An example of a challenge that has changed to where it has spawned a new challenge is that the task of supporting distributed employees has morphed to where a large percentage of those employees are now mobile.

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)

Many of the challenges listed above have become top of mind issues for the majority of IT organizations. For example, recent market research indicates that two thirds of IT organizations believe that getting better at managing internal SLAs is either very or extremely important.

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. In early 2010, 20% of IT organizations had virtualized the majority of their data center servers. Today, 32% of IT organizations have virtualized the majority of their data centers servers. In addition, market research that was recently conducted predicts that within a year, that 40% of IT organizations will have virtualized the majority of their data center servers.

One of the challenges associated with server virtualization is supporting the dynamic movement of VMs between data centers. Recent market research indicates that fifty percent of IT organizations find that it is either very or extremely important for them to get better at supporting the dynamic movement of VMs. Some of the other 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
- Over subscription of server resources
- 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

Desktop virtualization doesn't receive as much attention as server virtualization does in part because today the majority of IT organizations haven't virtualized any of their desktops. However, twenty five percent of IT organizations that haven't yet virtualized any desktops plan to do so in limited fashion in the next year.

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. Current market research indicates that by the end of the year, the vast majority of virtualized desktops will be utilizing server side virtualization.

While not widely deployed currently, IT organizations are showing a significant interest in implementing virtualized appliances. This interest is driven in part by the fact that a virtual appliance can help IT organizations respond to some of the challenges created by server and desktop virtualization. **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.

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. In addition, a software-based client can potentially leverage the functionality provided by the hypervisor's management system to provide a highly available system without having to pay for a second appliance.

One of the potential downsides of a virtual appliance is performance. The conventional wisdom in the IT 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. However, conventional wisdom is often wrong and IT organizations that are considering deploying a virtual appliance need to test the performance of that appliance in their production environment.

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

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.

The goal of cloud computing is to enable IT organizations to achieve a dramatic 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

There are three primary classes of cloud computing solutions. They are:

- Public
- Private
- Hybrid

Most public cloud based solutions are delivered over the Internet and no vendor will provide an end-to-end performance SLA for a service delivered over the Internet. As a result, many of the approaches to providing public cloud-based solutions will not be acceptable for the applications, nor for the infrastructure that supports the applications, for which enterprise IT organizations need to provide an SLA.

One way that an IT organization can improve the performance and availability of cloud computing solutions is to implement cloud balancing. Cloud balancing refers to routing service requests across multiple data centers based on myriad criteria. One way to think about cloud balancing is that it is the logical extension of global server load balancing (GSLB).

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.

Optimizing and Securing the Use of the Internet

The applications that are typically identified with public cloud computing are well known enterprise applications including CRM, SCM and ERP. While those applications will continue to be closely associated with cloud computing, it is becoming increasingly common for organizations to acquire a different category of application from a cloud computing service provider (CCSP). That class of applications is traditional network and infrastructure services such as VoIP, unified communications, management, optimization and security. Such applications will be referred to as a Cloud Networking Service (CNS) and recent market research indicates that over the next year that many IT organizations intend to make significant use of CNSs

The traditional application delivery solutions based on WAN optimization controllers (WOCs) and application delivery controllers (ADCs) 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 can’t be optimized. 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 the use of a CNS. Such a CNS would have to leverage 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.

Planning

Many planning functions are critical to the success of application delivery. One such function is identifying the company’s key applications and services and establishing SLAs for them.

Another key planning activity that will be elaborated on in the following chapter is Application Performance Engineering (APE). One of the characteristics of APE is that it is a life cycle approach to planning and managing application performance. 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.

For those organizations that run a large, complex network there often is a significant gap between network planning and network operations. A class of management tool that can facilitate the integration of planning and operations is typified by an IP route analytics solution. The goal of route analytics is to provide visibility, analysis and diagnosis of the issues that occur at the routing layer in complex, meshed networks.

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 a router's configuration by allowing the effect of a configuration change to be previewed before the change is actually implemented.

Most IT organizations that have already implemented either public or private cloud computing have not done so in a highly systematic fashion. 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.

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.

Application Performance Management

Successful APM requires a holistic approach based on integrated management of both the application and/or service itself as well as the end-to-end IT infrastructure. However, only a small percentage of IT organizations take such an approach.

A holistic approach to APM must focus on the experience of the end user of the application or service. Monitoring actual user transactions in production environments provides valuable insight into the end-user experience and provides the basis for an IT organization to be able to quickly identify, prioritize, triage and resolve problems that can affect business processes.

A holistic approach to APM must also address the following aspects of management:

- The adoption of a system of service level agreements (SLAs) at levels that ensure effective business processes and user satisfaction for at least a handful of key applications.
- Automatic discovery of all the elements in the IT infrastructure that support each service. This functionality provides the basis for an IT organization to be able 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.

Some of the challenges that make APM more difficult include:

- Port hopping
- Instant messaging
- Peer-to-peer networks
- The port 80 black hole
- Server virtualization
- Mobility
- Cloud computing

A concept that is closely related to APM is APE. APE is the practice of first designing for acceptable application performance and then testing, measuring and tuning performance throughout the application lifecycle. During the operational, or production phase of the lifecycle, APM is used to monitor, diagnose, and report on application performance. APM and APE are therefore highly complementary disciplines. For example, once an APM solution has identified that an application in production is experiencing systemic performance problems, an APE solution can be used to identify the root cause of the problem and to evaluate alternative solutions.

The key components of APE are:

- Setting Performance Objectives
- Discovery of the Network Topology
- Performance Modeling
- Performance Testing and Analysis
- Optimization

Another concept that is closely related to APM is route analytics. As noted, the goal of route analytics is to provide visibility, analysis and diagnosis of the issues that occur at the routing layer in complex, meshed networks. These issues are often referred to as logical issues in contrast to a physical issue such as an outage. One of the reasons why route analytics is a key component of APM is that recent market research has shown that in the vast majority of cases, logical factors cause as much or more business disruption than do physical factors. That same research also showed that in the vast majority of instances, logical errors take either somewhat more or notably more time to troubleshoot and repair than do physical errors.

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;

There are two principal categories of network and application optimization products. One category focuses on mitigating the negative effect that WAN services such as MPLS have on application and service performance. This category of products has historically included WAN optimization controllers (WOCs). However, due to some of the second generation of application and service delivery challenges, this category of products now also contains an emerging class of WAN optimization device - the Data Mobility Controller (DMC). As described in detail later in this section of the handbook, WOCs are focused primarily on accelerating end user traffic between remote branch offices and central data centers. In contrast, DMCs are focused on accelerating the movement of bulk data between data centers. This includes virtual machine (VM) migrations, storage replication, access to remote storage or cloud storage, and large file transfers. WOCs and DMCs are often referred to as ***symmetric solutions*** because they typically require complementary functionality at both ends of the connection. However, as is explained later in this section of the handbook, one way that IT organizations can accelerate access to a public cloud computing solution is to deploy WOCs in branch offices. The WOCs accelerate access by caching the content that a user obtains from the public cloud solution and making that content available to other users in the branch office. Since in this example there is not a WOC at the CCSP's site, this is an example of a case in which a WOC is an asymmetric solution. Roughly fifty percent of IT organizations have already deployed WOCs, although relatively few IT organizations have deployed them broadly within the organization. Over the next year, IT organizations plan to make a moderate increase in their deployment of WOCs.

The second category of optimization 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 on the remote end. 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.

Introduction

Background and Goals

Webtorials published the first edition of what became an annual series of application delivery handbooks in January 2007. The primary goal of the preceding handbooks was to help IT organizations ensure acceptable application delivery when faced with what is described in the next section of this document as the first generation of application delivery challenges.

Throughout the ***2011 Application and Service Delivery Handbook***, the phrase ***ensuring acceptable application and service delivery*** will refer to ensuring that the applications and services that an enterprise uses:

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

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

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

In part because the ongoing adoption of virtualization and cloud computing has created the concept of everything as a service (XaaS), this year's handbook will include more of a focus on managing and optimizing services. In addition to the concept of XaaS, services as discussed in the handbook will also include business services that involve multiple inter-related applications. To reflect this enhanced focus on services, this year's handbook is entitled the ***2011 Application and Service Delivery Handbook.***

The goal of the 2011 Application and Service Delivery Handbook is to help IT organizations ensure acceptable application delivery when faced with both the first generation, as well as the emerging generation of application delivery challenges.

Foreword to the 2011 Edition

As stated above, IT organizations are now encountering a new generation of application and service delivery challenges. So, while this year's edition of the application delivery handbook builds on the previous edition of the handbook, every section of the 2010 edition of the handbook was modified before being included in this document. For example, on the assumption that a number of the concepts that were described in previous editions of the handbook are by now relatively well understood, the description of those concepts was made more succinct in this year's handbook. In addition, all of the market research that was contained in the previous edition was deleted. To compensate for those changes, the 2010 Handbook of Application Delivery is still accessible at Webtorials.

Another change is that the section of the 2010 edition of the handbook that was entitled *Managed Service Providers* has been edited to focus on the task of optimizing and securing the Internet. In addition, the content that was contained in the section of the 2010 edition of the handbook that was entitled *Planning* as well as the content that was contained in the section that was entitled *Control* has been significantly reduced in size and included in other sections of the ***2011 Application and Service Delivery Handbook.***

In order to reflect the breadth of the movement to implement cloud computing, this year's handbook introduces the concept of a Cloud Networking Service (CNS). A CNS is a traditional network service, such as VoIP or optimization, which can now be acquired from a cloud computing service provider. The great interest in cloud computing drove a number of other additions to the handbook. This includes a discussion of cloud balancing – the advantages that it provides, the challenges that it presents and the role that application delivery controllers play in enabling cloud balancing. Because of the great interest in both virtualization and cloud computing, the handbook identifies the management and optimization challenges that are of most interest to IT organizations. The handbook also describes how virtualized application delivery appliances and cloud based optimization and management solutions can help IT organizations respond to these challenges. Given the extremely difficult management challenges associated with both virtualization and cloud computing, the management section of the handbook contains an added emphasis on application performance management and introduces the concept of application performance engineering.

In early 2011 two surveys were given to the subscribers of Webtorials. Throughout this document, the IT professionals who responded to the two surveys will be referred to as ***The Survey Respondents***. The results of surveys, such as the two that were given to the subscribers of Webtorials, which asked IT organizations about their plans, are always helpful because they enable IT organizations to see how their own plans fit with broad industry trends. Such survey results are particularly beneficial in the current environment when so much change is occurring.

One of the two surveys asked a broad set of questions relative to application delivery; e.g., how interested are IT organizations in emerging forms of virtualization such as desktop virtualization. The other survey focused on identifying the optimization and management tasks that are of most interest to IT organizations. With that later goal in mind, The Survey Respondents were given a set of twenty optimization tasks and twenty management tasks and asked to indicate how important it was to their IT organization to get better at these tasks over the next year. Because of the way the questions were worded, the responses highlight the aspects of management and optimization on which IT organizations are currently focused.

When asked about the optimization and management tasks that were important for their IT organization to get better at over the next year, The Survey Respondents 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 answers to both surveys will be used throughout the ***2011 Application and Service Delivery Handbook*** to demonstrate the breadth of application delivery challenges currently facing IT organizations.

Application and Service Delivery Challenges

This section of the handbook discusses some of the primary systemic challenges that are associated with ensuring acceptable application and service delivery. These challenges are grouped into two categories. The first category is the challenges that IT organizations have been responding to for the last several years and is referred to either as the first generation of application delivery challenges or alternatively as the traditional application delivery challenges. The other category is referred to either as the second generation of application and service delivery challenges, or alternatively as the emerging challenges that IT organizations are beginning to encounter.

As described below, an example of a systemic challenge to ensuring acceptable application and service delivery that is discussed in this section of the handbook is the limited focus that many IT organizations place on application performance during application development. More granular challenges to ensuring acceptable application and service delivery (e.g., ensuring acceptable performance of VoIP traffic, optimizing and/or managing in a hybrid cloud computing environment) will be discussed throughout the handbook.

One of the reasons why application and service delivery continues to be an important topic for IT organizations is the fact that approximately seventy five percent of The Survey Respondents indicated that when one of their company's key applications begins to degrade, that the degradation is typically noticed first by the end user and not by the IT organization.

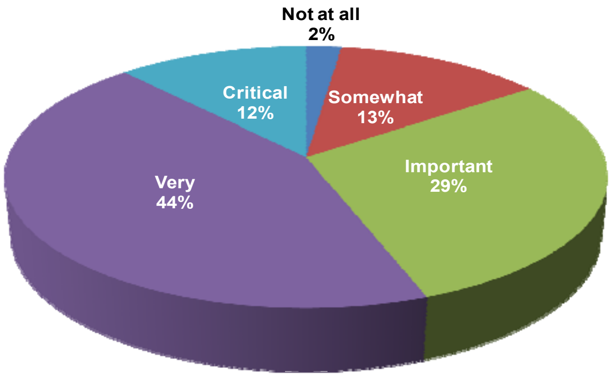
In the vast majority of instances, end users notice application degradation before the IT organization does.

The fact that it has been true for years that it is typically the end users that first notices application degradation makes it appear as if IT organizations are not getting better at ensuring acceptable application delivery. The reality is that most IT organizations do a better job today at ensuring acceptable application delivery than they did when the first handbook was published in 2007. Unfortunately, as is described below, the application delivery challenges facing IT organizations continue to get more formidable.

To further quantify the criticality of ensuring acceptable application delivery, the Survey Respondents were asked "If it is the end user who typically first notices application degradation, how important is that to senior management?" Their answers are shown in **Figure 1**.

Having the IT organization notice application degradation before the end user does is important to the vast majority of senior managers.

Figure 1: Importance of End User Noticing Degradation

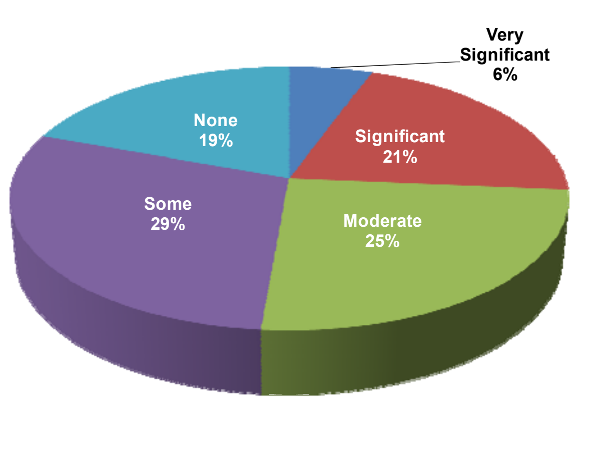


The Traditional Application Delivery Challenges

Limited Focus of Application Development

The Survey Respondents were asked “When your IT organization is in the process of either developing or acquiring an application, how much attention does it pay to how well that application will perform over the WAN?” Their answers are shown in **Figure 2**.

Figure 2: The Emphasis on Performance over the WAN



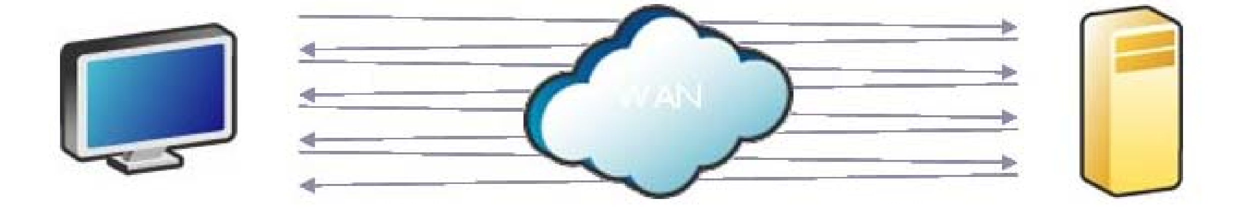
As is often the case with surveys, the data in **Figure 2** presents a classic good news – bad news situation. The good news is that the data in **Figure 2** indicates that just over a quarter of IT organizations place a significant or very significant emphasis on how an application performs over the WAN during application development or acquisition. The bad news is that almost three quarters of IT organizations don’t.

The vast majority of IT organizations don’t have any insight into the performance of an application 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[1] as illustrated in **Figure 3**.

Figure 3: 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 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.

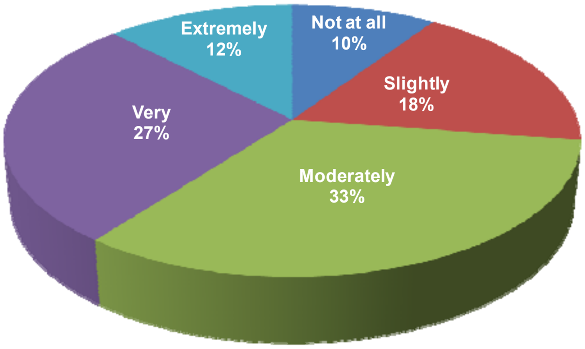
[1] Similar to a chatty protocol, a chatty application requires hundreds of round trips to complete a transaction.

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.

The Survey Respondents were asked how important it is for their IT organization over the next year to get better at optimizing the performance of chatty protocols such as CIFS. Their responses are shown in **Figure 4**.

Figure 4: The Importance of Optimizing Chatty Protocols



Optimizing chatty protocols such as CIFS was one of the primary challenges that gave rise to the first generation of WAN optimization products. In spite of the fact that IT organizations have been responding to this challenge for several years, over 80% of The Survey Respondents indicated that over the next year that it is at least moderately important for their organization to get better at optimizing these protocols.

Optimizing chatty protocols was one of a number of first generation application delivery challenges that are still important to IT organizations. For example, over 80% of The Survey Respondents also indicated that over the next year that it is at least moderately important for their organization to get better at optimizing the performance of TCP.

Responding to the first generation of application delivery challenges is still important to the majority of IT organizations.

Myriad Application Types

As described in the 2010 Application Delivery Handbook[2], the typical enterprise relies on hundreds of applications of different types, including applications that are business critical, enable other business functions, support communications and collaboration, are IT infrastructure-related (i.e., DNS, DHCP) or are recreational and/or malicious. In addition, an increasing amount of traffic results from social media. In some cases social media traffic is entirely recreational whereas in other cases it represents important, but typically not delay sensitive, business traffic.

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.

[2] The 2010 Handbook of Application Delivery, page 14

Over the last few years, IT organizations have begun to focus on the management and optimization of a small set of applications and services. The next sub-section of this document highlights one component of that trend - that component is the great interest that IT organizations have in getting better at providing SLAs for one or more business critical applications.

To further illustrate the trend of focusing on a small number of applications, The Survey Respondents were asked two questions. One question was how important it was over the next year for their IT organization to get better at optimizing the performance of specific applications such as SharePoint. The second question was how important it was over the next year for their IT organization to get better at optimizing the performance of a key set of applications that are critical to the success of the business. Their answers are shown **Table 1**.

Table 1: Importance of Optimizing Key Applications

	Optimizing Specific Applications such as SharePoint	Optimizing a Key Set of Business Critical Applications
Extremely Important	15.4%	25.9%
Very Important	34.6%	48.1%
Moderately Important	28.8%	16.7%
Slightly Important	13.5%	7.4%
Not at all Important	7.7%	1.9%

The data in **Table 1** demonstrates the importance that IT organizations place on optimizing both a specific application such as SharePoint as well as a small set of applications.

Over the next year, the most important optimization task facing IT organizations is optimizing a key set of business critical applications.

An example of an application that is time sensitive and important to most businesses is VoIP. Since the first application delivery handbook was published in 2007, a growing percentage of the traffic on the typical enterprise data network is VoIP traffic. To quantify the challenges associated with supporting a range of communications traffic, The Survey Respondents were asked to indicate how important it was over the next year for their IT organization to get better at managing the use of VoIP, traditional video traffic as well as telepresence. Their answers are shown in **Table 2**.

Table 2: Importance of Managing the Use of Communications Based Traffic

	VoIP	Traditional Video Traffic	Telepresence
Extremely Important	13.4%	6.8%	4.8%
Very Important	33.9%	20.3%	25.6%
Moderately Important	29.9%	29.7%	25.6%
Slightly Important	14.2%	28.0%	24.8%
Not at all Important	8.7%	15.3%	19.2%

The data in **Table 2** shows that almost 50% of The Survey respondents indicated that getting better at managing the use of VoIP traffic is either very or extremely important to their IT organization. This is a significant percentage and it is roughly the same percentage that indicated that optimizing specific applications such as SharePoint was either very extremely important. It is, however, notably less than the percentage of respondents who indicated that it was either very or extremely important to optimize a key set of business critical applications (**Table 1)** or to manage SLAs for one or more business critical applications (**Figure 10**). The data in **Table 2** also shows that in spite of all the discussion in the trade press about the growth in video traffic, that managing video traffic is notably less important to IT organizations than is managing VoIP.

Optimizing the performance of business critical data applications typically involves implementing techniques that will be described in a subsequent section of the handbook; e.g., protocol optimization, compression, de-duplication. While techniques such as these can make a minor difference in the performance of communications traffic, the primary way that IT organizations can ensure acceptable performance for this class of traffic is to identify the traffic and ensure that it is not interfered with by other traffic such as bulk file transfers.

The fact that IT organizations need to treat business critical traffic different than malicious traffic, than recreational traffic, than VoIP traffic leads to a number of conclusions.

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.

Security Vulnerabilities

In March 2011, IBM published its annual X-Force 2010 Trend and Risk Report[3]. The report documents a 27% increase in security vulnerabilities in 2010 vs. 2009 and stated that “This data points to an expanding threat landscape in which sophisticated attacks are being launched against increasingly complex computing environments.” In recognition of what this handbook refers to as the second generation of application delivery challenges, the IBM report dedicated a new section to the security trends and best practices that are associated with mobile devices and cloud computing.

The IBM report also made the following observations:

- ***Cloud Computing***
  IBM predicts that over time that the market will drive public cloud computing providers to provide access to security capabilities and expertise that is more cost effective than in-house implementations. IBM also stated that, “This may turn questions about cloud security on their head by making an interest in better security a driver for cloud adoption, rather than an inhibitor.”

[3] X-Force 2010 Trend and Risk Report

- ***Mobile Devices***
  The IBM report documented increases in the volume of vulnerabilities disclosed in mobile devices as well as the disclosure of exploits that target them. Nevertheless, the report concluded that malware is not yet common on the latest generation of mobile devices and that most IT professionals view the data stored on them and how that can be misused or lost as the main security threats associated with these devices.

- ***Cyber Crime***
  The IBM report stated that, "2010 is most remembered as a year marked by some of the most high profile, targeted attacks that the industry has ever witnessed. For example, the Stuxnet worm demonstrated that the risk of attacks against highly specialized industrial control systems is not just theoretical. These types of attacks are indicative of the high level of organization and funding behind computer espionage and sabotage that continues to threaten a widening variety of public and private networks."

- ***Web Applications***
  The IBM report stated that, "Web applications accounted for nearly half of vulnerabilities disclosed in 2010 -- Web applications continued to be the category of software affected by the largest number of vulnerability disclosures, representing 49 percent in 2010. The majority represented cross site scripting and SQL injection issues."

Server Consolidation

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

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

Server consolidation typically results in a chatty protocol such as Common Internet File System (CIFS), which was designed to run over the LAN, running over the WAN. As shown in Figure 4, getting better at optimizing the performance of chatty protocols such as CIFS is important to the majority of IT organizations.

Data Center Consolidation

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.

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

The reason why the preceding conclusion is so important is because, as previously discussed, even a small increase in network delay can result in a significant increase in application delay.

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.

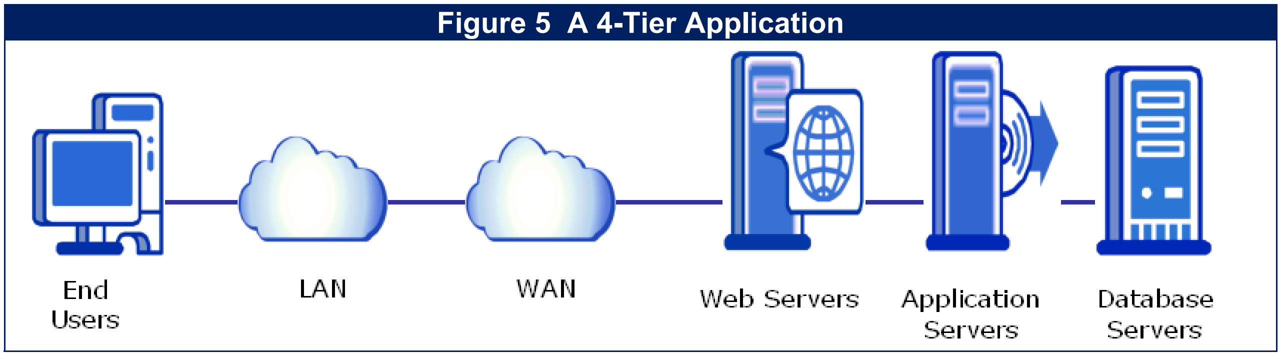
In the vast majority of situations, when people access an application they are accessing it over the WAN instead of the LAN.

The preceding discussion of chatty protocols exemplifies one of the challenges associated with accessing an application over a WAN. As that discussion showed, there are protocols and applications that perform in acceptable fashion when run over a LAN but which perform unacceptably when run over a WAN – particularly if the WAN exhibits even moderate levels of latency. The impact of that challenge is exacerbated by the fact that applications are typically developed over a LAN and as previously documented, during the application development process most IT organizations pay little if any attention to how well an application will run over the WAN.

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 5**) 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). Until recently, few, if any, of the servers were virtualized.

Figure 5 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.

As previously noted, few, if any, of the servers in the typical n-tier application are virtualized. However, as was also previously noted, virtualization and cloud computing are two of the emerging challenges that are complicating the task of ensuring acceptable application and service delivery. One example of how virtualization and cloud computing complicate application and service delivery is that managing an n-tier application becomes even more complex if one or more of the servers is virtualized and it becomes yet more complex if one or more of the servers is housed by a cloud computing service provider. A further complication that will be discussed in a subsequent sub-section of the handbook is that in the current environment it is common for a Web-based transaction to be supported by as many at ten different servers – most of which are virtualized.

Complexity

As noted in the preceding paragraph, the traditional 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:

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 was highlighted IBM's X-Force 2010 Trend and Risk Report, as the 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.

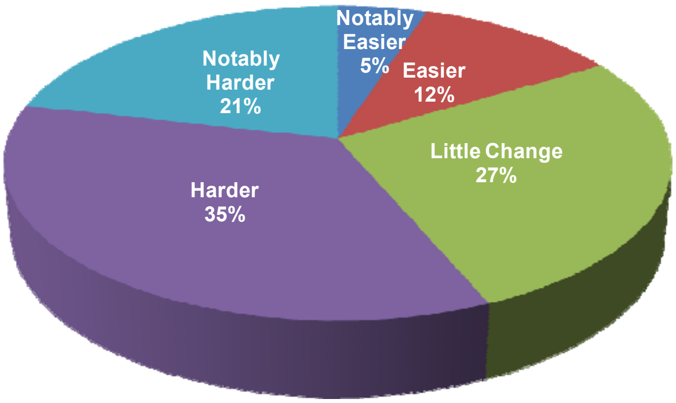
The Emerging Application and Service Delivery Challenges

While some of the emerging challenges described below are relatively new, some of them are the natural extension of the traditional challenges that were described above. For example, the challenges associated with supporting mobile workers are the natural extension of the challenges associated with supporting distributed employees that was previously discussed. In addition, the two new factors that are having the biggest impact on application and service delivery, virtualization and cloud computing will be discussed in subsequent sections of the handbook.

In order to get a snapshot relative to how much of an impact the emerging application and service delivery challenges will have on IT organizations, The Survey Respondents were asked "How will the ongoing adoption of mobile workers, virtualization and cloud computing impact the difficulty that your organization has with ensuring acceptable application performance?" Their responses are shown in **Figure 6**.

At the same time that most IT organizations are still responding to a traditional set of application and service delivery challenges, they are beginning to face a formidable set of new challenges.

Figure 6: Impact of Emerging Challenges on Application and Service Delivery



Growth of Mobile Workers and Mobile Devices

As previously noted, one of the traditional application delivery challenges was 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. The logical extension of that challenge is that most IT organizations now have to support a work force that is increasingly mobile. There are a number of key concerns relative to supporting mobile workers. One such concern is the number and types of devices that mobile workers use. As recently as a couple of years ago, many IT organizations tried to control the types of devices that their users could utilize; e.g., a blackberry. Now, the majority of IT organizations are in a position where they have to support a large and growing set of mobile devices from a range of vendors. It most cases mobile workers have two mobile devices (i.e., a laptop and a smartphone) and in a growing number of cases, mobile workers have three mobile devices; i.e., a laptop, a smartphone and a tablet.

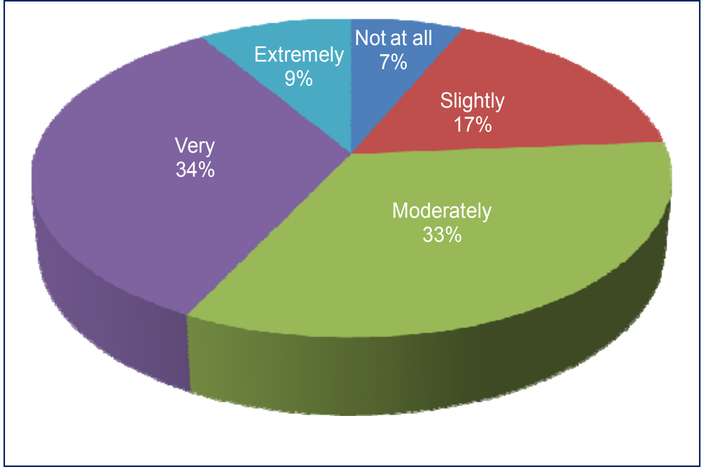
The security concerns associated with mobile workers was highlighted in IBM's X-Force 2010 Trend and Risk Report. Another key concern relative to supporting mobile workers is how the applications that these workers access has changed. At one time, mobile workers tended to primarily access either recreational 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. This shift in the applications accessed by mobile workers was highlighted by SAP's recent announcement[4] that it will leverage its Sybase acquisition to offer access to its business applications to mobile workers. 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.

In order to quantify the concern amongst IT organizations about ensuring acceptable application and service delivery to mobile workers, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at improving the performance of applications used by mobile workers. Their responses are shown in **Figure 7**.

One way to put the data in Figure 7 into context is to compare it to the data in Table 2 (The Importance of Managing the Use of Communications Based Traffic). Based on that comparison, it is reasonable to conclude that

improving the performance of applications used by mobile workers is less important than managing VoIP traffic, but more important than managing either traditional video or telepresence traffic.

Figure 7: Importance of Optimizing Mobile Applications

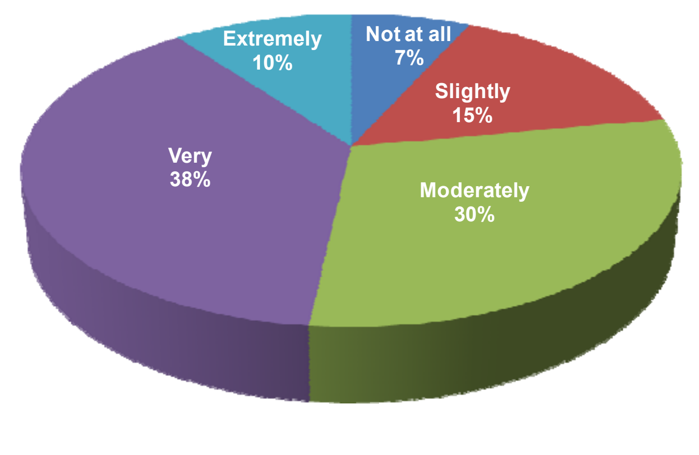


Webification of Applications

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. There are multiple challenges associated with this class of application. The security challenges associated with this class of application was highlighted in IBM's X-Force 2010 Trend and Risk Report. There are also performance challenges that are somewhat unique to this class of application. For example, unlike CIFS, HTTP is not a chatty protocol. However, HTTP is used to download web pages and 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.

[4] Wall Street Journal, May 17, 2011, page B7

Figure 8: Importance of Optimizing Protocols Other than TCP



The Survey Respondents were asked how important it was over the next year for their IT organization to get better at optimizing protocols other than TCP; e.g., HTTP and MAPI. Their answers, which are shown in **Figure 8**, demonstrate that the webification of applications and the number of round trips associated with downloading a web page is a traditional application delivery challenge that is still of interest to IT organizations.

An extension of the traditional problems associated with the webification of applications is that many organizations currently support Web-based applications that are accessed by customers. In many cases, customers abandon the application, and the company loses revenue, if the application performs badly. Unfortunately, according to recent market research[5], these Web-based applications have become increasingly complex. One result of that research is depicted in **Table 3**. As shown in that table, the number of hosts for a given user transaction varies around the world, but is typically in the range of six to ten.

Table 3: The Number of Hosts for a Web-Based Transaction

Measurement City	Number of Hosts per User Transaction
Hong Kong	6.12
Beijing	8.69
London	7.80
Frankfurt	7.04
Helsinki	8.58
Paris	7.08
New York	10.52

Typically several of the hosts that support a given Web-based transaction reside in disparate data centers. As a result, the negative impact of the WAN (i.e., variable delay, jitter and packet loss) impacts the Web-based transaction multiple times. The same research referenced above also indicated that whether or not IT organizations are aware of it, public cloud computing is having an impact on how they do business. In particular, that research showed that well over a third of Web-based transactions include at least one object hosted on Amazon EC2.

Web-based applications present a growing number of management, security and performance challenges.

[5] Steve Tack, Compuware, Interop Vegas, May 2011

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 another major 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 negative impact of the WAN (i.e., variable delay, jitter and packet loss) impacts 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.

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." [6]

The introduction of new technologies tends to further complicate the IT environment and leads to more security vulnerabilities. AJAX is a good example of that. AJAX is actually a group of interrelated web development techniques used on the 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[7] 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."

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

[6] Wikipedia on Rich Internet Applications
[7] Ajax Resource Center

definition of the term service will include the key characteristics of the ITIL definition of service[8]. 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).

In part because the ongoing adoption of virtualization and cloud computing has created the concept of everything as a service (XaaS), the term service as used in this handbook will sometimes refer to services that IT organizations acquired from a public cloud computing provider. These services include storage, compute and applications. Alternatively, the term service as used in this handbook will sometimes refer to business services that involve multiple inter-related applications. As is discussed in a subsequent section of the handbook part of the challenge in supporting effective service delivery is that on a going forward basis, a service will increasingly be supported by an infrastructure that is virtual. In addition, 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.

The Survey Respondents were asked to indicate how important it was over the next year for their IT organization to get better at monitoring and managing the services that they acquire from a public cloud computing vendor. Their answers are shown in **Table 4**.

Table 4: Importance of Monitoring and Managing Public Cloud Services

	Storage Services	Compute Services	Applications
Extremely Important	4.3%	10.6%	7.8%
Very Important	30.4%	21.3%	35.3%
Moderately Important	15.2%	17.0%	23.5%
Slightly Important	17.4%	23.4%	17.6%
Not at all Important	32.6%	27.7%	15.7%

As shown in **Table 4**, 32.6% of The Survey Respondents responded with “not at all important” when asked about the importance of getting better at monitoring and managing storage services that they acquire from a public cloud computing vendor. Vendors who supply storage and compute services are often referred to as being an Infrastructure as a Service (IaaS) vendor.

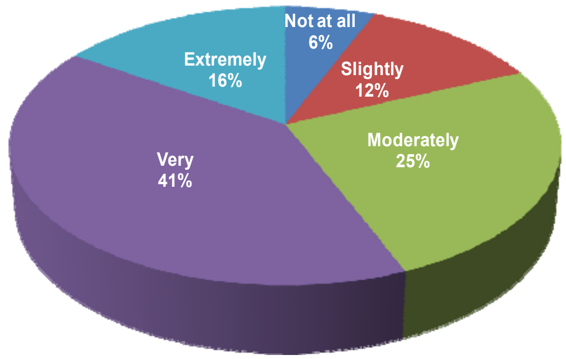
The 32.6% was the largest percentage to respond with “not at all important” for any of the twenty management tasks that were presented to The Survey Respondents. Given that, it is possible to conclude that monitoring and managing the services obtained from an IaaS vendor is not an important task. However, that conclusion is contradicted by the fact that just over a third of The Survey Respondents indicated that getting better at monitoring and managing storage services acquired from an IaaS vendor was either very or extremely important. A more reasonable conclusion is based on the observation that many companies don’t make any use of storage and compute services from an IaaS vendor and the ones that do often make only minor use of such services. Based on that observation, the data in **Table 4** suggests that if a company makes significant use of the services provided by an IaaS vendor, then monitoring and managing those services is indeed an important task.

[8] ITIL definition of service

The Survey Respondents were also asked to indicate how important it was over the next year for their organization to get better at managing a business service, such as CRM, that is supported by multiple, inter-related applications. Their responses are shown in **Figure 9**.

Getting better at managing a business service that is supported by multiple, inter-related applications is one of the most important tasks facing IT organizations over the next year.

Figure 9: Importance of Managing a Business Service



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 over the last couple of years 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:

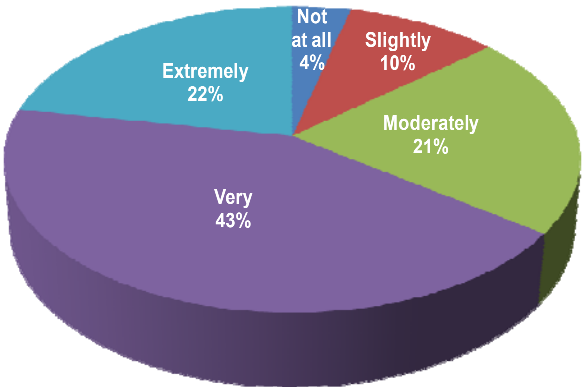
- Availability than for performance
- Networks than for applications
- A selected set of WAN links or applications rather than for all of the WAN or all applications

Most IT organizations, however, report that the internal SLAs that they offer are relatively weak and that they often don't have the tools and processes to effectively manage them.

The Survey Respondents were asked how important it is for their IT organization over the next year to get better at effectively managing SLAs for one or more business-critical applications. Their responses are shown in **Figure 10**.

The data in **Figure 10** leads to two related conclusions. The obvious conclusion is that managing internal SLAs is very important to the majority of IT organizations. The somewhat more subtle conclusion is that managing internal SLAs is difficult or else the majority of IT organizations would already be doing a good job of managing these SLAs and hence would not be striving to get better at the task. Unfortunately, as will be discussed in a subsequent section of the handbook, the movement to utilize public cloud computing services greatly increases the difficulty associated with managing an internal SLA.

Figure 10: The Important of Getting Better at Managing Internal SLAs



Virtualization

Server Virtualization

Interest in Server Virtualization

In order to quantify the interest that IT organizations have in server virtualization, The Survey Respondents were asked to indicate the percentage of their company's data center servers that have either already been virtualized or that they expected would be virtualized within the next year. Their responses are shown in **Table 5**.

Table 5: Deployment of Virtualized Servers

	None	1% to 25%	26% to 50%	51% to 75%	76% to 100%
Have already been virtualized	15%	33%	21%	18%	14%
Expect to be virtualized within a year	6%	25%	28%	20%	20%

In early 2010, a similar group of IT professionals was asked to indicate the percentage of their data center servers that had already been virtualized. Their responses are shown in Table 6.

Table 6: Deployment of Virtualized Servers as of Early 2010

	None	1% to 25%	26% to 50%	51% to 75%	76% to 100%
Have already been virtualized	30%	34%	17%	11%	9%

The data in **Table 1** and **Table 2** show the strength of the ongoing movement to virtualize data center servers. For example, in early 2010 20% of IT organizations had virtualized the majority of their data center servers. Today, 32% of IT organizations have virtualized the majority of their data centers servers. In addition, The Survey Respondents predict that within a year, that 40% of IT organizations will have virtualized the majority of their data center servers.

Another way to look at the data in **Table 1** and **Table 2** is that in early 2010 30% of IT organizations had not virtualized any data center servers. Today, only 15% of IT organizations have not virtualized any data center servers and The Survey Respondents predict that within a year, that only 6% of IT organizations will not have virtualized any of their data center servers.

The Fractal Data Center

As noted in a previous section of the handbook, in the current environment almost every component of IT can be virtualized. One way to think about the current generation of virtualized data centers, and the related management challenges, draws on the concept of a fractal[9]. A fractal is a geometric object that is similar to itself on all scales. If you zoom in on a fractal object

[9] PHA.JHU.EDU: Concept of a fractal

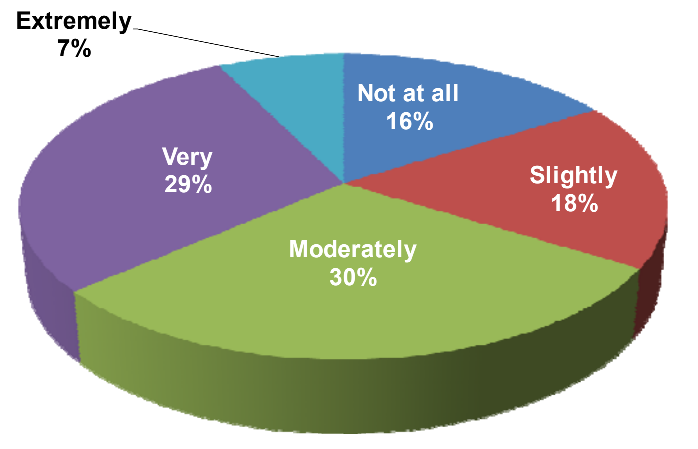
it will look similar or exactly like the original shape. This property is often referred to as self-similarity.

The relevance of fractals is that the traditional data center is comprised of myriad physical devices including servers, LAN switches, probes, WAN optimization controllers (WOCs) and firewalls. The virtualized data centers that most IT organizations are in the process of implementing are still comprised of physical servers, LAN switches and firewalls. However, as shown in **Table 5**, the vast majority of IT organizations have virtualized at least some of their data center servers. These virtualized data center servers are typically comprised of a wide range of functionality including virtual machines (VMs), a virtual LAN switch (vSwitch) that switches the traffic between the VMs and in many cases devices such as virtual probes, WOCs and firewalls. Hence, if you take a broad overview of the data center you see certain key pieces of functionality. If you were to then zoom inside of a virtualized data center server you would see most, if not all of that same functionality. Hence:

A virtualized data center can be thought of as a fractal data center.

One of the challenges that is introduced by the deployment of virtualized servers is that due to the limitations of vSwitches, once a server has been virtualized IT organizations loose visibility into the inter-VM traffic. This limits the IT organization's ability to perform functions such as security filtering or performance monitoring and troubleshooting. To quantify the impact of loosing visibility into the inter-VM traffic, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at managing the traffic that goes between virtual machines on a single physical server. Their responses are shown in **Figure 11**.

Figure 11: Manage the Inter-VM Traffic



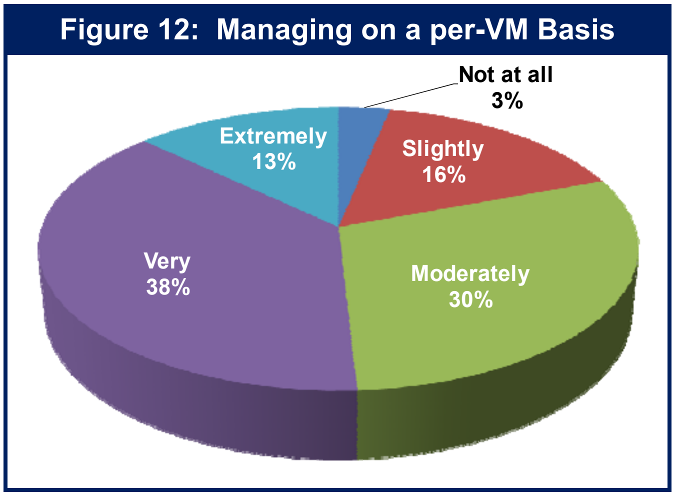
The data in **Figure 11** indicates that while there is significant interest in getting better at managing inter-VM traffic, the level of interest is less than the level of interest that The Survey Respondents indicated for many other management tasks

Because of the fractal nature of a virtualized data center, many of the same management tasks that must be performed in the traditional server environment need to be both extended into the virtualized environment and also integrated with the existing workflow and management processes. One example of the need to extend functionality from the physical server environment into the virtual server environment is that IT organizations must be able to automatically discover both the physical and the virtual environment and have an integrated view of both environments. This view of the virtual and physical server resources must stay current as VMs move from one host to another, and the view must also be able to indicate the resources that are impacted in the case of fault or performance issues.

To quantify the impact that managing on a per-VM basis is having on IT organizations, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at performing traditional management tasks such as troubleshooting and performance management on a per-VM basis. Their responses are shown in **Figure 12**.

One observation that can be drawn from the data in **Figure 12** is that unlike the situation with managing inter-VM traffic:

Figure 12: Managing on a per-VM Basis



Half of the IT organizations consider it to be either very or extremely important over the next year for them to get better performing management tasks such as troubleshooting on a per-VM basis.

To put the challenge of troubleshooting on a per-VM basis into perspective, consider a hypothetical 4-tier application that will be referred to as BizApp. For the sake of this example, assume that BizApp 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. One challenge that is associated with troubleshooting performance problems with BizApp is that each server has a different hypervisor management system and a different degree of integration with other management systems.

In order to manage BizApp in the type of virtualized environment described in the preceding paragraph, an IT organization needs to gather detailed information on each of the three VMs and the communications between them. For the sake of example, assume that the IT organization has deployed the tools and processes to gather this information and has been able to determine that the reason that BizApp sporadically exhibits poor performance is that the application server occasionally exhibits poor performance. However, just determining that it is the application server that is causing the application to perform badly is not enough. The IT organization also needs to understand why the application server is experiencing sporadic performance problems. The answer to that question might be that other VMs on the same physical server as the application server are sporadically consuming resources needed by the application server and that as a result, the application server occasionally performs poorly. A way to prevent one VM from interfering with the performance of another VM on the same physical server is to implement functionality such as VMotion[10] that would move a VM to another physical server if performance degrades. However, as discussed in the next sub-section, the dynamic movement of VMs creates a whole new set of challenges.

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

[10] VMotion

The next section of the handbook will make use of BizApp to discuss how cloud computing further complicates application and service delivery.

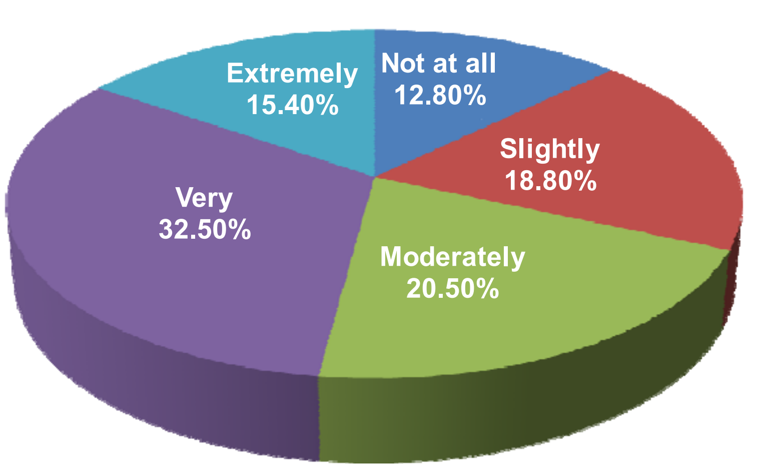
Challenges of Server Virtualization

The preceding sub-section mentioned some of the high level challenges created by server virtualization. Another high level challenge created by server virtualization is related to the dynamic nature of VMs. For example, a VM can be provisioned in a matter of seconds or minutes. However, in order for the VM to be useful, the IT organization must be able to establish management capabilities for the VM in the same timeframe – seconds or minutes.

In addition, one of the advantages of a virtualized server is that a production VM can be dynamically transferred to a different physical server, either to a server within the same data center or to a server in a different data center, without service interruption. The ability to dynamically move VMs between servers represents a major step towards making IT more agile. There is a problem, however, relative to supporting the dynamic movement of VMs that is similar to the problem with supporting the dynamic provisioning of VMs. That problem is that today the supporting network and management infrastructure is still largely static and physical. So while it is possible to move a VM between data centers in a matter of seconds or minutes, it can take days or weeks to get the network and management infrastructure in place that is necessary to enable the VM to be useful.

In order to quantify the concern that IT organization have with the mobility of VMs, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at supporting the movement of VMs between servers in different data centers. Their responses are shown in **Figure 13**.

Figure 13: Supporting the Mobility of VMs



Given that the data in **Table 5** indicates that IT organizations plan to increase their deployment of virtualized servers, one observation that can be drawn from the data in **Figure 13** is that:

Supporting the movement of VMs between servers in different data centers is an important issue today and will become more so in the near term.

Some of the other specific challenges created by server virtualization include:

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.

Contentious Management of the vSwitch

Each virtualized server includes at least one software-based 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.

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 a 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. This challenge is potentially alleviated by functionality such as VMotion.

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.

Meeting the Challenges of Server Virtualization

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 eliminates 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 probe11 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.

[11] This will be discussed in the subsequent analysis of virtual appliances.

Desktop[12] Virtualization

Interest in Desktop Virtualization

In order to quantify the interest that IT organizations have in desktop virtualization, The Survey Respondents were asked to indicate the percentage of their company's desktops that have either already been virtualized or that they expected would be virtualized within the next year. Their responses are shown in Table 7.

Table 7: Deployment of Virtualized Desktops

	None	1% to 25%	26% to 50%	51% to 75%	76% to 100%
Have already been virtualized	55%	36%	3%	1%	4%
Expect to be virtualized within a year	30%	51%	8%	4%	7%

Comparing the data in **Table 7** with the data in **Table 5** yields an obvious conclusion:

The deployment of virtualized desktops trails the deployment of virtualized data center servers by a significant amount.

Comparing the data in the first row of **Table 7** with the data in the second row of **Table 7** yields the following conclusion:

Over the next year, there will be a modest increase in the deployment of virtualized desktops.

The two fundamental forms of desktop virtualization are:

- Server-side virtualization
- Client-side 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).

There are two primary approaches to server-side 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

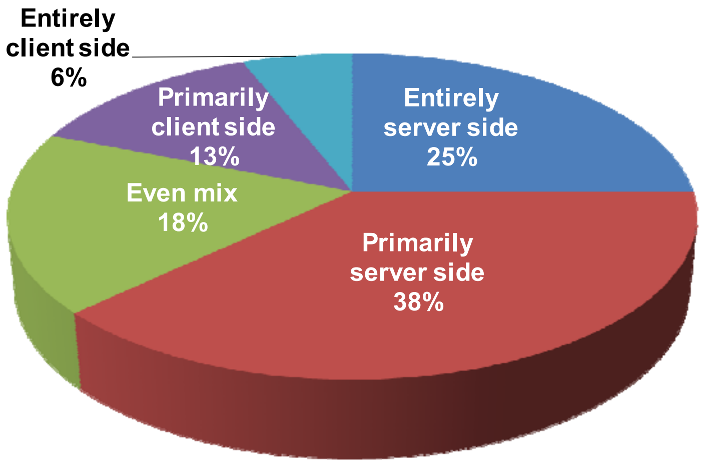
[12] In this context, the term 'desktop' refers to the traditional desktop as well as to various mobile devices including laptops and smartphones.

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.

The Survey Respondents whose company will have implemented desktop virtualization by the end of 2011 were asked to indicate which form(s) of desktop virtualization they will have implemented.   Their answers are shown in **Figure 14**.

Figure 14: Implementation of Desktop Virtualization



One conclusion that can be drawn from the data in **Figure 14** is:

By the end of the year, the vast majority of virtualized desktops will be utilizing server side virtualization.

Challenges of Desktop Virtualization

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

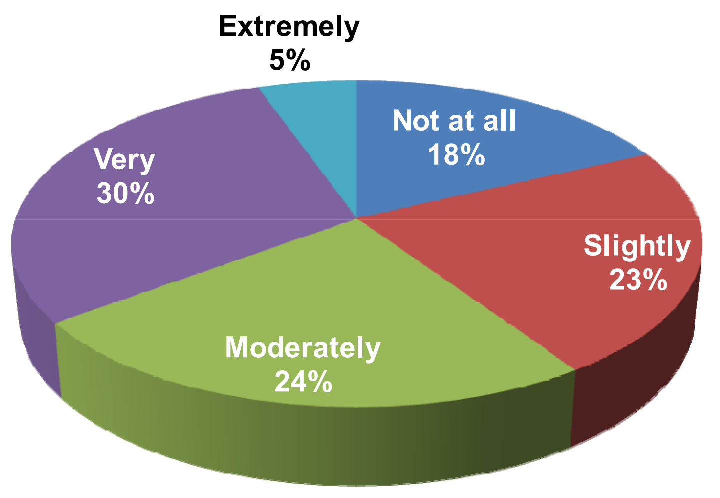
From a networking perspective, the primary challenge in implementing desktop virtualization is achieving adequate performance and an acceptable user experience for client-to-server connections over a WAN.

To quantify the concern that IT organizations have relative to supporting desktop virtualization, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at optimizing the performance of virtualized desktops. Their responses are shown in Figure 15.

One conclusion that could be drawn from the data in Figure 15 is that getting better at optimizing the performance of virtualized desktops is not that important to IT organizations. However, given that in the current environment there is a very limited deployment of virtualized desktops, and that the forecast is for only a modest increase in deployment, a more viable conclusion is that IT organizations who are implementing virtualized desktops realize the importance of optimizing performance.

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 in which there is a significant amount of packet loss.

Figure 15: Optimizing the Performance of Virtualized Desktops



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 8** show estimates for the amount of WAN bandwidth required by XenDesktop as documented in an entry in The Citrix Blog[13].

Table 8: 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 8** depicts one possible scenario of what fifty simultaneous branch office users are doing and identifies that the total WAN bandwidth that is required by this scenario is just less than 16 Mbps.

[13] Community.Citrix.com: How Much Bandwidth Do I Need?

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

While PCoIP resolves some challenges, it also creates others. For example, a document the Teradici published[14] 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 uniformly supported by the current generation of WOCs. Another concern relative to 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 network and application optimization solutions.

Some of the relevant optimization techniques 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

[14] Teradici.com: PCoIP WAN brief

- QoS and traffic shaping

Although virtually all 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 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 and this form of communications is best optimized by byte-level caching that is not supported by all 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.

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

Interest in 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, routers and performance monitoring solutions among others.

The deployment of multiple classes of virtual appliances can create some significant synergies. For example, one of the challenges associated with migrating a VM between physical servers is replicating the VM's networking environment in its new location. However, unlike a physical appliance, virtual appliances can be easily migrated along with the VM. This makes it easier for the IT organization to replicate the VMs' networking environment in its new location.

In a branch office, a suitably placed virtualized server could potentially host a virtual WOC appliance as well as other virtual appliances forming what is sometimes referred to as a Branch Office Box (BOB). 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.

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[15]. 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[16].

As noted in a previous section of the handbook, one approach to monitoring and troubleshooting inter-VM traffic is to deploy a virtual performance management appliance or probe (vProbe). 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. 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.

A virtual firewall can help IT organizations meet some of the challenges associated with server virtualization. That follows because virtual firewalls 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 virtual firewall would use the same software as the physical firewalls already in use in the data center. In addition to firewall functionality, the virtual appliance may

[15] The actual price difference between a hardware-based appliance and a software-based appliance will differ by vendor.

[16] This statement makes a number of assumptions, including the assumption that the vendor does not charge for the backup software-based appliance.

provide other security functionality including anti-malware, IDS/IPS, integrity monitoring (e.g., registry changes), and log inspection functionality.

One of the potential downsides of a virtual appliance is performance. The conventional wisdom in the IT 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.

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 further increases 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

Cloud Computing

Within the IT industry there is not an agreed to definition of exactly what is meant by the phrase ***cloud computing***. This handbook 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.

The goal of cloud computing is to enable IT organizations to achieve a dramatic 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 as described in a following sub-section of the handbook:

The SLAs that are associated with public cloud computing services such as Salesforce.com or Amazon's Simple Storage System are generally weak both in terms of the goals that they set and the remedies they provide when those goals are not met.

As a result, the 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[17]. 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 reports: *A Guide for Understanding Cloud Computing*[18] and *Cloud Computing: A Reality Check and Guide to Risk Mitigation*[19].

[17] The Google-ization of Bechtel, Carolyn Duffy Marsan, Network World, October 28, 2008

[18] Webtorials: A Guide for Understanding Cloud Computing

[19] Webtorials: Cloud Computing - A Reality Check Guide to Risk Migration

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.

- ***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.*** Roughly a year and a half ago, Cisco announced its Unified Computing System[20] (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. Several other vendors either already have, or soon will, announce similar solutions.

- ***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.

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

- ***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 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.

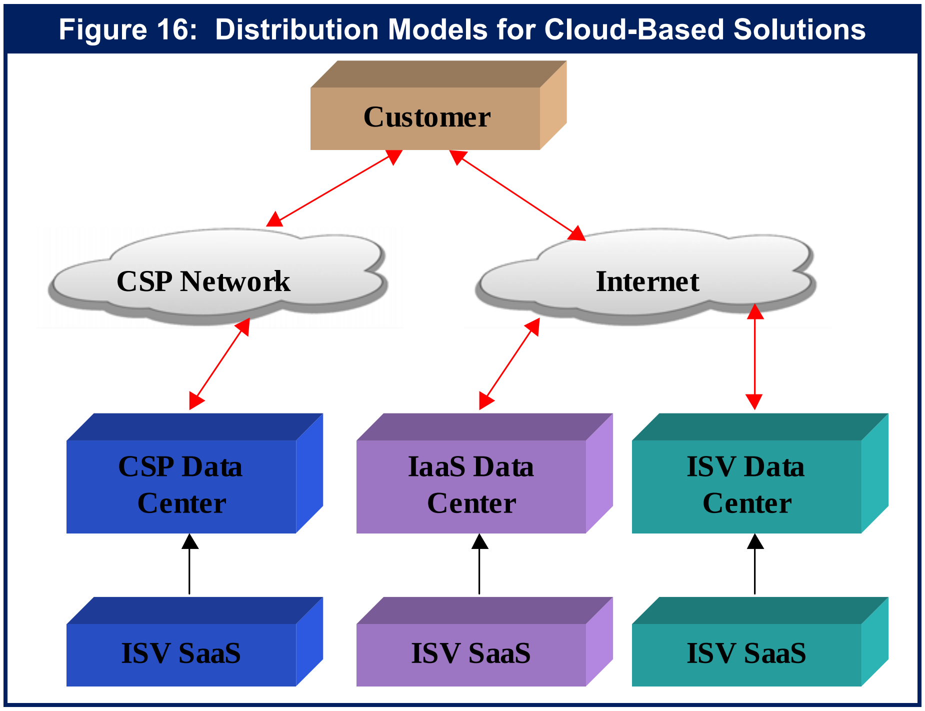
Public Cloud Computing

Background

CCSPs that provide their services either over the public Internet or over other WAN services such as MPLS are offering a class of solution that is often referred to as the *public cloud* or *public cloud computing.* One form of public cloud computing is referred to as Platform-as-a-Service (PaaS). Platform services provide software development environments, including application programming interfaces (APIs) and middleware that abstract the underlying infrastructure in order to support rapid application development and deployment. SalesForce.com provides one of the initial PaaS offerings: Force.com[21]. In April 2011, VMware announced its intention to provide a PaaS offering[22].

The two categories of public cloud computing solutions the handbook will focus on are Software-as-a-Service (SaaS) and Infrastructure-as-a-Service (IaaS). **Figure 16** shows some of the common distribution models for SaaS and IaaS solutions. As shown in **Figure 16**, one approach to providing public cloud-based solutions is based on the solution being delivered to the customer directly from an independent software vendor's (ISV's) data center via the Internet. This is the distribution model currently used for Salesforce.com's CRM application. Another approach is for an ISV to leverage an IaaS provider such as Amazon to host their application on the Internet. Lawson Software's Enterprise Management Systems (ERP application) and Adobe's LiveCycle Enterprise Suite are two examples of applications hosted by Amazon EC2.

Figure 16: Distribution Models for Cloud-Based Solutions



[21] Salesforce.com: Force.com
[22] CtoEdge: PaaS

Both of the two approaches described in the preceding paragraph rely on the Internet and it is not possible to provide end-to-end quality of service (QoS) over the Internet. As a result, neither of these two approaches lends itself to providing an SLA that includes a meaningful commitment to critical network performance metrics such as delay, jitter and packet loss. As was described in a preceding section of the handbook, over the last couple of years IT organizations have begun to focus on providing an internal SLA for at least a handful of key applications.

Many of the approaches to providing public cloud-based solutions will not be acceptable for the applications, nor for the infrastructure that supports the applications, for which enterprise IT organizations need to provide an SLA.

An approach to providing Cloud-based solutions that does lend itself to offering SLAs is based on a Communications Service Provider (CSP) providing these solutions to customers from the CSP's data center and over the CSP's MPLS network.

SaaS and IaaS

As previously mentioned, the classes of public cloud computing solutions that this section of the handbook will focus on are SaaS and IaaS.

SaaS

One of the key characteristics of the SaaS marketplace is that:

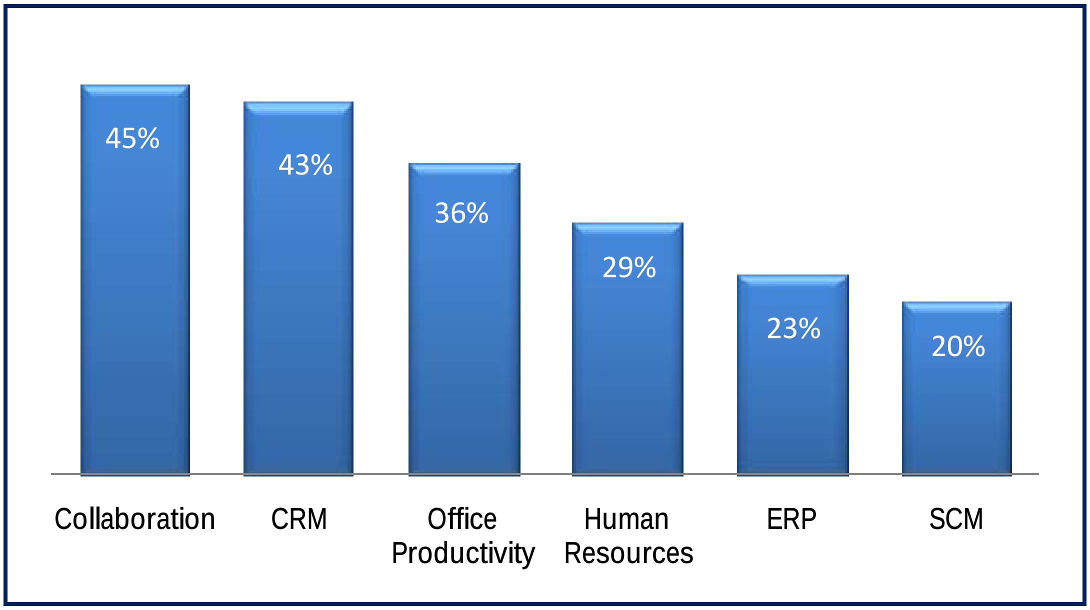
The SaaS marketplace is comprised of a small number of large players such as Salesforce.com, WebEx and Google Docs as well as thousands of smaller players.

According to IDC[23], the Software as a Service (SaaS) market had worldwide revenues of $13.1 billion in 2009 and is projected to reach $40.5 billion by 2014.

The Survey Respondents were asked about their company's use of SaaS-based applications. Figure 7 shows the percentage of respondents whose company either currently acquires, or is likely to acquire within the next year, various categories of applications from a SaaS provider.

Figure 17: Popular Categories of SaaS-Based Applications

[23] BusinessWire.com



The functionality provided by each of the six categories of applications listed in **Figure 7** can be quite extensive and is sometimes overlapping. ERP, for example, can encompass myriad functionality including product lifecycle management, supply chain management (e.g. Purchasing, Manufacturing and Distribution), warehouse management, customer relationship management (CRM), sales order processing, online sales, financials, human resources, and decision support systems.

For each category of application shown in **Figure 17**, there are tens, and sometimes hundreds, of SaaS-based solutions currently available[24]. **Table 9** contains a listing of some representative SaaS providers for each category.

Table 9: Representative SaaS Providers

Collaboration	CRM	Office Productivity	Human Resources	ERP	SCM
WebEx	Salesforce.com	Google Docs	Subscribe-HR	SAP	ICON-SCM
Zoho	NetSuite	Microsoft's Office Web Apps	ThinMind	Workday	E2open
clarizen	Update	feng office	Greytip Online	Lawson Software	Northrop Grumman

IaaS

Infrastructure services are comprised of the basic compute and storage resources that are required to run applications. The barrier to enter the IaaS marketplace is notably higher than is the barrier to enter the SaaS marketplace. That is one of the primary reasons why there are

[24] Saas-showplace.com

fewer vendors in the IaaS market than there are in the SaaS market. Representative IaaS vendors include Amazon, AT&T, CSC, GoGrid, IBM, Joyent, NaviSite, NTT Communications, Orange Business Services, Rackspace, Savvis, Terremark (recently acquired by Verizon) and Verizon. The IaaS market is expected to exhibit significant growth in the next few years. For example, Gartner[25] estimates that the IaaS market will grow from $3.7 billion in 2011 to $10.5 billion in 2014.

Table 6 provides a high level overview of some of the services offered by IaaS vendors. The data in **Table 6** is for illustration purposes only. That follows because it is extremely difficult, if not impossible, to correctly summarize in a table the intricate details of an IaaS solution; e.g., how the solution is priced, the SLAs that are provided and the remedies that exist for when the SLAs are not met. For example, consider the availability of an IaaS solution. On the surface, availability appears to be a well-understood concept. In fact, vendors often have differing definitions of what constitutes an outage and hence, what constitutes availability. For example, within Amazon's EC2 offering an outage is considered to have occurred only when an instance[26] is off line for 5 minutes and a replacement instance cannot be launched from another Availability Zone[27] within Amazon's geographical region. Not all IaaS providers have a similar definition of availability.

Table 10: Representative IaaS Providers

	Amazon AWS	RackSpace	GoGrid
Cloud Server (Virtual Machine (VM) with 2-4 vCPUs and ~8 GB RAM)	34¢/hour	40¢/hour	40¢-$1.53//hour *
Data Transfer	In 10¢/GB Out 15¢/GB	In 8¢/GB Out 18¢/GB	In free Out 7-29¢/GB
Load Balancer	2.5¢//hour 0.8¢/GB in/out	1.5¢/hour/LB 1.5¢/hour/100 connections	Included with server
VM Storage	(Elastic Block Store) 10¢/GB/month 10¢/million I/O requests/month	320 GB included with server	Included with server 400GB per 8 GB RAM
Cloud Storage	5.5-14¢/GB/month	15¢/GB/month	15¢/GB/month over 10 GB
Hypervisors	Xen plus VMware import	Xen (Linux) CitrixXenServer (Windows)	Xen
Server availability SLA	99.95%	100%	100%
Server SLA Remedy	10% of monthly charge/incident	5% of monthly charge/30 minutes downtime	100x hourly rate for downtime period

*=includes O/S licenses and some other items and depends on a variety of pre-payment plans

[25] Qas.com
[26] Amazon.com EC2 Instance types
[27] AWSEC2 UserGuide

Table 6 illustrates that:

There are significant differences amongst the solutions offered by IaaS providers, especially when it comes to the SLAs they offer.

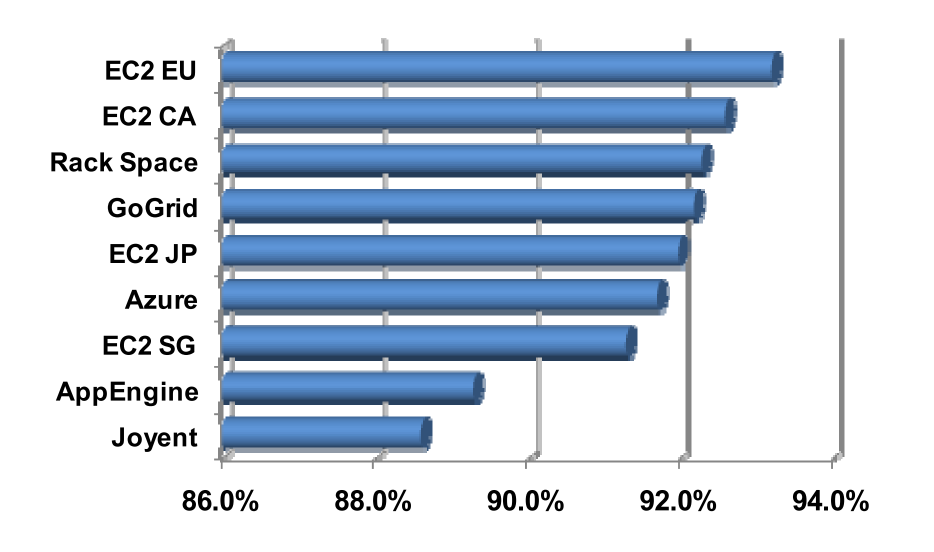
It is important to realize that the value of an availability SLA is only partially captured by the number of 9s it features. A number of factors can cause an SLA that promises four or more 9s of availability to become notably less meaningful. One such factor was previously mentioned – how the vendor defines what constitutes an outage. Another such factor is the remedy that the vendor provides for those instances in which the service it offers doesn't achieve the promised availability. In those cases in which the SLA remedies are weak, the IaaS provider can provide a fairly low level of availability and not suffer a significant loss of revenue. This can have the affect of minimizing the incentive that the vendor has to take the necessary steps to ensure high availability. A related factor is the degree of difficulty that an IT organization has in gathering the documentation that is required to establish that the service was unavailable and to apply for the service credits that are specified in the SLA. As the difficulty of this process increases, the meaningfulness of the SLA decreases.

Insight into the availability of a number of IaaS solutions was provided by Cedexis at the Interop conference in May, 2011[28]. Cedexis presented data that represented roughly 17 billion measurements that were taken between March 15, 2011 and April 15 2011. As shown in Figure 8, none of the IaaS providers that were monitored delivered availability that was greater than 95% (*Source: Cedexis*)

Figure 18 illustrates that:

The availability of IaaS solutions can vary widely.

Figure 18: Availability of Server Instances at Various IaaS Providers

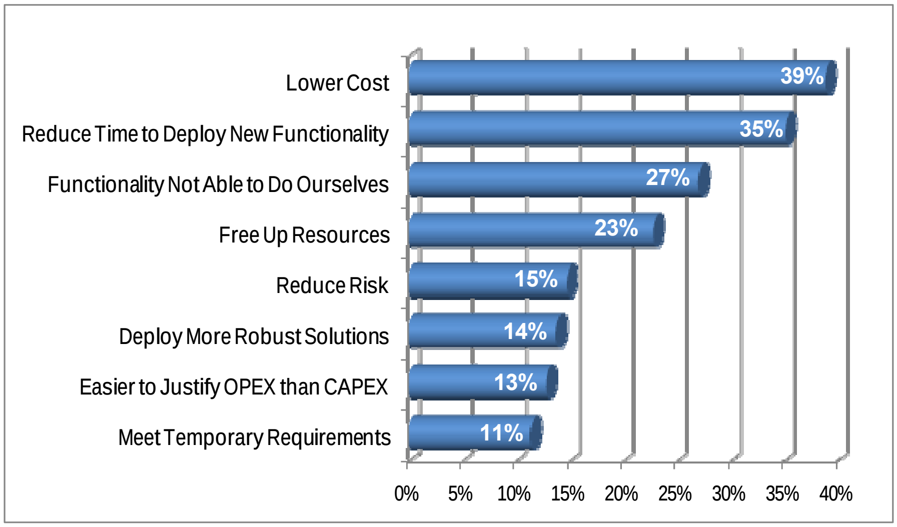


[28] Comparing Public Clouds: The State of On-Demand Performance, Marty Kagan, President and Co-Founder, Cedexis

The Drivers of Public Cloud Computing

The Survey Respondents were asked to indicate the two primary factors that are driving, or would likely drive their company to use public cloud computing services. Their responses are shown in **Figure 19**.

Figure 19: The Drivers of Public Cloud Computing



One of the observations that can be drawn from **Figure 19** is that:

The primary factors that are driving the use of public cloud computing solutions are the same factors that drive any form of out-tasking.

The next sub-section of this handbook analyzes a topic that is frequently discussed – the risks that are associated with public cloud computing. However, as shown in **Figure 19**, almost 15% of The Survey Respondents indicated that reducing risk was a factor that would cause them to use a public cloud computing solution. For the most part, their reasoning was that acquiring and implementing a large software application (e.g., ERP, CRM) presents considerable risk to an IT organization and one way to minimize this risk is to acquire the functionality from a SaaS provider.

In some cases, the use of a public cloud computing solution reduces risk.

A previous section of this handbook referenced IBM's X-Force 2010 Trend and Risk Report. In that report IBM predicts that over time that the market will drive public cloud computing providers to provide access to security capabilities and expertise that is more cost effective than in-house implementations. IBM also stated that, "This may turn questions about cloud security on their head by making an interest in better security a driver for cloud adoption, rather than an inhibitor."

The Risks Associated with Public Cloud Computing

One of the risks associated with public cloud computing is performance. One of the 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 who are close to the data-center can be subject to unacceptable performance as a result of sub-optimal routing and the inefficient protocols used within the Internet.

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 dramatic 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 with a traditional approach to IT.

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.

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 undergo 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 IT organization get its data back in the same format that it was in when it was provided to the CCSP?

- 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 ADCs and WOCs does the CCSP support?

- Does the CCSP allow a customer to incorporate their own WOC or ADC as part of the service provided by the CCSP?

- 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?

Managing and Optimizing Public Cloud Computing

As previously noted, in the current environment there are significant limitations to the steps that an IT organization can take to either manage or optimize a solution that involves one or more CCSPs. In spite of that limitation, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at monitoring and managing storage, compute and application services that they acquire from a CCSP. Their responses are shown in Table 11.

Table 11: The Importance of Managing Public Cloud Services			
	Storage	**Compute**	**Applications**
Extremely	2.9%	8.1%	9.4%
Very	20.0%	20.7%	30.8%
Moderately	28.6%	25.2%	23.9%
Slightly	20.0%	22.5%	18.8%
Not at All	28.6%	23.4%	17.1%

The Survey Respondents were also asked how important it is for their IT organization over the next year to get better at optimizing the storage, compute and application services that they acquire from a CCSP. Their responses are shown in **Table 12**.

Table 12: The Importance of Optimizing Public Cloud Services			
	Storage	**Compute**	**Applications**
Extremely	4.4%	3.1%	7.8%
Very	16.7%	17.7%	28.2%
Moderately	23.3%	26.0%	25.2%
Slightly	28.9%	26.0%	20.4%
Not at All	26.7%	27.1%	18.4%

There are many conclusions that can be drawn from the data in **Table 11**.and **Table 12**. One of which is that getting better at managing and optimizing SaaS solutions is more important to IT organizations than is getting better at managing and optimizing IaaS solutions. One reason for that situation is that IT organizations make more use of SaaS solutions than they do IaaS solutions. Another observation is that getting better at managing and optimizing SaaS and IaaS solutions is less important to IT organizations than is getting better at many other management and optimization tasks. One reason for that situation is that in the current environment it is often impossible to effectively manage and/or optimize a SaaS or an IaaS solution.

Private and Hybrid Cloud Computing

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 sub-section of this handbook.

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 BizApp. As that section pointed out, it is notably more difficult to troubleshoot BizApp in a virtualized environment than it would be to troubleshoot the same application in a traditional environment. Now assume that BizApp 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.***

To quantify the concerns that IT organizations have in managing cloud computing environments, The Survey Respondents were asked to indicate how important it was over the next year for their organization to get better at managing private, hybrid and public cloud computing solutions. Their responses are shown in **Table 13**.

Table 13: Importance of Managing Cloud Solutions

	Private Cloud	Hybrid Cloud	Public Cloud
Extremely	16.5%	9.2%	5.3%
Very	35.7%	31.1%	23.9%
Moderately	21.7%	25.2%	23.9%
Slightly	11.3%	15.1%	23.9%
Not at All	14.8%	19.3%	23.0%

One observation that can be drawn from the data in **Table 13** is that managing a private cloud is more important than managing a hybrid cloud which is itself more important than managing a public cloud. One of the primary reasons for this phenomenon is that as complicated as it is to manage a private cloud, it is notably more doable than is managing either a hybrid or public cloud and IT organizations are placing more emphasis on activities that have a higher chance of success.

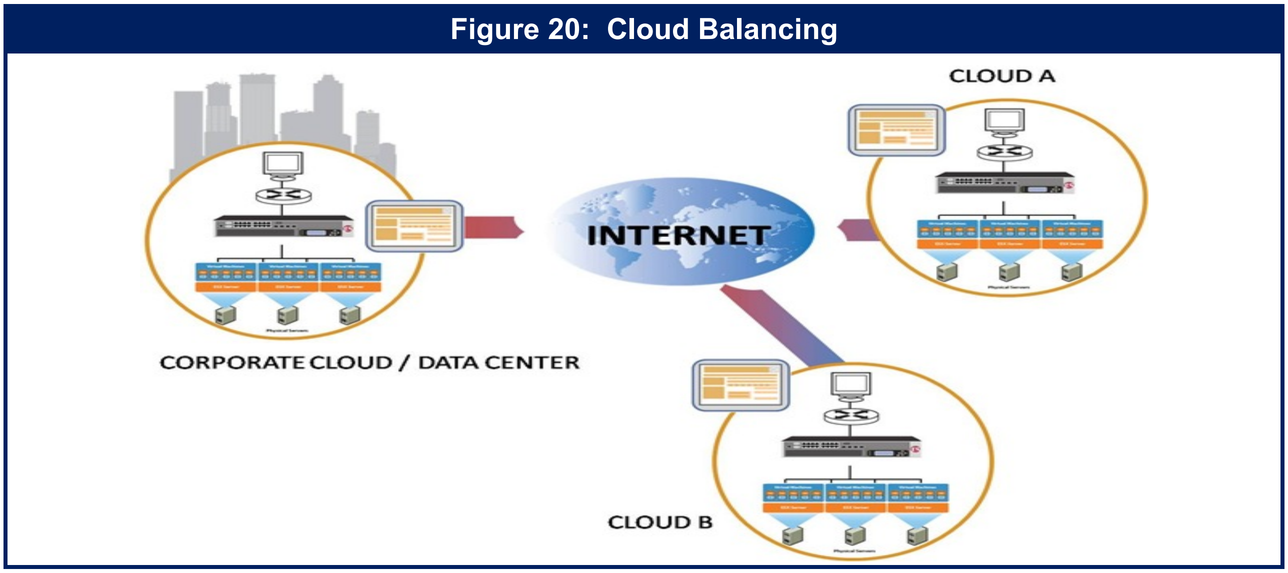
Cloud Balancing

Background

Cloud balancing refers to routing service requests across multiple data centers based on myriad criteria. As shown in **Figure 20**, cloud balancing involves one or more corporate data centers and one or more public cloud data centers. Cloud balancing is an example of hybrid cloud computing.

Cloud balancing can be thought of as the logical extension of global server load balancing (GSLB).

Figure 20: Cloud Balancing



The goal of a GSLB solution is to support high availability and maximum performance. In order to do this, a GSLB solution typically makes routing decisions based on criteria such as the application response time or the total capacity of the data center. A cloud balancing solution may well have as a goal supporting high availability and maximum performance and may well make routing decisions in part based on the same criteria as used by a GSLB solution. However, a cloud balancing solution extends the focus of a GSLB solution to a solution with more of a business focus. Given that extended focus, a cloud balancing solution includes in the criteria that it uses to make a routing decision the:

- Performance currently being provided by each cloud
- Value of the business transaction
- Cost to execute a transaction at a particular cloud
- Relevant regulatory requirements

Some of the benefits of cloud balancing include the ability to:

- ***Maximize Performance***
  Routing a service request to a data center that is close to the user and/or to one that is exhibiting the best performance results in improved application performance.

- ***Minimize Cost***
  Routing a service request to a data center with the lowest cost helps to reduce the overall cost of servicing the request.

- ***Minimize Cost and Maximize Service***
  Cloud balancing enables a service request to be routed to a data center that provides a low, although not necessarily the lowest cost while providing a level of availability and performance that is appropriate for each transaction.

- ***Comply with Data Privacy Regulations***
  The right to personal privacy is a highly developed area of law in parts of the world such as Europe. For example, all the member states of the European Union have data privacy laws that regulate the transfer of personal data to countries outside the European Union. In general, personal data may only be transferred to a country that is deemed to provide an adequate level of protection. Where such regulations come into play, it may be possible to execute data access portions of a web services application in a cloud data center located in the same country or regulatory domain as the data itself.

- ***Ensure Other Regulatory Compliance***
  For compliance with regulations such as PCI, it may be possible to partition a web services application such that the PCI-related portions remain in the PCI-compliant enterprise data center, while other portions are cloud balanced. In this example, application requests are directed to the public cloud instance unless the queries require the PCI-compliant portion, in which case they are directed to the enterprise instance.

- ***Managing Risk***
  Hosting applications and/or data in multiple clouds increases the availability of both. Balancing can be performed across a number of different providers or, as described below, it can be performed across multiple independent locations of a single cloud service provider. In view of the Cedexis data cited earlier, cloud balancing across two or more independent IaaS sites may be required in order to achieve acceptable availability for the public portion of a hybrid cloud solution.

  The global infrastructures of large cloud providers provide an opportunity for cloud balancing without the complexity of dealing with multiple providers. For example, Amazon EC2 locations are composed of Regions and Availability Zones. Availability Zones are distinct locations that are engineered to be insulated from failures in other Availability Zones and are provided with low latency network connectivity to other Availability Zones in the same Region. In theory, cloud balancing across Availability Zones or Regions can greatly reduce the probability of outages within the Amazon AWS global cloud. However, an outage that Amazon suffered in April 2011 gave the indication that the Availability Zones didn't provide the promised protection[29].

Enabling Cloud Balancing

One of the goals of cloud balancing is to have the collection of individual data centers appear to both users and administrators as a single cloud data center, with the physical location of application resources as transparent as possible. The goal of having the location of application resources be transparent has a number of implications, including

[29] TheRegister.co.uk

- ***VLAN Extension***
  The VLANs within which VMs are migrated must be extended over the WAN between the private and public data centers. This involves the creation of an overlay network that allows the Layer 2 VLAN traffic to be bridged or tunneled through the WAN.

- ***Secure Tunnels***
  These tunnels must provide an adequate level of security for all the required data flows over the Internet. For the highest level of security, this would typically involve both authentication and encryption, such as that provided by IPsec tunnels.

- ***Universal Access to Central Services***
  All application services, such as load balancing, DNS, and LDAP should be available and function transparently throughout the hybrid cloud. This approach allows these application services to be provisioned from the private enterprise data center and it also eliminates the need for manual intervention to modify server configurations as the application and the associated VM are transferred from the private cloud to the public cloud.

- ***Application Performance Optimization***
  Application performance must meet user expectations regardless of the location of the user or the servers. This means that the public cloud data center extensions must provide effective network and application optimization functionality. In addition, high throughput WAN optimization controllers on each end of the bridged connection between the enterprise private cloud data center and the public cloud data center can accelerate VM migration, system backups, and other bulk data transfers between these data centers.

- ***Application Delivery Controller (ADC) Virtual Appliances***
  One way to maintain a consistent architecture across private and public clouds is to use virtual versions of WAN optimization controllers (vWOCs) and ADCs (vADCs). These virtual appliances can be installed in virtual machines in the various clouds that comprise the global hybrid cloud infrastructure. This allows the enterprise to standardize on a single architecture across the entire cloud balancing environment as long as the virtual appliances support the hypervisors employed by the relevant IaaS providers. One of the advantages of this architectural consistency is that it insures that each cloud site will be able to provide the information needed to make global cloud balancing routing decisions.

- ***Interoperability Between Local and Global ADC Functions***
  Cloud balancing is based on making routing decisions based on a combination of local and global variables. This requires interoperability between local and global ADC functions. Standards-based APIs may eventually emerge that will facilitate the cross-vendor exchange of cloud balancing variables. In the mean time, in those situations in which multiple ADC vendors are involved, IT organizations will need to take advantage of the APIs supported by each vendor in order to achieve an integrated set of variables to use to make routing decisions. Another option that IT organizations have is to adopt a single vendor strategy for both local and global ADC functions. The feasibility of implementing a single vendor strategy across the enterprise and one or more IaaS providers is enhanced if the ADC is available in a virtual appliance form factor.

- ***Synchronizing Data between Cloud Sites***
  In order for an application to be executed at the data center that is selected by the cloud balancing system, the target server instance must have access to the relevant data. In some cases, the data can be accessed from a single central repository. In other cases, the data needs to co-located with the application. The co-location of data can be achieved by migrating the data to the appropriate data center, a task that typically requires highly effective optimization techniques. In addition, if the data is replicated for simultaneous use at multiple cloud locations, the data needs to be synchronized via active-active storage replication, which is highly sensitive to WAN latency.

Optimizing and Securing the Use of the Internet

Figure 16 in the preceding section of the handbook highlights some of the common distribution models for cloud based services. As was discussed in that section, it isn't possible to provide end-to-end quality of service over the Internet. The inability to ensure the end-to-end performance of the Internet can be a problem, however, whether or not it is a cloud based service that is being supported.

Figure 17 in the preceding section of the handbook identifies some of the popular categories of applications that can be acquired from a CCSP. The applications identified in that figure are well known enterprise applications including CRM, SCM and ERP. For the last few years, this is the class of applications that has been most closely associated with cloud computing. While those applications will continue to be closely associated with cloud computing, it is becoming increasingly common for organizations to acquire a different category of application from a CCSP. That class of applications is traditional network and infrastructure services such as VoIP, unified communications, management, optimization and security. Throughout this document, if such a service is provided by a CCSP it will be referred to as a Cloud Networking Service (CNS).

One goal of this section is to quantify the interest that IT organizations have in using CNSs. Other goals of this section are to describe some of the performance and security challenges associated with using the Internet and to also describe how CNSs can mitigate these challenges.

The Interest in CNSs

The Survey Respondents were asked to indicate how likely it was over the next year that their company would acquire a CNS. Their responses are shown in **Table 10**.

Table 14: Interest in Cloud Networking Services

	Will Not Happen	Might Happen	50/50 Chance	Will Likely Happen	Will Happen
Application Hosting	18.4%	23.4%	12.8%	19.1%	26.2%
VoIP	34.3%	17.5%	12.6%	15.4%	20.3%
Unified Communications	26.1%	26.8%	16.9%	14.8%	15.5%
Network and Application Optimization	33.8%	22.1%	14.7%	14.0%	15.4%
Disaster Recovery	30.8%	23.8%	20.0%	11.5%	13.8%
Security	39.0%	16.9%	16.9%	14.0%	13.2%
Network Management	38.8%	26.6%	7.2%	17.3%	10.1%
Application Performance Management	35.8%	28.4%	15.7%	12.7%	7.5%

Table 14: Interest in Cloud Networking Services

	Will Not Happen	Might Happen	50/50 Chance	Will Likely Happen	Will Happen
Virtual Desktops	40.7%	24.4%	18.5%	9.6%	6.7%
High Performance Computing	41.9%	24.8%	16.3%	10.1%	7.0%

The data in **Table 14** shows that there is strong interest in a number of CNSs – most notably application hosting and VoIP. The interest in CNSs, however, is quite broad as over twenty-five percent of The Survey Respondents indicated that over the next year that each of the services listed in the top seven rows of **Table 14** would either likely be acquired or would be acquired. That represents the beginning of what could be a fundamental shift in terms of how IT services are provisioned.

Over the next year IT organizations intend to make a significant deployment of Cloud Networking Services.

The factors that are driving IT organizations to consider CNSs are the same factors that are driving companies to consider any cloud computing service. Those factors were identified in **Figure 19** of the preceding section of the handbook and include:

- Lowering cost
- Reducing the time it takes to deploy new functionality
- Being able to acquire functionality that was not previously available
- Freeing up resources

The Use of the Internet

As was quantified in a market research report entitled *The 2010 Guide to Cloud Networking*, the two most commonly used WAN services are MPLS and the Internet. As that report also documented, while other WAN services (e.g., Frame Relay, ATM) are losing in popularity, the Internet is gaining in popularity.

The growing attractiveness of the Internet is due in part to the fact that it is a lower cost alternative to WAN services such as Frame Relay and MPLS, and in part 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 an 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 will continue to increase at a significant rate.

Over the last few years that IT organizations have focused on ensuring acceptable 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[30].

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 a CNS that provides that functionality. These Internet optimization services are based primarily on proprietary application acceleration and WAN optimization servers located at points of presence (PoPs) distributed across the Internet and like most cloud based services, they don't 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 Internet optimization service with complementary application acceleration technologies provided by WAN optimization controllers (WOCs) or application delivery controllers (ADCs) deployed in the data center or at remote sites.

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 – whether those applications are provided by the company's IT organization or acquired from a CCSP.

The primary reason for the limitation of the Internet is that as pointed out by Wikipedia[31], 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, service level agreements (SLAs) for the availability and performance of the Internet are currently not available and are unlikely to ever be 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 the most common business critical applications) behave as good network citizens. This means that these applications react to a lost packet by reducing the offered load by halving the transmission window size and then by 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.

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

With UDP-based applications, such as VoIP, Videoconferencing, and streaming video, there is not a congestion control mechanism. 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 that 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, securing application sessions and transactions using SSL authentication and encryption provides extra security. However the 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. The Survey Respondents recognized this fact. As previously mentioned, over 80% of The Survey Respondents indicated that over the next year it was at least moderately important to their organization to get better at optimizing the performance of TCP.

One of the characteristics of TCP that can lead to poor performance 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 has been widely analyzed[32]. 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:

Figure 21: Factors that Impact Throughput

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

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

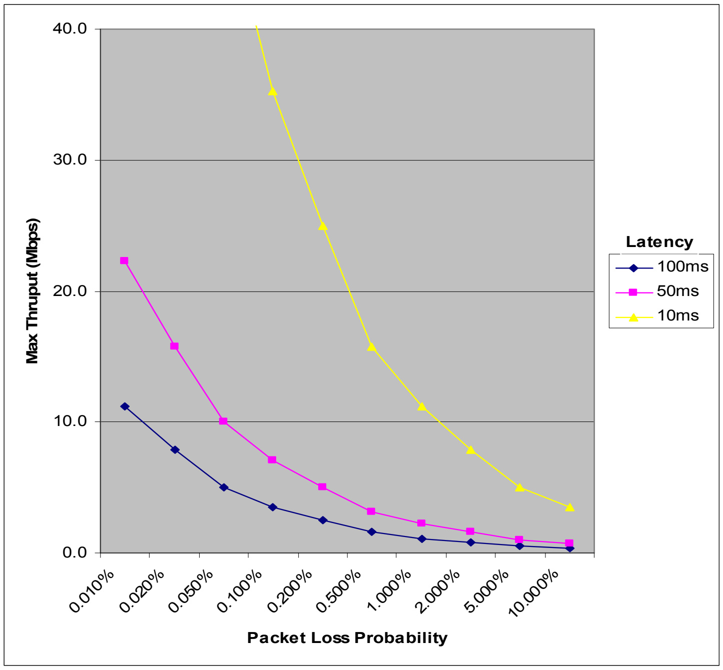
The preceding equation shows that throughput decreases as either the RTT or the packet loss rate increases. To illustrate the impact of packet loss, assume that MSS is 1,420 bytes, RTT is 100 ms. and p is 0.01%. Based on the formula, the maximum throughput is 1,420 Kbytes/second. If however, the loss were to increase to 0.1%, the maximum throughput drops to 449 Kbytes/second. **Figure 22** 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.

One conclusion we can draw from **Figure 22** 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.

Figure 22: Impact of Packet Loss on Throughput

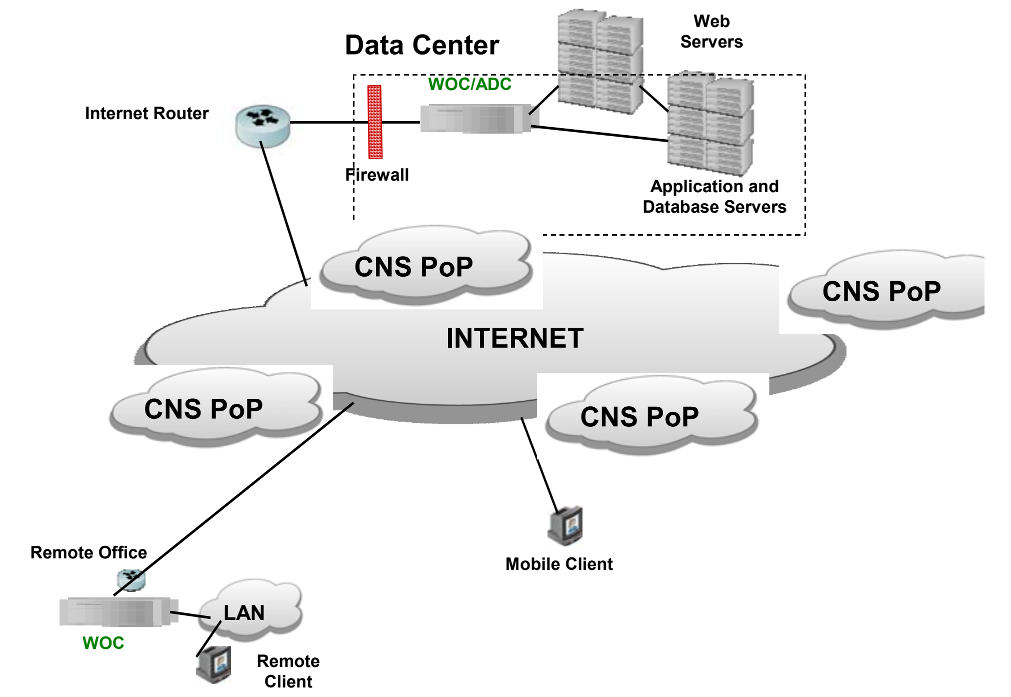


[32] The macroscopic behavior of the TCP congestion avoidance algorithm by Mathis, Semke, Mahdavi & Ott in Computer Communication Review, 27(3), July 1997

Internet-Based Application Delivery Optimization

The traditional classes of application delivery solutions (e.g., ADC, WOC) that are described in detail in a subsequent chapter of the handbook 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 are not capable of being optimized 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 the use of a CNS. Such a CNS would out of necessity leverage 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 23**, all client requests to the application's origin server in the data center are redirected via DNS to a server in a nearby PoP that is part of the CNS. This edge server then optimizes the traffic flow to the CNS server closest to the data center's origin server.

Figure 23: The Internet Infrastructure for a CNS



The servers at the CNS 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 functions provided by the CNS 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 CNS servers that are deployed in the service provider's PoPs to measure the performance of multiple paths through the Internet

and to choose 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 CNS servers in the ingress and egress PoPs. The edge CNS servers can also apply asymmetrical TCP optimization to the transport between the subscriber sites and the PoPs that are associated with the CNS. 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 CNS servers. By controlling both ends of the long-haul Internet connection with symmetric CNS 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 CNS 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 CNS server with the cache performing intelligent pre-fetching from the origin. With pre-fetching, the CNS 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 CNS 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 CNS 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. CNS content offload complements ADC functionality to further enhance the scalability of the data center.

- ***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 CNS 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 previously described, one of the characteristics of the current environment 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.

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 use of a CNS to provide Web application firewall functionality that is distributed throughout the Internet. This means that Web application functionality is close to the source of security attacks and hence can prevent many security attacks from reaching the organization. The distribution of security functionality as part of a CNS is analogous to the distribution of optimization functionality as part of a CNS that was discussed in the preceding subsection.

In the current environment, high-end DDoS attacks can generate 100 Gbps of traffic or more[33]. Attacks of this magnitude cannot be prevented by onsite solutions. They can, however, be prevented by implementing a CNS that includes security functionality analogous to what is provided by a Web application firewall and that can identify and mitigate 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 CNS-based Web application firewall service needs to be deployed as broadly as possible, preferably in tens of thousands of locations. When responding to an attack, the 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.

[33] DDoS-attacks-growing-in-size

- 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[34], 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 a CNS that supports a Web application firewall service. That follows because while a CNS-based 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.

[34] Wikipedia Tarpit(networking)

Planning

In the classic novel *Alice in Wonderland,* English mathematician Lewis Carroll first explained part of the need for why planning is important to application and service delivery (though he may not have known it at the time). In the novel, 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 and service delivery requires careful planning.

Many planning functions are critical to the success of application delivery. One planning function that has been previously discussed, and will be discussed again in the next sub-section of this handbook, is identifying the company's key applications and services and establishing SLAs for them. As described in the next sub-section, it is not sufficient to just establish SLAs for the company's key applications and services. IT organizations must also identify the key elements (e.g., specific switches and routers, WAN links, servers, virtual machines) that support each of the applications. Other key steps include:

- Baselining the performance of each of the organization's critical applications.
- Baselining the performance of each of the key elements that support each of the critical applications and identifying at what levels of utilization and delay the performance of each of the elements has an unacceptable impact on the performance of the application.
- Establishing SLAs for each of the key elements.

Another key planning activity that is discussed in the next sub-section of this handbook is Application Performance Engineering (APE).

The goal of APE is to help IT organizations reduce risk and build better relationships with the company's business unit managers.

APE achieves this goal by anticipating, and wherever possible, eliminating performance problems at every stage of the application lifecycle.

Another key planning activity is performing a pre-deployment assessment of the current environment to identify any potential problems that might affect an IT organization's ability to deploy a new application. One task that is associated with this activity is to either create or update the IT organization's inventory of the applications running on the network. Part of the value of this task is to identify unauthorized use of the network; i.e., on-line gaming and streaming radio or video. Blocking unauthorized use of the network can free up additional WAN bandwidth. Another part of the value of this task is to identify business activities, such as downloads of server patches that are being performed during peak times. Moving these activities to an off-peak time also releases additional bandwidth.

Another task associated with performing a pre-deployment assessment is to create a current baseline of the network and the key applications. Relative to baselining the network, IT organizations should modify how they think about baselining to focus not just on utilization, but also on delay. In some instances, however, even measuring delay is not enough. If, for example, a company is about to deploy an application such as telepresence then the pre-assessment baseline must also measure the current levels of jitter and packet loss. Relative to baselining the company's key 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 key applications.

Integrating Network Planning and Network Operations

As noted, the next section of the handbook discusses APE. One of the characteristics of APE is that it is a life cycle approach to planning and managing application performance. 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 planning and operations 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 the goal of which is to reduce complexity and maximize the productivity of the IT staff.

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. These solutions, however, cannot identify how the intended changes would impact the overall performance of the network.

Route Analytics

A class of management tool that can facilitate the integration of planning and operations is typified by an IP route analytics solution[35].

> ***The goal of route analytics is to provide visibility, analysis and diagnosis of the issues that occur at the routing layer in complex, meshed networks.***

A route analytics appliance draws its primary data directly from the network in real time by participating in the IP routing protocol exchanges. This allows the route analytics solution 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. As described in the next section of the handbook, a logical issue such as route flapping typically causes notably more business disruption than does a physical issue and a logical issue typically takes notably longer to troubleshoot and repair than does a physical issue.

Route analytics is also valuable because it can be used to eliminate problems stemming from human errors in a router's 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:

[35] More information on this topic can be found at: Webtorials.com

- 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 Cloud Computing Service Provider (CCSP).

- What levels of service are good enough for each class of application and for the myriad storage and compute requirements.

- How the IT organization will evolve over time the twelve characteristics of a cloud computing solutions that were discussed in a previous section of the handbook; e.g., virtualization, automation, simplification.

- How the IT organization will evolve its data center LAN architecture to support cloud computing.

- How the IT organization 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.

Application Performance Management[36]

Background

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. In spite of the newness of APM, over a quarter of The Survey Respondents said that APM was currently important to their organization and another one third of The Survey Respondents said that it is important to their organization to get better at APM. In addition to the fact that APM in general is important to IT organizations, some specific components of APM are particularly important. For example, as described below, a critical component of APM is the adoption of service level agreements (SLAs). As described in a preceding section of the handbook, two thirds of The Survey Respondents indicated that over the next year it is either very important or extremely important for their organization to get better at managing SLAs for one or more business critical applications.

Successful APM requires a holistic approach based on integrated management of both the application and/or service itself as well as the end-to-end IT infrastructure. However, only about 15% of The Survey Respondents indicated that their organization's approach to APM was both top down and tightly coordinated.

Only a small minority of IT organizations has a top down, tightly coordinated approach to APM.

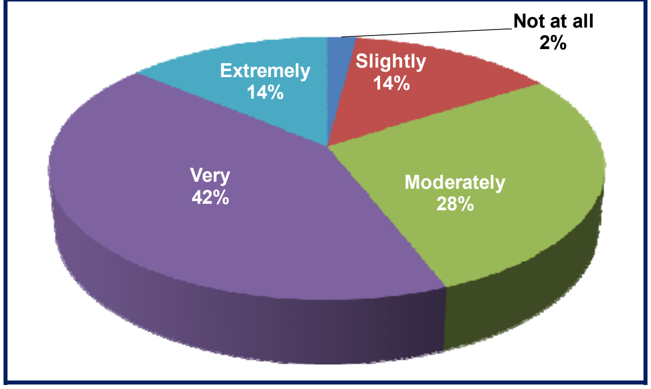
A holistic approach to APM that is based on the integrated management of both the application itself as well as the end-to-end IT infrastructure must focus on the experience of the end user of the application or service. Monitoring actual user transactions in production environments provides valuable insight into the end-user experience and provides the basis for an IT organization to be able to quickly identify, prioritize, triage and resolve problems that can affect business processes.

To quantify the interest that IT organizations have in this task, The Survey Respondents were asked how important it was over the next year for their organization to get better at monitoring the end user's experience and behavior. Their responses are shown in Figure 24.

.

Over the next year, getting better at monitoring the end user's experience and behavior is either very or extremely important to the majority of IT organizations.

Figure 24: Getting Better at Monitoring End User Behavior



[36] The phrase APM will be used to apply both to an application and to a service as previously defined in this handbook.

A holistic approach to APM must also address the following aspects of management:

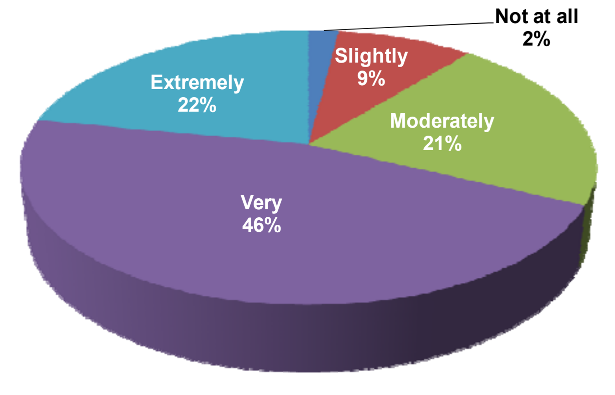
- The adoption of a system of service level agreements (SLAs) at levels that ensure effective business processes and user satisfaction for at least a handful of key applications.

- Automatic discovery of all the elements in the IT infrastructure that support each service. This functionality provides the basis for an IT organization to being able 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.

The Survey Respondents were asked how important it was over the next year for their organization to get better at identifying the components of the IT infrastructure that support the company's critical business applications. Their responses are shown in Figure 25.

Getting better at identifying the components of the IT infrastructure that support the company's critical business applications and services is one of the most important management tasks facing IT organizations.

Figure 25: Importance of Identifying Components of Critical Applications



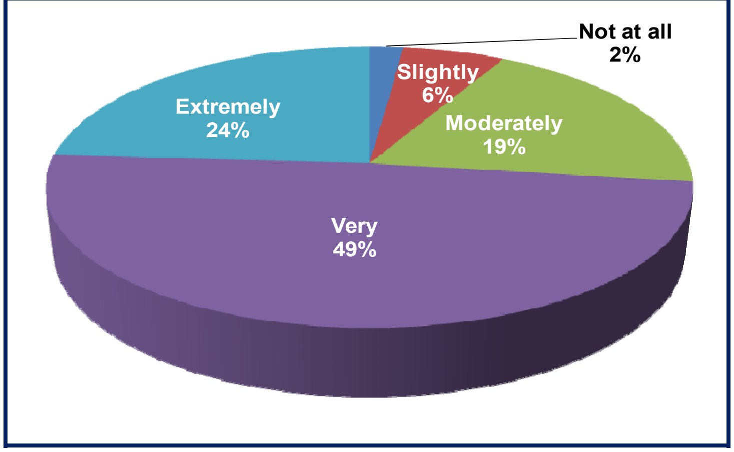
If IT organizations can effectively identify which components of the infrastructure support a particular application or service, monitoring can much more easily identify when services are about to begin to degrade due to problems in the infrastructure. As part of this monitoring, predictive techniques 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. In addition, outages and other incidents that generate alerts can be prioritized based on their 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.

Once the components of the infrastructure that support a given application or service has been identified, triage and root cause analysis can be applied at both the application and the 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.

The Survey Respondents were asked how important it was over the next year for their organization to get better at rapidly identifying the causes of application degradation. Their responses are shown in Figure 26.

Getting better at rapidly identifying the causes of application degradation is the most important management task facing IT organizations.

Figure 26: The Importance of Getting Better at Root Cause Analysis



As part of an effective approach to APM, the automated generation of performance dashboards and historical reports allows both IT and business managers to gain insight into SLA compliance and performance trends. The insight that can be gleaned from these dashboards and reports can be used to enhance the way that IT supports key business processes, help the IT organization to perform better capacity and budget planning, and identify where the adoption of new technologies can further improve the optimization, control and management of application and service performance. Ideally, the dashboard is a single pane of glass that can be customized to suit different management roles; e.g., the individual contributors in the Network Operations Center, senior IT management as well as senior business management.

Challenges for Application Management

Below is a discussion of some of the technical factors that complicate the ability of IT organizations to perform the APM related tasks that were described in the preceding sub-section of the handbook. While the technical factors present a significant challenge, an equally significant challenge is organizational – the difficulty of actually taking a top down, tightly integrated approach to APM.

Port Hopping

As previously noted, identifying the applications and services that are running on a network is a critical part of managing application performance. TCP and UDP ports are frequently used by routers, firewalls and other network devices to identify the application that generated a particular packet. 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.

Some applications have been designed to use port hopping to avoid detection and blocking by firewalls. Applications that do port hopping create significant management and security challenges. 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. 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. Analogously, network management might see that there is no traffic on ports 5190 – 5193 and assume that AIM is not being used. That may or may not be the case because if these ports are blocked AIM will use port 80 in an effort to circumvent the firewall via the Port 80 black hole described below.

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 – a company that Microsoft recently acquired. 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.

The Port 80 Black Hole

Many enterprise applications are accessed via browsers over port 80. Therefore, a firewall can't block port 80 without eliminating much of the traffic on which a business may depend. As mentioned, some applications will port-hop to port 80 when their normally assigned ports are blocked by a firewall. In addition, the port number 80 can't be used as a means of identifying individual web based enterprise applications and port 80 becomes a black hole unless firewalls and other devices are capable of deep packet inspection to identify Layer 7 application signatures.

Lack of visibility into the traffic that transits port 80 is a significant management and security challenge for most IT organizations.

The port 80 black hole can have four primary effects on an IT organization. It can cause increased:

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

Server Virtualization

Server virtualization presents a number of challenges relative to APM. For example, the VMs that reside on a given physical server communicate with each other using a vSwitch within the server's hypervisor software. As discussed in the section of this handbook entitled Virtualization, unlike the typical physical switch, a vSwitch usually provides limited visibility for the traffic that is internal to the physical server. In addition, prior to virtualization, most server platforms were dedicated to a single application. With server virtualization, virtual machines share the server's CPU, memory and I/O resources. Over-subscription of VMs on a physical server can result in application performance problems due to factors such as limited CPU cycles or memory or I/O bottlenecks. One way to mitigate the impact of the over-subscription of VMs is to implement functionality such as VMotion[37] in an automated fashion. However, automated VMotion creates additional challenges.

While the problems discussed in the preceding paragraph can occur in a traditional physical server, they are more likely to occur in a virtualized server due to the consolidation of multiple applications onto a single shared physical server. In addition, as described in the section of this handbook entitled Virtualization, it is notably more difficult to troubleshoot a performance problem in a virtualized environment than it is in a traditional physical environment. That is why, as is also pointed out in that section of the handbook, half of the IT organizations consider it to be either very or extremely important over the next year for them to get better performing management tasks such as troubleshooting on a per-VM basis.

Mobility

Another factor that is making APM more complex is that most IT organizations have to support a growing number of mobile employees[38]. As described in the section of the handbook entitled Application and Service Delivery Challenges, at one time mobile workers tended to primarily access either recreational applications or business applications that were not very delay sensitive; e.g., email. However, mobile workers now need to access an increasingly 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. As such, there is a significant risk that an application that performs well when accessed over a wired network will run poorly when accessed over a wireless network.

[37] VMWare.com VMotion

[38] One analyst firm has predicted that there will be one billion mobile workers worldwide by year end 2011 - FindArticles.com

The challenges associated with supporting mobility are why, as highlighted in the section of the handbook entitled Application and Service Delivery Challenges, two thirds of The Survey Respondents indicated that over the next year it is either moderately or very important for their IT organization to get better at managing the performance of applications delivered to mobile users.

Cloud Computing

There are many ways that the adoption of cloud computing adds to the complexity of APM. For example, assume that the 4-tier application BizApp that was described in the section of this report that is entitled *Virtualization*, is moved to a cloud computing service provider's data center. Without the appropriate tools and processes it is impossible to tell in advance what impact that move will have on application performance. However, the fact that BizApp will run on different servers, which are most likely virtualized, and is accessed over different WAN links than it had been previously, means that the application performance will be different. This lack of ability to understand in advance how a change in the IT environment will impact the performance of an application is one of the factors driving the need for Application Performance Engineering which is described below.

As was described in the section of the handbook entitled *Virtualization*, troubleshooting any performance degradation exhibited by BizApp is complex even if each tier of the application is hosted by an enterprise IT organization. However, if one or more tiers of the application are hosted by a CCSP troubleshooting becomes notably more complex because management data must now be gathered from multiple organizations.

APM in the Private Enterprise Network[39]

Enterprise IT organizations can choose among several types of tools for monitoring and managing application performance over a private enterprise network. These include: application agents, monitoring of real and synthetic transactions, network flow and packet capture, analytics, and dashboard portals for the visualization of results.

At a high level, there are two basic classes of tools. The first class of tool monitors global parameters such as user response time or transaction completion time and provides alerts when thresholds are exceeded. These tools include agents on end user systems and monitoring appliances in the data center. The second class of tool supports triage by monitoring one or more of the components that make up the end-to-end path of the application. These tools include devices that capture application traffic at the flow and packet levels, agents on database, application, and web servers, as well as agents on various network elements.

The ultimate goal of APM is have a single screen that integrates the information from all of the tools in both categories. The idea being that a dashboard on the screen would indicate when user response time or transaction completion time begins to degrade. Then, within a few clicks, the administrator could determine which component of the infrastructure was causing the degradation and could also determine why that component of the infrastructure was causing degradation; e.g., high CPU utilization on a router.

Each type of individual tool has its strengths and weaknesses. For example, agents can supply the granular visibility that is required for complex troubleshooting but they represent an additional maintenance burden while also adding to the load on the servers and on the network. Monitoring appliances have more limited visibility, but they don't require modification of server configurations and don't add traffic to the network. Taking into consideration these trade-offs, IT organizations need to make tool decisions based on their goals for APM, their application and network environment as well as their existing infrastructure and network management vendors.

A complete discussion of APM tools and methodology is outside the scope of this section of the handbook. That said, the remainder of this section is devoted to the following topics that are of particular importance for APM within the private enterprise network:

- **Application Performance Engineering** that deals with the processes of optimizing the performance of applications over their lifecycles.

- **End-to-End Visibility** of all aspects of all the infrastructure components that can have an effect of application performance.

- **Route Analytics** that deals with mitigating the logical issues within the routed IP network that can negatively impact application performance.

[39] This refers to managing the performance of applications that are delivered over WAN services such as Frame Relay, ATM and MPLS.

Application Performance Engineering

Ideally the issue of application performance would be addressed at all stages of an application's lifecycle, including multiple iterations through the design/implement/test/operate phases as the application versions are evolved to meet changing requirements. However, as discussed in a preceding section of the handbook, the vast majority of IT organizations don't have any insight into the performance of an application until after the application is fully developed and deployed. In addition, the vast majority of IT organizations have little to no insight into how a change in the infrastructure, such as implementing server virtualization, will impact application performance prior to implementing the change.

Application Performance Engineering (APE) is the practice of first designing for acceptable application performance and then testing, measuring and tuning performance throughout the application lifecycle.

During the operational, or production phase of the lifecycle, APM is used to monitor, diagnose, and report on application performance. APM and APE are therefore highly complementary disciplines. For example, once an APM solution has identified that an application in production is experiencing systemic performance problems, an APE solution can be used to identify the root cause of the problem and to evaluate alternative solutions. Possible solutions include modifying the application code or improving application performance by making changes in the supporting infrastructure, such as implementing more highly performing servers or deploying WAN Optimization Controllers (WOCs). Throughout this section of the handbook, implementing products such as WOCs will be referred to as a Network and Application Optimization (NAO) solution. Independent of which remedial option the IT organization takes, the goal of APE can be realized – performance bottlenecks are identified, root causes are determined, alternative remedies are analyzed and bottlenecks are eliminated.

An IT organization could decide to ignore APE and just implement NAO in a reactive fashion in an attempt to eliminate the sources of the degraded application performance. Since this approach is based on the faulty assumption that NAO will resolve all performance problems, this approach is risky. This approach also tends to alienate the company's business unit managers whose business processes are negatively impacted by the degraded application performance that is not resolved until either WOCs are successfully deployed or some other solution is found. A more effective approach was described in the preceding paragraph. This approach calls for NAO to be a key component of APE – giving IT organizations another option to proactively eliminate performance problems before they impact key business processes.

The key components of APE are described below. The components are not typically performed in a sequential fashion, but in an iterative fashion. For example, as a result of performing testing and analysis, an IT organization may negotiate with the company's business unit managers to relax the previously established performance objectives.

- ***Setting Performance Objectives***
  This involves establishing metrics for objectives such as user response time, transaction completion time and throughput. A complex application or service, such as unified communications, is comprised of several modules and typically different objectives need to be established for each module.

- ***Discovery***

Performance modeling and testing should be based on discovering and gaining a full understanding of the topology and other characteristics of the production network.

- ***Performance Modeling***
  APE modeling focuses on creating the specific usage scenarios to be tested as well as on identifying the performance objectives for each scenario. A secondary focus is to identify the maximum utilization of IT resources (e.g., CPU, memory, disk I/O) and the metrics that need to be collected when running the tests.

- ***Performance Testing and Analysis***
  Test tools can be configured to mimic the production network and supporting infrastructure, as well as to simulate user demand. Using this test environment, the current design of the application can be tested in each of the usage scenarios against the various performance objectives. The ultimate test, however, is measured performance in the actual production network or in a test environment that very closely mimics the actual production environment.

- ***Optimization***
  Optimization is achieved by identifying design alternatives that could improve the performance of the application and by redoing the performance testing and analysis to quantify the impact of the design alternatives. In conjunction with the testing, an ROI analysis can be performed to facilitate cross-discipline discussion of the tradeoffs between business objectives, performance objectives, and cost. This component of APE is one of the key ways that APE enables an IT organization to build better relationships with the company's business unit managers.

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 until they receive a response back from the 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 symptoms of the degradation and as a result enables the IT organization to reduce the amount of time it takes to identify and remove the causes of the degraded application performance.

- 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 manage the recreational use of the network.

- Allows the IT organization to measure the performance of a critical application before, during and after a change is made. These changes could be infrastructure upgrades, configuration changes or the adoption of a cloud computing delivery model. 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.

Visibility can enable better cross-functional collaboration if two criteria are met. One criterion is that all members of the IT organization use the same tool or set of tools. The second criterion is that the tool(s) are detailed and accurate enough to identify the sources of application degradation. One factor that complicates achieving this goal is that so many tools from so many types of vendors (e.g., APM, NAO) all claim to provide the necessary visibility.

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 from a wide range of vendors. 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.

The section of the handbook entitled Virtualization highlighted a visibility challenge created by server virtualization. That problem is that in most cases once a server is virtualised the IT organization looses visibility into the inter-VM traffic on a given server. There are a number of solutions for this problem. One of these solutions is based on configuring one of the ports on the virtual switch inside the server as a SPAN port or mirror port. This allows a monitor to capture flow and packet information within the physical server. The monitoring device can be a virtual appliance installed on the physical server. Transaction and response time monitors are also available as virtual appliances. While changes in the virtual topology can be gleaned from flow analysis, a more direct approach is for the APM tool to access data in the hypervisor's management system via supported APIs. Gathering data from this source also provides access to granular performance information such as a VM's utilization of allocated CPU and memory resources.

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 application layer protocols; e.g. FTP, Telnet, HTTPS and SSH.
- Services, where a service is comprised of multiple inter-related applications.
- Applications provided by a third party; e.g., Oracle, Microsoft, SAP.
- 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.
- 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.
- 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.

Route Analytics

Background

The use of route analytics for planning purposes was discussed in the preceding section of the handbook. 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 in a traditional IT environment 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.

As previously noted in this handbook, the majority of IT organizations have already implemented server virtualization and the amount of server virtualization is expected to increase over the next year. Once an IT organization has implemented server virtualization, or a private cloud computing solution that includes server virtualization, VMs can be transferred without service interruption from a given physical server to a different physical server. This can make it difficult for the network operations team to know the location of an application at any given point in time – a fact that makes troubleshooting a problem that much more difficult.

To exemplify a related management challenge, assume that an IT organization has implemented a type of hybrid cloud computing solution whereby the IT organization hosts the application and data base tiers in one of their data centers and that the relevant servers have been virtualized. Further assume that a CCSP hosts the application's web tier and that all of the CCSP's physical servers have been virtualized. All of the users access the application over the Internet and the connectivity between the web server layer and the application server layer is provided by an MPLS service.

Since the web, application and database tiers can be moved, either dynamically or manually, it is extremely difficult at any point in time for the IT operations organization to know the exact routing between the user and the web tier, between the Web tier and the application tier or between the application tier and the database tier. This difficulty is compounded by that fact that as previously discussed, not only does the location of the tiers of the application change, but the path that traffic takes to go from point A to point B also changes regularly.

The dynamic movement of VMs will increase over the next few years in part because organizations will increase their use of virtualization and cloud computing and in part because organizations will begin to deploy techniques such as cloud bursting. Cloud bursting refers to taking an application that currently runs in a data center controlled by an IT organization and dynamically deploying that application and the subtending storage in a data center controlled by a CCSP. Techniques such as cloud bursting will enable organizations to support peak demands while only deploying enough IT infrastructure internally to support the average demand. These techniques, however, will further complicate the task of understanding how traffic is routed end-to-end through a complex, meshed network.

The operational challenges that are created due to a lack of insight into the router layer are greatly exacerbated by the adoption of server virtualization and cloud computing.

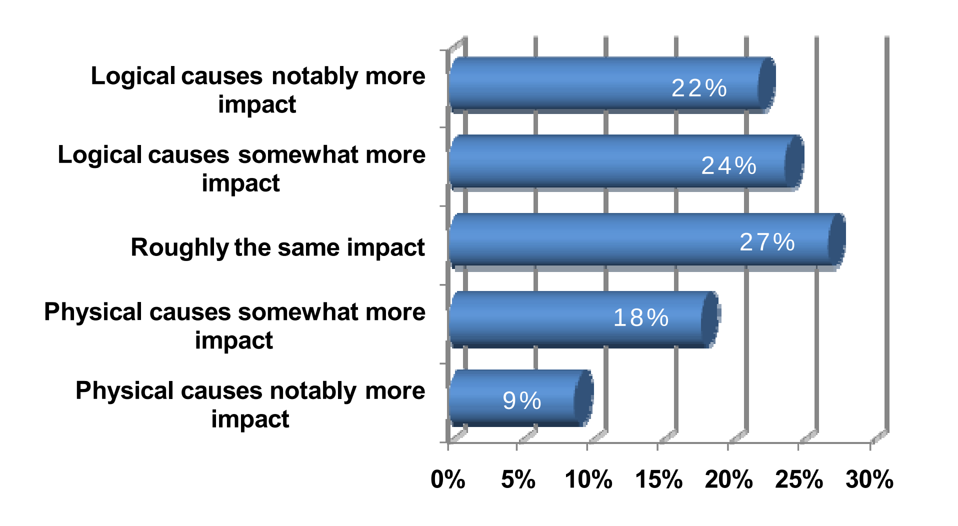
Logical vs. Physical Factors

Factors such as route flapping can be classified as logical as compared to a device specific factor such as a link outage, which is considered to be a physical factor. Both logical and physical factors impact application performance. In simple networks, such as small hub and spoke networks, logical factors are typically not a significant source of application degradation. However, in large complex networks that is not the case.

To quantify the relative impact of logical and physical factors, The Survey Respondents were asked two questions.

One question asked The Survey Respondents to indicate the relative impact of logical and physical factors on the business disruption they cause. Their answers are shown in **Figure 27**.

Figure 27: Impact of Logical and Physical Factors on the Business

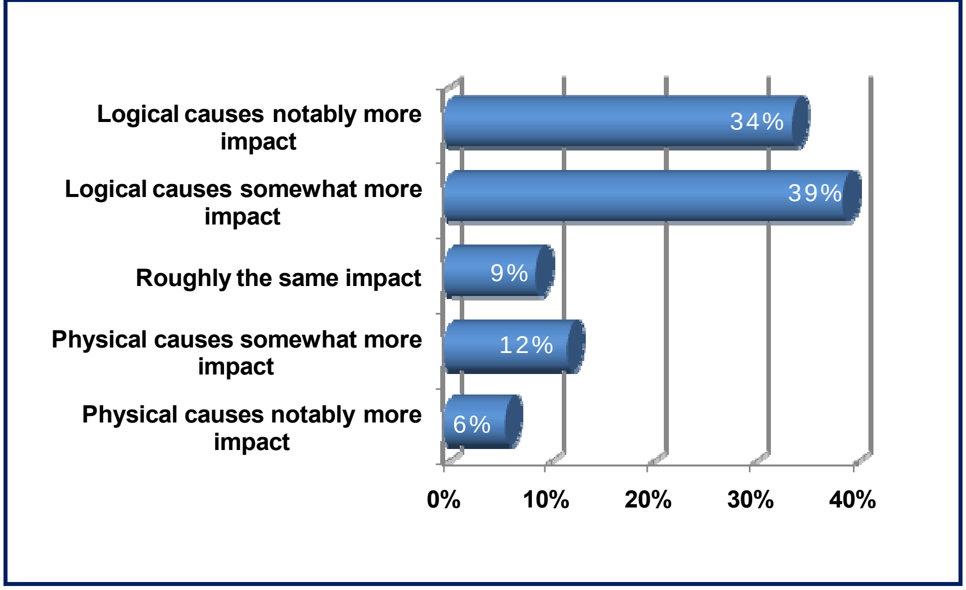


In the vast majority of cases, logical factors cause as much or more business disruption than do physical factors.

The other question asked The Survey Respondents to indicate the relative amount of time it takes to troubleshoot and repair a physical error vs. a logical error. Their answers are shown in **Figure 28**.

In the vast majority of instances, logical errors take either somewhat more or notably more time to troubleshoot and repair than do physical errors.

Figure 28: Impact of Logical and Physical Factors on Troubleshooting



SNMP-based management systems can discover and display the individual network elements and their physical or Layer 2 topology. However, these systems 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.

As noted in the preceding section, 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 the 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 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.

APM in Public and Hybrid Clouds

As is widely known, IT organization have begun to make significant use of public and hybrid cloud computing solutions and the use of those solutions is expected to increase significantly. Once enterprise applications are partially or completely hosted outside of private data centers, IT organizations will need to make some adjustments in their approach to APM. In particular, public clouds have a significant impact on each the topics discussed in the preceding section

- ***APE***
  While an enterprise IT organization might hope that a SaaS provider would use APE as part of developing their application, they typically can't cause that to happen. IT organizations can, however, use APE to quantify the impact of taking an application, or piece of an application, that is currently housed internally and hosting it externally. IT organizations can also use APE for other cloud related activities, such as quantifying the impact on the performance of a SaaS based application if a change is made within the enterprise. For example, APE can be used to measure the impact of providing mobile users with access to a SaaS-based application that is currently being used by employees in branch offices.

- ***End-to-End Visibility***
  The visibility necessary for effective APM can be compromised by the dynamic nature of cloud environments and by the difficulty of extending the enterprise monitoring solutions for application servers, Web servers, databases into a public IaaS cloud data center. Part of this challenge is that many IaaS providers have an infrastructure that has often been optimized based on simplicity, homogeneity and proprietary extensions to open source software.

- ***Route Analytics***
  As noted in the preceding section, both hosting enterprise assets at a CCSP's premise, and using services provided by a CCSP creates a more complex network topology. This fact combined with the potential for the dynamic movement of those assets and services increases the probability of a logical error. As such, the adoption of cloud based service increases the need for route analytics.

There are a number of possible ways that an IT organization can adjust their APM strategies in order to accommodate accessing services hosted by a CCSP. These include:

- Extend the Enterprise APM Monitoring solutions into the public cloud using agents on virtual servers and by using virtual appliances. This option assumes that the CCSP offers the ability to install multiple virtual appliances (e.g., APM monitors, WOCs, and ADCs) and to configure the virtual switches to accommodate these devices.

- Focus on CCSPs that offer either cloud resource monitoring or APM as a service as described in the section of the handbook entitled Cloud Networking Services. Basic cloud monitoring can provide visibility into resource utilization, operational performance, and overall demand patterns. This includes providing metrics such as CPU utilization, disk reads and writes and network traffic. The value of cloud monitoring is increased where it is tied to other capabilities such as automated provisioning of instances to maintain high availability and the elastic scaling of capacity to satisfy demand spikes. A possible issue with this option is integrating the cloud monitoring and enterprise monitoring and APM solutions.

- Increase the focus on service delivery and transaction performance by supplementing existing APM solutions with capabilities that provide an outside-in service delivery view from the perspective of a client accessing enterprise applications or cloud applications over the Internet or mobile networks. Synthetic transactions against application resources located in public clouds are very useful when other forms of instrumentation cannot be deployed. One option for synthetic transaction monitoring of web applications is a third party performance monitoring service with end user agents distributed among numerous global ISPs and mobile networks.

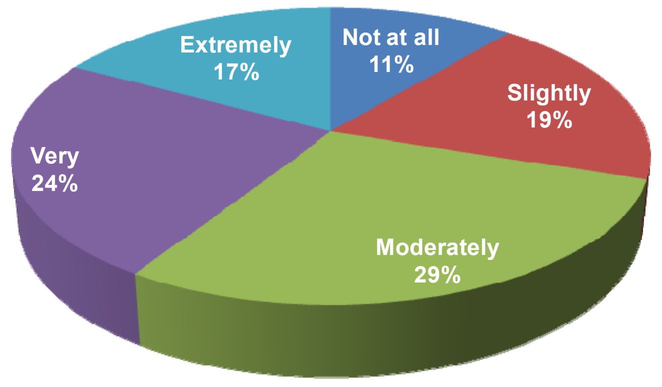
Network and Application Optimization

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 networked applications and services as part of assuring acceptable application performance while also controlling WAN bandwidth expenses. 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 networks that are typically over-subscribed;
- Mitigate the inefficiencies of protocols;
- Offload computationally intensive tasks from client systems and servers;
- Direct traffic to the most appropriate server based on a variety of metrics.

The functionality described in the preceding bullets is intended primarily to improve the performance of applications and services. However, another factor driving the use of optimization techniques is the desire to reduce cost. To quantify the impact of that factor, The Survey Respondents were asked to indicate how important it was to their organization over the next year to get better at controlling the cost of the WAN by reducing the amount of WAN traffic by techniques such as compression. Their responses are shown in **Figure 29**.

Figure 29: Importance of Using Optimization to Reduce Cost



The data in **Figure 29** indicates that improving performance is not the only reason why IT organizations implement optimization functionality.

The value proposition of network and application optimization is partly to improve the performance of applications and services and partly to save money.

As described in a previous section of the handbook, some optimizations tasks, such as optimizing the performance of a key set of business critical applications, has become extremely important to the vast majority of IT organizations. Because of that importance, The Survey Respondents were asked to indicate their company's approach to optimizing network and application optimization. Their responses are shown in **Table 15**.

Table 15: How IT Organizations Approach Network and Application Optimization

Response	Percentage
We implement very little if any functionality specifically to optimize network and application performance	27.4%
We implement optimization functionality on a case-by-case basis in response to high visibility problems	45.7%
We have implemented optimization functionality throughout our environment	21.3%
Other	5.5%

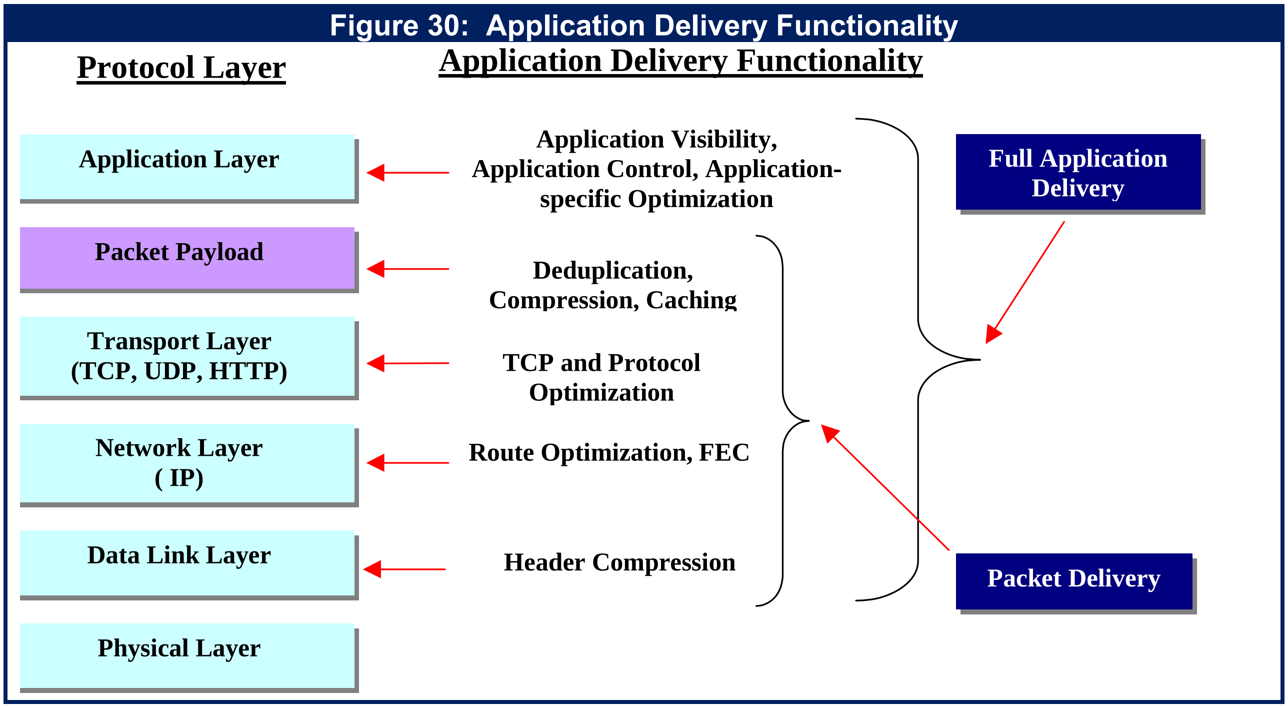
The most common way that IT organizations approach implementing optimization functionality is on a case-by-case basis.

As was previously explained in handbook, some Cloud Computing Service Providers (CCSPs) offer network and application optimization as a service. It is also possible for an IT organization to acquire and implement network and application optimization products such as WOCs and ADCs. In many cases, these two approaches are complimentary.

There are two principal categories of network and application optimization products. One category focuses on mitigating the negative effect that WAN services such as MPLS have on application and service performance. This category of products has historically included WAN optimization controllers (WOCs). However, due to some of the second generation of application and service delivery challenges, this category of products now also contains an emerging class of WAN optimization device - the Data Mobility Controller (DMC). As described in detail later in this section of the handbook, WOCs are focused primarily on accelerating end user traffic between remote branch offices and central data centers. In contrast, DMCs are focused on accelerating the movement of bulk data between data centers. This includes virtual machine (VM) migrations, storage replication, access to remote storage or cloud storage, and large file transfers. WOCs and DMCs are often referred to as *symmetric solutions* because they typically require complementary functionality at both ends of the connection. However, as is explained later in this section of the handbook, one way that IT organizations can accelerate access to a public cloud computing solution is to deploy WOCs in branch offices. The WOCs accelerate access by caching the content that a user obtains from the public cloud solution and making that content available to other users in the branch office. Since in this example there is not a WOC at the CCSP's site, this is an example of a case in which a WOC is an asymmetric solution.

When WOCs were first deployed they often focused on improving the performance of a protocol such as TCP or CIFS. As discussed in a preceding section of the handbook, optimizing those protocols is still important to the majority of IT organizations. However, as WOCs continue to evolve, much more attention is being paid to the application layer. As shown in **Figure 30**, WOC solutions that leverage application layer functionality focus on recognizing application layer signatures of end user applications and mitigating application-specific inefficiencies in communicating over the WAN. In contrast to WOCs, DMCs are focused primarily on the packet delivery functionality at the transport layer and with the packet payload. However, like the WOC, the DMC can use QoS functionality at the transport and application layers to classify traffic and identify traffic that requires optimization.

Figure 30: Application Delivery Functionality


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

In order to choose the most appropriate optimization solution, IT organizations need to understand their environment, including traffic volumes and the characteristics of the traffic they wish to accelerate. For example, the degree of data reduction experienced will depend on a number of factors including the degree of redundancy in the data being transferred over the WAN link, the effectiveness of the de-duplication and compression algorithms and the processing power of the WAN optimization platform. If the environment includes applications that transfer data that has already been compressed, such as the remote terminal traffic (a.k.a. server-side desktop virtualization), VoIP streams, or jpg images transfers, little improvement in performance will result from implementing advanced compression. In some cases, re-compression can actually degrade performance.

The second category of optimization 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 on the remote end. 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.

Because a network and application optimization solution will provide varying degrees of benefit to an enterprise 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 31**) is a variation of the application response time model created by Sevcik and Wetzel[40]. 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 a number of factors including the amount of data being transmitted (Payload), the goodput which is the actual throughput on a WAN link, 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 31: Application Response Time Model

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

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

[40] Why SAP Performance Needs Help

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. The Survey Respondents were asked to indicate their company's current deployment of WOCs. They were also asked whether or not they have currently deployed WOCs, to indicate their company's planned deployment of WOCs over the next year. Their responses are shown in **Table 16** and **Table 3** respectively.

Table 16: Current Deployment of WOCs

Response	Percentage
No deployment	50.6%
Employed in test mode	10.4%
Limited production deployment	18.9%
Broadly deployed	15.9%
Other	4.3%

Roughly half of IT organizations have not made any deployment of WOCs.

Table 17: Planned Deployment of WOCs

Response	Percentage
No plans	45.1%
Will deploy in test mode	8.5%
Will make limited production deployment	20.7%
Will deploy broadly	20.7%
Other	4.9%

Comparing the data in **Table 16** and **Table 17** yields the conclusion that:

Over the next year, IT organizations plan to make a moderate increase in their deployment of WOCs.

WOC Functionality

Table 18 lists some of WAN characteristics that impact application delivery and identifies WAN optimization techniques that a WOC can implement to mitigate the impact of the WAN.

Table 18: Techniques to Improve Application Performance

WAN Characteristics	WAN Optimization Techniques
Insufficient Bandwidth	Data Reduction: • Data Compression • Differencing (a.k.a., de-duplication) • Caching

Table 18: Techniques to Improve Application Performance

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 by WOCs:

- ***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. Either data compression or byte caching can potentially facilitate loading the remote object into the cache.

- ***Compression***
  The role of compression is to reduce the size of a file prior to transmitting it 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 and to create 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 and De-Duplication*
  The same basic LZ data compression algorithms discussed above and proprietary de-duplication algorithms can also be applied to individual blocks of data rather than entire files. This approach results in smaller dynamic dictionaries that can reside in memory rather than on disk. As a result, the processing required for compression and de-compression introduces only a relatively small amount of delay, allowing the technique to be applied to real-time, streaming data. Real time de-duplication applied to small chunks of data at high bandwidths requires a significant amount of memory and processing power.

- ***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[41].

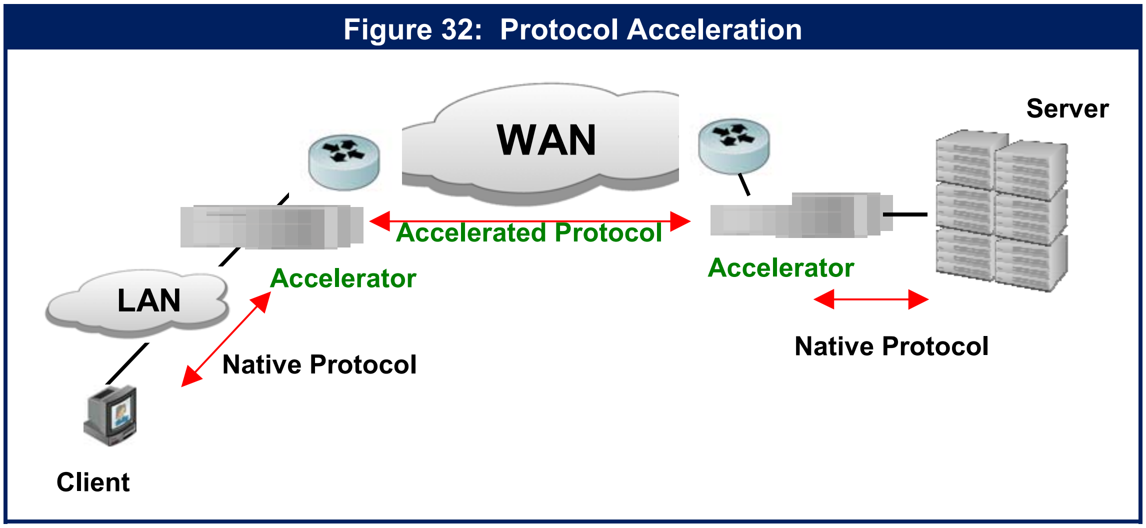
- ***Forward Error Correction (FEC)***
  FEC is typically used at the physical layer (Layer 1) of the OSI stack. FEC can also be applied at the network layer (Layer 3) whereby an extra packet is transmitted for every *n* packets sent. This extra packet is used to recover from an error and hence avoid having to retransmit packets. A subsequent subsection will discuss some of the technical challenges

[41] Transmission_Control_Protocol

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 32**. 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 32: Protocol Acceleration



  - *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 Form Factors and WOC Selection Criteria

The preceding sub-section described the wide range of techniques implemented by WOCs. In many cases, these techniques are evolving quite rapidly. For this reason, almost all WOCs are software based and are offered in a variety of form factors. The range of form factors include:

- ***Standalone Hardware/Software Appliances***
  These are typically server-based hardware platforms that are based on industry standard CPUs with an integrated operating system and WOC software. The performance level they provide depends primarily on the processing power of the server's multi-core architecture. The variation in processing power allows vendors to offer a wide range of performance levels.

- ***Integrated Hardware/Software Appliances***
  This form factor corresponds to a hardware appliance that is integrated within a device such as a LAN switch or WAN router via a card or other form of sub-module.

- ***Virtual Appliances***
  The operating system and WOC software can be optimized to run in a virtual machine on a virtualized server. The performance of the resulting virtual appliance is largely determined by the processing power of the underlying physical server and in part by the functionality provided by the hypervisor. Ideally, the performance of a virtual appliance would be identical to the performance of a standalone appliance assuming the same underlying server hardware. One of the primary advantages of a virtual WOC appliance is the ease of deployment and the centralized provisioning via the hypervisor management system. Another advantage is that in many cases a virtual WOC costs considerably less than a hardware-based WOC. Virtual appliances can also be deployed in support of public and private cloud projects, with the virtual WOCs deployed at the IaaS or SaaS cloud service provider sites.

- ***Client software***
  WOC software can also be provided as client software for a PC, PDA, or Smartphone to provide optimized connectivity for mobile and SOHO workers.

The recommended criteria for evaluating WAN Optimization Controllers are listed in **Table 19**. This list is intended as a fairly complete compilation of all possible criteria, so a given organization may want to 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 19** 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 hypothetical 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 19: 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			
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 contained in **Table 19** is explained below.

- ***Performance***
  Third party tests of an optimization 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.

  A preceding section of the handbook highlighted the fact that the most important optimization task currently facing IT organizations is optimizing a small set of business critical applications. Because of that, IT organizations must test the degree to which a WOC optimizes the performance of those solutions.

- ***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 for 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***
  An IT organization 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 highly effective because it can leverage detailed information about how the application performs. However, this approach is by definition application specific and so it might be negatively impacted by changes made to the application.

- ***Capability to Perform or Support Application Monitoring***
  Some WOCs provide significant application monitoring functionality. That functionality might satisfy the monitoring needs of an IT organization. If it does not, it is important that the WOC doesn't interfere with other tools that an IT organization uses for monitoring. For example, 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. A more important metric is how much throughput the box can actually support with the desired optimization functionality activated. 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 individual laptops and/or wireless devices.

- ***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.
- ***Module vs. Application Optimization***
  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.

- ***Support for Virtualization***
  This criterion includes an evaluation of the support that virtual appliances have for different hypervisors, hypervisor management systems, and VM migration.

- ***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 previously discussed, 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.

- ***Ease 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***
  Support for bulk data transfers between branch offices and central data center is a WOC requirement, but in most cases the volume of bulk traffic per branch is quite low compared to the volume of bulk data traffic over WAN links connecting large data centers. The DMC is the type of product focused on the latter problem.

  There are exceptions to the statement that the volume of bulk transfer per branch is small. For example, in those cases in which there are virtualized servers at the branch office that run applications locally, a key benefit of having virtualized the branch office servers 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. These images have to transit the WAN in and out of the branch office and can constitute very large file transfers. Client-side application virtualization also involves high volume data transfers from the data center to the remote site.

- ***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. One such factor is the ongoing deployment of VoIP. 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. 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. 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.

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 such as VoIP, video or telepresence. Traffic management can be provided by a WOC or alternatively by a router.

To gain insight into the interest that IT organizations have in traffic management and QoS, The Survey Respondents were asked how important it was over the next year for their organization to get better at ensuring acceptable performance for VoIP, traditional video and telepresence. Their responses are shown in **Table 20**.

Table 20: Importance of Optimizing Communications Based Traffic

	VoIP	Traditional Video Traffic	Telepresence
Extremely Important	18.7%	8.6%	4.4%
Very Important	42.3%	23.3%	25.4%
Moderately Important	23.6%	30.2%	27.2%
Slightly Important	8.1%	24.1%	23.7%
Not at all Important	7.3%	13.8%	19.3%

One of the conclusions that can be drawn from the data in Table 6 is:

Optimizing VoIP traffic is one of the most important optimization tasks facing IT organizations.

The section of the handbook that is entitled “Application and Service Delivery Challenges” discussed the importance of managing communications based traffic. In that discussion the observation was made that it is notably more important to IT organizations to get better at managing VoIP than it is for them to get better at managing either traditional video traffic or telepresence. The data in Table 6 indicates that a similar comment applies to the importance of getting better at optimizing VoIP vs. getting better at optimizing traditional video and telepresence.

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 that the solution needs to have detailed Layer 7 knowledge of the application. This follows because, as previously discussed, many applications share the same port or 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.

Hybrid WAN Optimization

The traditional approach to providing Internet access to branch office employees is to carry the Internet traffic on the organization's enterprise network (e.g., their MPLS network) to a central site where the traffic is handed off to the Internet. The primary advantage of this approach is that it enables IT organizations to exert more control over the Internet traffic.

A 2010 market research report entitled *Cloud Networking* reported on the results of a survey in which the survey respondents were asked to indicate how they currently route their Internet traffic and how that is likely to change over the next year. Their responses are contained in **Table 21**.

Table 21: Routing of Internet Traffic

Percentage of Internet Traffic	Currently Routed to a Central Site	Will be Routed to a Central Site within a Year
100%	39.7%	30.6%
76% to 99%	24.1%	25.4%
51% to 75%	8.5%	13.4%
26% to 50%	14.2%	14.2%
1% to 25%	7.1%	6.7%
0%	6.4%	9.7%

One of the disadvantages of a centralized approach to Internet access is performance. For example, an IT organization might use WOCs to optimize the performance of the traffic as it flows from the branch office to the central site. However, once the traffic is handed off to the Internet, the traffic is not optimized.

The vast majority of IT organizations have a centralized approach to Internet access.

The preceding section of the handbook entitled *Optimizing and Securing the Use of the Internet* described the use of a Cloud Networking Service to optimize the performance of applications that transit the Internet. One approach that IT organizations can take to optimize the end-to-end performance of Internet traffic is to implement WOCs to optimize the performance of that traffic as it transits the enterprise WAN. This WOC-based solution is then integrated with a CNS that optimizes the performance of the traffic as it transits the Internet. Since this solution is a combination of a private optimization and a public optimization solution, it will be referred to as hybrid optimization solution.

Data Mobility Controllers (DMCs)

Background

Most large IT organizations are experiencing dramatic increases in the volume of business critical traffic traversing the WAN between data centers. The growth in traffic volume stems not only from data proliferation, but also from an increased emphasis on improving IT support for business agility and business continuity.

Inter-data center traffic is generated by a number of server-to-server utilities and applications, including:

- ***Storage Replication***
  Replicating data on storage arrays and NAS filers is a critical aspect of many disaster recovery strategies because of the need to preserve the critical data that enables business continuity. Disk array and filer volumes are expanding rapidly driven by the combination of business requirements and advances in disk technology. Volume sizes of up to 500 TB –1 PB are becoming increasingly common. Like most inter-data center traffic, storage replication traffic is characterized by a relatively small number of flows or connections, but very high traffic volume per flow.

- ***System Backups***
  Backups are an important component of a disaster recovery strategy because they focus on the continuity of physical and virtual application servers as well as data base servers. The increased complexity of operating systems and application software is causing server image sizes to grow dramatically. Backup solutions that can minimize the backup window and allow for more frequent backups can help make a backup strategy more effective.

- ***Virtual Machine Migration***
  Enabled by the wide spread adoption of virtualization and cloud computing, the migration of VMs between data centers is becoming increasingly common. Live migration of production VMs between physical servers provides tremendous value. For example, it allows for the automated optimization of workloads across resource pools. VM migration also makes it possible to transfer VMs away from physical servers that are undergoing maintenance procedures or that are either experiencing faults or performance issues. During VM migration the machine image, which typically runs 10 Gigabytes or more, the active memory, and the execution state of a virtual machine is transmitted over a high-speed network from one physical server to another. For VM migrations that go between data centers, the virtual machine's disk space may either be asynchronously replicated to the new data center or accessed from the original data center over the WAN. A third approach is to use synchronous replication between the data centers. This approach allows the data to reside at both locations and to be actively accessed by VMs at both sites; a.k.a., active-active storage. In the case of VMotion, VMware recommends that the network connecting physical servers involved in a VMotion transfer have at least 622 Mbps of bandwidth and no more than 5 ms of end-to-end latency42.

[42] VMWare.com

- ***High Performance Computing***
  A significant portion of supercomputing is performed on very large parallel computing clusters resident at R&D labs, universities, and Cloud Computing Service Providers that offer HPC as a service. Transferring HPC jobs to these data centers for execution often involves the transfer of huge files of data over the WAN. In some cases, HPC applications can be executed in parallel across servers distributed across two or more data centers. For these relatively loosely coupled applications (e.g. those based on Hadoop or MapReduce), inter-processor communications may involve large data transfers of interim results.

The Interest in DMCs

In order to quantify the interest that IT organizations have in some of the challenges that were described in the preceding sub-section of the handbook, The Survey Respondents were asked two questions. One question was how important was it to their organization over the next year to get better at optimizing the transfer of storage data between different data centers. The other question was how important was it to their organization over the next year to get better at optimizing the transfer of virtual machines. Their responses are shown in **Figure 33** and **Figure 34**.

Figure 33: The Interest in Optimizing the Transfer of Storage Data

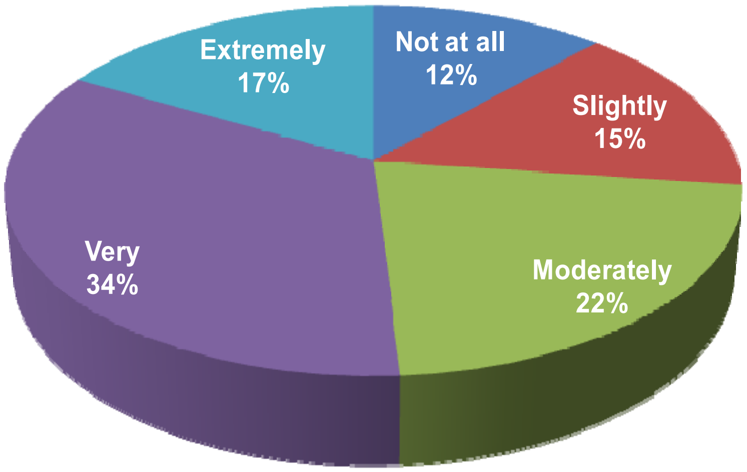
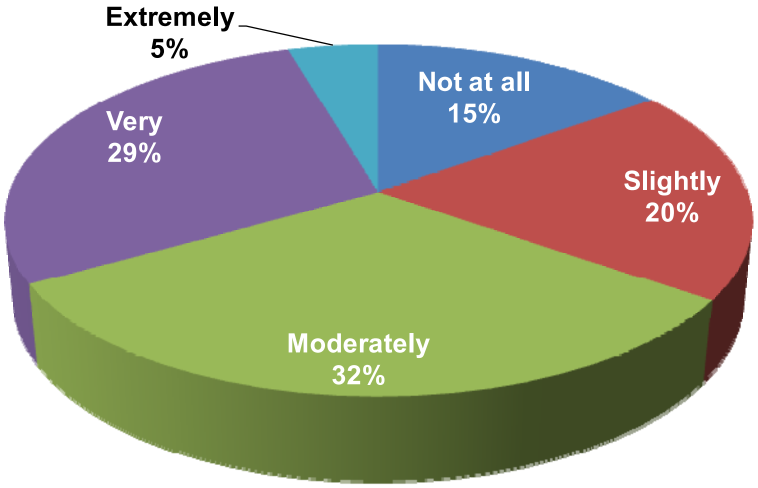


Figure 34: The Interest in Optimizing the Transfer of VMs



For the Majority of IT organizations, getting better at optimizing the transfer of data between data centers is either very or extremely important.

The data in **Figure 34** indicates that getting better at optimizing the transfer of VMs between data centers is of at least moderate importance to the majority of IT organizations. That importance should grow as IT organizations make increased use of virtualized servers and as they make increased use of functionality such as VMotion.

DMC Functionality

One way to support the huge inter-data center traffic flows described above is to connect the data centers with WAN links running at speeds of 10 Gbps or higher. One limitation of this approach is that these WAN links are not always available. Another more fundamental limitation is that these WAN links can be inordinately expensive.

A far more practical way to support the huge inter-data center traffic flows is to implement techniques that reduce the amount of data that gets transferred over the WAN and that guarantee performance for critical traffic. Over the last few years, many IT organizations have implemented WAN optimization controllers (WOCs). Two of the primary functions of a WOC are to reduce the amount of data that gets transferred over the WAN and to guarantee performance for business critical traffic. Unfortunately, the traditional WOC may not be able to effectively optimize the huge inter-data center traffic flows. That follows because while the functionality they provide appears to be what is needed, WOCs were designed to support traffic between branch offices and a data center. This traffic is comprised of tens, if not hundreds, of slow-speed connections. As previously mentioned, inter-data center traffic is comprised of a small number of very high-speed connections.

DMCs focus on a subset of the capabilities listed in **Table 18.** This includes QoS and TCP optimization. The TCP optimization provided by DMCs often includes functionality such as packet level FEC and the ability to recovery from out of order packets. It also includes de-duplication, compression and support for specific backup applications and specialized transfer protocols. The primary challenge for this class of device is to be able to perform high levels of data reduction while filling high bandwidth WAN pipes (e.g., 1 Gbps or more) with non-conflicting data flows.

Some DMC implementations are based on narrowing the functionality of an existing high end WOC hardware appliance to focus on the set of capabilities required for bulk data transfers. In addition, software enhancements may be added to provide support for specific replication and backup applications. When these devices are placed in DMC mode, the TCP buffers and other system resources may be tuned to support replication and backup applications. These hardware appliance devices typically rely on multi-core CPUs to achieve elevated performance levels. When the performance limit of a single appliance is reached, it may be necessary to load balance the inter-data center traffic across a cluster of DMC appliances.

Another approach to DMC design is to use network processors and other programmable logic to provide specialized hardware support for compute-intensive functions, such as transport optimization, de-duplication and compression. Hardware support for some of these functions may be the only cost-effective way to fill very high bandwidth (multi-gigabit) WAN links with highly optimized bulk data traffic.

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

Techniques for Coping with Packet Loss and Out of Order Packets

The section of the handbook that is entitled "Optimizing and Securing the Use of the Internet" discussed in detail the impact of packet loss on TCP throughput. As discussed in that section, small amounts of packet loss can significantly reduce the maximum throughput of a single TCP session. In addition, while packet loss affects throughput for any TCP stream, it particularly affects throughput for high-speed streams, such as those associated with bulk data transfers.

As a result of the well-known impact of packet loss on throughput, numerous techniques have been developed to mitigate the impact of packet loss and out-of-order packets. This includes Performance Enhancing Proxies (PEPs), which are intended to mitigate link-related degradations and Forward Error Correction (FEC), which is intended to offset the impact of dropped packets.

PEPs are often employed in transport optimization solutions that use a proprietary transport protocol to transfer data between the DMCs. The proprietary transport can support very aggressive behaviors in order to expand the window sizes as new connections are formed. This allows the PEP to efficiently fill a link even though that link has both high bandwidth and high delay. With transparent proxies, TCP is used as the transport protocol between the DMCs and the end systems, allowing the end systems to run unmodified. The DMC's PEPs can intercept and terminate the TCP connections from the end systems. They typically use large buffers to isolate the end systems from having an awareness of packet loss and therefore they eliminate the need for retransmissions and adjustments in the TCP window sizes of the end systems.

FEC[43] has long been used at the physical level to ensure error free transmission with a minimum of re-transmissions. 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%. To put this in the context of application performance, 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 preceding example demonstrates the value of FEC in a TCP environment. The technique, however, applies equally well to any application regardless of transport protocol. A negative factor that is associated with FEC is that the use of FEC introduces overhead which itself can reduce throughput. One way to avoid this is to implement a FEC algorithm that dynamically adapts to packet loss. For example, if a WAN link is not experiencing packet loss, no extra packets are transmitted. When loss is detected, the algorithm begins to carry extra packets and it increase the amount of extra packets as the amount of loss increases.

A DMC can also perform re-sequencing of packets at both ends of the WAN link to eliminate the re-transmissions that occur when packets arrive out of order. Packet re-ordering is applicable to all IP data flows regardless of transport protocol.

[43] RFC 2354, Options for Repair of Streaming Media

The criteria for evaluating DMC solutions presented below in **Table 22** is to a large degree a subset of the criteria for evaluating WOC solutions summarized in **Table 18**.

Table 22: Criteria for WAN Optimization Solutions

Criterion	Weight Wi	Score for Solution "A" Ai	Score for Solution "B" Bi
Performance: Throughput and Latency			
Transparency			
Solution Architecture			
Capability to Perform Application Monitoring			
Scalability			
Cost-Effectiveness			
QoS and Traffic Management			
Data Reduction Efficiency			
Security			
Ease of Deployment and Management			
Change Management			
High Availability Features			
Total Score		Σ WiAi	Σ WiBi

Trends in WOC and DMC Evolution

One of the most significant trends in the WAN optimization market is in the development of functionality that support enterprise IT organizations that are implementing either private cloud strategies or strategies to leverage public and hybrid clouds as extensions of their enterprise data centers. Some recent and anticipated developments include:

- ***Cloud Optimized WOCs***
  This is a purpose-built virtual WOC (vWOC) appliance that was designed with the goal of it being deployed in public and/or hybrid cloud environments. One key feature of this class of device is compatibility with cloud virtualization environments including the relevant hypervisor(s). Other key features include SSL encryption and the acceleration and the automated migration or reconfiguration of vWOCs in conjunction with VM provisioning or migration.

- ***Cloud Storage Optimized WOCs***
  This is a purpose-built virtual or physical WOC appliance that was designed with the goal of it being deployed at a cloud computing site that is used for backup and/or archival storage. Cloud optimized features include support for major backup and archiving tools, sophisticated de-duplication to minimize the data transfer bandwidth and the storage capacity that is required, as well as support for SSL and AES encryption.

- ***DMC Enhancements***
  As discussed, DMCs facilitate the transfer of high volume data between data centers owned either by enterprise IT organizations or by CCSPs. DMC products are still in an early stage of evolution and a number of developments can be expected in this space in the near term. This includes enhanced hardware support for various functions including encryption and

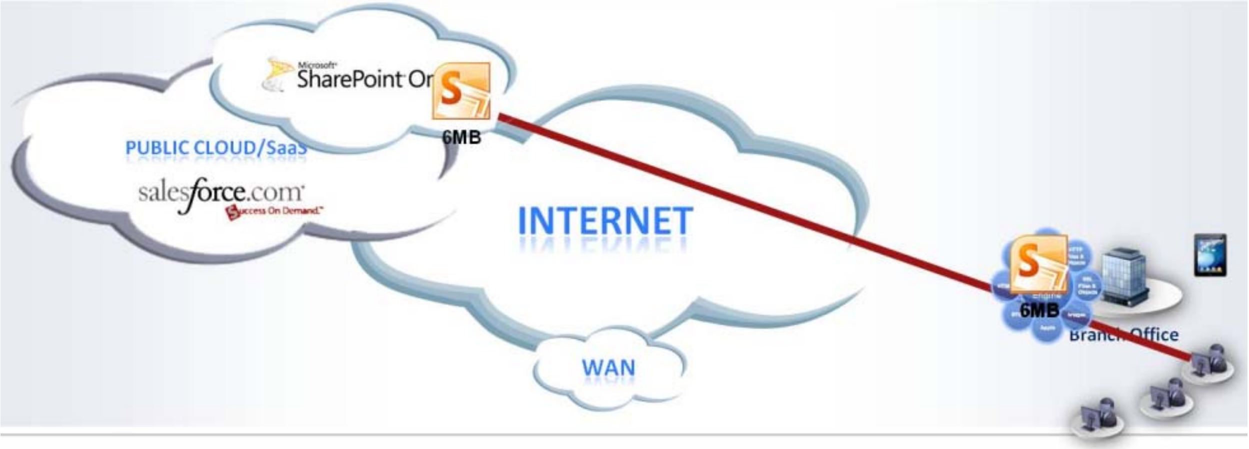
higher speed WAN and LAN interfaces (10 GbE and higher) to support a combination of highly efficient data reduction and high bandwidth WAN services.

- ***Cloud Networking Services***
  Not all CCSPs will either provide WOC functionality themselves nor support vWOC instances being hosted at their data centers. A previous section of the handbook described the growing use of Cloud Networking Services (CNSs). Using a CNS that provides optimization can accelerate the performance of services acquired from these CCSPs.

Asymmetric WOCs

Another technique that IT organizations can utilize in those instances in which the CCSP doesn't provide WOC functionality themselves nor do they support vWOC instances being hosted at their data centers is to implement WOCs in an asymmetric fashion. As shown in **Figure 35**, content is downloaded to a WOC in a branch office. Once the content is stored in the WOC's cache for a single user, subsequent users who want to access the same content will experience accelerated application delivery. Caching can be optimized for a range of cloud content, including Web applications, streaming video (e.g., delivered via Flash/RTMP or RTSP) and dynamic Web 2.0 content.

Figure 35: Asymmetric WOC Deployment



IPV6 Application Acceleration

Now that the industry has depleted the IPv4 address space, there will be a gradual transition towards IPv6 and mixed IPV4/ IPV6 environments. As applications transition to IPV6 from IPV4, application level optimizations such as those for CIFS, NFS, MAPI, HTTP, and SSL will need to be modified to work in the mixed IPV4/ IPV6 environment.

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 to balance the load over a server farm.

While an ADC still functions as a SLB, the ADC has assumed, and will most likely continue to assume, a wider range of more sophisticated roles that enhance server efficiency and provide asymmetrical functionality to accelerate the delivery of applications from the data center to individual remote users. 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 and hence can reduce the number of servers that are required for a given level of business activity.

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[44]. 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.

Like the WOC, the ADC is available in a number of form factors including virtual appliances, hardware appliances, and as line card modules for switches and routers. Hardware implementations can be based on multi-core CPUs (possibly with specialized co-processors for compute–intensive operations such as encryption/decryption) or with network processors.

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

- ***<u>Traditional SLB</u>***
  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.

- ***<u>SSL Offload</u>***
  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

[44] This example ignores the impact of server virtualization.

significant increase in the performance of secure intranet or Internet Web sites. SSL offload frees up server resources which allows 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 the 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.

- ***Denial of Service (DOS) Attack Prevention***
  ADCs can provide an additional line of defense against DOS attacks, isolating servers from a range of Layer 3 and Layer 4 attacks that are aimed at disrupting data center operations.

- ***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 the 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 for the additional capacity that is required in order to maintain service levels.

- ***Support for Server Virtualization***
  Once a server has been virtualized, 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 external agent could be a start-up script inside of the VM or it could be the provisioning or management agent that initiated the provisioning of the VM. The provisioning or 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. It is preferable if the process of configuring the network elements, including the ADCs, to support new VMs and the movement of VMs within a data center can readily be automated and integrated within the enterprise's overall architecture for managing the virtualized server environment.

When a server administrator adds a new VM to a load balanced cluster, the integration between the hypervisor management system and the ADC manager can modify the configuration of the ADC to accommodate the additional node and its characteristics. When a VM is de-commissioned a similar process is followed with the ADC manager taking steps to ensure that no new connections are made to the outgoing VM and that all existing sessions have been completed before the outgoing VM is shut down.

For a typical live VM migration, the VM remains within the same subnet/VLAN and keeps its IP address. As previously described, a live migration can be performed between data centers as long as the VM's VLAN has been extended to include both the source and destination physical servers and other requirements regarding bandwidth and latency are met.

In the case of live migration, the ADC does not need to be reconfigured and the hypervisor manager ensures that sessions are not lost during the migration. Where a VM is moved to a new subnet, the result is not a live migration, but a static one involving the creation of a new VM and decommissioning the old VM. First, a replica of the VM being moved is created on the destination server and is given a new IP address in the destination subnet. This address is added to the ADC's server pool, and the old VM is shut down using the process described in the previous paragraph to ensure session continuity.

ADC Selection Criteria

The ADC evaluation criteria are listed in **Table 23**. 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 23: 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 also pay attention to the ability of the ADC to have all features turned on and still support the peak traffic load.

- ***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. As noted, 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 be able to be extended 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.

  As functional integration continues to evolve, the traditional ADC can begin to assume a broader service delivery role in enterprise data center by incorporating additional functions, such as global server load balancing (GSLB), inter-data center WAN optimization, multi-site identity/access management and enhanced application visibility functions.

- ***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. With proper integration, vADCs can be managed along with VMs by the hypervisor management console. It is also important to know how the ADC supports the creation and movement of VMs within a dynamic production environment. One option is to pre-provision VMs as members of ADC server pools. For dynamic VM provisioning data center orchestration functionality, based on plug-ins or APIs can automatically add new VMs to resource pools.

  The preceding section of the handbook entitled "Virtualization" described one way of virtualizing an ADC. That was as a virtual appliance in which the ADC software runs in a VM. Partitioning a single physical ADC into a number of logical ADCs or ADC contexts is another way to virtualize an ADC. 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. A third way that an ADC can be virtualized is that two or more ADCs can be made to appear to be one larger ADC.

- ***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 in those instances in which 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.

Trends in ADC Evolution

As noted earlier, one trend in ADC evolution is increasing functional integration with more data center service delivery functions being supported on a single platform. As organizations continue to embrace cloud computing models, service levels need to be assured irrespective of where applications run in a private cloud, hybrid cloud or public cloud environment. As is the case with WOCs, ADC vendors are in the process of adding enhancements that support the various forms of cloud computing. This includes:

- ***Hypervisor–based Multi-tenant ADC Appliances***
  Partitioned ADC hardware appliances have for some time allowed service providers to support a multi-tenant server infrastructure by dedicating a single partition to each tenant. Enhanced tenant isolation in cloud environments can be achieved by adding hypervisor functionality to the ADC appliance and dedicating an ADC instance to each tenant. Each ADC instance then is afforded the same type of isolation as virtualized server instances, with protected system resources and address space. ADC instances differ from vADCs installed on general-purpose servers because they have access to optimized offload resources of the appliance. A combination of hardware appliances, virtualized hardware appliances and virtual appliances provides the flexibility for the cloud service provider to offer highly customized ADC services that are a seamless extension of an enterprise customer's application delivery architecture. Customized ADC services have revenue generating potential because they add significant value to the generic load balancing services prevalent in the first generation of cloud services. If the provider supplies only generic load balancing services the vADC can be installed on a service provider's virtual instance, assuming hypervisor compatibility.

- ***Cloud Bursting and Cloud Balancing ADCs***
  Cloud bursting refers to directing user requests to an external cloud when the enterprise private cloud is at or near capacity. Cloud balancing refers to routing user requests to applications instances deployed in the various different clouds within a hybrid cloud. Cloud balancing requires a context-aware load balancing decision based on a wide range of business metrics and technical metrics characterizing the state of the extended infrastructure. By comparison, cloud bursting can involves a smaller set of variables and may be configured with a pre-determined routing decision. Cloud bursting may require rapid activation of instances at the remote cloud site or possibly the transfer of instances among cloud sites. Cloud bursting and balancing can work well where there is consistent application delivery architecture that spans all of the clouds in question. This basically means that the enterprise application delivery solution is replicated in the public cloud. One way to achieve this is with virtual appliance implementations of GSLBs and ADCs that support the range of variables needed for cloud balancing or bursting. If these virtual appliances support the cloud provider's hypervisors, they can be deployed as VMs at each

cloud site. The inherent architectural consistency insures that each cloud site will be able to provide the information needed to make global cloud balancing routing decisions. When architectural consistency extends to the hypervisors across the cloud, the integration of cloud balancing and/or bursting ADCs with the hypervisors' management systems can enable the routing of application traffic to be synchronized with the availability and performance of private and public cloud resource. Access control systems integrated within the GSLB and ADC make it possible to maintain control of applications wherever they reside in the hybrid cloud.

Conclusions

Throughout the ***2011 Application and Service Delivery Handbook***, a number of conclusions were reached. Those conclusions are that:

- In the vast majority of instances, end users notice application degradation before the IT organization does.

- Having the IT organization notice application degradation before the end user does is important to the vast majority of senior managers.

- The vast majority of IT organizations don't have any insight into the performance of an application until after the application is fully developed and deployed.

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

- Responding to the first generation of application delivery challenges is still important to the majority of IT organizations.

- Over the next year, the most important optimization task facing IT organizations is optimizing a key set of business critical applications.

- 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.

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

- One of the effects of data center consolidation 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.

- 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.

- At the same time that most IT organizations are still responding to a traditional set of application and service delivery challenges, they are beginning to face a formidable set of new challenges.

- Improving the performance of applications used by mobile workers is less important than managing VoIP traffic, but more important than managing either traditional video or telepresence traffic.

- Web-based applications present a growing number of management, security and performance challenges.

- Getting better at managing a business service that is supported by multiple, inter-related applications is one of the most important tasks facing IT organizations over the next year.

- A virtualized data center can be thought of as a fractal data center.

- Half of the IT organizations consider it to be either very or extremely important over the next year for them to get better performing management tasks such as troubleshooting on a per-VM basis.

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

- Supporting the movement of VMs between servers in different data centers is an important issue today and will become more so in the near term.

- The deployment of virtualized desktops trails the deployment of virtualized data center servers by a significant amount.
- Over the next year, there will be a modest increase in the deployment of virtualized desktops.

- By the end of the year, the vast majority of virtualized desktops will be utilizing server side virtualization.

- 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.

- Packet loss can have a very negative impact on the performance of desktop virtualization solutions

- IT organizations that are implementing virtualized desktops should analyze the viability of implementing network and application optimization solutions.

- 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.

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

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

- The SLAs that are associated with public cloud computing services such as Salesforce.com or Amazon’s Simple Storage System are generally weak both in terms of the goals that they set and the remedies they provide when those goals are not met.

- Many of the approaches to providing public cloud-based solutions will not be acceptable for the applications, nor for the infrastructure that supports the applications, for which enterprise IT organizations need to provide an SLA.

- The SaaS marketplace is comprised of a small number of large players such as Salesforce.com, WebEx and Google Docs as well as thousands of smaller players.

- There are significant differences amongst the solutions offered by IaaS providers, especially when it comes to the SLAs they offer.

- The availability of IaaS solutions can vary widely.

- The primary factors that are driving the use of public cloud computing solutions are the same factors that drive any form of out-tasking.

- In some cases, the use of a public cloud computing solution reduces risk.

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

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

- Cloud balancing can be thought of as the logical extension of global server load balancing (GSLB).

- Over the next year IT organizations intend to make a significant deployment of Cloud Networking Services.

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

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

- Hope is not a strategy. Successful application and service delivery requires careful planning.

- The goal of APE is to help IT organizations reduce risk and build better relationships with the company’s business unit managers.

- The goal of route analytics is to provide visibility, analysis and diagnosis of the issues that occur at the routing layer in complex, meshed networks.

- 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.

- Only a small minority of IT organizations has a top down, tightly coordinated approach to APM.

- Over the next year, getting better at monitoring the end user's experience and behavior is either very or extremely important to the majority of IT organizations.

- Getting better at identifying the components of the IT infrastructure that support the company's critical business applications and services is one of the most important management tasks facing IT organizations.

- Getting better at rapidly identifying the causes of application degradation is the most important management task facing IT organizations.

- Lack of visibility into the traffic that transits port 80 is a significant management and security challenge for most IT organizations.

- Application Performance Engineering (APE) is the practice of first designing for acceptable application performance and then testing, measuring and tuning performance throughout the application lifecycle.

- 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 until they receive a response back from the application.

- The operational challenges that are created due to a lack of insight into the router layer are greatly exacerbated by the adoption of server virtualization and cloud computing.

- In the vast majority of cases, logical factors cause as much or more business disruption than do physical factors.

- In the vast majority of instances, logical errors take either somewhat more or notably more time to troubleshoot and repair than do physical errors.

- The value proposition of network and application optimization is partly to improve the performance of applications and services and partly to save money.

- The most common way that IT organizations approach implementing optimization functionality is on a case-by-case basis.

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

- Roughly half of IT organizations have not made any deployment of WOCs.

- Over the next year, IT organizations plan to make a moderate increase in their deployment of WOCs.

- Optimizing VoIP traffic is one of the most important optimization tasks facing IT organizations.

- The vast majority of IT organizations have a centralized approach to Internet access.
- For the Majority of IT organizations, getting better at optimizing the transfer of data between data centers is either very or extremely important.
- 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.

About the Webtorials® Editorial/Analyst Division

The Webtorials® Editorial/Analyst Division, a joint venture of industry veterans Steven Taylor and Jim Metzler, is devoted to performing in-depth analysis and research in focused areas such as Metro Ethernet and MPLS, as well as in areas that cross the traditional functional boundaries of IT, such as Unified Communications and Application Delivery. The Editorial/Analyst Division's focus is on providing actionable insight through custom research with a forward looking viewpoint. Through reports that examine industry dynamics from both a demand and a supply perspective, the firm educates the marketplace both on emerging trends and the role that IT products, services and processes play in responding to those trends.

Jim Metzler has a broad background in the IT industry. This includes being a software engineer, an engineering manager for high-speed data services for a major network service provider, a product manager for network hardware, a network manager at two Fortune 500 companies, and the principal of a consulting organization. In addition, he has created software tools for designing customer networks for a major network service provider and directed and performed market research at a major industry analyst firm. Jim's current interests include cloud networking and application delivery.

For more information and for additional Webtorials® Editorial/Analyst Division products, please contact Jim Metzler at jim@webtorials.com or Steven Taylor at taylor@webtorials.com.

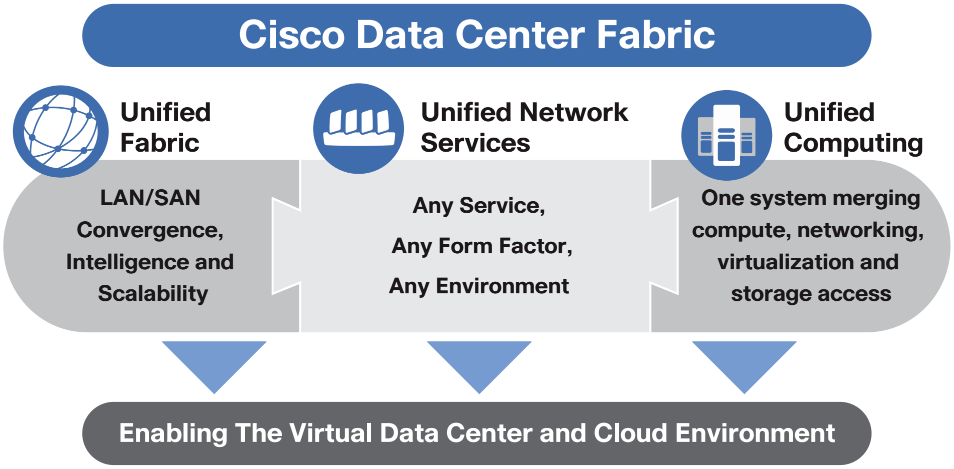
Published by Webtorials Editorial/Analyst Division
www.Webtorials.com

Division Cofounders:
Jim Metzler
jim@webtorials.com
Steven Taylor
taylor@webtorials.com

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

For editorial and sponsorship information, contact Jim Metzler or Steven Taylor. The Webtorials Editorial/Analyst Division is an analyst and consulting joint venture of Steven Taylor and Jim Metzler.

Cisco Unified Network Services


Cisco Data Center Fabric
Unified Fabric
Unified Network Services
Unified Computing
LAN/SAN Convergence, Intelligence and Scalability
Any Service, Any Form Factor, Any Environment
One system merging compute, networking, virtualization and storage access
Enabling The Virtual Data Center and Cloud Environment

Highly virtualized data center and cloud environments impose enormous complexity on the deployment and management of network services. Provisioning dynamic services and accommodating mobile workloads present challenges for layered services, such as security and application controllers, that traditionally have required in-line deployment and static network topologies. Cisco® Unified Network Services meets these challenges with integrated application delivery and security solutions for highly scalable, virtualized data center and cloud environments.

Any Service: Cisco Unified Network Services is a critical component of the Cisco Data Center Business Advantage architecture. It consists of Cisco Application Control Engine (ACE) application controllers, Cisco Wide Area Application Services (WAAS) WAN acceleration products, Cisco Adaptive Security Appliances (ASA) data center security solutions, Cisco Virtual Security Gateway (VSG), Cisco Network Analysis Module (NAM), and associated management and orchestration solutions.

Any Form Factor: Cisco Unified Network Services provides consistency across physical and virtual services for greater scalability and flexibility. One element of the Cisco Unified Network Services approach is the concept of a virtual service node (VSN), a virtual form factor of a network service running in a virtual machine. Cisco VSG for Cisco Nexus® 1000V Series Switches and Cisco Virtual WAAS (vWAAS) are examples of VSNs that enable service policy creation and management for individual virtual machines and individual applications.

Outstanding Scalability: In addition to virtualization-aware policies and services, Cisco Unified Network Services supports greater data center scalability and cloud deployments, with the services themselves being virtualized. The application and security services can be provisioned and scaled on demand and can be easily configured to support the needs of dynamically deployed and scalable virtual applications.

Integrated Management Model: Cisco Unified Network Services enables consistency of management across different services and across physical and virtual form factors. Cisco Unified Network Services is thus a critical component of a fabric-centered data center architecture that is well integrated with the virtual servers and applications to readily enable scalable public and private cloud environments.

Application Delivery Controllers

Enhanced web application performance, availability, and server scalability

WAN Optimization

Reduce branch IT costs and enhanced application performance for the distributed enterprise

Network Analysis and Monitoring

Simplifies application performance monitoring

Data Center Security

Physical and virtual solutions remove multi-tenant security risks and external threats

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices**.