The 2012 Application & Service Delivery Handbook

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

Platinum Sponsors:

cisco

ipanema Technologies

NETSCOUT

Gold Sponsors:

A10 Networks

Akamai

Aryaka

Blue Coat

agility made possible™
ca technologies

certeon Accelerate & Broaden Application Access

radware

riverbed

VCE

VISUAL NETWORK SYSTEMS

Produced by:

Webtorials

2012 Application and Service Delivery Handbook

Executive Summary

Throughout the ***2012 Application and Service Delivery Handbook***, (The 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

There is a growing relationship between the requirements listed above. For example, in order to implement an appropriate level of security, an IT organization may implement encryption. However, the fact that the information flow is encrypted may preclude the IT organization from implementing the optimization techniques that are required to ensure acceptable performance. In addition, IT organizations don't want to optimize the performance of malware and spyware. IT organizations must identify this traffic and eliminate it.

IT organizations need to plan for optimization, security and management in an integrated fashion.

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 such as supporting chatty protocols or transmitting large files between a branch office and a data center, a second 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

Webtorials published the first edition of what became an annual series of application delivery handbooks in January 2007. Until last year, the primary goal of the handbook was to help IT organizations ensure acceptable application delivery when faced with the first generation of application delivery challenges. Beginning last year, the goal of the handbook changed in response to both the growing emphasis on services as well as the increasing impact of the second generation of application delivery challenges.

The goal of the 2012 Application and Service Delivery Handbook is to help IT organizations ensure acceptable application and/or service delivery when faced with both the first generation, as well as the emerging second generation of application and service delivery challenges.

To help to achieve this goal, in early 2012 three surveys were given to the subscribers of Webtorials. Throughout this document, the IT professionals who responded to the surveys will be referred to as ***The Survey Respondents***.

The Survey Respondents were given a set of outcomes that could result from poor application performance. They were asked to indicate the type of impact that typically occurs if one or more of their company's business critical applications are performing badly, and they were allowed to indicate multiple impacts. The impacts that were mentioned the most often are shown in **Table 1.**

Table 1: Impact of Poor Application Performance

Impact	Percentage
The Company Loses Revenue	62.0%
IT Teams Are Pulled Together	59.8%
Company Loses Customers	45.1%
CIO Gets Pressure from his/her Boss	45.1%
Harder for IT to get Funding	44.6%
CIO Gets Other Pressure	42.9%

If a business critical application is performing poorly, it has a very significant business impact and it also has a very significant impact on the IT organization.

First Generation Application and Service Delivery Challenges

There are a number of fairly well understood challenges that have over the years complicated the task of ensuring acceptable application and service delivery. Those challenges are listed below and are described in detail in The Handbook.

- Limited Focus on Application Development
- Network Latency
- Availability
- Bandwidth Constraints
- Packet Loss
- Characteristics of TCP
- Chatty Protocols and Applications
- Myriad Application Types
- Webification of Applications
- Expanding Scope of Business Critical Applications
- Server Consolidation
- Data Center Consolidation
- Server Overload
- Distributed Employees
- Distributed Applications
- Complexity
- Increased Regulations
- Security Vulnerabilities

Second Generation Application and Service Delivery Challenges

There are a number of emerging challenges that are beginning to complicate the task of ensuring acceptable application and service delivery. Those challenges are listed below and are described in detail in The Handbook.

- **Mobility and BYOD**
  The majority of organizations have adopted BYOD. Unfortunately, the BYOD movement has resulted in a loss of control and policy enforcement.

- **The Mandate for Agility**
  Because of their awareness of the technology that is available to them in their homes and from the Internet, a growing number of business and functional managers have increased expectations of the IT organization. As a result, IT organizations are under more pressure for agility than they ever have been in the past.

- **IT Organizations as Service Brokers**
  IT organization need to modify their traditional role of being the primary provider of IT services and adopt a role in which they provide some IT services themselves and act as a broker between the company's business unit managers and cloud computing service providers for other services.

- **The Increasing Number of Business Critical Applications**
  Over a quarter of The Survey Respondents indicated that their company has over 20 business critical applications.

- **New Application Architectures: SOA, Web 2.0 and Rich Internet Applications**
  These new application architectures tend to be more susceptible to performance problems due to WAN impairments than do traditional application architectures. In addition, the introduction of technologies such as AJAX creates significant security vulnerabilities.

- **Internal SLAs**
  Roughly half of IT organizations provide internal SLAs and that percentage is expected to grow. According to The Survey Respondents, getting better at managing internal SLAs over the next year is one of their most important management tasks.

- **Server Virtualization**
  One of the challenges associated with server virtualization comes from the fact that in most cases, data centers with virtualized servers will have different hypervisors that each has their own management capabilities. Another challenge is the need to integrate the management of virtual servers into the existing workflow and management processes. In addition, half of The Survey Respondents indicated that they consider it to be either very or extremely important over the next year for their organization to get better at performing management tasks such as troubleshooting on a per-VM basis.

- **Desktop Virtualization**
  Based on the responses of The Survey Respondents, over the next year the number of IT organizations who have implemented at least some desktop virtualization will increase dramatically. 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.

- **Private Cloud Computing**
  Some of the primary challenges associated with private cloud computing are the same challenges that are associated with server virtualization. Another challenge that is associated with private cloud computing is supporting the dynamic movement of virtual machines between physical servers, both within a data center and between disparate data centers.

- **Public and Hybrid Cloud Computing**
  Managing server virtualization is also a challenge for providers of both public and hybrid cloud computing. However, the adoption of those forms cloud computing creates a new set of management challenges for enterprise IT organizations. Some of these new challenges stem from the fact that IT organizations are typically held responsible for the performance of these public and hybrid cloud solutions even though in most cases they don't have the same access to the enabling IT infrastructure that they would have if the application was entirely intra-company. Other new management challenges stem from the sheer complexity of the public and cloud environments. What this complexity means is that in order to manage end-to-end in either a public cloud or a hybrid cloud environment, management data must be gathered from the enterprise, one or more Network Service Providers (NSPs) and one or more cloud computing service providers.

Network and Application Optimization

As shown in Figure 1, 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 1: Application Response Time Model

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

The WOCs, Cloud-based optimization services, Internet optimization services and ADCs that are described in this section of the handbook are intended to mitigate the impact of those factors.

WAN Optimization Controllers (WOCs)

Table 2 lists some of WAN characteristics that impact application delivery and identifies WAN optimization techniques that a WOC can implement to mitigate the impact of those characteristics. These techniques are described in detail in The Handbook and The Handbook also provides a suggested approach for evaluating WOCs.

Table 2: Techniques to Improve Application Performance

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

WOCs come in a variety of form factors including:

- **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.

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

- **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 WOCs**
  The phrase virtual WOC refers to optimizing the operating system and the WOC software to run in a VM on a virtualized server.

Performing tasks such as moving VMs or doing storage replication between data centers greatly increase the demand for inter-data center bandwidth. While it is possible to just continually add more WAN bandwidth, a more practical solution is to focus on increasing the *Effective Bandwidth* of WAN links. Effective bandwidth is determined by two factors. One factor is the *Bandwidth Efficiency*, which is how completely the WAN link bandwidth can be utilized, even when faced with high WAN latency and a relatively small number of high volume flows. The second factor is the *Bandwidth Multiplication Factor,* which is the gain in link throughput that is derived from implementing techniques such as data compression and de-duplication. The formula for Effective Bandwidth is given by:

Effective BW = BW Efficiency x BW Multiplication Factor x Physical BW

In order to optimize workload migration an inter-data center WAN Optimization solution should have the following functionality:

- **High Throughput**
  The inter-data center WAN Optimization solution should be capable of saturating a multi-gigabit WAN link and hence provide a bandwidth efficiency of 1.0, even if the number of current flows between data centers is quite small.

- **Transport Optimization**
  The congestion control mechanism for TCP needs to be very aggressive in its control of window sizes in order to achieve high bandwidth efficiency and consume all of the bandwidth allocated to each type of traffic flow.

- **Low Latency**
  For synchronous storage replication any significant amount of WOC device latency reduces the inter-data center distance over which synchronous replication is feasible. WAN Optimization device internal latency can also be a significant factor affecting the inter-data center distances over which VM migration can be reliably performed.

- **Maximal Data Reduction**
  Storage replication and backup applications typically send only those blocks of data that have changed since the previous transfer. In these cases, good WOC de-duplication ratios depend on identifying patterns that are far smaller than the typical data block addressed by disk systems that are typically 4 KB.

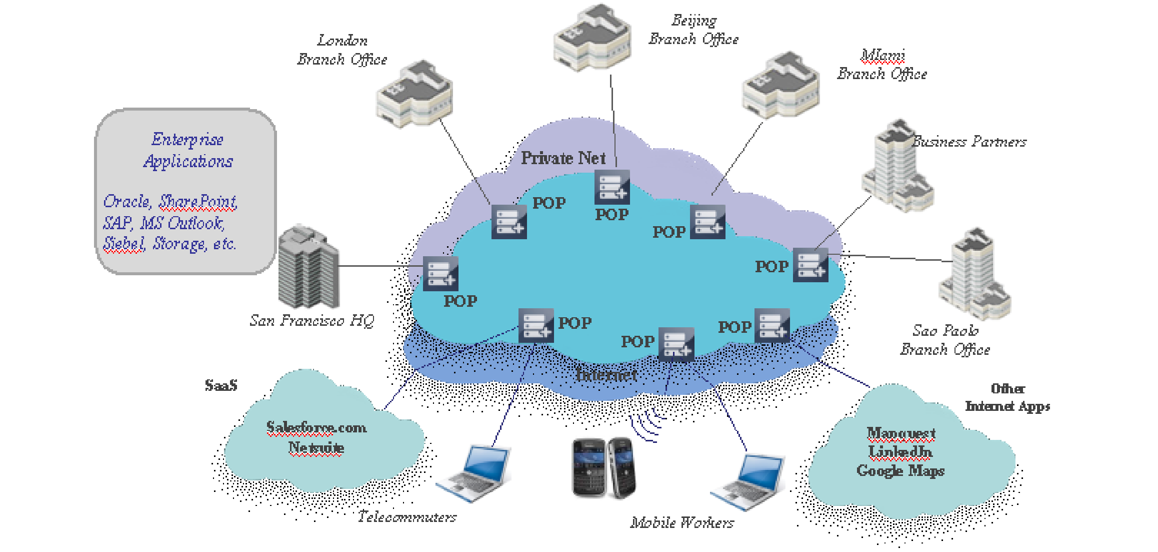
- **QoS and Traffic Management**
  The WAN Optimization system must have a hardware-based QoS/traffic management system that can classify and prioritize traffic at multi-gigabit line rates and allocate bandwidth in accordance with configured QoS policies.

- **High Availability**
  In addition to providing a number of internal high availability features, such as redundant power supplies, these solutions should support high availability network designs based on in-line or out-of-path redundant configurations.

Cloud-Based Optimization Solutions

As shown in **Figure 2**, it is now possible to acquire a number of IT-centric functions, such as network and application optimization from a cloud service provider.

Figure 2: Network and Application Optimization Solution



As shown in **Figure 2**, a variety of types of users (e.g., mobile users, branch office users) access WAN optimization functionality at the service provider's points of presence (POPs). Ideally these POPs are inter-connected by a dedicated, secure and highly available network. To be effective, the solution must have enough POPs so that there is a POP in close proximity to the users. In addition, the solution should support a wide variety of WAN access services.

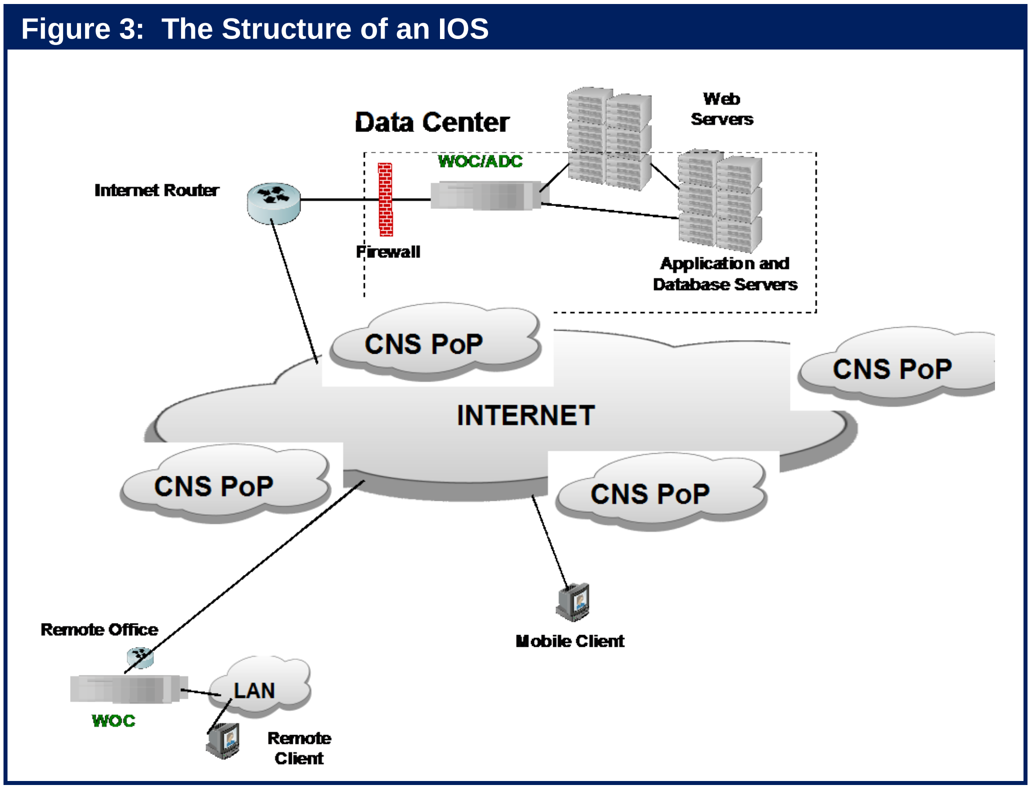
There are at least three distinct use cases for the type of solution shown in **Figure 2.** One such use case is that this type of solution can be leveraged to solve the type of optimization challenges that an IT organization would normally solve by deploying WOCs. The second use case is the ongoing requirement that IT organizations have to support mobile workers. The third use case for utilizing a solution such as the one shown in **Figure 2** is the expanding requirement that IT organizations have to support access to public cloud services.

The Optimization of Internet Traffic

WOCs 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 such as 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. Throughout The Handbook, a service that optimizes Internet traffic will be referred to as an Internet Optimization Service (IOS).

An IOS 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 3**, all client requests to the application's origin server in the data center are redirected via DNS to a server in a nearby point of presence (PoP) that is part of the IOS. This edge server then optimizes the traffic flow to the IOS server closest to the data center's origin server.

Figure 3: The Structure of an IOS



The servers at the IOS provider's PoPs perform a variety of optimization functions. Some of the functions provided by the IOS include:

- **Route Optimization**
  A route optimization solution leverages the intelligence of the IOS 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.

- **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.

- **HTTP Protocol Optimization**
  HTTP inefficiencies can be eliminated by techniques such as compression and caching at the edge IOS server with the cache performing intelligent pre-fetching from the origin.

- **Content Offload**
  Static content can be offloaded out of the data-center to caches in IOS servers and through persistent, replicated in-cloud storage facilities.

Hybrid WAN Optimization

The traditional approach to providing Internet access to branch office employees has been to backhaul that Internet traffic on the organization's enterprise network (e.g., their MPLS network) to a central site where the traffic was handed off to the Internet. The advantage of this approach is that it enables IT organizations to exert more control over their Internet traffic and it simplifies management in part because it centralizes the complexity of implementing and managing security policy. One disadvantage of this approach is that it results in extra traffic transiting the enterprise's WAN, which adds to the cost of the WAN. Another disadvantage of this approach is that it usually adds additional delay to the Internet traffic.

One way to minimize the degradation in application performance is based on the use of an IOS. One way that an IOS would add value is if the organization used the IOS to carry traffic directly from the branch office to the SaaS provider. In this case, in addition to providing optimization functionality, the IT organization is relying on the security functionality provided by the IOS to compensate for the security functionality that was previously provided in the corporate data center. Another way that an IOS would add value is if the solution enabled IT organizations to keep its current approach to backhauling traffic. However, in this case, the IT organization would use WOCs to optimize the performance of the Internet traffic as it transits the enterprise WAN. This WOC-based solution would then have to be integrated with the IOS 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 WAN optimization solution.

Application Delivery Controllers

Among the functions users can expect from an ADC are the following:

- **Traditional SLB**
  ADCs can provide traditional load balancing across local servers or among geographically dispersed data centers based on Layer 4 through Layer 7 intelligence.

- **SSL Offload**
  One of the primary new roles played by an ADC is to offload CPU-intensive tasks from data center servers. A prime example of this is SSL offload, where the ADC terminates the SSL session by assuming the role of an SSL Proxy for the servers.

- **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.

- **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.

- **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 Handbook describes the techniques used within ADCs and also provides a suggested approach for evaluating ADCs.

IPv6 has the potential to affect almost any component of IT that is used for application and service delivery. The most obvious change occurs on networking devices including routers, LAN switches, firewalls and Application Delivery Controllers. With that in mind, The Handbook contains a detailed description of the varying options available to IT organizations relative to IPv6 migration.

As described in The Handbook, there are multiple ways to implement a virtualized ADC including:

- **General Purpose VM Support**
  A specialized network O/S along with ADC software that have been modified to run efficiently in a general purpose virtualization environment including VMWare's vSphere, Citrix's XenServer and Microsoft's Hyper-V.

- **Network Appliance O/S Partitioning**
  This involves the implementation of a lightweight hypervisor in a specialized network O/S by partitioning critical memory and I/O ports for each ADC instance, while also maintaining some memory and I/O ports in common.

- **Network Appliance with OEM Hypervisor**
  A general-purpose virtualization solution is adapted to run on a network appliance and provides the ability to run multiple ADCs on a single device. Since the hypervisor is based on an OEM product, other applications can be run on the device as it can participate in an enterprise virtualization framework such as VMWare's vCenter, Citrix's Xencenter or Microsoft's System Center.

- **Network Appliance with Custom Hypervisor**
  General-purpose hypervisors are designed for application servers and not optimized for network service applications. To overcome these limitations, custom hypervisors optimized for network O/Ss have been added to network appliances.

One area of innovation relative to ADCs is the implementation of Web content optimization (WCO). WCO refers to efficiently optimizing and streamlining Web page delivery. WCO is available in a number of form factors, including being part of an ADC.

Some of the techniques that are used in a WCO solution include:

- Image spriting: A number of images are merged onto a single image reducing the number of image requests.

- JPEG resampling: An image is replaced with a more compact version of the image by reducing the resolution to suit the browser.

- HTTP compression: Compress HTTP, CSS and JavaScript files.

- URL versioning: Automatically refresh the browser cache when the content changes.

Planning, Management and Security

Many planning functions are critical to the success of application delivery. One planning function that is discussed in length in The Handbook is identifying the company's key applications and services and establishing SLAs for them. Another key planning activity that is discussed in detail in The Handbook is Application Performance Engineering (APE). The primary 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 that is discussed in The Handbook 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.

The Handbook discussed the importance of integrating network planning and network operations and provides some suggestions for how that can be accomplished. The Handbook also provides an outline of a plan that IT organizations can use to plan for the ongoing deployment of cloud computing.

The Handbook identifies the weaknesses of traditional network management, traditional application performance management and synthetic transactions. Also identified are a set of challenges (e.g., server virtualization, cloud computing, delay sensitive traffic, converged infrastructure, mobile device management) that will further complicate the task of being able to ensure acceptable application and service delivery. With that as a background, The Handbook provides a detailed outline of an approach that IT organizations can use to get better at managing application and service delivery. One of the key components of that approach is a single unified view of all of the components that support a service. This includes the highly visible service components such as servers, storage, switches and routers. It also includes the somewhat less visible network services such as DNS and DHCP, which are significant contributors to application degradation. Since on an increasing basis going forward one or more network service providers (NSPs) and one or more cloud computing service providers (CCSPs) will provide some or all of these service components, management data must be gathered from the enterprise, one or more NSPs and one or more CCSPs. In addition, in order to help relate the IT function with the business functions, IT organizations need to be able to understand the key performance indicators (KPIs) for critical business processes such as supply chain management and relate these business level KPIs to the performance of the IT services that support the business processes.

As discussed in The Handbook, IT organizations must also be able to provide a common and consistent view of both the network and the applications that ride on the network to get to a service-oriented perspective. The level of granularity provided needs to vary based on the requirements of the person viewing the performance of the service or the network. For example, a business unit manager typically wants a view of a service than is different than the view wanted by the director of operations, and that view is often different than the view wanted by a network engineer. The Handbook also makes detailed suggestions for how IT organizations should evaluate the management capabilities of cloud service providers.

The Handbook contains a detailed discussion of the technologies and governance models that IT organizations are currently using to respond to the traditional threats as well as the threats brought on by BYOD. The Handbook contains a description of Cloud-based security solutions including the identification of the value proposition of these solutions and the identification of some of the most important use cases.   Also, similar to the discussion of optimizing the performance of the Internet, there is a discussion of using an IOS to provide security functionality such as Web application firewalls. The Handbook also contains a set of security focused criteria that IT organizations can use to evaluate the services offered by cloud service providers.

Introduction

Background and Goals of the *2012 Application and Service Delivery Handbook*

Throughout the ***2012 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

There is a growing relationship between the requirements listed above. For example, in order to implement an appropriate level of security, an IT organization may implement encryption. However, the fact that the information flow is encrypted may preclude the IT organization from implementing the optimization techniques that are required to ensure acceptable performance. In addition, IT organizations don't want to optimize the performance of malware and spyware. IT organizations must identify this traffic and eliminate it.

IT organizations need to plan for optimization, security and management in an integrated fashion.

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 such as supporting chatty protocols or transmitting large files between a branch office and a data center, a second 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

Webtorials published the first edition of what became an annual series of application delivery handbooks in January 2007. Until last year, the primary goal of the handbook was to help IT organizations ensure acceptable application delivery when faced with the first generation of application delivery challenges. Beginning last year, the goal of the handbook changed in response to both the growing emphasis on services as well as the increasing impact of the second generation of application delivery challenges.

The goal of the 2012 Application and Service Delivery Handbook is to help IT organizations ensure acceptable application and/or service delivery when faced with both the first generation, as well as the emerging second generation of application and service delivery challenges.

Foreword to the 2012 Edition

While this year's edition of the application delivery handbook builds on the previous edition of the handbook, every section of the 2011 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. To compensate for those changes, the 2011 Handbook of Application Delivery is still accessible at Webtorials[1].

In early 2012 three surveys were given to the subscribers of Webtorials. Throughout this document, the IT professionals who responded to the surveys will be referred to as ***The Survey Respondents***. Two of the 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 third 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. 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 all of surveys will be used throughout the ***2012 Application and Service Delivery Handbook*** to demonstrate both the challenges facing IT organizations as well as the relative importance that IT organizations place on a wide variety of optimization and management tasks. Because many of the same questions were asked of the same survey base a year ago, the ***2012 Application and Service Delivery Handbook*** will also identify those instances in which there was a significant change in the response of the survey base over the last year. The results of surveys that ask 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.

The Importance of Ensuring Successful Application and Service Delivery

Over the past few decades, an extremely wide variety of tasks have been automated and now run on varying forms of compute devices supported by a variety of types of networks. From the hot dog vendor on the street who needs to place his orders for the next day on a website, to the Fortune 50 Company with hundreds of thousands of network-connected devices, applications and the networks that support them are critical to all businesses.

If the applications and networks that support an organization's business processes are not running well, neither are those business processes.

[1] http://www.webtorials.com/content/2011/08/2011-application-service-delivery-handbook.html

In addition to the fact that the success of a company's key business processes depends on the performance of a wide variety of applications and the networks that support them, another reason why application and service delivery continues to be an important topic for IT organizations is the fact that approximately sixty 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, the application delivery challenges facing IT organizations continue to become more formidable.

The Survey Respondents were given a set of outcomes that could result from poor application performance. They were asked to indicate the type of impact that typically occurs if one or more of their company's business critical applications are performing badly, and they were allowed to indicate multiple impacts. The impacts that were mentioned the most often are shown in **Table 3**.

Table 3: Impact of Poor Application Performance

Impact	Percentage
The Company Loses Revenue	62.0%
IT Teams Are Pulled Together	59.8%
Company Loses Customers	45.1%
CIO Gets Pressure from his/her Boss	45.1%
Harder for IT to get Funding	44.6%
CIO Gets Other Pressure	42.9%

If a business critical application is performing poorly, it has a very significant business impact and it also has a very significant impact on the IT organization.

To illustrate the importance that IT organizations place on improving application performance The Survey Respondents were asked 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 4**.

Table 4: Importance of Optimizing Business Critical Applications	
	Percentage
Extremely Important	21.6%
Very Important	38.9%
Moderately Important	28.9%
Slightly Important	8.9%
Not at all Important	1.6%

As shown in **Table 4**, 90% of The Survey Respondents indicated that getting better at optimizing the performance of a key set of business critical applications is at least moderately important to their organization.

> ***Over the next year, the most important optimization task facing IT organizations is optimizing the performance of 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. 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 and they were also asked to indicate the importance of ensuring acceptable performance for VoIP traffic. Their answers are shown in **Table 5**.

Table 5: Importance of Managing and Optimizing VoIP		
	Managing	**Ensuring Acceptable Performance**
Extremely Important	14.1%	19.8%
Very Important	31.7%	34.5%
Moderately Important	32.2%	24.4%
Slightly Important	14.6%	15.7%
Not at all Important	7.5%	5.6%

The data in **Table 5** shows that over half of The Survey respondents indicated that getting better ensuring acceptable performance for VoIP traffic is either very or extremely important to their IT organization.

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 such as VoIP, 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.

Traditional Application Delivery Challenges

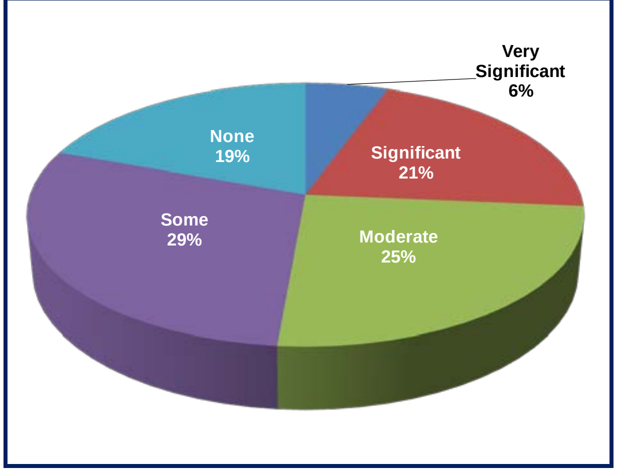
Limited Focus of Application Development

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 4.

As is often the case with surveys, the data in **Figure 4** presents a classic good news – bad news situation. The good news is that the data in **Figure 4** 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.

Figure 4: The Emphasis on Performance over the WAN



The situation depicted in **Figure 4** is unlikely to improve in the near term. That follows because with the ongoing global recession and economic challenges, many organizations are under pressure to cut costs. In many cases this business pressure results in a significant effort to trim the application development costs. This often results in a reduction in performance testing and a corresponding increase in the likelihood that an application will run badly when deployed over a WAN.

Network Latency

Network latency refers to the time it takes for data to go from the sender to the receiver and back. Since the speed of data flow is basically constant[2], WAN latency is directly proportional to the distance between the sender and the receiver. **Table 6** contains representative values for network latency; both for a LAN as well as for a private WAN[3].

[2] There are slight variations in the speed of data flow in copper vs. the speed of date flow in fiber optics.
[3] The phrase *private WAN* refers to services such as Frame Relay and MPLS that are intended primarily to interconnect the sites within a given enterprise.

Table 6: Network Latency Values

Network Type	Typical Latency
LAN	5 ms
East coast of the US to the West coast of the US	80 ms – 100 ms
International WAN Link	100 ms – 450 ms
Satellite Link	Over 500 ms

As described by Moore's Law of Internet Latency[4], Internet latency is typically greater than the latency in a private WAN. That law references the business model used by the Internet and it states, "As long as Internet users do not pay for the absolute (integrated over time) amount of data bandwidth which they consume (bytes per month), Internet service quality (latency) will continue to be variable and often poor."

Availability

Despite the Internet's original intent to provide communication even during a catastrophic event, application availability over the Internet is somewhat problematic. The Internet is not a single network, but rather millions of networks interconnected to appear as a single network. The individual networks that compose the Internet exchange information between each other that describes what IP address ranges they contain (a.k.a., routes). Within a single network - called a routing domain - a specialized networking protocol is used to communicate IP address ranges to all the routers within the individual network. Routing protocols within a network can detect a network link failure and update the routing table on all routers within a few seconds when properly designed. For the exchange of information between networks - called inter-domain routing - a special routing protocol, the Border Gateway Protocol (BGP), is used. The size and complexity of the Internet as well as the inherent characteristics of BGP mean that a failed network link and the resulting routing path change may take several minutes before all routing tables are updated. In contrast, traditional voice circuits take milliseconds to reroute voice calls when a network link fails.

The impact of a network link failure and the time it takes for the Internet to update its routing table and to find an alternative path varies according to type of application involved. For a simple web application, a brief outage may go unnoticed if users are not loading the web page during the outage. For real-time applications like VoIP or IP Video, an outage of several seconds may cause interrupted calls and video sessions. In addition, there are two primary types of communication over the Internet: TCP and UDP. With TCP communication, lost packets are retransmitted until the connection times out. With UDP communication, there is no built-in mechanism to retransmit lost data and UDP applications tend to fail rather than recover from brief outages.

Bandwidth Constraints

Unlike the situation within a LAN, within a WAN there are monthly recurring charges that are generally proportional to the amount of bandwidth that is provisioned. For example, the cost of T1/E1 access to an MPLS network varies from roughly $450/Mbps/month to roughly $1,000/Mbps/month. In similar fashion, the cost for T1/E1 access to a Tier 1 ISP varies from

[4] http://www.tinyvital.com/Misc/Latency.htm

roughly $300/Mbps/month to roughly $600/Mbps/month. The variation in cost is largely a function of geography. WAN costs tend to be the lowest in the United States and the highest in the Asia-Pacific region.

To exemplify how the monthly recurring cost of a WAN leads to bandwidth constraints, consider a hypothetical company that has fifty offices and each one has on average a 2 Mbps WAN connection that costs on average $1,000/month. Over three years, the cost of WAN connectivity would be $1,800,000. Assume that in order to support an increase in traffic, the company wanted to double the size of the WAN connectivity at each of its offices. In most cases there wouldn't be any technical impediments to doubling the bandwidth. There would, however, be financial impediments. On the assumption that doubling the bandwidth would double the monthly cost of the bandwidth, it would cost the company over a three-year time frame an additional $1,800,000 to double the bandwidth. Because of the high costs, very few if any companies provision either their private WAN or their Internet access to support peak loads. As such, virtually all WANs, both private WANs and the Internet, exhibit bandwidth constraints which result in packet loss.

Packet Loss

Packet loss can occur in either a private WAN or the Internet, but it is more likely to occur in the Internet. Part of the reason for that was previously mentioned - 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. Another part of the reason for why there is more packet loss in the Internet than there is in a private WAN is the previously mentioned Internet business model. One of the affects of that business model is that there tend to be availability and performance bottlenecks at the peering points.

If packet loss occurs, TCP will re-transmit packets. In addition, the TCP slow start algorithm (see below) assumes that the loss is due to congestion and takes steps to reduce the offered load on the network. Both of the actions have the affect of reducing throughput on the WAN.

Characteristics of TCP

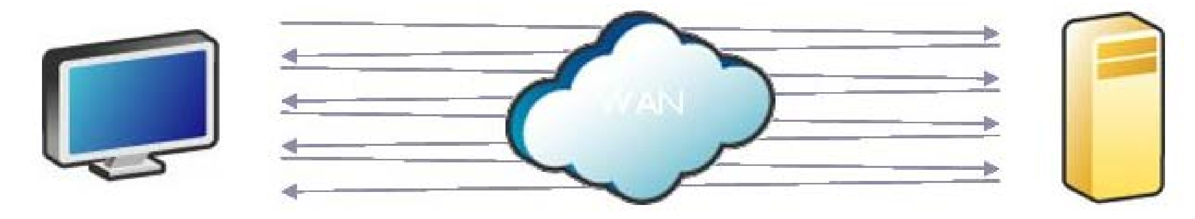
TCP is the most commonly used transport protocol and it causes missing packet(s) to be re-transmitted based on TCP's retransmission timeout parameter. This parameter controls how long the transmitting device waits for an acknowledgement from the receiving device before assuming that the packets were 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 TCP parameter that impacts performance is the 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 if there are no problems with the communications. In addition to the initial communications between two devices, the slow start algorithm is also applied in those situations in which a packet is dropped.

Chatty Protocols and Applications

The lack of emphasis on an application's performance over the WAN during application development is one of the factors that can result in the deployment of chatty applications[5] as illustrated in **Figure 5**.

Figure 5: 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 5 milliseconds, 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 1 second, which is generally not noticeable. However, the delay over the WAN is 20 seconds, which is very noticeable.

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

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

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 and the responses of last year's survey respondents are shown in **Table 7.**

Table 7: Importance of Optimizing Chatty Protocols

Level of Importance	2011 Responses	2012 Responses
Extremely	12%	6%
Very	27%	21%
Moderately	33%	33%
Slightly	18%	24%
Not at all	10%	16%

Optimizing chatty protocols such as CIFS was one of the primary challenges that gave rise to the first generation of WAN optimization products. The data in **Table 7** indicates that optimizing chatty protocols is becoming somewhat less important to IT organizations. That said, the data

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

in **Table 7** indicates that for 60% of The Survey Respondents it is at least moderately important for their organization to get better at optimizing these protocols over the next year.

Optimizing chatty protocols was only one of a number of first generation application delivery challenges that are still important to IT organizations. For example, 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

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. As discussed below, the typical social media site contains a wide variety of categories of content.

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.

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. Web-based applications is a mainstream model of computing in which an application is accessed over the Internet or an Intranet and the user interface is a browser. Web-based applications are popular in part due to the ubiquity of web browsers. Another reason for the popularity of Web-based applications is that in contrast to traditional client/server applications, an upgrade to the server-side code does not require that changes be made to each client. Browser functionality is a key enabler that allows businesses to adopt BYOD[6], and hence avoid the capital investment it takes to refresh end user devices, but still have the requisite functionality to enable users to successfully access applications.

There are, however, 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[7]. That 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."

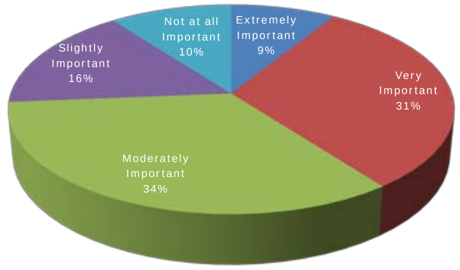
There are also performance challenges that are somewhat unique to this class of application. For example, unlike CIFS, HTTP is not a chatty protocol. However, HTTP is used to download

[6] Bring Your Own Device to Work (BYOD) will be explained in more detail in a subsequent section.
[7] http://www-07.ibm.com/businesscenter/au/services/smbservices/include/images/Secure_mobility.pdf

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.

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 6**, 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 the vast majority of IT organizations.

Figure 6: Importance of Optimizing Protocols Other than TCP



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 market research[8], these Web-based applications have become increasingly complex. One result of that research is depicted in **Table 8**. 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 8: 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.

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

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.

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.

Server Overload

A server farm is a group of servers that are networked together with the goal of meeting requirements that are beyond the capability of a single server. One of the challenges associated with implementing a server farm is to ensure that a request for service is delivered to the most appropriate server. There are many ways to define what the phrase *most appropriate server* means. Certainly the server has to be available. Ideally, the most appropriate server is the server that is processing the lightest load of any member of the server farm.

In addition to the situation in which there are more requests for service than can be handled by a single server, another way that a server can become overloaded is by having to process computationally intense protocols such as SSL.

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.

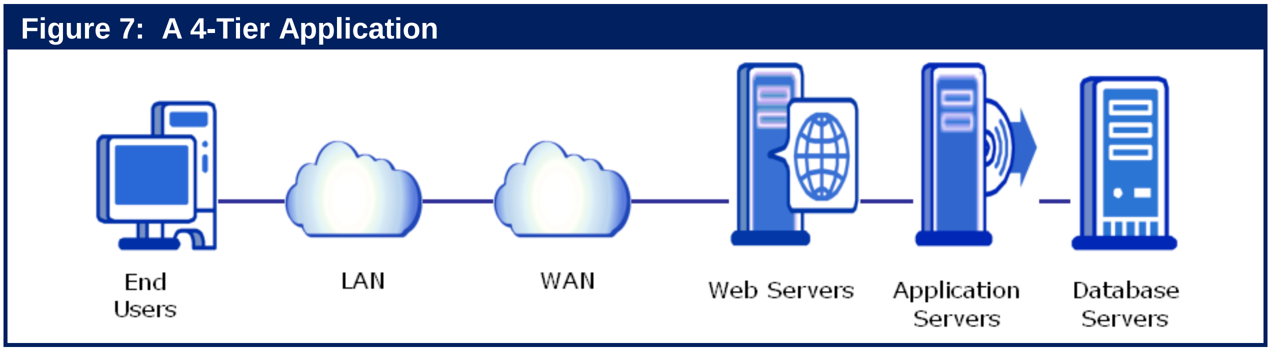
In the vast majority of situations, when people access an application they are accessing it over the WAN instead of the LAN.

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 7**) 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 7: 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 recently as a few years ago, few, if any, of the servers in the typical n-tier application were virtualized. However, in the current environment it is becoming increasingly common to have these servers be virtualized. The unique challenges that server virtualization and cloud computing bring to managing and optimizing n-tier applications is discussed in a subsequent section of the handbook.

Complexity

The overall complexity of both private WANs and the Internet tends to increase the impact of the previously described application delivery challenges. For example, as the number of links that the data has to transit between origin and destination increases, so does the delay. As delay increases, the negative impact of a chatty protocol or application is magnified.

It is not, however, just the number of links and the complex topologies that complicate application delivery, it is also the use of complex protocols such as TCP and BGP. The Internet uses BGP to determine the routes from one subtending network to another. When choosing a route, BGP strives to minimize the number of hops between the origin and the destination. BGP doesn't, however, strive to choose a route with the optimal performance characteristics; i.e., lowest delay, lowest packet loss. Given the complex, dynamic nature of the Internet, a given network or a particular peering point router can go through periods where it exhibits 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.

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.

As the complexity increases, so does the vulnerability to security attacks.

Expanding Scope of Business Critical Applications

A decade or so ago when business critical applications were deployed, the scope of the application was intra-company; e.g., all of the users of the application were employees of the company. In the current environment, virtually all organizations use their applications and networks to interact with myriad people outside of the company including suppliers, business partners and customers. This presents some significant challenges for IT organizations, as they are typically held responsible for the performance and management of these applications even though in many cases they don't have access to the enabling IT infrastructure that they would have if the application was entirely intra-company. Also, as described in subsequent sub-section, it creates some very significant security challenges in part due to the growing sophistication of cyber attacks.

Increased Regulations

Most governments and regulators are aware of the growing criticality of IT in general, and of the increased importance of the Internet in particular. This awareness has led to increased legislation and regulation as governments attempt to exert control over how businesses operate. These regulations range from law enforcement (i.e., Communications Assistance for Law Enforcement Act – CALEA) to privacy (i.e., Health Insurance Privacy and Accountability Act – HIPPA) to fraud protection (i.e., Payment Card Industry Data Security Standard – PCI DSS) and to hundreds of other regulations. Legislative processes, however, operate considerably slower than the technology industry does and so regulations are often out of date with, or inappropriate for the current technology. Another application and service delivery challenge is that with a growing body of laws and regulations at both the national and local level, it is often difficult for an IT organization to know if they are in compliance with regulations.

Security Vulnerabilities

Security vulnerabilities can be classified as both a first and a second-generation application and service delivery challenge. The distinction between a first and a second-generation security challenge is based on factors such as who is doing the attack, what are they attacking, what tools and techniques are they using and what is their motivation.

For example, until recently the majority of security attacks were caused by individual hackers, such as Kevin Mitnick, who served five years in prison in the late 1990s for computer- and communications-related hacking crimes. The goal of this class of hacker is usually to gain notoriety for themselves and they often relied on low-technology techniques such as dumpster diving.

However, over the last few years a new class of hacker has emerged and this new class of hacker has the ability in the current environment to rent a botnet or to develop their own R&D lab. This new class includes crime families and hactivists such as Anonymous. In addition, some national governments now look to arm themselves with Cyber Warfare units and achieve their political aims via virtual rather physical means. Examples include China's attack on Google and the Stuxnet attack on Iran's nuclear program.

In addition to the types of attacks mentioned in the preceding paragraph, one of the ways that the sophistication of the new generation of attackers has been manifested is just the sheer scale of the attacks. As recently as a decade ago, the peak rate of Distributed Denial of Service (DDoS) attacks was roughly 500 Mbps. In the current environment, the peak rate is more than 50 Gbps. This means that over the last decade the peak rate of a DDoS attack has increased by at least a factor of one hundred. Another example of the sophistication of the current generation of hacker is the growing number of attacks based on SQL injection. In this type of attack, malicious code is inserted into strings that are later passed to an instance of SQL Server for parsing and execution. The primary form of SQL injection consists of direct insertion of code into user-input variables that are concatenated with SQL commands and executed.

In March 2012, IBM published its annual X-Force 2011 Trend and Risk Report[9]. That report highlighted the fact that new technologies such as mobile and cloud computing continue to create challenges for enterprise security. Some of the key observations made in that report are:

- **Mobile Devices**
  The report stated that in 2011 there was a 19 percent increase over 2010 in the number of exploits publicly released that can be used to target mobile devices such as those that are associated with the movement to Bring your Own Device (BYOD) to work. The report added that there are many mobile devices in consumers' hands that have unpatched vulnerabilities to publicly released exploits, creating an opportunity for attackers.

- **Social Media**
  With the widespread adoption of social media platforms and social technologies, this area has become a target of attacker activity. The IBM report commented on a surge in phishing emails impersonating social media sites and added that the amount of information people are offering in social networks about their personal and professional

[9] X-Force 2011 Trend and Risk Report

lives has begun to play a role in pre-attack intelligence gathering for the infiltration of public and private sector computing networks.

- **Cloud Computing**
  According to the IBM report, in 2011, there were many high profile cloud breaches affecting well-known organizations and large populations of their customers. IBM recommended that IT security staff should carefully consider which workloads are sent to third-party cloud providers and what should be kept in-house due to the sensitivity of data. The IBM X-Force report also noted that the most effective means for managing security in the cloud may be through Service Level Agreements (SLAs) and that IT organizations should pay careful consideration to ownership, access management, governance and termination when crafting SLAs.

The Blue Coat Systems 2012 Web Security Report[10], focused on a number of topics including malnets and social networking. A malware network, or malnet, gathers users, most frequently when they are visiting trusted sites and routes them to malware. According to the Blue Coat Report, "In 2011, malnets emerged as the next evolution in the threat landscape. These infrastructures last beyond any one attack, allowing cybercriminals to quickly adapt to new vulnerabilities and repeatedly launch malware attacks. By exploiting popular places on the Internet, such as search engines, social networking and email, malnets have become very adept at infecting many users with little added investment."

The report noted the increasing importance of social networking and stated that, "Since 2009, social networking has increasingly eclipsed web-based email as a method of communications." The report added that, "Now, social networking is moving into a new phase in which an individual site is a self-contained web environment for many users – effectively an Internet within an Internet." For example, according to the Blue Coat report 95% of content types that are found on the Internet are also found within social networking sites. The five most requested subcategories of content that were requested from social networking sites, and the percentage of times that they were requested are shown in **Table 9**.

Table 9: Most Requested Content from Social Media Sites

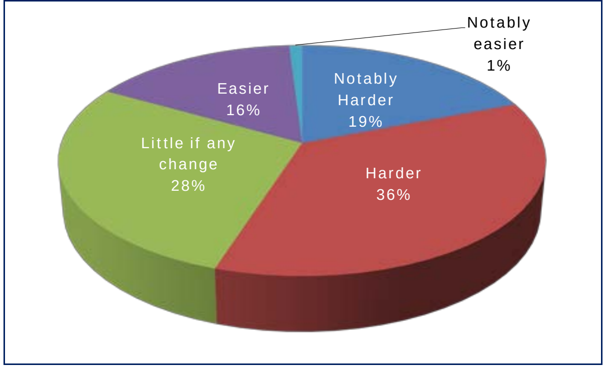
Subcategory of Content	Percentage of Times it was Requested
Games	37.9%
Society/Daily Living	23.8%
Personal Pages/Blogs	6.4%
Pornography	4.9%
Entertainment	4.2%

Part of the challenge that is associated with social network sites being so complex is that IT organizations cannot just look at a social media site as one category and either allow or deny access to it. Because these sites contain a variety of classes of content, IT organizations need the granular visibility and control to respond differently to requests at the same social media site for different types of content.

[10] http://www.bluecoat.com/sites/default/files/documents/files/BC_2012_Security_Report-v1i-optimized.pdf

The Emerging Application and Service Delivery Challenges

Figure 8: Impact of Emerging Challenges on Application and Service Delivery



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 8**. The data in **Figure 8** indicates that ensuring acceptable application and service delivery continues to become increasingly challenging.

IT organizations are beginning to face a set of new challenges which is expected to significantly complicate the task of ensuring acceptable application and service delivery.

Mobility and BYOD

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 else; 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 concerns relative to supporting mobile workers. One such concern is that up through 2010, the most common device used by a mobile worker was a PC. In 2011, however, more tablets and smartphones shipped than PCs[11]. Related to the dramatic shift in the number and types of mobile devices that are being shipped, many companies have adopted the BYOD (Bring Your Own Device to work) concept whereby employees use their own devices to access applications.

The Survey Respondents were asked to indicate the types of employee owned devices that their organization allows to connect to their branch office networks and which of these devices is actively supported, Their responses are shown in **Table 10**.

[11] http://gizmodo.com/5882172/the-world-now-buys-more-smartphones-than-computers

Table 10: Support for Employee Owned Devices

	Not Allowed	Allowed but not Supported	Allowed and Supported
Company managed, employee owned laptop	22%	24%	54%
Employee owned and managed laptop	38%	38%	25%
Blackberry	17%	24%	58%
Apple iPhone	14%	30%	55%
Android phone	19%	33%	48%
Windows mobile phone	26%	40%	34%
Apple iPad	18%	40%	52%
Android based tablet	28%	37%	35%
Windows based tablet	28%	36%	37%

The data in **Table 10** indicates that there is wide acceptance BYOD. As a result, the typical branch office network now contains three types of end user devices that are all accessing business critical applications and services. This includes PCs as well as the new generation of mobile devices; i.e., smartphones and tablet computers. Because of their small size, this new generation of mobile devices doesn't typically have wired Ethernet ports and so they are typically connected via what is hopefully a secure WiFi network in the branch office.

This new generation of mobile devices, however, doesn't run the Windows O/S and the existing security and management services for PCs must be extended for mobile devices or alternatively, additional products added to perform these functions. Similar to PCs, smartphone and tablet computers are subject to malware and network intrusion attacks. On PCs, there are mature, robust products for malware protection (e.g. anti-virus software) and network intrusion protection (e.g., personal firewall), but these protections are just now emerging for smartphones and tablet computers[12]. Similarly, inventorying and updating installed software on smartphone and tablet computers are emerging capabilities and a critical area for Mobile Device Management solutions.

The BYOD movement has resulted in a loss of control and policy enforcement.

These new mobile devices are more mobile than is the traditional laptop and this causes some changes relative to how users remotely access corporate applications. For example, with the new generation of mobile devices, end users utilize remote access services more frequently and for more total time. The adoption of BYOD also results in a doubling or tripling of the number of operating systems that must be supported. This requires the expansion of remote access solutions. Often times, however, the existing remote access gateways cannot support the new mobile O/S platforms and parallel remote access solutions must be added. Avoiding this expansion is one advantage of using thin client access for the new generation of mobile devices.

[12] http://www.computerworld.com/s/article/9224244/5_free_Android_security_apps_Keep_your_smartphone_safe)

Unfortunately, this new generation mobile devices were architected and designed primarily for consumer use which is an environment in which the IT security risk is lower than it is in a corporate environment. A compromised consumer device typically exposes the consumer to loss in the range of hundreds to thousands of dollars. A compromise in a corporate setting can result in a loss of tens of thousands to millions of dollars. Unfortunately, as noted, the new generation of end user devices cannot currently match the security and manageability of PCs. This creates security and management challenges in general and can prevent these devices from being used where strict security regulations must be adhered to; e.g., the Healthcare Insurance Portability and Accountability Act (HIPPA) and the Payment Card Industry Data Security Standard (PCI DSS).

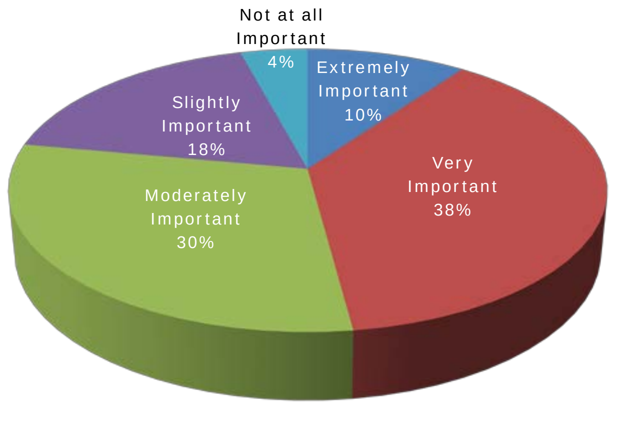
Adopting BYOD increases a company's vulnerability to security breaches.

Another key concern relative to supporting mobile workers is how the applications that these workers access have 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 announcement[13] 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, as previously discussed, 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 9**.

One conclusion that can be drawn from the data in **Figure 9** is that roughly half of all IT organizations consider it to be either extremely or very important to get better at improving the performance of applications used by mobile workers.

Figure 9: Importance of Optimizing Mobile Applications



Mandate for Agility

BYOD is a key component of the overall consumerization of IT movement that has been evolving slowly over the last ten years.   A decade ago the vast majority of the company's employees didn't regard themselves as being technology savvy. However, today the

[13] Wall Street Journal, May 17, 2012, page B7

environment is dramatically different. It has become very common for a company's employees to have wifi in their apartments and homes as well as high-speed access to the Internet. They usually have email accounts from companies such as Google or Yahoo that allow them to save a huge volume of emails and most have devices at home that can print, scan and copy documents. They also have smartphone and tablets that allow them to quickly download, either for free or for very little money, an application that will tell them whatever they need to know, from the artist who recorded a song that is playing in the background to the location of the closest frozen yogurt store to their arrival gate at virtually any airport.

Because of their awareness of the technology that is available to them in their homes and from the Internet, a growing number of business and functional managers don't want to be told that it will takes months for the IT organization to implement the functionality they need. This pressure is pushing IT organizations to become much more agile than they ever have been.

IT organizations are under more pressure for agility than they ever have been in the past.

In many cases, if IT organizations don't become more agile, the business and functional managers that they support will increasingly turn to public cloud providers and the value provided by IT organizations will steadily diminish.

IT Organizations as Service Brokers

In the traditional IT environment, the IT organization is the primary provider of IT services. Part of the challenge that is associated with the IT organization being the primary provider of IT services is that sometimes the IT organization can't meet the needs of the business units in a timely fashion. In the past, the way that business unit managers have dealt with this lack of support was by having their own shadow IT organization whereby the business unit managers have some people on their staff whose role is to provide the IT services that the business unit manager can't get from the IT organization.

In the current environment, public cloud providers sometimes play the role of a shadow IT organization.

Public cloud providers play this role when a company's business and functional managers go around the company's IT organization to obtain services or functionality that they either can't get from their IT organization or they can't get in a timely or cost effective manner. In some instances the IT function is in a position to stop the non-sanctioned use of public cloud computing once they find out about it. However, in many other instances they either are unaware that public cloud computing solutions are being used or they aren't in a position to stop that from happening.

Instead of trying to prevent business unit managers from acquiring public cloud services, a better role for an IT organization is to modify their traditional role of being the primary provider of IT services and to adopt a role in which they provide some IT services themselves and act as a broker between the company's business unit managers and cloud computing service providers for other services. In addition to contract negotiations, the IT organization can add value by ensuring that the acquired application or service doesn't create any security or compliance issues, can perform well, can be integrated with other applications as needed, can scale, is cost effective and can be managed.

IT organizations can provide a lot of value by acting as a broker of services provided both internally and externally.

Increasing Number of Business Critical Applications

As automation and web technologies have been embraced, the functional size of the typical application has decreased. A few decades ago, software packages were bought and customized. The software package – the *application* – typically provided comprehensive functionality to multiple business units within a company and it was often customized to present bite size units of functionality to each department. For example, a financial application would often include functionality for accounts receivable, fixed assets, accounts payable, inventory tracking, purchase orders, order entry, etc. The application was customized to slice down the options to present just the right amount of functionality to the various departments of the business. In the vast majority of cases, large companies ran the majority of their key business processes on just a handful of these large, comprehensive business critical applications.

As web technologies evolved, the original application became a database with web services in front of the database and the functionality provided to a given department became the new application. In essence, the large software package became a series of smaller web applications. In most cases, the sum of the functionality provided by the smaller web applications equaled or surpassed the functionality of the original application. The introduction of smartphones and tablet computers further narrows the focus of an application – now called an *app* – to just a portion of the transactions done on the departmental website. The transformation of large software packages into smaller web applications and the deployment and use of smartphones and tablets are two of the reasons driving the increase in the number of business critical applications.

Within most organizations the number of business critical applications is increasing dramatically.

In order to better understand the trend to have an increasing number of business critical applications, The Survey Respondents were asked, “How many applications does your company consider to be business critical?” Their responses are shown in **Table 11**.

Table 11: Number of Business Critical Applications

Number of Business Critical Applications	All Companies	Large Companies
1- 5	32.1%	5.7%
6 – 10	22.6%	7.5%
11 – 20	16.5%	15.1%
21 – 100	18.4%	37.7%
> 100	10.4%	34.0%

The middle column of **Table 11** shows the responses of all of The Survey Respondents while the right hand column shows the responses of just The Survey Respondents who work in a company that has 10,000 or more employees. One observation that can be drawn from Table 11 is that currently over a quarter of all companies have more than 20 business critical applications. The trend to have a growing number of business critical applications is even more pronounced in large companies.

Over a third of large companies have more than 100 business critical applications.

Services Oriented Architectures (SOA) with Web Services

The movement to adopt 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 more than it does 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; a.k.a.; a mashup.

Another industry movement often associated with Web 2.0 is the deployment of Rich Internet Applications (RIA). In a traditional Web application all processing is done on the server, and a new Web page is downloaded each time the user clicks. In contrast, an RIA can be viewed as “a cross between Web applications and traditional desktop applications, transferring some of the processing to a Web client and keeping (some of) the processing on the application server.” [14]

The introduction of new technologies tends to further complicate the IT environment and leads to more security vulnerabilities. AJAX (Asynchronous JavaScript and XML) 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[15] 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.”

[14] Wikipedia on Rich Internet Applications
[15] Ajax Resource Center

The Increased Focus on Services

Just as IT organizations are getting somewhat comfortable with managing the performance of applications they are being tasked with managing the performance of services. IT professionals use the term ***service*** in a variety of ways. Throughout this handbook, the definition of the term ***service*** will include the key characteristics of the ITIL (Information Technology Infrastructure Library) definition of service[16]. 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 acquire 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 and 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 12**.

Table 12: Importance of Monitoring and Managing Public Cloud Services

	Storage Services	Compute Services	Applications
Extremely Important	7.2%	6.4%	15.6%
Very Important	16.3%	20.3%	21.7%
Moderately Important	26.5%	23.8%	29.4%
Slightly Important	25.3%	25.0%	19.4%
Not at all Important	24.7%	24.4%	13.9%

The data in **Table 12** indicates that IT organizations are notably more interested in managing the applications that they acquire from a Software-as-a-Service (SaaS) provider than they are the services that they acquire from an Infrastructure-as-a-Service (IaaS) provider. Unfortunately the task of managing SaaS-based applications is significantly harder than managing solutions from an IaaS provider. That follows because it is relatively easy for an IT organization to host some virtualized management, security or optimization functionality at the IaaS provider's facility and almost impossible for an IT organization to host virtualized management, security or optimization functionality at a SaaS provider's facility.

[16] ITIL definition of service

As shown in **Table 12**, 24.7% 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. The 24.7% 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 storage services obtained from an IaaS vendor is not an important task. However, that conclusion is contradicted by the fact that almost a quarter 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 12** 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.

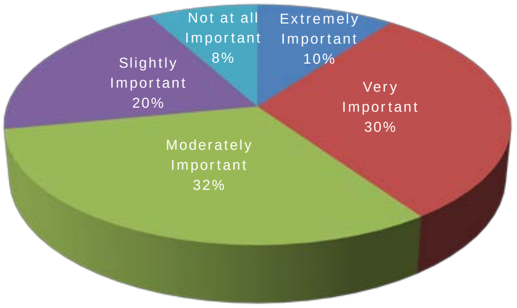
It is also insightful to realize the in last year's surveys, 32.6% of The Survey Respondents responded with "not at all important" when asked about the importance of getting better at monitoring and managing storage services that they acquire from a public cloud computing vendor.

The interest on the part of IT organizations to manage services that they acquire from an IaaS vendor has increased over the last year.

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 10**.

Getting better at managing a business service that is supported by multiple, inter-related applications is an important task for the vast majority of IT organizations.

Figure 10: Importance of Managing a Business Service



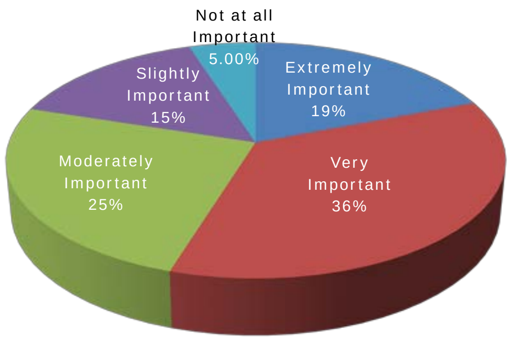
Internal Service Level Agreements (SLAs)

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

Figure 11: The Important of Getting Better at Managing Internal SLAs



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 11**.

The data in **Figure 11** leads to two related conclusions. The obvious conclusion is that managing internal SLAs is either very or extremely 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 subsection of the handbook, the movement to utilize public cloud computing services greatly increases the difficulty associated with managing an internal SLA.

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 13**.

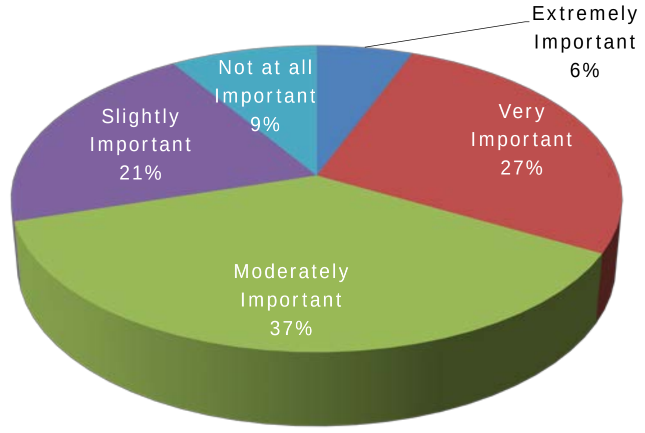
Table 13: Deployment of Virtualized Servers

	None	1% to 25%	26% to 50%	51% to 75%	76% to 100%
Have already been virtualized	18%	30%	25%	16%	11%
Expect to be virtualized within a year	11%	28%	24%	25%	12%

The data in **Table 13** indicates that the vast majority of organizations have made at least some deployment of server virtualization and that the deployment of server virtualization will increase over the next year.

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 lose visibility into the inter-VM traffic. This limits the IT organization's ability to perform functions such as security filtering, performance monitoring and troubleshooting. To quantify the impact of losing 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 12**.

Figure 12: Manage the Inter-VM Traffic



The data in **Figure 12** 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

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.

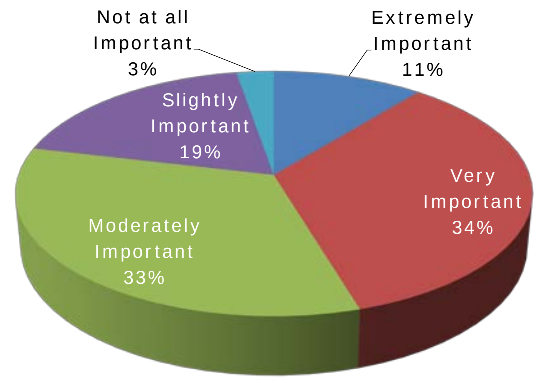
To quantify the impact that managing on a per-VM basis is having on IT organizations,

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 13**.

One observation that can be drawn from the data in **Figure 13** is that unlike the situation with managing inter-VM traffic:

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.

Figure 13: Managing 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 has 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[17] that would move a VM to

[17] VMotion

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.

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

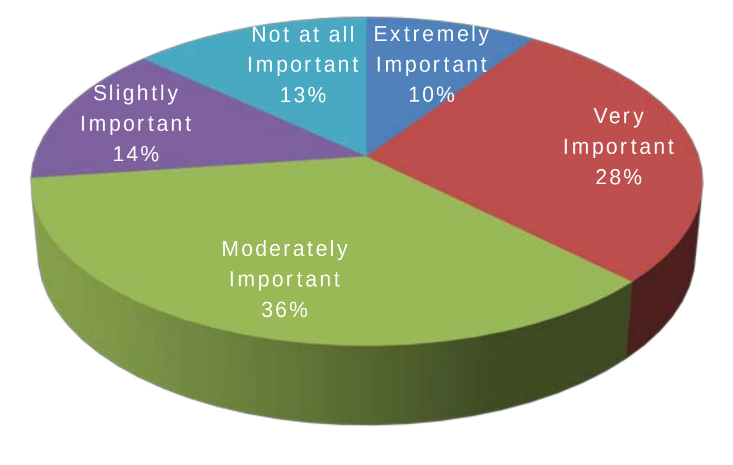
Challenges of Server Virtualization

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 and, as previously discussed, becoming more agile is a critical goal for IT organizations. 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 14**.

Figure 14: Supporting the Mobility of VMs



Given that the data in **Table 13** indicates that IT organizations plan to increase their deployment of virtualized servers, one observation that can be drawn from the data in **Figure 14** 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 data 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.

Desktop 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 14**.

Table 14: Deployment of Virtualized Desktops

	None	1% to 25%	26% to 50%	51% to 75%	76% to 100%
Have already been virtualized	44%	49%	6%	1%	0%
Expect to be virtualized within a year	24%	53%	20%	2%	1%

Comparing the data in **Table 14** with the data in **Table 13** 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 14** with the data in the second row of **Table 14** yields the following conclusion:

Over the next year, the number of IT organizations who have implemented at least some desktop virtualization will increase dramatically.

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. VDI is a relatively new form of server-side virtualization in which a VM on a central server is dedicated to host a single virtualized desktop.

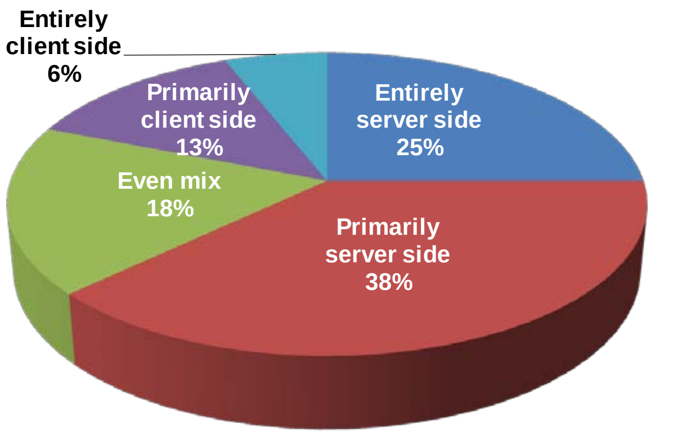
Client-side application virtualization is based on a model in which applications are streamed on-demand from central servers to client devices over a LAN or a WAN. On the client-side, streamed applications are isolated from the rest of the client system by an abstraction layer inserted between the application 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 were asked to indicate which form(s) of desktop virtualization they will have implemented within twelve months. Their answers are shown in **Figure 15**.

One conclusion that can be drawn from the data in **Figure 15** is:

The vast majority of virtualized desktops will be utilizing server side virtualization.

Figure 15: Implementation of Desktop Virtualization



Challenges of Desktop Virtualization

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

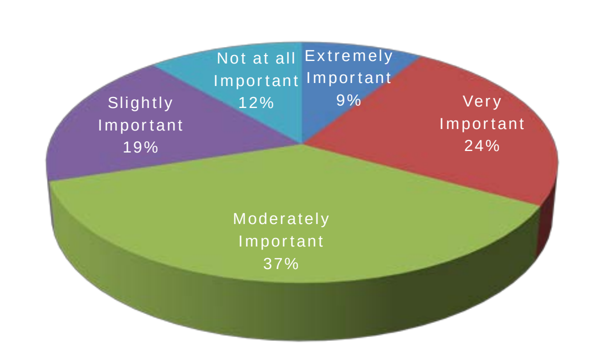
From a networking perspective, the primary challenge in implementing desktop virtualization is achieving adequate performance and an acceptable user experience for client-to-server connections over a WAN.

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 16**.

One conclusion that could be drawn from the data in **Figure 16** 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. In addition, 70% of The Survey Respondents indicated that getting better at optimizing the performance of virtualized desktops was at least moderately important to their organization. A year ago, only 59% of The Survey Respondents gave that indication.

> ***Improving the performance of virtualized desktops is becoming increasingly important to IT organizations.***

Figure 16: Optimizing the Performance of Virtualized Desktops



Ensuring acceptable performance for desktop virtualization 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.

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.

Some of the specific changes that result from implementing VDI include:

- **Network Traffic Changes** – The original network traffic between the branch office and the corporate data center now stays within the data center and is replaced with display data to the thin clients.

- **Application Performance Changes** – The end user's application performance experience is now determined by the performance of display data across the network rather than by the performance of data from the server sent across the network.

- **Expanding End User Devices** – Using thin clients to access applications gives end users more choices with end user devices. As long as a compatible thin client is available on the device, it can be used. Some corporate applications will work better with the smaller displays on new generation mobile devices than other applications. Each corporate application should be tested against a variety of devices to ensure acceptable results.

- **Device Security Posture Changes** – Thin client access to corporate applications reduces the amount of data stored locally on the end user's devices and thus lessens the need to secure and encrypt local data.

- **No off-line working** – Using a thin client requires network connectivity to the corporate data center and so if there is no connectivity, access to applications and services is not available.

- **Segmentation of Personal and Business use** – Using a thin client to access a corporate application on a BYOD device allows a segmentation of personal and business use of the device by keeping the business-related data separate from the personal data. This limits the business IT security risk to just the display.

- **Reduced End User Device Costs** – Thin client access to corporate applications requires only enough computing power to process display operations and eliminates any database, business logic computation and other calculations. Low compute power devices with meager storage needs cost less than powerful computers.

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 15** show estimates for the amount of WAN bandwidth required by XenDesktop as documented in an entry in The Citrix Blog[18].

Table 15: 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 15** 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.

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.

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

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.

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.

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. More detail on these characteristics can be found in the 2011 Application and Service Delivery Handbook.

- ***Centralization*** of applications, servers and storage resources.
- Extensive ***virtualization*** of every component of IT.
- ***Standardization*** of the IT infrastructure.
- ***Simplification*** of the applications and services provided by IT.
- ***Technology convergence*** such as the integration of servers, networks and computing.
- ***Service orchestration*** to automate provisioning and controlling the IT infrastructure.
- ***Automation*** of as many tasks as possible.

- ***Self-service*** to enable end users to select and modify their use of IT resources.

- ***Usage sensitive chargeback*** on a user and/or departmental basis.

- The ***dynamic movement of resources*** such as virtual machines and the associated functionality.

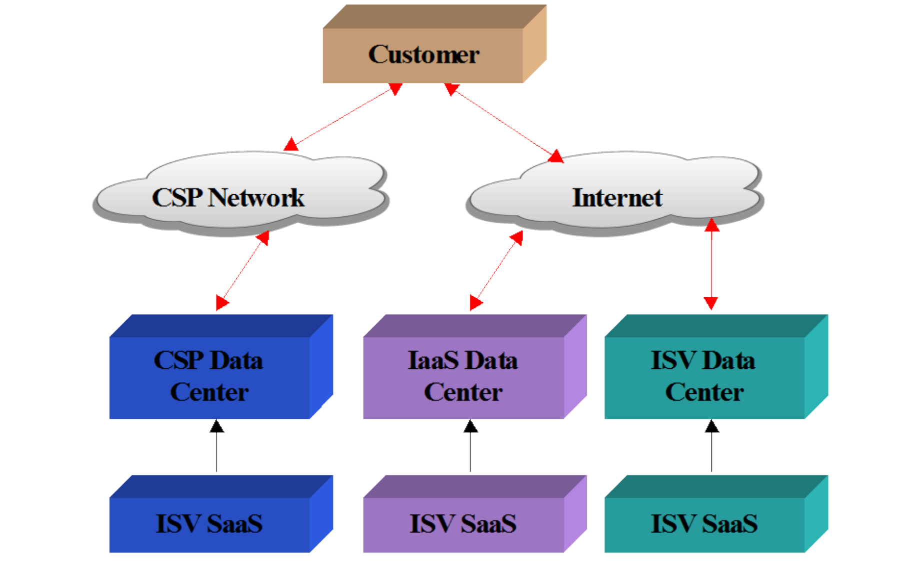
Public Cloud Computing

Background

Cloud Computing Service Providers (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). PaaS solutions 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.

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 17 shows some of the common distribution models for SaaS and IaaS solutions. As shown in Figure 17, 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 17: Distribution Models for Cloud-Based Solutions



Both of the 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. As was also previously mentioned, getting better at managing internal SLAs is either very or extremely important to the majority of IT organizations.

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.

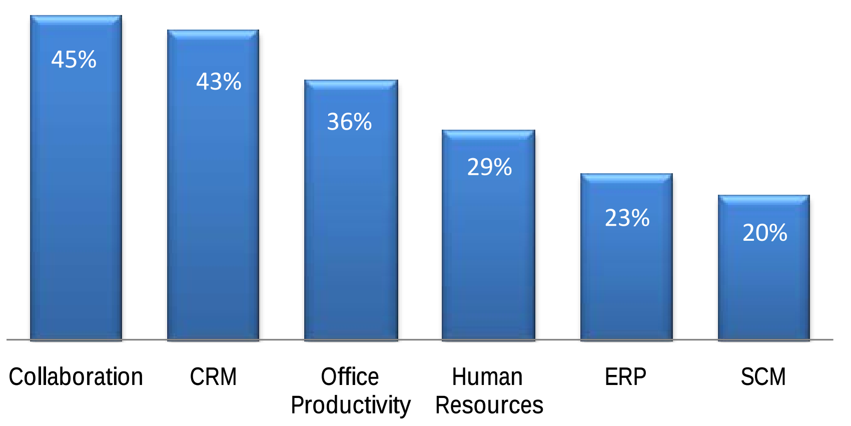
SaaS

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.

The Survey Respondents were asked about their company's use of SaaS-based applications. Figure 18 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 18: Popular Categories of SaaS-Based Applications



The functionality provided by each of the six categories of applications listed in Figure 18 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 18**, there are tens, and sometimes hundreds, of SaaS-based solutions currently available[19]. **Table 16** contains a listing of some representative SaaS providers for each category.

Table 16: 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 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 (acquired by Time Warner), NTT Communications, Orange Business Services, Rackspace, Savvis (acquired by CenturyLink), Terremark (acquired by Verizon) and Verizon.

The Survey Respondents were asked how likely it was over the next year that their company would acquire some of the traditional services provided by an IaaS supplier. Their responses are shown in **Table 17**.

Table 17: Interest in Traditional IaaS Services

	Will Not Happen	Might Happen	50/50 Chance	Will Likely Happen	Will Happen
Application Hosting	19.4%	26.4%	17.6%	17.6%	19.0%
Disaster Recovery	27.0%	28.8%	16.7%	13.0%	14.4%
High Performance Computing	44.5%	23.9%	16.3%	9.6%	5.7%

[19] Saas-showplace.com

Given that high performance computing (HPC) is somewhat of a niche application, it was not surprising that there was relatively little interest in acquiring HPC from an IaaS supplier. That said, over a third of The Survey Respondents indicated that over the next year that their company either would or would likely acquire application hosting services from an IaaS. In addition, over a quarter of The Survey Respondents indicated that over the next year their company either would or would likely acquire disaster recovery services from an IaaS.

With the exception of collaboration, the solutions that organizations have acquired from CCSPs have typically been enterprise applications such as CRM or the basic compute and storage resources that are required to run applications. Recently, a new class of solutions has begun to be offered by CCSPs. These are solutions that have historically been provided by the IT infrastructure group itself and include network and application optimization, VoIP, Unified Communications (UC), security, network management and virtualized desktops.

The Survey Respondents were asked how likely it was over the next year that their company would acquire a traditional IT service from an IaaS provider. Their responses are shown in **Table 18**.

Table 18: Interest in Obtaining IT Services from an IaaS Provider

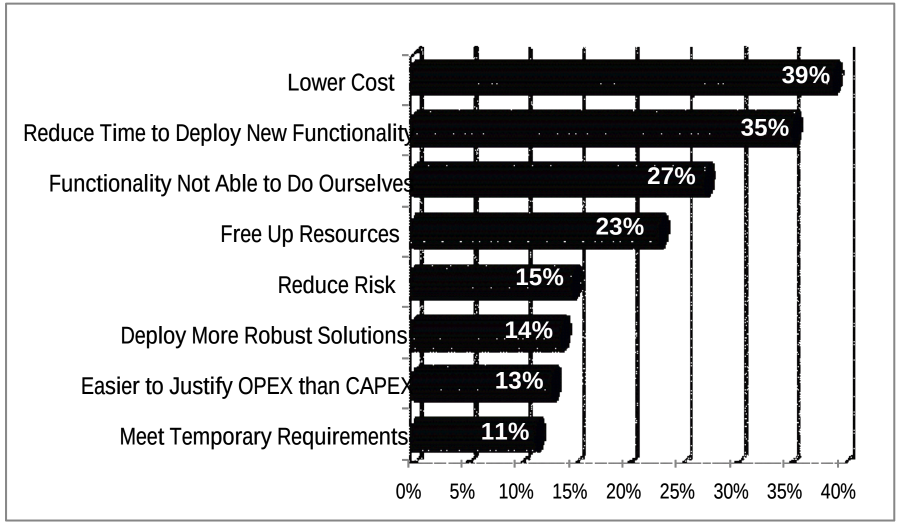
	Will Not Happen	Might Happen	50/50 Chance	Will Likely Happen	Will Happen
VoIP	32.6%	18.6%	15.3%	13.5%	20.0%
Unified Communications	30.2%	22.8%	20.5%	14.9%	11.6%
Security	42.6%	17.1%	14.4%	11.6%	14.4%
Network and Application Optimization	32.1%	28.8%	16.0%	14.6%	8.5%
Network Management	41.4%	22.3%	13.5%	13.5%	9.3%
Application Performance Management	37.9%	26.5%	15.6%	11.4%	8.5%
Virtual Desktops	38.8%	28.0%	15.9%	12.1%	5.1%

The data in **Table 18** indicates that IT organizations have a strong interest in acquiring a wide range of IT functionality from IaaS providers.

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.

Part of the conventional wisdom in the industry is that one of the inhibitors to the adoption of public cloud computing solutions is the associated risk of using these solutions. 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[20]. 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."

Managing and Optimizing Public Cloud Computing

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 19**.

Table 19: The Importance of Managing Public Cloud Services

	Storage	Compute	Applications
Extremely	7.2%	6.4%	15.6%
Very	16.3%	20.3%	21.7%
Moderately	26.5%	23.8%	29.4%
Slightly	25.3%	25.0%	19.4%
Not at All	24.7%	24.4%	13.9%

[20] http://www-07.ibm.com/businesscenter/au/services/smbservices/include/images/Secure_mobility.pdf

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 20**.

Table 20: The Importance of Optimizing Public Cloud Services

	Storage	Compute	Applications
Extremely	3.9%	5.2%	6.0%
Very	11.8%	14.8%	25.9%
Moderately	26.1%	26.5%	24.7%
Slightly	34.0%	31.0%	27.7%
Not at All	24.2%	22.6%	15.7%

There are many conclusions that can be drawn from the data in **Table 19** and **Table 20**. 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 optimizing and managing SaaS solutions is somewhat to very important to IT organizations. As previously mentioned, unlike the situation with an IaaS provider, it generally will not be possible for an IT organization to place management, security or optimization functionality at a SaaS provider's facility. Hence, other types of solutions are necessary in order to improve the management, security and performance of SaaS-based applications.

Private and Hybrid Cloud Computing

IT organizations that implement the characteristics of a cloud computing solution (e.g., virtualization, automation, centralization) within their own environment are implementing what is usually referred to as ***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 cloud.

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 end-to-end private, hybrid and public cloud computing solutions. Their responses are shown in **Table 21**.

Table 21: Importance of Managing Cloud Solutions

	Private Cloud	Hybrid Cloud	Public Cloud
Extremely	15.6%	10.7%	9.1%
Very	25.1%	25.3%	19.3%
Moderately	24.6%	27.5%	23.3%
Slightly	25.1%	23.6%	29.0%
Not at All	9.5%	12.9%	19.3%

One observation that can be drawn from the data in **Table 21** 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.

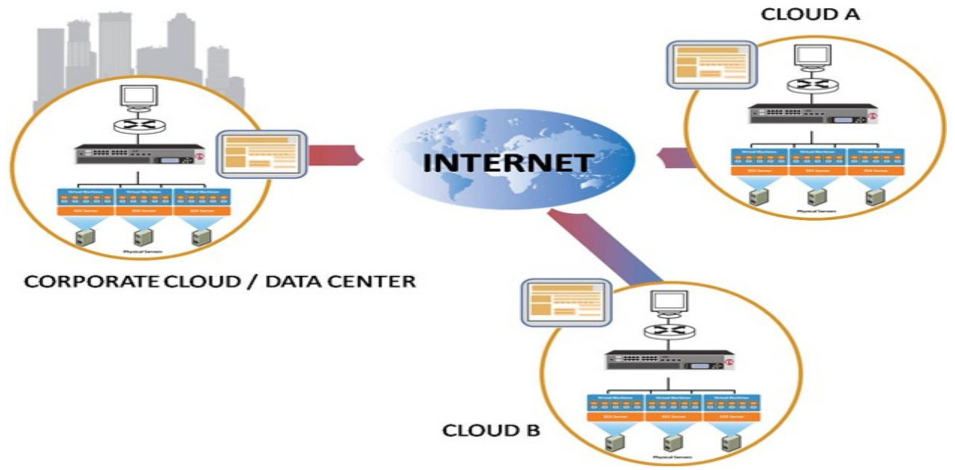
Cloud Balancing

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.

  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[21].

  As beneficial as cloud balancing is, it makes significant demands on the network. This includes the demand for a more effective level of optimization than has historically been required to support branch office to data center communications.

[21] TheRegister.co.uk

Network and Application Optimization

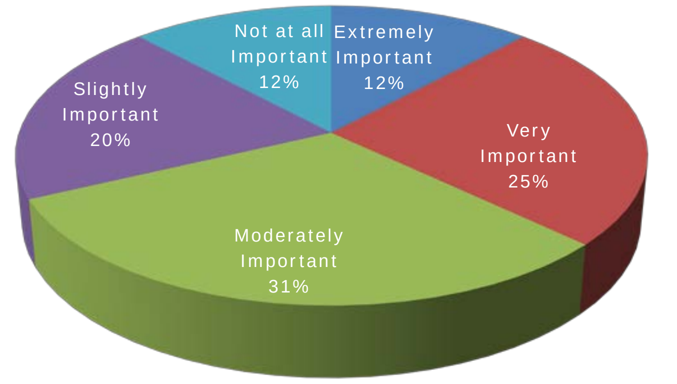
Background

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 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 and applications;
- 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, as mentioned, 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 21.**

Figure 21: Importance of Using Optimization to Reduce Cost



The data in **Figure 21** 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.

The Survey Respondents were asked to indicate their company's approach to optimizing network and application optimization. Their responses are shown in **Table 22**.

Table 22: 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 currently approach implementing optimization functionality is on a case-by-case basis.

The Survey Respondents were given a set of ten viable factors and were asked to indicate the two factors that would likely have the most impact on the evolution of their company's WAN over the next two years. The five factors that were mentioned the most frequently are shown in **Table 23**.

Table 23: Factors Driving WAN Evolution

Factor	Percentage of Respondents
Reduce Cost	34.3%
Improve Application Performance for Business Critical Applications	32.6%
Support video and/or telepresence	20.4%
Support mobile users	18.3%
Provide access to public cloud computing services	17.0%

The data in **Table 23** reflects the responses of all of the 230 IT professionals who responded to the survey. In general, there are only minor differences in the responses of the IT professionals who work for large companies; i.e., 10,000 or more employees. A notable exception to that statement is that whereas the most common factor driving WAN evolution for all companies is reducing cost, which is not the case for large companies. For them it is improving application performance for business critical applications[22].

While historically IT organizations have primarily implemented WAN optimization on a case-by-case basis, that situation is likely to change. One of the key drivers of that change is that as previously explained, the number of business critical applications that the typical business has

[22] Of The Survey Respondents who work for large companies, 46.0% indicated that improving application performance for business critical applications was one of the factors driving WAN evolution and 38.1% indicated that reducing cost was one of the factors.

to support has increased dramatically in the last couple of years. The importance of that driver is enhanced by the fact that, as previously discussed, the most likely impact of poor performance of a business critical application is that the company loses revenue.

The growing importance of improving the performance of a growing number of business critical applications is underscored by the data in **Table 23**. That importance will make it increasingly burdensome to implement optimization functionality on a case-by-case basis. In addition, developments that are discussed later in this document, such as the virtualization of WAN Optimization Controllers and the growing deployment of integrated WOCs, will make it easier for IT organization to implement WOC functionality more broadly.

The deployment of WAN optimization is evolving from being narrowly focused to being broadly focused.

There are two principal categories of network and application optimization products: WAN optimization controllers (WOCs) and Application Delivery Controller (ADCs). There are also services that an IT organization can utilize that provide a wide and growing range of optimization functionality.

The role of a WOC is to mitigate the negative effect that the characteristics of WAN services, such as packet loss, have on application and service performance. The affect of packet loss on TCP has been widely analyzed[23]. 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 22: Factors that Impact Throughput

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

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 23** 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.

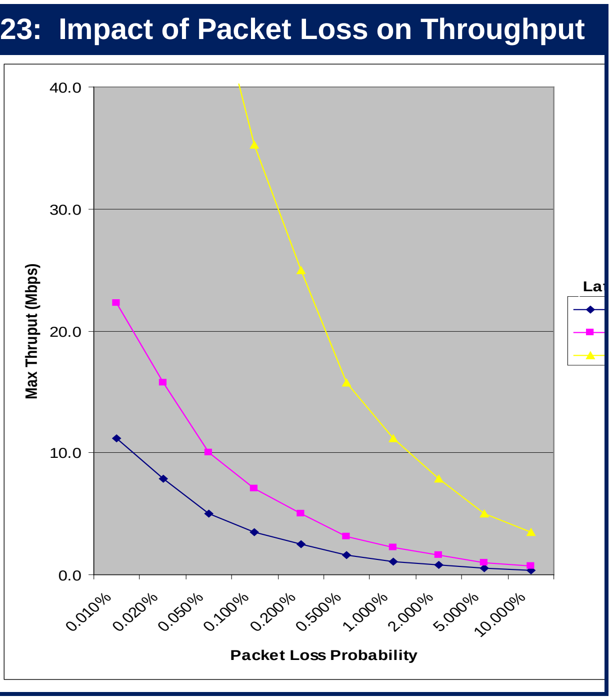
[23] The macroscopic behavior of the TCP congestion avoidance algorithm by Mathis, Semke, Mahdavi & Ott in Computer Communication Review, 27(3), July 1997

One conclusion we can draw from **Figure 23** 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 23: Impact of Packet Loss on Throughput



As described in the next subsection of the handbook, WOCs traditionally focused on accelerating end user traffic between remote branch offices and central data centers. Recently a trend has developed whereby IT organizations use WOCs to accelerate 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.

The Survey Respondents were asked to indicate how important it was to their organization over the next year to get better at optimizing the transfer of storage data between different data centers. Their responses are shown in **Table 24.**

Table 24: The Importance of Optimizing Storage Data

Extremely Important	13%
Very Important	32%
Moderately Important	27%
Slightly Important	15%
Not at all Important	13%

Getting better at optimizing the transfer of storage data between different data centers is one of the most important optimization tasks facing IT organizations.

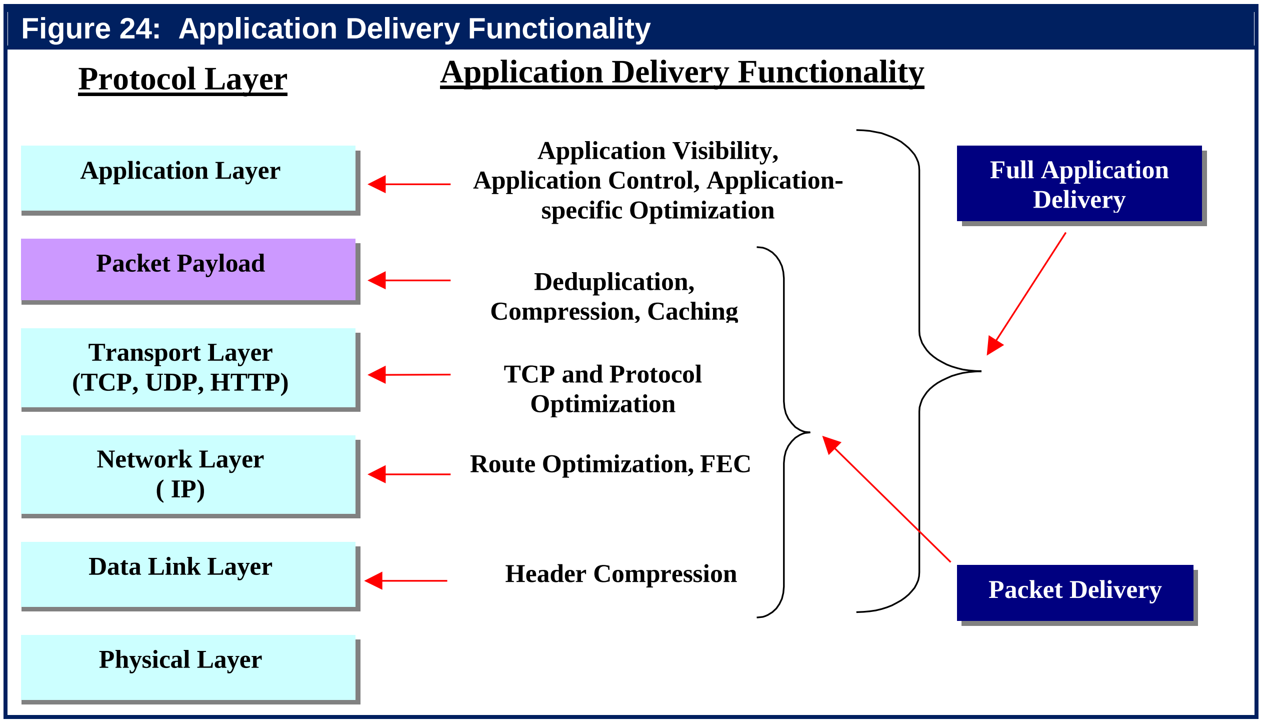
In the vast majority of cases, IT organizations acquire and implement WOCs on a do-it-yourself (DIY) basis. It is also possible for IT organizations to acquire WOC functionality from a managed service provider (MSP). In that scenario, the MSP is responsible for designing, implementing and managing the WOCs. IT organizations have a third option, because as was previously explained in handbook, some Cloud Computing Service Providers (CCSPs) offer network and application optimization as a service. Cloud-based optimization services are discussed in detail in a subsequent section.

IT organizations have a variety of options for how they acquire WOC functionality.

WOCs are often referred to as *symmetric solutions* because they typically require complementary functionality at both ends of the connection. However, as is elaborated upon 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 just 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 24,** many WOCs that are available in the marketplace can recognize the application layer signatures of applications and can leverage optimization techniques to mitigate the application-specific inefficiencies that sometimes occur when these applications communicate over a WAN.

Figure 24: 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 the anticipated traffic volumes by application and the characteristics of the traffic they wish to accelerate. For example, the amount of data reduction 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 production 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 25)** is a variation of the application response time model created by Sevcik and Wetzel[24]. Like all models, the following is only an approximation and it 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 25: Application Response Time Model

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

The WOCs, Cloud-based optimization services and ADCs that are described in this section of the handbook are intended to mitigate the impact of the factors in the preceding equation.

[24] Why SAP Performance Needs Help

WAN Optimization Controllers (WOCs)

WOC Functionality

Table 25 lists some of WAN characteristics that impact application delivery and identifies WAN optimization techniques that a WOC can implement to mitigate the impact of those characteristics.

Table 25: Techniques to Improve Application Performance

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

Below is a description of some of the key techniques used 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*

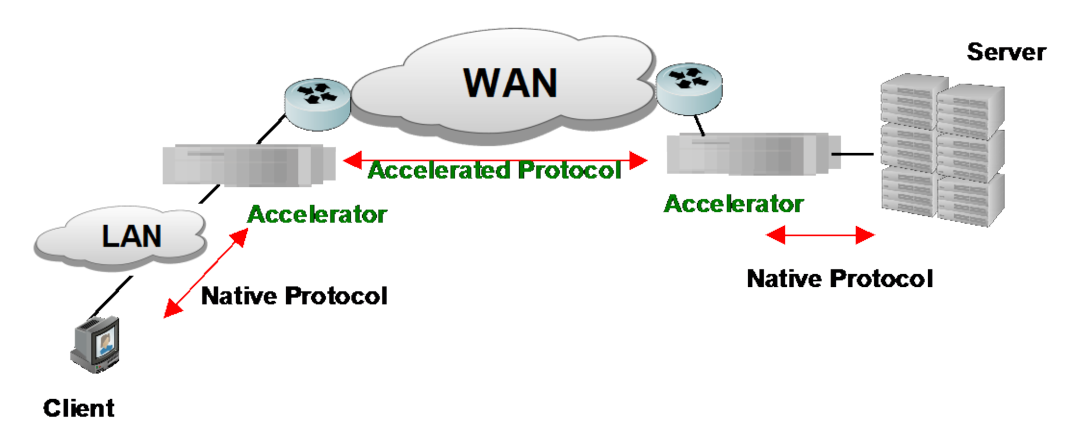
    The same basic LZ data compression algorithms discussed above and proprietary de-duplication algorithms can also be applied to individual blocks of data rather than entire files. This approach results in smaller dynamic dictionaries that can reside in memory rather than on disk. As a result, the processing 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[25].

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

- ***Protocol Acceleration***
  Protocol acceleration refers to a class of techniques that improves application performance by circumventing the shortcomings of various communication protocols. Protocol acceleration is typically based on per-session packet processing by appliances at each end of the WAN link, as shown in **Figure 26**. 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 26: 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

[25] Transmission_Control_Protocol

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

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.

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

- ***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.

The Survey Respondents were told that the phrase *integrated WAN optimization controller (WOC)* refers to running network and application optimization solutions that are integrated within another device such a server or router. They were then asked to indicate whether their IT organization had already implemented, or they expected that they would implement an integrated WOC solution within the next twelve months. Slightly over a third of The Survey Respondents responded *yes* - indicating that they either already had or would. The Survey Respondents who responded *no* were asked to indicate the primary factor that is inhibiting their organization from implementing an integrated WOC. By over a two to one margin, the most frequently mentioned factor was that they had not yet analyzed integrated WOCs.

There is a significant and growing interest on the part of IT organizations to implement integrated WOCs.

The WOC form factor that has garnered the most attention over the last year is the virtual WOC (vWOC). The phrase virtual WOC refers to optimizing the operating system and the WOC software to run in a VM on a virtualized server. One of the factors that are driving the deployment of vWOCs is the growing interest that IT organizations have in using Infrastructure-as-a-Service (IaaS) solutions. IaaS providers typically don't want to install custom hardware such as WOCs for their customers. IT organizations, however, can bypass this reluctance by implementing a vWOC at the IaaS provider's site.

Another factor that is driving the deployment of vWOCs is the proliferation of hypervisors on a variety of types of devices. For example, as previously discussed the majority of IT organizations have virtualized at least some of their data center servers and it is becoming increasingly common to implement disk storage systems that have a storage hypervisor. As a result, in most cases there already are VMs in an enterprise's data center and these VMs can be used to host one or more vWOCs. In a branch office, a suitably placed virtualized server or a router that supports router blades could host a vWOC as well as other virtual appliances forming what is sometimes referred to as a Branch Office Box (BOB). Virtual appliances can therefore support branch office server consolidation strategies by enabling a single device (i.e., server, router) to perform multiple functions typically performed by multiple physical devices.

To understand the interest that IT organizations have in virtual appliances in general, and virtual WOCs in particular, The Survey Respondents were asked, "Has your organization already implemented, or do you expect that you will implement within the next year, any virtual functionality (e.g., WOC, firewall) in one or more of your branch offices." Just under half responded *yes*. The Survey Respondents that responded *yes* were also given a set of possible IT functionality and asked to indicate the virtual functionality that they have already implemented or that they expected to implement within the next year. Their responses are shown in **Table 26**.

Table 26: Implementation of Virtual Functionality

Functionality	Percentage of Respondents
Virtual Firewall	41.7%
Virtual WOC	27.2%
Virtual IDS/IPS	19.4%
Virtual Gateway Manager	19.4%
Virtual Wireless Functionality	17.5%
Virtual Router	15.5%
Other	4.9%

There is broad interest in deploying a wide range of virtual functionality in branch offices.

One advantage of a vWOC is that some vendors of vWOCs provide a version of their product that is completely free and is obtained on a self-service basis. The relative ease of transferring a vWOC also has a number of advantages. 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 hardware-based WOC, a vWOC 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.

Many IT organizations choose to implement a proof-of-concept (POC) trial prior to acquiring WOCs. The purpose of these trials is to enable the IT organization to quantify the performance improvements provided by the WOCs and to understand related issues such as the manageability and transparency of the WOCs. While it is possible to conduct a POC using a hardware-based WOC, it is easier to do so with a vWOC. This follows in part because a vWOC can be downloaded in a matter of minutes, whereas it typically takes a few days to ship a hardware-based WOC.   Whether it is for a POC or to implement a production WOC, the difference between the amount of time it takes to download a vWOC and the time it takes to ship a hardware-based appliance is particularly acute if the WOC is being deployed in a part of the world where it can take weeks if not months to get a hardware-based product through customs.

In addition to the criterion discussed in the next subsection, when considering vWOCs, IT organizations need to realize that there are some significant technical differences in the solutions that are currently available in the marketplace. These differences include the highest speed LAN and WAN links that can be supported as well as which hypervisors are supported; e.g., hypervisors from the leading vendors such as VMware, Citrix and Microsoft as well as proprietary hypervisors from a cloud computing provider such as Amazon. Another key consideration is the ability of the vWOC to fully leverage the multi-core processors being developed by vendors such as Intel and AMD in order to continually scale performance.

In addition to technical considerations, IT organizations need to realize that there are some significant differences in terms of how vendors of vWOCs structure the pricing of their products. One option provided by some vendors is typically referred to as *pay as you go*. This pricing option allows IT organizations to avoid the capital costs that are associated with a perpetual license and to acquire and pay for a vWOC on an annual basis. Another option provided by some vendors is typically referred to as *pay as you grow*. This pricing option provides investment protection because it enables an IT organization to get stared with WAN optimization by implementing vWOCs that have relatively small capacity and are priced accordingly. The IT organization can upgrade to a higher-capacity vWOC when needed and only pay the difference between the price of the vWOC that it already has installed and the price of the vWOC that it wants to install.

WOC Selection Criteria

The recommended criteria for evaluating WAN Optimization Controllers are listed in **Table 27**. 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 27** 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 27: 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 27** 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.

  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 28.**

Table 28: Importance of Optimizing Communications Based Traffic

	VoIP	Traditional Video Traffic	Telepresence
Extremely Important	19.8%	5.4%	3.4%
Very Important	34.5%	22.0%	25.0%
Moderately Important	24.4%	30.1%	29.5%
Slightly Important	15.7%	25.8%	25.6%
Not at all Important	5.6%	16.7%	16.5%

One of the conclusions that can be drawn from the data in **Table 28** is:

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

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 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.

Transferring Storage Data

Background

As previously mentioned, transferring storage data between data centers is an area of growing interest for IT organizations. Transferring storage data between data centers, however, greatly increases the demand for inter-data center bandwidth. While it is possible to just continually add more WAN bandwidth, a more practical solution is to focus on increasing the *Effective Bandwidth* of WAN links. Effective bandwidth is determined by two factors. One factor is the *Bandwidth Efficiency*, which is how completely the WAN link bandwidth can be utilized, even when faced with high WAN latency and a relatively small number of high volume flows. The second factor is the *Bandwidth Multiplication Factor,* which is the gain in link throughput that is derived from implementing techniques such as data compression and de-duplication. The formula for Effective Bandwidth is given by:

Effective BW = BW Efficiency x BW Multiplication Factor x Physical BW

For example, assume that a company has a 1 Gbps link between its two data centers and assume that it implements techniques that allow it to fill 75% of the link on average. Also assume that the company implements optimization techniques on both ends of the link that on average provide a 5:1 improvement in link utilization. Then:

Average Effective BW = 0.75 x 5 x 1 Gbps = 3.75 Gbps

The Challenges of Moving Workflows Among Cloud Data Centers

A majority of IT organizations see tremendous value in being able to move workflows between and among data centers. However, as is described in this section, one of the key challenges that currently limits the movement of workloads is the sheer volume of data that must be moved. In some cases, gigabytes or even terabytes must be moved in a very short amount of time.

- ***Virtual Machine Migration***
  With the previously discussed adoption of varying forms of cloud computing, the migration of VMs between and among disparate data centers is gaining ever-increasing importance. The live migration of production VMs between physical servers can allow for the automated optimization of workloads across resource pools spanning multiple data centers. VM migration also makes it possible to transfer VMs away from physical servers that are experiencing maintenance procedures, faults, or performance issues. During VM migration, the machine image, which is typically ~10+ GB per VM, the active memory and the execution state of a virtual machine are transmitted over a high speed network from one physical server to another. As this transfer is being made, the source VM continues to run, and any changes it makes are reflected to the destination. When the source and destination VM images converge, the source VM is eliminated, and the replica takes its place as the active VM. The VM in its new location needs to have access to its virtual disk (vDisk). For inter-data center VM migrations, this means one of three things:

  - The SAN or other shared storage system must be extended to the new site;

  - The virtual machine disk space must be migrated to the new data center;

- The vDisk must be replicated between the two sites.

In the case of VMotion, VMware recommends that the network connecting the physical servers involved in a VMotion live transfer to have at least 622 Mbps of bandwidth and no more than 5 ms of end-to-end latency[26] [27]. Another requirement is that the source and destination physical servers need to be on the same Layer 2 virtual LAN (VLAN). For inter-data center VM migration, this means that the Layer 2 network must be extended over the WAN.

MPLS/VPLS offers one approach to bridging remote data center LANs together over a Layer 3 network. Another alternative is to tunnel Layer 2 traffic through a public or private IP network using Generic Router Encapsulation (GRE). A more general approach that addresses some of the major limitations of live migration of VMs across a data center network is the IETF draft Virtual eXtensible LAN (VXLAN). In addition to allowing VMs to migrate transparently across Layer 3 boundaries, VXLAN provides support for virtual networking at Layer 3, circumventing the 802.1Q limitation of 4,094 VLANs, which is proving to be inadequate for VM-intensive enterprise data centers and multi-tenant cloud data centers.

VXLAN is a scheme to create a Layer 2 overlay on a Layer 3 network via encapsulation. The VXLAN segment is a Layer 3 construct that replaces the VLAN as the mechanism that segments the network for VMs. Therefore, a VM can only communicate or migrate within a VXLAN segment. The VXLAN segment has a 24 bit VXLAN Network identifier, which supports up to 16 million VXLAN segments within an administrative domain. VXLAN is transparent to the VM, which still communicates using MAC addresses. The VXLAN encapsulation and other Layer 3 functions are performed by the hypervisor virtual switch or by the Edge Virtual Bridging function within a physical switch or possibly by a centralized server, The encapsulation allows Layer 2 communications with any end points that are within the same VXLAN segment even if these end points are in a different IP subnet, allowing live migrations to transcend Layer 3 boundaries.

NVGRE is a competing virtual networking proposal before the IETF. It uses GRE as a method to tunnel Layer 2 packets across an IP fabric, and uses 24 bits of the GRE key as a logical network identifier or discriminator, analogous to a VXLAN segment.

The development of schemes such as VXLAN and NVGRE address many of the networking challenges that are associated with migrating VMs between and among data centers. The primary networking challenge that remains is ensuring that the LAN-extension over the WAN is capable of high bandwidth and low latencies. Schemes such as VXLAN and NVGRE do, however, create some additional challenges because they place an extra processing burden on appliances such as WAN Optimization Controllers (WOCs) that are in the network path between data centers. In instances where the WOCs are software-based, the extra processing needed for additional packet headers can reduce throughput and add latency that cuts into the 5ms end-to-end delay budget.

[26] http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns836/white_paper_c11-557822.pdf
[27] It is expected that these limitations will be relaxed somewhat by the end of 2012.

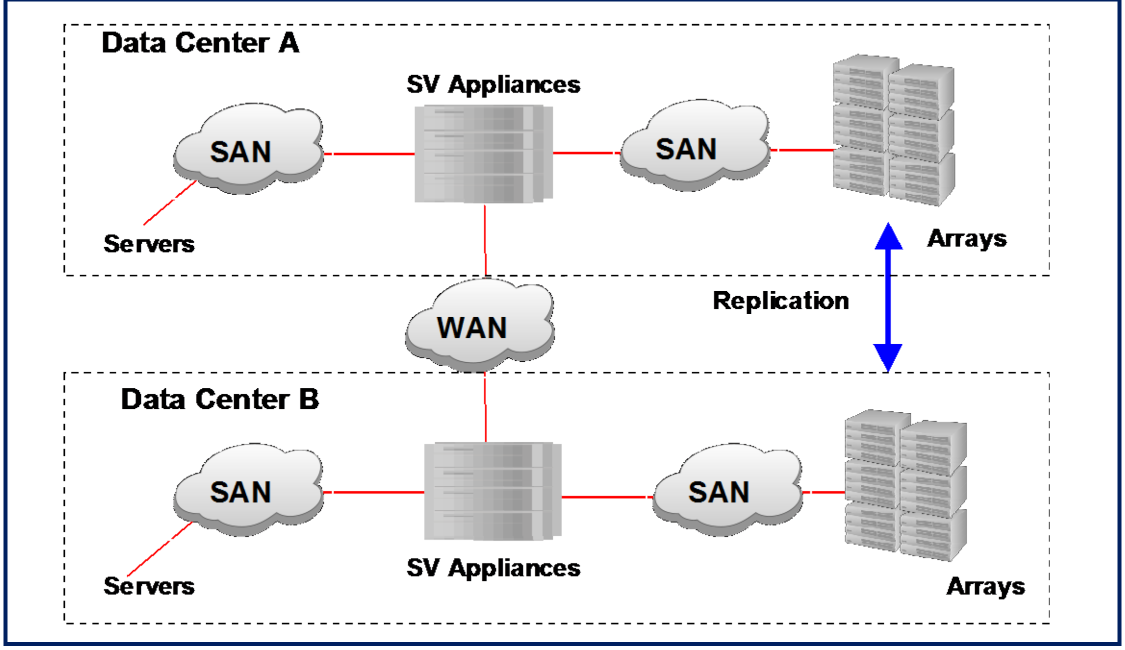
- ***Maintaining VM Access to its vDisk***
  When a VM is migrated, it must retain access to its vDisk. For VM migration within a data center, a SAN or NAS system provides a shared storage solution that allows the VM to access its vDisk both before and after migration. When a VM is migrated to a remote data center, maintaining access to the vDisk involves some form of data mobility across the WAN. The technologies that are available to provide that mobility are: SAN Extension, Live Storage Migration by the hypervisor, and Storage Replication.

- ***SAN Extension***
  If the vDisk stays in its original location, the SAN that it resides on must be extended to the destination data center. Technologies that are available for SAN extension include SONET, dense wave division multiplexing (DWDM) and Fibre Channel over IP (FCIP). Where there is a significant amount of SAN traffic over the WAN, the only transmission technologies with the required multi-gigabit bandwidth are DWDM or 10/40 GbE over fiber. However, the cost of multi-gigabit WAN connections is likely to prove to be prohibitive for most IT departments. An additional problem is that application performance would suffer because of high latency due to propagation delay over the WAN.

- ***Live Storage Migration***
  Storage migration (e.g., VMware Storage VMotion) can be performed by the server's hypervisor, which relocates the virtual machine disk files from one shared storage location to another shared storage location. The transfer can be completed with zero downtime, with continuous service availability, and complete transaction integrity. VMotion works by using a bulk copy utility in conjunction with synchronization functionality, such as I/O Mirroring, which mirrors all new writes from the source to the destination as the bulk copying proceeds. Once the two copies are identical, the operational VM can be migrated and directed to use the destination copy of the virtual disk. The challenge with this type of storage migration is that the VM cannot be moved until the vDisk copy is completed. Since the vDisk may contain many gigabytes or terabytes of data, the VM migration is delayed by the bulk copy time, which is inversely proportional to the effective WAN bandwidth between the two sites. WAN bandwidth of 1 Gbps is typically the minimum amount that is recommended in order to support storage migration. Even with this large amount of WAN bandwidth, delays of many minutes or even hours can occur. Delays of this magnitude can impede the ability of organizations to implement highly beneficial functionality such as Cloud Balancing.

- ***Storage Replication***
  One way to migrate VMs without the delays associated with storage migration's bulk copy operation is to identify the VMs that are likely to need migration and to replicate the vDisks of those VMs at the remote site in anticipation of an eventual VM migration. **Figure 27** shows in-line server virtualization (SV) appliances performing storage replication over the WAN. Note that storage replication can also be performed by utilities included with some storage devices. In addition to supporting VM migration, storage replication facilitates recovery from data center failures or catastrophic events.

Figure 27: Storage Replication via Storage Virtualization (SV) Appliances



Synchronous replication guarantees zero data loss by means of an atomic write operation, in which the write is not considered complete until acknowledged by both local and remote storage. Most applications wait for a write transaction to complete before proceeding with further processing, so a remote write causes additional delay to the application of twice the WAN round trip time (RTT). In practice, the RTT delay has the affect of limiting the distance over which synchronous replication can be performed to approximately 100 km. It is generally recommended that there should be a minimum of 1 Gbps of WAN bandwidth in order to support synchronous replication. Synchronous replication between sites allows the data to reside simultaneously at both locations and to be actively accessed by VMs at both sites, which is commonly referred to as active-active storage.

Asynchronous replication does not guarantee zero data loss and it is not as sensitive to latency as is synchronous replication. With asynchronous replication, the write is considered complete once acknowledged by the local storage array. Application performance is not affected because the server does not wait until the write is replicated on the remote storage array. There is no distance limitation and typical asynchronous replication applications can span thousands of kilometers or more. As with synchronous replication, at least 1 Gbps of WAN bandwidth is recommended.

The primary networking challenge of storage migration and replication is to maximize the effective bandwidth between cloud data centers without incurring the excessive costs of very high bandwidth WAN connectivity. This approach will minimize the delays associated with bulk storage transfers and replications, optimizing the dynamic transfer of workloads between cloud sites.

Resolving the Challenges of Workload Migration

Many of the previously described challenges can be at least partially addressed by the deployment of appropriate WOC functionality at each data center. Due to the special characteristics of VM migration, storage migration, and storage replication, the requirements for data center-to-data center WAN optimization differ significantly from those for WOCs designed for accelerating end user traffic between branch offices and a central data center. In order to optimize workload migration an inter-data center WAN Optimization solution should have the following functionality:

- ***High Throughput***
  The inter-data center WAN Optimization solution should be capable of saturating a multi-gigabit WAN link and hence provide a bandwidth efficiency of 1.0, even if the number of current flows between data centers is quite small. For example if replication of a large storage array is the only active flow, the device should ideally have the processing power and TCP protocol optimization functionality needed to fill a multi-gigabit pipe with traffic, eliminating any significant amount of stranded bandwidth. In addition to improving the utilization of expensive high bandwidth WAN links, high throughput improves the efficiency of operations such as storage replication, backup, and VM migration. Ideally, high throughput can be achieved without the high cost and complexity of a number of load balanced WOCs at each data center.

- ***Transport Optimization***
  The congestion control mechanism for TCP needs to be very aggressive in its control of window sizes in order to achieve high bandwidth efficiency and consume all of the bandwidth allocated to each type of traffic flow. In addition, the WAN Optimization solution needs dynamically tuned, and potentially very large buffers, in order to shield the end systems at each data center from the effects of WAN propagation latency and any WAN packet loss.

- ***Low Latency***
  As previously described, a number of inter-data center operations are improved if the inter-data center WAN Optimization device has very low internal (processing) latency. For example, for synchronous storage replication any significant amount of WOC device latency reduces the inter-data center distance over which synchronous replication is feasible. WAN Optimization device internal latency can also be a significant factor affecting the inter-data center distances over which VM migration can be reliably performed. In addition, operations such as virtual machine migration across data centers have strict latency requirements, so high levels of latency for WAN optimization processing would not be workable.

- ***Maximal Data Reduction***
  Data Reduction based on de-duplication and compression decreases WAN bandwidth consumption and reduces the time-to-completion of inter-data center tasks, such as storage replication, backups, and large file transfers. Data reduction essentially provides a bandwidth multiplication factor that can dramatically increase the effective bandwidth of the WAN link. Storage replication and backup applications typically send only those blocks of data that have changed since the previous transfer. In these cases, good WOC de-duplication ratios depend on identifying patterns that are far smaller than the typical data block addressed by disk systems that are typically 4 KB. Ideally, for maximal

data reduction, the WOC de-duplication implementation should be able to find repetitions all the way down to sub-10 byte packet segments both within and across individual streams or flows. The efficiency of the de-duplication process should be independent of throughput, ideally scaling to speeds in the range of 10 Gbps. This means that the de-duplication engine has to have the processing power to look for short duplicate strings even at very high data rates. Data compression should ideally also occur after de-duplication has occurred in order to make the data reduction function more efficient.

- ***QoS and Traffic Management***
  Inter-data center WAN links typically carry a number of different traffic types with varying requirements for low latency and bandwidth. Therefore, the WAN Optimization system must have a hardware-based QoS/traffic management system that can classify and prioritize traffic at multi-gigabit line rates and allocate bandwidth in accordance with configured QoS policies. Leveraging these policies, the appropriate acceleration techniques and priorities can be applied to business critical traffic. Traffic that does not need acceleration/optimization can be classified as such and allowed to bypass the WOC functionality.

- ***High Availability***
  Given the business critical nature of accelerating inter-data center traffic, WAN Optimization systems should be capable of high availability deployments. In addition to providing a number of internal high availability features, such as redundant power supplies, the WAN Optimization system should support high availability network designs based on in-line or out-of-path redundant configurations.

Trends in WOC 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:

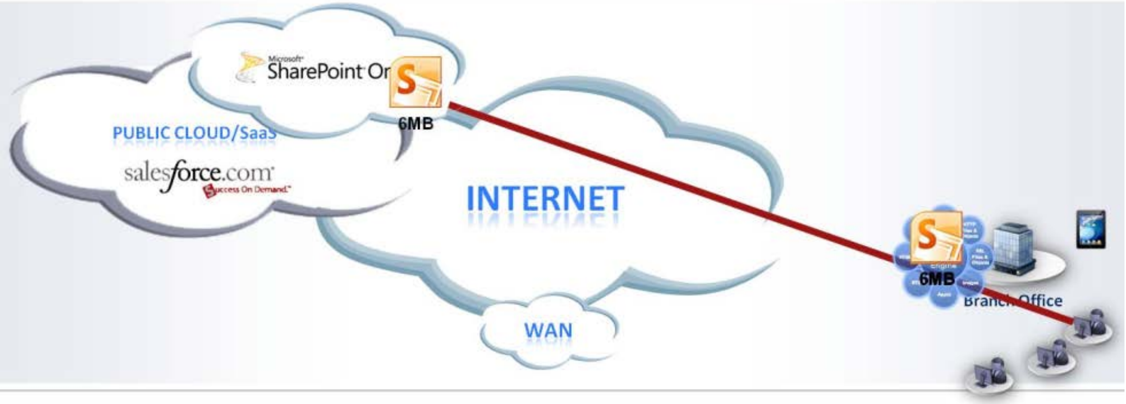
- ***Cloud Optimized WOCs***
  This is a purpose-built virtual WOC (vWOC) appliance that was designed with the goal of it being deployed in public and/or hybrid cloud environments. One key feature of this class of device is compatibility with cloud virtualization environments including the relevant hypervisor(s). 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.

- ***Cloud-Based Optimization Services***
  In the current environment, there are few Cloud-based optimization services. It is reasonable to expect that the overall use of these services will increase and that the number of available services will also increase.

- ***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 28**, 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 28: 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. The impact that the adoption of IPv6 has on ADCs will be discussed in detail in the next section of the handbook.

Cloud-Based Optimization Solutions

Background

The preceding section of this handbook discussed a new class of solutions that has recently begun to be offered by CCSPs. These are solutions that have historically been provided by the IT organization itself and include network and application optimization, VoIP, Unified Communications, security, network management and virtualized desktops. As pointed out in the preceding section, roughly a quarter of The Survey Respondents indicated that within a year, their organization would either adopt, or would likely adopt, a network and application optimization solution provided by a CCSP.

The preceding section also mentioned some of the factors that are both driving and inhibiting the adoption of public cloud services in general. The primary drivers are:

- Lower cost
- Reduce time to deploy new functionality
- Provide functionality that the IT organization could not provide itself

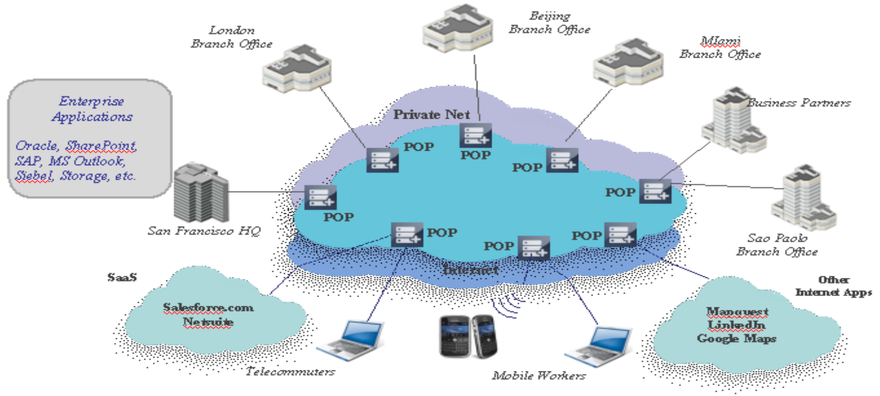
The primary inhibitors are:

- Performance
- Management
- Security

Use Cases

As noted, The Survey Respondents demonstrated significant interest in a network and application optimization solution, such as the one depicted in **Figure 29** that is provided by a CCSP.

Figure 29: WAN and Application Optimization Solution



In **Figure 29**, a variety of types of users (e.g., mobile users, branch office users) access WAN optimization functionality at the service provider's points of presence (POPs). Ideally these POPs are inter-connected by a dedicated, secure and highly available network. To be effective, the solution must have enough POPs so that there is a POP in close proximity to the users. In addition, the solution should support a wide variety of WAN access services. Additional evaluation criteria are described below.

There are at least three distinct use cases for the type of solution shown in **Figure 29**. One such use case is that this type of solution can be leveraged to solve the type of optimization challenges that an IT organization would normally solve by deploying WOCs; e.g., optimizing communications between branch office users and applications in a corporate data center or optimizing data center to data center communications. In this case, the factors that would cause an IT organization to use such a solution are the same factors that drive the use of any public cloud based services; e.g., cost savings, reduce the time it takes to deploy new functionality and provide functionality that the IT organization could not provide itself

The second use case is the ongoing requirement that IT organizations have to support mobile workers. Some IT organizations will resolve the performance challenges associated with supporting mobile users by loading optimization software onto all of the relevant mobile devices. There are two primary limitations of that approach. One limitation is that it can be very cumbersome. Consider the case in which a company has 10,000 mobile employees and each one uses a laptop, a smartphone and a tablet. Implementing and managing optimization software onto those 30,000 devices is very complex from an operational perspective. In addition, the typical smartphone and tablet doesn't support a very powerful processor. Hence, another limitation is that it is highly likely that network and application optimization software running on these devices would not be very effective.

The third use case for utilizing a solution such as the one shown in **Figure 29** is the expanding requirement that IT organizations have to support access to public cloud services. As previously mentioned, in some instances it is possible for an IT organization to host a soft WOC at an IaaS provider's site. However, that is generally not possible at a SaaS provider's site. In those instances in which it is not possible to host a soft WOC at the CCSP's site, a Cloud-based optimization solution can improve the users access to cloud services by providing to the users the type of functionality typically provided in a WOC.

Evaluating Solutions

The use of Cloud-based network and application optimization solution is just the latest example of IT organizations using a third party to provide needed functionality; a.k.a., out-tasking. Hence, IT organizations that are evaluating these solutions should evaluate these solutions the same way that they would evaluate any form of out-tasking. For example, IT organizations that are evaluating these solutions need to understand whether or not these solutions meet the requirements and whether or not they meet the requirements in a more effective manner than an internally provided solution would.

Evaluating whether or not a given solution provides the required functionality is standard operating procedure for IT organizations. In addition, there is not much difference in terms of how an IT organization would evaluate the functionality provided by a premise based WOC-based solution vs. how it would evaluate the functionality provided by a Cloud-based solution. What is different about evaluating the later class of solution stems from the fact that they are cloud-based. As such, IT organizations need to closely look at how well the service provider has dealt with the impediments to the use of public cloud computing solutions that were also previously discussed; e.g., the performance of the solution. The performance of a Cloud-based optimization solution is similar to the performance of a standard WOC-based solution – it will vary somewhat based on the requirements of each IT organization. Hence, as was the case with WOC-based solutions, the best way to understand the performance gains that result from using a Cloud-based optimization solution is to test that solution with production traffic.

However, just as important as whether or not the Cloud-based optimization solution mitigates the issues that IT organizations have with public cloud based solutions is whether or not the solution actually provides the benefits (e.g., cost savings) that drive IT organizations to use public cloud computing solutions. While it can be a little tricky to compare the usage sensitive pricing of a Cloud-based optimization solution with the fully loaded cost of a premise based WOC solution[28], the cost information that the IT organization receives from the solution provider should enable the IT organization to do the requisite analysis. The key financial advantages of a Cloud-based solution are that it enables IT organizations to avoid the CAPEX costs that are associated with a typical WOC based solution and also enables IT organizations to migrate away from expensive WAN services such as MPLS.

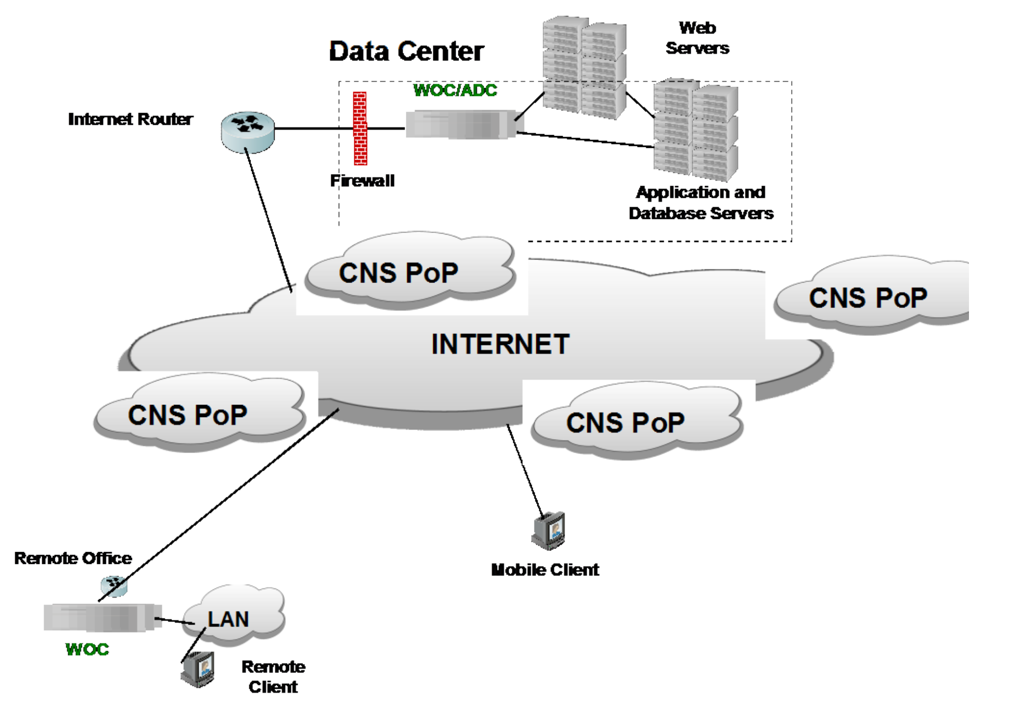
[28] The tricky part is determining the totality of the labor costs associated with the premise based solution.

The Optimization of Internet Traffic

As previously described, WOCs 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 such as 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. Throughout the handbook, a service that optimizes Internet traffic will be referred to as an Internet Optimization Service (IOS).

An IOS 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 30**, all client requests to the application's origin server in the data center are redirected via DNS to a server in a nearby point of presence (PoP) that is part of the IOS. This edge server then optimizes the traffic flow to the IOS server closest to the data center's origin server.

Figure 30: The Structure of an IOS



The servers at the IOS provider's PoPs perform a variety of optimization functions. Some of the functions provided by the IOS include:

- ***Route Optimization***
  Route optimization is a technique for circumventing the previously discussed 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 IOS

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 by a pair of IOS servers in the ingress and egress PoPs. The edge IOS servers can also apply asymmetrical TCP optimization to the transport between the subscriber sites and the PoPs that are associated with the IOS. 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 IOS servers. By controlling both ends of the long-haul Internet connection with symmetric IOS servers, a high performance transport protocol can eliminate most of the previously discussed inefficiencies associated with TCP, including the three-way handshake for connection setup and teardown, the slow start algorithm and the re-transmission timer issues. For subscriber traffic flowing between IOS 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 IOS server with the cache performing intelligent pre-fetching from the origin. With pre-fetching, the IOS 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 IOS 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 IOS servers and through persistent, replicated in-cloud storage facilities. Offloading content and storage to the Internet 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. IOS 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.

Visibility and Security

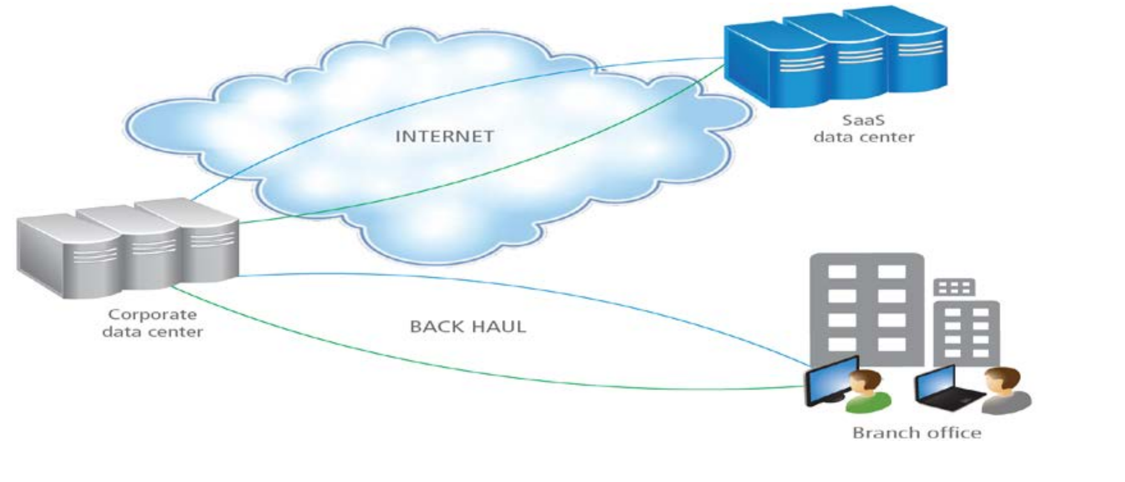
Intelligence within the IOS servers can also be leveraged to provide extensive monitoring, configuration control and SLA monitoring of a subscriber's application with performance metrics, analysis, and alerts made visible to the subscriber via a Web portal.

In many cases, in addition to providing optimization of Internet traffic, an IOS can also provide security functionality. This will be discussed in more detail in the next section of the handbook.

Hybrid WAN Optimization

As shown in **Figure 31**, the traditional approach to providing Internet access to branch office employees has been to backhaul that Internet traffic on the organization's enterprise network (e.g., their MPLS network) to a central site where the traffic was handed off to the Internet. The advantage of this approach is that it enables IT organizations to exert more control over their Internet traffic and it simplifies management in part because it centralizes the complexity of implementing and managing security policy. One disadvantage of this approach is that it results in extra traffic transiting the enterprise's WAN, which adds to the cost of the WAN. Another disadvantage of this approach is that it usually adds additional delay to the Internet traffic.

Figure 31: Backhauled Internet Traffic



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 29.**

Table 29: 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%

Although the vast majority of IT organizations currently have a centralized approach to Internet access, IT organizations are continually adopting a more decentralized approach.

Because backhauling Internet traffic adds delay, one of the disadvantages of this approach to providing Internet access is degraded performance. For example, in the scenario depicted in **Figure 31** (Backhauled Internet Traffic), the delay between users in a branch office and the SaaS application is the sum of the delay in the enterprise WAN plus the delay in the Internet. In order to improve performance, an IT organization might use WOCs to optimize the performance of the traffic as it flows from the branch office to the central site over their enterprise WAN. However, once the traffic is handed off to the Internet, the traffic is not optimized and the organization gets little value out of optimizing the traffic as it flows over just the enterprise WAN.

One way to minimize the degradation in application performance is to not backhaul the traffic but hand it off locally to the Internet. For this approach to be successful, IT organizations must be able to find another way to implement the security and control that it has when it backhauls traffic. This can be done either by putting appropriate functionality at the branch office, acquiring the appropriate functionality from a CCSP or some combination of those approaches.

Another way to minimize the degradation in application performance is based on the previous discussion of an IOS. One way that an IOS would add value is if the organization used the IOS to carry traffic directly from the branch office to the SaaS provider. In this case, in addition to providing optimization functionality, the IT organization is relying on the security functionality provided by the IOS to compensate for the security functionality that was previously provided in the corporate data center. Another way that an IOS would add value is if the solution enabled IT organizations to keep its current approach to backhauling traffic. However, in this case, the IT organization would use WOCs to optimize the performance of the Internet traffic as it transits the enterprise WAN. This WOC-based solution would then have to be integrated with the IOS 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 WAN optimization solution.

Application Delivery Controllers (ADCs)

Background

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[29]. 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.

ADC Functionality

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

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

- ***SSL Offload***
  One of the primary new roles played by an ADC is to offload CPU-intensive tasks from data center servers. A prime example of this is SSL offload, where the ADC terminates the SSL session by assuming the role of an SSL Proxy for the servers. SSL offload can provide a significant increase in the performance of secure intranet or Internet Web

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

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

ADC evaluation criteria are listed in **Table 30**. **Figure 24** is intended to describe standard ADC functionality. Subsequent subsections describe in detail how to evaluate an ADCs ability to enable a migration to IPv6 and how to characterize the varying ways to virtualize an ADC. 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 30: 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.

- ***<u>Features</u>***
  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.

- ***<u>Performance</u>***
  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.

IPv6 and ADCs

Background

June 6th, 2012 was World IPv6 Launch day (http://www.worldipv6launch.org/) and IPv6 is now a permanent part of the Internet. While it won't happen for several years, IPv6 will replace IPv4 and the entire Internet will be IPv6 only. Gartner, Inc. estimates that 17% of the global Internet users and 28% of new Internet connections will use IPv6 by 2015.[30] This is creating an imperative for enterprises to develop an IPv6 strategy and migration plan. A key component of that strategy and migration plan is ensuring that devices such as firewalls and ADCs that you are implementing today, fully support IPv6.

Developing a strategy for IPv6 involves examining how your organization uses the Internet and identifying what will change as IPv6 usage grows. While developing an IPv6 strategy, it can be safely assumed that your customers, business partners and suppliers will start to run IPv6. It is also a good assumption that your mobile workers will use IPv6 addresses in the future when accessing corporate applications via the Internet. This creates a challenge for businesses and other organizations to establish an IPv6 presence for application accessed by customers, business partners, suppliers and employees with IPv6 devices and networks.

IPv6 was created as an improvement over IPv4 for addressing, efficiency, security, simplicity and Quality of Service (QoS). IPv6's addressing scheme is the centerpiece of its achievement and the main driver behind IPv6 implementation. IPv4 uses 32 bits for IP addresses which allows for a maximum of 4 billion addresses. While this is a large number, rapid increases in Internet usage and growth in Internet devices per person have depleted almost all of the available IPv4 addresses. Network Address Translation (NAT) and use of private IP addresses (IETF RFC 1918) have raised the efficiency of IPv4 addressing, but have also limited Internet functionality. IPv6 addresses quadruples the number of bits used in the network addressing to 128 bits which provides 4.8 x 10^{28} addresses (5 followed by 28 zeros) for each person on the Earth today. IPv6 eliminates the need to use NAT for IP addresses preservation.

[30] http://www.verisigninc.com/assets/preparing-for-ipv6.pdf

NAT will likely continue to be used for privacy or security, but it is not needed for address conservation in IPv6.

IPv6 has the potential to affect almost everything used for application and service delivery. The most obvious change occurs on networking devices including routers, LAN switches, firewalls and Application Delivery Controllers/Load Balancers. IPv6 also affects servers and end user devices that connect to the network. Applications, platforms, DNS servers, service provision and orchestration systems, logging, systems management, monitoring systems, service support systems (e.g. incident management), network and application security systems are also affected.

While complete migration to IPv6 is a daunting task, it is not as difficult as it first seems. IPv6 is not “backwards compatible” with IPv4, but there are a number of standards and technologies that help with IPv6 migration. These include:

- **Tunneling** – Transporting IPv6 traffic in IPv4 areas and vice versa.

- **Network Address Translation** – Translating between IPv4 and IPv6 addresses, including DNS support.

- **Dual Stack** – Both IPv4 and IPv6 packets are processed by devices simultaneously.

The IETF recommends Dual Stack as the best approach to IPv6 migration, but different situations and individual requirements will dictate a variety of migration paths. For most organizations, they will use a combination of IPv6 migration technologies - usually in concert with their service providers and suppliers.

Enabling Standards and Technologies

IPv6/IPv4 Tunneling

Tunneling permits the Internet Service Providers (ISPs) flexibility when implementing IPv6 by carrying the traffic over their existing IPv4 network or vice versa. There are various approaches to IPv6 tunneling, they may include:

- **6rd**– Mostly used during initial IPv6 deployment, this protocol allows IPv6 to be transmitted over an IPv4 network without having to configure explicit tunnels. 6rd or “IPv6 **R**apid **D**eployment”, is a modification to 6to4 that allows it to be deployed within a single ISP.

- **6in4**–Tunnels are usually manually created and use minimal packet overhead (20 bytes) to minimize packet fragmentation on IPv4 Networks.

- **Teredo**–Encapsulates IPv6 traffic in IPv4 UDP packets for tunneling. Use of UDP allows support of IPv4 Network Address Translation (NAT44 or NAT444) when carrying the IPv6 traffic. This is similar to encapsulating IPSec traffic in UDP to support NAT devices for remote access VPNs.

- **Dual-stack Lite (DS-Lite) –** Encapsulates IPv4 traffic over an IPv6 only network allowing retirement of older IPv4 equipment while still allowing IPv4 only devices a connection to the IPv4 Internet.

6rd and DS-Lite will mostly be used by ISPs and not corporate IT groups, but it is important to understand which IPv6 tunneling technologies are supported when creating your IPv6 migration strategy.

Network Address Translation (NAT)

Network Address Translation (NAT) has been used for several decades with IPv4 networks to effectively extend the amount of available IPv4 addresses. Each IP address can have up to 65,535 connections or ports, but it is rare for this limit to be reached – especially for devices used by end users. In reality, the number of active connections is usually under 100 for end user devices, however behind a home CPE device it may be from 200-500 with multiple devices connected. In addition, connections are typically initiated by the end user device, rather than from the application or server to the end user device. Taking advantage of end user initiated connections with a low connection count, it is quite common to multiplex multiple end user devices' IP addresses together into a few IP addresses and increase the number of connections per IP address. This is accomplished by translating the end user IP address and port number to one of a few IP addresses in each outgoing and returning packet. This is usually accomplished using a network firewall or ADC and this hides the original end user's IP address from the Internet. Since the end user's original IP address is hidden from the public Internet, end user IP addresses can be duplicated across different networks with no adverse impact. Multiple networks behind firewalls can use the same IP subnets or "private IP subnets", as defined in IETF RFC 1918. NAT has been used extensively in IPv4 to preserve the IPv4 address space and since it translates both IPv4 address and the TCP/UDP port numbers is more correctly called Network Address and Port Translation (NAPT). When NAT is used to translate an IPv4 address to an IPv4 address, it is referred to as NAT44 or NAT444 if these translations are done twice.

One of the fundamental problems with NAT is that it breaks end-to-end network connectivity, which is a problem for protocols such as FTP, IPsec, SIP, Peer-to-Peer (P2P) and many more. One way to deal with this is to implement an Application Layer Gateway (ALG), which can manipulate the IP addresses in the Layer 7 portion of the IP packet to ensure the applications still work.

In addition to effectively extending the use of the limited IPv4 address space, NAT is an important technology for migrating to IPv6. NAT for IPv6 has gone through several revisions and today, a single standard providing both stateless (RFC 5145) and stateful (RFC 6146) bidirectional translation between IPv6 and IPv4 addresses. This allows IPv6 only devices and servers to reach IPv4 devices and servers. Three earlier protocols in IPv6, Network Address Translation/Protocol Translation (NAT-PT), Network Address Port Translation/Protocol Translation (NAPT-PT) and Stateless IP/ICMP Translation (SIIT) have been replaced by NAT64. Stateless NAT64 allows translation between IPv6 and IPv4 addresses without needing to keep track of active connections, while stateful NAT64 uses an active connection table. Stateless NAT64 has the ability to work when asymmetric routing or multiple paths occur, but also consumes more precious IPv4 addresses in the process. Stateful NAT64 consumes a minimum amount of IPv4 addresses, but requires more resources and a consistent network path.

Network addresses are very user unfriendly and the Domain Naming System (DNS) translates between easy to remember names like www.ashtonmetzler.com and its IPv4 addresses of 67.63.55.3. IPv6 has the same need for translating friendly names to IPv6 and IPv4 addresses and this is accomplished with DNS64. When a DNS64 server is asked to provide the IPv6 address and only an IPv4 address exists, it responds with a virtual IPv6 address (an "AAAA" record in DNS terms) that works together with NAT64 to access the IPv4 address. DNS64 in conjunction with NAT64 provides name level transparency for IPv4 only servers and helps provide access to the IPv4 addresses from IPv6 addresses.

Carrier Grade NAT (CGN)

Carrier Grade NAT (CGN) is also known as Large Scale NAT (LSN) as it is not just a solution for carriers. Many vendors provide basic NAT technology; it is necessary for a load-balancer feature for example, but what some vendors define as CGNAT technology as it relates to the true CGN standard is often lacking. The premise that legacy NAT at increased volumes is carrier-grade, and therefore equals Carrier Grade NAT, is incorrect. Service providers and enterprises wanting to replace aging NAT devices, are increasingly requiring true CGN as a solution to IPv4 exhaustion due to the standardized, non-propriety implementation and also the advanced features not in standard NAT. The true IETF reference [draft-nishitani-cgn-05*]* clearly differentiates from legacy NAT with many more features such as:

- Paired IP Behavior
- Port Limiting
- End-point Independent Mapping and Filtering (full-cone NAT)
- Hairpinning

True Carrier-Grade NAT involves much more than basic IP/port translation. Because there are so many subscribers, with multiple end-devices (smart phones, tablets, and laptops for example), it is imperative for a network administrator to be able to limit the amount of ports that can be used by a single subscriber. This is in order to guarantee connectivity (available ports) for other subscribers. DDoS attacks are notorious for exhausting the available ports. If just a few subscribers are (usually unknowingly) participating in a DDoS attack, the port allocations on the NAT gateway increases exponentially, quickly cutting off Internet connectivity for other subscribers.

The CGN standard also includes a technology called "Hairpinning". This technology allows devices that are on the "inside" part of the CGN gateway to communicate with each other, using their peers' "outside" addresses. This behavior is seen in applications such as SIP for phone calls, or online gaming networks, or P2P applications such as BitTorrent.

Another essential element to consider when implementing CGN is the logging infrastructure. Because the IP addresses used inside the carrier network are not visible to the outside world, it is necessary to track what subscriber is using an IP/port combination at any given time. This is important not only for troubleshooting, but also it is mandated by local governments and by law enforcement agencies. With so many concurrent connections handled by a CGN gateway, the logging feature itself and the logging infrastructure require a lot of resources. To reduce and simplify logging, there are smart solutions available such as port batching, Zero-Logging, compact logging and others.

Dual Stack

Early on, the IETF recognized that both IPv4 and IPv6 would exist side-by-side for some time on the Internet. It would be clumsy and costly to have two of everything, one with an IPv4 address and one with an IPv6 address on the Internet. For example, it would be impractical to switch between two laptops depending upon whether or not you wanted to browse to an IPv4 or IPv6 web site. The IETF provided a simple approach to this problem by encouraging devices to simultaneously have both IPv4 and IPv6 addresses. In essence, this creates two networking stacks on a device, similar to having both IP and IPX protocols stacks on the same device. One stack runs IPv4 and the other stack runs IPv6, thus creating a Dual Stack approach to IPv6 migration. Eventually, as IPv4 usage dwindles, the IPv4 stack could be disabled or removed from the device.

The Dual Stack approach provides high functionality, but has some disadvantages. Chief among the disadvantages is that every device running Dual Stack needs both an IPv4 and IPv6 address and with a rapidly growing number of devices on the Internet there are simply not enough IPv4 addresses to go around.

Creating an IPv6 Presence and Supporting Mobile IPv6 Employees

Armed with an understanding of IPv6 and migration, technologists can now turn to applying this knowledge to solve business problems. Two main business-needs arise from IPv6: Create an IPv6 presence for your company and its services as well as support mobile IPv6 employees.

Inside corporate IT, as IPv6 is adopted, it is imperative to make sure that the general public, customers, business partners and suppliers can continue to access a company's websites. This typically includes not only the main marketing website that describes a company's products, services and organization, but also e-mail systems, collaboration systems (e.g. Microsoft SharePoint, etc.), and secure data/file transfer systems. Depending upon the type and methods used to conduct business, there could also be sales, order entry, inventory and customer relationship systems that must be accessible on the Internet. The objective is to make sure that a customer, business partner, supplier or the general public can still access your company's application when they are on an IPv6 or a Dual Stack IPv6/IPv4 device. In theory, a Dual Stack IPv6/IPv4 device should work just like an IPv4 only device to access your company's applications, but this should be verified with testing.

To a greater or lesser extent, every company has some form of mobile worker. This could be anything from remote access for IT support staff on weekends and holidays to business critical access for a mobile sales staff or operating a significant amount of business processes over mobile networks. As the IPv4 address supply dwindles further, it is inevitable that your employees will have IPv6 addresses on their devices. This is likely to happen on both corporate managed laptops as well as Bring-Your-Own-Devices (BYOD) since they are both subject to the constraints of mobile wireless and wired broadband providers. Preparation and testing for this inevitability will prevent access failures to business critical applications.

Faced with the objective of establishing an IPv6 presence, there are two main decisions to be made. First, should the IPv6 presence be established separate from the IPv4 presence – a so called “dual legged” approach or alternatively should a Dual Stack approach be used? Second, in what section or sections of the IT infrastructure should an IPv6 be established?

Using a dual legged approach instead of a Dual Stack IPv6 approach provides the least risk to existing applications and services, but is the highest cost and most difficult to implement. With a dual legged approach, a separate IPv6 Internet connection, IPv6 network firewall, IPv6 application servers and related infrastructure are built in the corporate data center. IPv6 Application servers have data synchronized with their IPv4 application counterparts to create a cohesive application. This can be accomplished with multiple network cards where one network card runs only IPv6 and one network card runs only IPv4. This approach is high cost due to hardware duplication and requires implementing IPv6 in the several sections of the data center including the ISP connection, Internet routers, LAN switches, data center perimeter firewalls, network and system management services, IDS/IPS systems, Application Delivery Controllers/Load Balancers and application servers. The dual legged approach is appropriate where the lowest risk levels are desired and there are fewer constraints on the IT budget.

In contrast, a Dual Stack approach to IPv6 migration uses the ability of network devices and servers to simultaneously communicate with IPv6 and IPv4, thus eliminating the need to purchase duplicate hardware for the IPv6 presence. There is some additional risk with Dual Stack in that implementing Dual Stack code on an existing production device may cause problems. Dual Stack should be carefully evaluated, tested and implemented to avoid a decrease in reliability. Dual stack is the recommended approach for IPv6 migration from the IETF, but each situation should be evaluated to validate this approach.

After choosing dual legged or Dual Stack to create your IPv6 presence, IPv6 can be implemented in one of several sections of the IT infrastructure. First, IPv6 to IPv4 services can be purchased via the ISP. Minimal changes are needed to the existing IT infrastructure since the ISP creates a “virtual” IPv6 presence from your IPv4 IT infrastructure. Second, IPv6 can be implemented on the data center perimeter firewalls and translated to the existing IPv4 infrastructure. Third, Application Delivery Controllers/Load Balancers in front of application servers can translate between IPv6 and IPv4 for application servers.

Each of the three approaches above has advantages and disadvantages. Relying on the ISP to create a virtual IPv6 presence from your IPv4 setup is perhaps the simplest and least costly approach, but also offers the lowest amount of flexibility and functionality. Using the data center perimeter firewalls or ADCs for IPv6 migration provides more flexibility and functionality but also raises project costs and complexity. After reviewing their options, organizations may choose to progress through each option in three or more stages, starting with relying on the ISP for IPv6 presence and then progressing into using data center perimeter firewalls, ADCs and finally native IPv6 on application servers.

When reviewing your IPv6 migration strategy, a natural place to start is your current ISP or ISPs if you have more than one connection. For example, your ISPs may support:

- 6to4, 6rd, 6in4, DS-Lite and Teredo tunneling
- NAT64 and DNS64
- Dual Stack Managed Internet Border Routers

- Dual Stack Managed Firewall Services
- IPv6 addressing, including provider independent IPv6 addressing
- IPv6 BGP
- Network monitoring and reporting for IPv6, including separate IPv6 and IPv4 usage

If you are coming close to the end of your contract for ISP services, consider doing an RFI or RFP with other providers to compare IPv6 migration options.

Once the ISP's IPv6 migration capabilities have been assessed, examination of the data center perimeter firewall capabilities is needed. IPv6 and IPv4 (Dual Stack) is typically used on the external firewall or ADC interface and IPv4 for internal/DMZ interfaces. Keep in mind that by simply supporting IPv6 on the external interface of the firewall, the number of firewall rules is at least doubled. If these devices are managed by your ISP or another outsourced provider, you will want to assess both what the devices are capable of as well as what subset of IPv6 functionality the provider will support. Firewalls capabilities can be assessed on:

- Dual Stack IPv6/IPv4
- How IPv6 to IPv4, IPv6 to IPv6 and IPv4 to IPv6 firewall rules are created and maintained
- Network monitoring and reporting on the firewall for IPv6, including separate IPv6 and IPv4 usage statistics
- Ability to NAT IPv6 to IPv6 for privacy (NAT66)
- Support for VRRP IPv6 (e.g. VRRPv3 RFC 5798) and/or HSPR IPv6 for redundancy
- If the same firewalls are used to screen applications for internal users, then IPv6 compatibility with IF-MAP (TCG's Interface for Metadata Access Points) should be checked if applicable.
- Support for IPv6 remote access VPN (IPsec or SSL or IPsec/SSL Hybrid) termination on firewall

Using the data center perimeter firewall to create an IPv6 presence and support remote mobile workers provides more flexibility than just using your ISP to provide IPv6 support, but this approach will require more effort to implement. This arrangement provides the capability to start supporting some native IPv6 services within the corporate data center.

Once the data center perimeter firewall supports IPv6, attention can now turn to Application Delivery Controllers (ADCs) that provide load balancing, SSL offloading, WAN optimization, etc. When establishing an IPv6 presence for customers, business partners and suppliers, there are architectures with two or more data centers that benefit from IPv6 ADCs with WAN optimization. ADCs can have the following IPv6 capabilities[31]:

[31] http://www.a10networks.com/news/industry-coverage-backups/20120213-Network_World-Clear_Choice_Test.pdf

- Ability to provide IPv6/IPv4 Dual Stack for Virtual IPs (VIP)

- Server Load Balancing with port translation (SLB-PT/SLB-64) to IPv4 servers (and the ability to transparently load balance a mix of IPv4 and IPv6 servers)

- 6rd

- NAT64 and DNS64 (to provide IPv6 name resolution services for IPv4-only servers)

- Dual-stack Lite (DS-lite)

- SNMP IPv4 and IPv6 support for monitoring, reporting and configuration

- Ability to provide utilization and usage statistics separated by IPv4 and IPv6

Using the ADC to implement your IPv6 migration gives you the ability to insert Dual Stack IPv6/IPv4 or IPv6 only servers transparently into production. This is a critical first step to providing a low risk application server IPv6 migration path, which in turn is needed to gain access to a larger IP address pool for new and expanded applications. Just using the ISP or data center perimeter firewall for IPv6 does not provide the scalability nor the routing nor security benefits of IPv6.

Supporting Areas

In addition to ISP, network firewall and ADCs IPv6 support, there are usually several supporting systems that need to support IPv6 in the data center. First among these are remote access VPN gateways. Ideally, a remote access VPN gateway that supports IPv4 SSL and/or IPSec connections should work unaltered with 6to4, NAT64 and DNS64 ISP support for an end user device with an IPv6 Internet address. Having said that, statically or dynamically installed software on the end user devices may not work correctly with the end user device's IPv6 stack and this should be tested and verified.

Most organizations also have Intrusion Detection/Protection Systems (IDS/IPS), Security Information Event Monitoring (SIEM), reverse proxies and other security related systems. These systems, if present, should be checked IPv6 Dual Stack readiness and tested as part of a careful IPv6 migration effort.

Last, but not least, there will probably be an myriad of IT security policies, security standards, troubleshooting and operating procedures that need to be updated for IPv6. At a minimum, the format of IP addresses in IT documents should be updated to include IPv6.

Virtual ADCs

Background

A previous section of the handbook outlined a number of the application and service delivery challenges that are associated with virtualization. However, as pointed out in the preceding discussion of WOCs, the emergence of virtualized appliances can also mitigate some of those challenges. As discussed in this subsection of the handbook, there are many ways that an organization can implement a virtual ADC.

In order to understand the varying ways that a virtual ADC can be implemented, it is important to realize that server virtualization technology creates multiple virtual computers out of a single computer by controlling access to privileged CPU operations, memory and I/O devices for each of the VMs. The software that controls access to the real CPU, memory and I/O for the multiple VMs is called a hypervisor. Each VM runs its own complete operating system (O/S) and in essence the hypervisor is an operating system of operating systems. Within each VM's O/S, multiple applications, processes and tasks run simultaneously.

Since each VM runs its own operating system, different operating systems can run in different VMs and it is quite common to see two or more operating systems on the same physical machine. The O/S can be a multi-user O/S where multiple users access a single VM or it can be a single user O/S where each end user gets their own VM. Another alternative is that the O/S in the VM can be specialized and optimized for specific applications or services.

Computers can have more than one CPU that shares memory and I/O ports on a machine and most operating systems can take advantage of multiple CPUs by controlling access to memory blocks with semaphores. Computers with multiple CPUs – sometimes referred to as cores – that share memory and I/O ports are called tightly coupled computing systems. Computers that do not share memory nor I/O ports but which are interconnected by high-speed communications are called loosely coupled. Several CPUs running appropriate operating systems can cooperate together to form a loosely coupled cluster of CPUs and appear as a single computer. Similarly, hypervisors used for VM technology can take advantage of multiple CPU systems in either tightly coupled or loosely coupled arrangement.

VM technology has many benefits including:

- **Consolidation of Computers**
  Multiple systems can be combined onto one system providing CAPEX and OPEX savings.

- **Running Multiple Software Versions**
  When upgrading either an operating system or a business critical application, VM technology allows both versions to be run simultaneously eliminating the need for extra hardware just to enable these upgrades.

- **IT Infrastructure Agility**
  New virtual machines can be added quicker than installing a physical machine and this shortens the time to implement new systems.

- **Security Compartmentalization**
  Each VM is segmented from every other VM and this helps – but does not prevent - security issues from spreading between computers.

The Evolution of Network Appliances

Over the last decade, driven by the need to more securely and reliably deliver applications and services, the network has become increasingly sophisticated. For example, routers and firewalls that were once run on general-purpose servers, now run on specialized appliances. Additional network functionality moved from application servers to network devices. This includes encryption, data compression and data caching. In addition, network services running on servers also moved to specialized network appliances; i.e., DNS and RADIUS authentication servers.

As shown in **Figure 24**, as network functionality grew, the network evolved from a *packet delivery* service to an *application and service delivery* service. Network appliances evolved from general purpose servers to become the standard building block of the Application and Service Delivery Network. Network appliances improved upon server technology in two important ways. First, the O/S was changed from a general purpose O/S to one optimized for network operations and processing. Second, the server hardware was updated to include specialized co-processors (e.g. SSL operations and encryption) and network adapters for high performance network operations. This simplified IT operations, as typically only one IT group (e.g. Networks Operations) was involved in changes as opposed to two IT groups (e.g., Network Operations and Server Operations). In general, software updates and security patches are less frequent on network appliances than for general purpose O/Ss and this further reduces the IT operations effort.

Virtualization and Cloud Computing technology challenged network appliances in two important ways and this resulted in a split evolutionary path of the network appliance. The rise of public cloud offerings caused network equipment manufacturers to update their specialized network appliance operating systems to run under general-purpose hypervisors in CCSP locations. This allowed CCSPs to run specialized network and security functions on their low cost, virtualized server infrastructure filling a much needed functionality gap for their offerings.

Data center and branch office network consolidation also pushed network manufacturers to add VM technology to their appliances to run multiple network functions on fewer appliances. To keep performance and cost levels in line, specialized network appliance hypervisors where developed that not only partitioned CPU, memory and I/O, but also partitioned other hardware resources such as network bandwidth and encryption coprocessors. Many of the specialized network hypervisors developed were capable of using loosely coupled systems across multiple appliances and multiple chassis.

Network appliances such as ADCs are evolving along two paths. One path is comprised of general-purpose hardware, a general-purpose hypervisor and a specialized O/S. The other path is comprised of specialized network hardware, specialized network hypervisors and a specialized O/S.

The Types of ADC Virtualization

This two-path evolution of network appliances has resulted in a wide array of options for deploying ADC technology. These options include:

- **General Purpose VM Support**
  A specialized network O/S along with ADC software that have been modified to run efficiently in a general purpose virtualization environment including VMWare's vSphere, Citrix's XenServer and Microsoft's Hyper-V.

- **Network Appliance O/S Partitioning**
  This involves the implementation of a lightweight hypervisor in a specialized network O/S by partitioning critical memory and I/O ports for each ADC instance, while also maintaining some memory and I/O ports in common.

- **Network Appliance with OEM Hypervisor**
  A general-purpose virtualization solution is adapted to run on a network appliance and provides the ability to run multiple ADCs on a single device. Since the hypervisor is based on an OEM product, other applications can be run on the device as it can participate in an enterprise virtualization framework such as VMWare's vCenter, Citrix's Xencenter or Microsoft's System Center. Support for loosely couple systems (e.g. VMWare's VMotioin and Citrix's XenMotion) is common.

- **Network Appliance with Custom Hypervisor**
  General-purpose hypervisors are designed for application servers and not optimized for network service applications. To overcome these limitations, custom hypervisors optimized for network O/S have been added to network appliances. Depending on implementation, these specialized network hypervisors may or may not support loosely coupled systems.

Each of these approaches has advantages and disadvantages that effect overall scalability and flexibility. General purpose VM support has the most flexibility, but when compared to network appliance hardware, general purpose VM support gives the lowest level of performance and reliability. Network appliances with custom hypervisors can provide the greatest performance levels, but provide the least flexibility with limited co-resident applications and virtualization framework support.

High Availability and Hardware Options

ADCs have several options for high availability and scalability configurations. This usually involves a combination of dual instance arrangements on the same LAN and Global Server Load Balancing (GSLB) across data centers. Two ADC devices or instances on a LAN segment can act as single ADC instance using VRRP (RFC 5798) or HSRP and sharing session state information. When one ADC instance fails, the other ADC instance takes control of the virtual MAC address and uses its copy of the synchronized session state data to provide a continuous service. For ADC instances across data centers, GSLB services can redirect traffic to alternative ADC pairs when an ADC pair is unavailable. Hypervisors that support loosely coupled systems (e.g. VMWare's VMotion and Citrix's XenMotion) provide additional high availability options by moving ADC instances to alternative hardware either for maintenance operations or backup.

High availability mechanisms not only provide better access to a business's applications, but these mechanisms can also be used for load sharing to boost overall scalability. The computing hardware of the network appliance also plays a significant role in overall scalability. Two popular form factors include self-contained units and chassis based devices. Self-contained units contain all the components including power supply, I/O devices, ports and network connections. They have a limited ability to increase capacity without being replaced, but are generally lower cost than an entry-level chassis system.

Chassis systems consist of a chassis and a number of expansions cards that can be added to scale capacity. The chassis usually provides common power, internal bus and network connections to each expansion card. Fully populated chassis systems are usually more cost effective than self-contained devices, but a failure of a common chassis component (e.g. power supply) will affect the entire chassis rather as compared to a single device failure in an array of self-contained devices.

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:

- ***<u>Hypervisor–based Multi-tenant ADC Appliances</u>***
  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.

- ***<u>Cloud Bursting and Cloud Balancing ADCs</u>***
  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.

- ***Web Content Optimization (WCO)***
  Two of the challenges that are associated with delivering Web pages are the continually growing number of objects per page, which result in a continually increasing number of round trips per page and the continually growing size of Web pages. Another challenge is the wide range of browsers and mobile devices that access Web pages. Having a range of browsers and mobile devices makes it very time consuming to manually optimize the Web page for delivery to all the users. WCO refers to efficiently optimizing and streamlining Web page delivery. WCO is available in a number of form factors, including being part of an ADC.

Some of the techniques that are used in a WCO solution include:

- Image spriting: A number of images are merged onto a single image reducing the number of image requests.

- JPEG resampling: An image is replaced with a more compact version of the image by reducing the resolution to suit the browser.

- HTTP compression: Compress HTTP, CSS and JavaScript files.

- URL versioning: Automatically refresh the browser cache when the content changes.

Developing your ADC Strategy

As with developing any IT strategy, the process begins with understanding the organization's overall strategy, business drivers and applications. If the mission of the network is to deliver applications, not just packets, and an understanding of the organizations applications is a must. Some, but not all, of the things to consider when creating your ADC strategy are:

- **Current ADC or Server Load Balancing (SLB) Deployment** – Current ADC or SLB deployments provide an opportunity to understand the organization's application characteristics as well as save costs by reusing or trading in existing devices.

- **Use or planned use of Cloud Computing and other outsourcing** – Understand if there is a private, public or hybrid Cloud Computing strategy or specific CCSP in place. If a specific CCSP is in place and unlikely to change, it is important to under which ADCs products the CCSP supports and what virtualization management frameworks the CCSP uses.

- **Application availability and reliability requirements and preferences**– To scale ADC deployment you need both the average and peak requirements for all of the applications using ADC services.

- **New application acquisition plans** – The application portfolio is dynamic and the ADC strategy should consider the current application portfolio as well as planned and possible expansions.

- **Application performance constraints** – An ADC strategy needs to handle the performance and load requirements of the applications it supports. To scale the ADC strategy, the application speeds need to be considered. At a minimum, average and peak connections per second and the bandwidth consumed should be known.

- **Data center spare capacity, power density and cabling capacities** - Different physical sizes, rack airflow, power consumption and network cabling for ADC products can create deployment problems in data centers. Data center preferences and constraints should be taken into account.

- **IPv4 to IPv6 migration plans** – ADCs are a key point where IPv6 to IPv4 transitions occur as part of an overall IPv6 migration. An organizations IPv6 migration strategy and plans affect the ADC strategy.

- **Established IT architecture principles** – Many IT organizations have created a list of IT architecture principles that should be adhered to. Some IT organizations may have an IT architecture principle approval process as well as an architecture principle exception process or tracking system.

Perhaps the biggest factor from the above list in developing your ADC strategy is the use of Cloud Computing. Using a CCSP or other outsourcing constrains your ADC options and this helps narrow the field of choices. If your CCSP choice is established and will not change, then you are constrained to use the ADC products and technologies supported by the CCSP. If you are or will use a hybrid cloud or cloud bursting arrangement, the CCSP's ADC choices can also constrain the ADC choices in the private data center. With a hybrid or cloud bursting approach, you may also be constrained to certain virtualization management frameworks, which in turn will influence your ADC choice.

After considering your Cloud Computing strategy, next consider the availability and reliability needed for the applications. As the need for application availability rises, this will drive the requirements for single or multiple devices for resiliency as well as the choice of single or multiple chassis. Multiple devices and/or chassis will provide high levels of availability and reliability. Chassis can usually provide greater performance scaling than devices, but can also increase initial costs. Chassis usually have a higher capacity connection between loosely coupled systems than devices that are LAN/WAN interconnected.

After your ADC strategy is developed, an ADC product set needs to be chosen. Based on your ADC strategy, you may be able to reduce to possible product selection to reduce the number of candidate ADC suppliers. This will lower project costs and improve implementation times.

Some requirements to consider adding to your ADC product selection criteria include:

- Feature Parity between Network Appliance, Virtualized Network Appliance and Virtual products.

- Number of processors and speeds available for network appliance models. Consider any encryption coprocessors and bandwidth (NIC card) partitioning capabilities as well.

- Availability of chassis hardware for scaling and speeds between blades in the chassis as well as external speeds between chassis.

- Ability to virtualize across network appliances, network hardware chassis and virtual instances both locally and across WAN links.

- Aggregate scaling with network appliances, chassis and virtual instances.

- Completeness and flexibility of IPv6 support.

- Ability to support hybrid and cloud bursting deployments

- Flexibility to integrate with virtualization management frameworks including VMware vCenter, Citrix's Xencenter and Microsoft's System Center.

- Overall functionality including load balancing, load detection flexibility, SSL offloading, security processing, proxy support, TCP optimization, WAN Optimization and reporting.

In addition to these suggestions, there are selection criteria that are common across most products including support options, delivery times, hardware maintenance options, service and account reviews, legal terms, etc.

Planning, Management and Security

Planning

Background

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 a subsequent sub-section of this handbook, is identifying the company's key applications and services and establishing SLAs for them. As described in that 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 a subsequent sub-section of this handbook is Application Performance Engineering (APE).

The primary 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 caused an unacceptable impact on the company's other 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[32].

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:

[32] 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 solution 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 private cloud computing.

- How the IT organization will evolve its use of WAN services to support all forms of cloud computing.

- How the IT organization will minimize the security and confidentiality risks associated primarily with public cloud computing services, but also with private cloud computing.

- 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.

Management

Background

As will be discussed in this section of the handbook, in order to respond to the myriad challenges facing them, IT organizations need to adopt an approach to service management that is comprised of the following four components:

- Either a multi-tier application or multiple applications
- Supporting protocols
- Enabling network services; e.g., DNS, DHCP
- The end-to-end network

Limitations of the Traditional Approaches to Management

The almost total reliance that organizations have on networked applications and services has pushed the task of ensuring acceptable application and service performance up to the critical level for almost all organizations. Despite this high priority, the traditional network and application performance management systems have lagged behind.

Traditional Network Performance Management Systems

Most Network Performance Management Systems (NPMS) had their origins in monitoring the performance of telecommunication carriers to verify that organizations were getting the services they paid for. These systems are based on a combination of Simple Network Management Protocol (SNMP) and Internet Control Message Protocol (ICMP, also known as “ping”). Traditional NPMS measured how long it took a packet to travel from the data center to the branch office network and back - thus determining the Round Trip Time (RTT). If the return packet did not arrive within a few seconds, the original packet was deemed lost and this is how packet loss was measured.

These early NPMS solution worked acceptably well for traditional client/server applications and other centrally hosted applications. However, as technology and applications evolved, the limitations of these systems became apparent. Those limitations include the fact that early NPMS systems:

- Only measured from the central data center to the edge of the branch office network. Problems inside the branch office network went unreported until end users complained.
- Had difficulty measuring network paths outside of the data center, such as those used by VoIP, IP Video and other peer-to-peer communication traffic.
- Measured performance across the entire path but did not isolate which network segments have performance issues.

Traditional Application Performance Management

As application architectures evolved from client/server to n-tier web-based applications, application functionality on the server was usually divided up into two or three segments. These segments are the web front-end (presentation tier or tier 1), business logic processes (logic tier or tier 2) and database operations (data tier or tier 3).

In an n-tier web based application, the user interacts with the presentation tier and the presentation tier in turn communicates to the logic tier, which in turn communicates to the data tier. Each tier uses servers that are optimized to the characteristics of their tier. A presentation tier server, for example, is optimized for network I/O and web traffic; e.g. multiple network cards, large network buffers, etc. A logic tier server is optimized for logic computations; e.g. high speed CPUs, large memory size, etc. A data tier server is optimized for database operations; e.g. multiple disk I/O controllers, large disk cache, large memory size, etc.)

Traditional application performance management (APM) was typically performed separately from network performance management. For example, when application degradation occurs, the triage process typically assigns the incident to either the network or server areas for resolution. Each area then examines their basic internal measurements of network and server performance and a pronouncement is made that the source of the issue is either the network or the application server or both or neither. Since these tasks are typically done by different parts of the IT organization using different toolsets and management frameworks, it is quite possible that conflicting answers are given for the source of application performance issues.

Similar to traditional NPMS, traditional APM solutions have limitations. Those limitations include the fact that that traditional APM solutions:

- Only describe the performance within a single server, not the combined performance across all tiers of an application.
- Cannot attribute CPU, disk I/O, network I/O nor memory utilization to specific classes of transactions. Only aggregate server performance information is available.
- Do not integrate network performance data between tiers to monitor and analyze application performance problems.

Synthetic Transactions

Synthetic transactions provide a somewhat more realistic measurement of application performance than traditional NPMS and APM solutions. While synthetic transactions have the advantage of being a better representation of the end user’s experience, they also have several disadvantages, including:

- The application being monitored has to be constructed to allow transactions that have no business impact. For example, a banking application would have to have a special account so that when money was added or subtracted from this special account, it would not count towards the banks total assets.

- Synthetic transactions frequently originate from the same data center in which the application servers reside and are not subject to the typical network latencies and availabilities that are present in branch office networks.

- Frequently exercising a synthetic transaction can cause the transaction to perform notably differently than a real production transaction would. For example, a frequently exercised transaction may have its related data in cache all the time and not loaded from disk. As a result, the synthetic transaction would occur notably quicker than a production transaction would.

Forces Driving Change

Previous sections of this handbook detailed the traditional and emerging service and application delivery challenges. This subsection will identify how some of those challenges are forcing a change in terms of how IT organizations perform services.

Server Virtualization

Until recently, IT management was based on the assumption the IT organizations performed tasks such as monitoring, baselining and troubleshooting on a server-by-server basis. Now, given the widespread adoption of server virtualization, the traditional approach to IT management must change to enable management tasks to be performed on a virtual machine (VM)-by-VM basis. Another assumption that underpinned the traditional approach to IT management was that the data center environment was static. For example, it was commonly assumed that an application resided on a given server, or set of servers, for very long periods of time. However, part of the value proposition that is associated with server virtualization is that it is possible to migrate VMs between physical servers, both within the same data center and between disparate data centers.

IT organizations need to adopt an approach to management that is based on the assumption that the components of a service, and the location of those components, can and will change frequently.

Cloud Balancing

IT management has historically been based on the assumption that users of an application accessed that application in one of the enterprise's data centers and that the location of that data center changed very infrequently over time. The adoption of IaaS solutions in general, and the adoption of cloud balancing in particular demonstrates why IT organizations need to adopt an approach to IT management that is based on gathering management data across myriad data centers, including ones that are owned and operated by a third party. The adoption of cloud balancing is also another example of why IT organizations need to adopt an approach to management that is based on the assumption that the components of a service, and the location of those components, can and will change frequently.

Delay Sensitive Traffic

Voice and video are examples of applications that have high visibility and which are very sensitive to transmission impairments. As part of the traditional approach to IT management it is common practice to use network performance measurements such as delay, jitter and packet loss as a surrogate for the performance of applications and services. A more effective approach is to focus on aspects of the communications that are more closely aligned with ensuring acceptable application and service delivery. This includes looking at the application payload and measuring the quality of the voice and video communications. In the case of UC, it also means monitoring the signaling between the components of the UC solution.

In addition to having a single set of tools and more of a focus on application payload, IT organizations need to implement management processes that understand the impact that each application is having on the other applications and that can:

- Analyze voice, video, UC and data applications in consort with the network
- Support multi-vendor environments
- Support multiple locations

Converged Management

As mentioned in the section of the report that focused on the emerging application and service delivery challenges, one of the characteristics of cloud computing is the integration of networking, servers and computing in the data center. While a converged data center infrastructure offers a number of benefits, it does create a number of management challenges. These challenges generally fall into the following categories:

- Implementing control of the infrastructure in a significantly more efficient manner than is possible with the traditional labor-intensive tool sets that are not well integrated or automated.
- Ensuring compliance with enterprise policies, as well as best-practice guidelines for service deployment in the converged infrastructure.
- Implementing granular, near real-time provisioning and change management for virtualized infrastructure services.
- Expediting resolution of any issues that could compromise the performance or availability of infrastructure services.

Meeting these challenges will involve a number of changes in the way the data center is managed. In particular, the converged infrastructure requires a management system and management processes that have the same level of integration and cross-domain convergence that the infrastructure has. For example, in order to support the requirement for the dynamic provisioning and re-allocation of resources to support a given IT service, the traditional manual processes for synchronizing the required server, network and storage resources will have to be replaced with integrated, automated processes. In order to enable this change, the provisioning and change management processes will need to be integrated and will need to feature the automatic configuration of network and storage resources when additional infrastructure services are deployed or when additional physical or virtual servers are brought on line or are moved. In a similar fashion, operations management needs to be consolidated and automated to keep service quality in line with user expectations.

IT departments can take a do-it-yourself (DIY) approach to implementing an integrated, cross-domain management system. They could, for example, leverage the available element manager plug-ins and APIs and piece together a management system that integrates at least some of the overall components of the infrastructure. However, this type of ad hoc automation and integration across the end-to-end, cross-domain infrastructure is quite time-consuming and involves considerable specialized programming expertise and detailed technical knowledge. It is also expensive and time consuming to support over time. There is an alternative to the DIY approach. IT departments that are evaluating converged infrastructure solutions can expect to see highly effective pre-packaged integrated management solutions that negate the need for a DIY approach. In addition, IT organizations can use the breadth, efficiency, and effectiveness of the converged management solutions as one of the key decision criteria to evaluate the converged infrastructure solutions that are offered by competing vendors.

While not a requirement, the cross-domain integrated management of a converged infrastructure will bring the greatest benefit in those instances in which a single administrator has the authority to initiate and complete cross-domain management tasks, such as provisioning and modifying infrastructure services. For example, the use of a single administrator can eliminate the considerable delays that typically occur in a traditional management environment where the originating administrator must request other administrators to synchronize the configuration of elements within their domains of responsibility. However, in many cases the evolution from the current approach of having separate administrators for each technology domain to an approach in which there is a single administrator will involve organizational challenges. As a result, many IT organizations will evolve to this new approach slowly over time.

Mobile Device Management

In the current environment, it is possible to find a Cloud-based service that provides almost any possible form of management functionality. For example, there are multiple Cloud-based services currently available in the marketplace that manage network equipment such as routers, firewalls and IPSs.

The section on the handbook that described the first generation of application and service delivery challenges discussed the fact that the IBM X-Force 2011 Trend and Risk Report[33] highlighted the fact that new trends such as the ongoing adoption of mobile devices creates challenges for enterprise management and security. For example, the IBM stated that in 2011 there was a 19 percent increase over 2010 in the number of exploits publicly released that can be used to target mobile devices such as those that are associated with the movement to Bring your Own Device (BYOD) to work. The report added that there are many mobile devices in consumers' hands that have unpatched vulnerabilities to publicly released exploits, creating an opportunity for attackers.

One way to respond to the challenges associated with mobile devices is by using a Cloud-based mobile device management service. In order to be an effective, such a service should:

- Offer a portal that provides a comprehensive view into the mobile device environment.

- Given the ongoing consumerization of IT, the portal should provide visibility into both corporate and employee-owned mobile devices.

[33] X-Force 2011 Trend and Risk Report

- Enable the organization to implement management security policies through simple click configuration and provisioning.

- Control access to key network resources and applications by specific user groups or by lines of business.

- Detect if a mobile device has been lost and if so, delete the data from the device.

- Detect when roaming, data, voice, or SMS usage thresholds have been reached, and provide real-time alerts to prevent unwanted billing overages.

Evaluating Cloud Computing Service Providers

When It organizations are evaluating the adoption of public cloud computing solutions, they need to evaluate the CCSP's management capabilities. This sub-section contains two sets of criteria that can be used to evaluate those capabilities. The first set is focused on traditional management functionality and the second set focuses on managing the performance of applications and services.

Traditional Management Functionality

- What is the ability of the CCSP to manage the challenges associated with virtualization that were previously discussed in 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?

Managing Application and Service 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 and/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?

Application Performance Management

Background

This section of the handbook will outline an approach that IT organizations can utilize to better manage application and service delivery, where the term *service* was previously defined. However, in an effort to not add any more confusion to an already complex topic, instead of using a somewhat new phrase *application and service delivery management*, this section will use the more commonly used phrase *application performance management (APM)*.

APM is a relatively new management discipline. 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, over half 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.

Since any component of a complex service can cause service degradation or a service outage, IT organizations need a single unified view of all of the components that support a service. This includes the highly visible service components such as servers, storage, switches and routers, in both their traditional stand-alone format as well as in their emerging converged format; i.e., Cisco's UCS and VCE's Vblock platforms. It also includes the somewhat less visible network services such as DNS and DHCP, which are significant contributors to application degradation. Multiple organizational units within the IT organization have traditionally provided all of these service components. On an increasing basis, however, one or more network service providers and one or more cloud computing service providers will provide some or all of these service components and so in order to achieve effective service delivery management, management data must be gathered from the enterprise, one or more Network Service Providers (NSPs) and one or more CCSPs. In addition, in order to help relate the IT function with the business functions, IT organizations need to be able to understand the key performance indicators (KPIs) for critical business processes such as supply chain management and relate these business level KPIs to the performance of the IT services that support the business processes.

IT organizations must also be able to provide a common and consistent view of both the network and the applications that ride on the network to get to a service-oriented perspective. The level of granularity provided needs to vary based on the requirements of the person viewing the performance of the service or the network. For example, a business unit manager typically wants a view of a service than is different than the view wanted by the director of operations, and that view is often different than the view wanted by a network engineer.

In spite of the importance of providing a holistic approach to APM, 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.

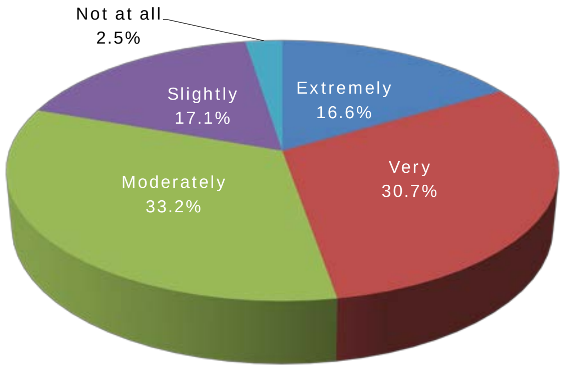
One of the reasons why it is important to get better at managing the user's experience was previously mentioned in the handbook. That reason being that in spite of all of the effort and resources that have gone into implementing IT management to date, it is the end user, and not the IT organization who typically is the first to notice when the performance of an application begins to degrade.

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 32**.

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

Figure 32: Getting Better at Monitoring End User Behavior



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

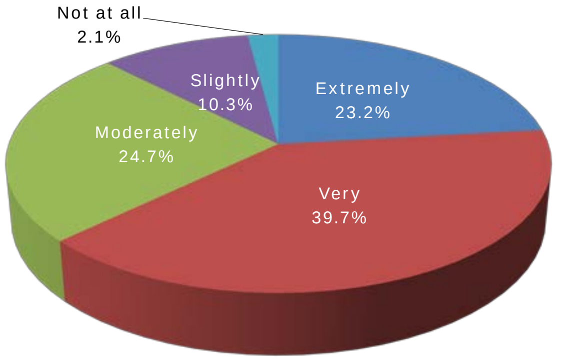
- The adoption of a system of service level agreements (SLAs) at levels that ensure effective business processes and user satisfaction for at least a handful of key applications.

- 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 33**.

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 33: Importance of Identifying Components of Critical Applications



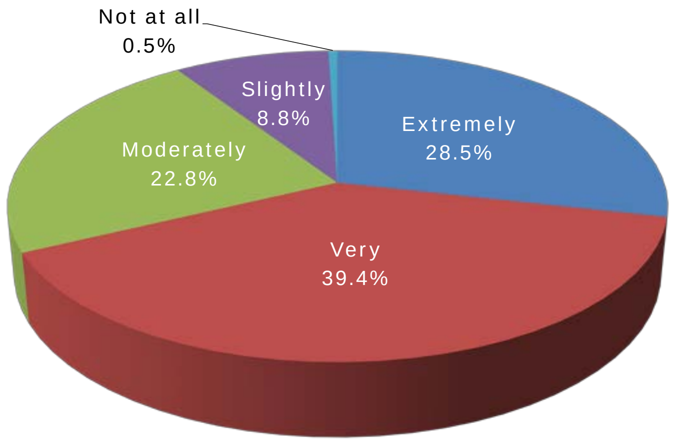
If IT organizations can effectively identify which components of the infrastructure support a particular application or service, monitoring can much more easily identify when services are about to begin to degrade due to problems in the infrastructure. As part of this monitoring, predictive techniques 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 34**.

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

Figure 34: 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 effective APM. 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.

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[34] 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. 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.

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

[34] VMWare.com VMotion

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.

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 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

APM in the Private Enterprise Network[35]

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

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

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.

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. This functionality 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 sub-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 virtualized the IT organization loses 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.

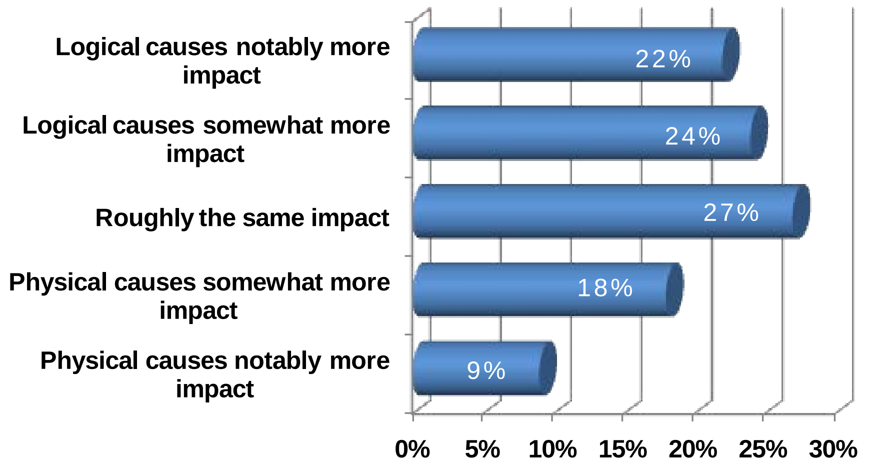
Logical vs. Physical Factors

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 35**.

Figure 35: Impact of Logical and Physical Factors on the Business

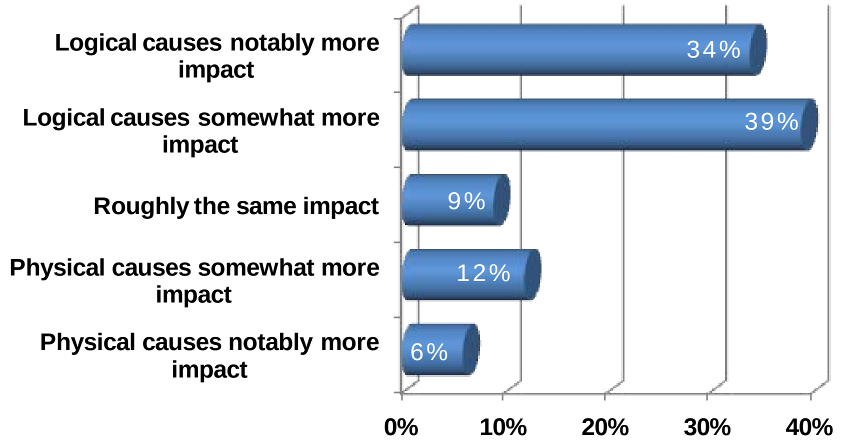


In the vast majority of cases, logical factors cause as much or more business disruption than do physical factors.

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 36**.

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 36: 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.

Security

The section of this handbook that is entitled “First Generation Application and Service Delivery Challenges” referenced a number of industry reports in order to describe the current security environment and to identify the types of attacks that are becoming increasingly common.

For example, that section of the handbook referenced IBM’s X-Force 2011 Trend and Risk Report[36]. Some of the key observations made in that report are:

- **Mobile Devices**
  The report stated that in 2011 there was a 19 percent increase over 2010 in the number of exploits publicly released that can be used to target mobile devices such as those that are associated with the BYOD movement. The report added that there are many mobile devices in consumers' hands that have unpatched vulnerabilities to publicly released exploits, creating an opportunity for attackers.

- **Social Media**
  With the widespread adoption of social media platforms and social technologies, this area has become a target of attacker activity. The IBM report commented on a surge in phishing emails impersonating social media sites and added that the amount of information people are offering in social networks about their personal and professional lives has begun to play a role in pre-attack intelligence gathering for the infiltration of public and private sector computing networks.

- **Cloud Computing**
  The report stated that there were many high profile cloud breaches affecting well-known organizations and large populations of their customers. IBM recommended that IT security staff should carefully consider which workloads are sent to third-party cloud providers and what should be kept in-house due to the sensitivity of data. The IBM X-Force report also noted that the most effective means for managing security in the cloud may be through Service Level Agreements (SLAs) and that IT organizations should pay careful consideration should be given to ownership, access management, governance and termination when crafting SLAs.

That section also referenced Blue Coat Systems’ 2012 Web Security Report[37]. According to the Blue Coat report, “In 2011, malnets emerged as the next evolution in the threat landscape. These infrastructures last beyond any one attack, allowing cybercriminals to quickly adapt to new vulnerabilities and repeatedly launch malware attacks. By exploiting popular places on the Internet, such as search engines, social networking and email, malnets have become very adept at infecting many users with little added investment.” That report also noted the increasing importance of social networking and stated that, “Since 2009, social networking has increasingly eclipsed web-based email as a method of communications.” and that, “Now, social networking is moving into a new phase in which an individual site is a self-contained web environment for many users – effectively an Internet within an Internet.”

[36] X-Force 2011 Trend and Risk Report

[37] http://www.bluecoat.com/sites/default/files/documents/files/BC_2012_Security_Report-v1i-optimized.pdf

This section of the handbook will discuss the technologies and services that IT organizations can use to respond to these security challenges.

How IT Organizations are Implementing Security

As previously described in this handbook, the security landscape has changed dramatically in the last few years. In the very recent past, the typical security hacker worked alone, relied on un-sophisticated techniques such as dumpster diving, and was typically motivated by the desire to read about their hack in the trade press. In the current environment, sophisticated cyber criminals have access to malware networks and R&D labs and can use these resources to launch attacks whose goal is usually to make money for the attacker. National governments and politically active hackers (hacktivists) are engaging in cyber warfare for a variety of politically motivated reasons. Unfortunately terms such as Distributed Denial-of-Service (DDoS), Advanced Persistent Attacks (APTs) and SQL injections attacks are becoming common used as illicit activity continues.

In many aspects security is both a first and a second-generation application and service delivery challenge and it will remain a significant challenge for the foreseeable future. Rapid changes in IT, such as those created by the adoption of cloud computing, social networking and the new generation of mobile devices, combined with the ongoing evolution of regulations pose a spate of new challenges for IT security systems and policies in much the same manner that they present challenges to the IT infrastructure.

IT security systems and policies have evolved and developed around the traditional application delivery architecture in which branch offices are connected to application servers in a central corporate data centers. In this architecture, the central corporate data center is a natural location to implement IT security systems and policies that provide layered defenses as well a single, cost efficient location for a variety of IT security functions. With the adoption of public cloud computing, applications and services are moving out of the central corporate data center and there is no longer a convenient single location for security policies and systems.
IT security systems and policies have traditionally distinguished between people who were using IT services for work versus those who were using it for personal use. The use of an employer provided laptop was subject to the employer's IT security policies and systems. In this environment, the use that employees made of personal laptops was generally outside of the corporate IT security policy. With the arrival of smartphones and tablet computers, the ownership, operating systems and security capabilities of the end user devices have changed radically. IT security policies and standards that were developed for PCs are no longer effective nor optimal with these devices. Most corporations have embraced the BYOD movement and end users are less willing to accept strict corporate security policies on devices they own. Additionally, strict separation of work and personal usage for security on an employee owned device is impractical.

The demands of governments, industry and customers have historically shaped IT security systems and policies. The wide diversity of organizations that create regulations and standards can lead to conflicts. For example, law enforcement requires access to network communications (Communications Assistance for Law Enforcement Act – CALEA) which may in turn force the creation of locations in the network that do not comply with the encryption requirements of other standards (e.g. Health Insurance Portability Accountability Act – HIPPA). In order to determine how IT organizations are responding to the traditional and emerging security challenges, The Survey Respondents were asked a series of questions. For example, to get a high level view of how IT organizations are providing security, The Survey Respondents were asked to indicate which of a number of network security systems their organization supports. The Survey Respondents were asked to check all of the alternatives that applied in their environment. Their responses are shown in **Table 31**.

Table 31: The Network Security Systems in Use

Network Security Systems	Percentage
Remote Access VPN	86.30%
Network Access Control	73.50%
Intrusion Detection/Protection Systems (IDS/IPS)	65.70%
Next Generation Firewalls (Firewall+IPS+Application Control)	56.90%
Secure Web Gateways	46.10%
Web Application and/or XML Firewalls	36.30%
Mobile Device Security/Protection	36.30%
Security Information Event Management	31.40%
Data Loss Prevention	24.50%
Password Vault Systems (either local or portal based)	12.70%
SAML or WS-Federation Federated Access Control	8.80%

One obvious conclusion that can be drawn from **Table 31** is that IT organizations use a wide variety of network security systems. A slightly less obvious conclusion is that on average, IT organizations use 4.8 of the network security systems listed in the preceding table.

The Survey Respondents were asked to indicate the approach that best describes how their company uses data classification to create a comprehensive IT security environment. Their responses are shown in **Table 32**.

Table 32: Approach to Comprehensive IT Security

Approach	Percentage
We have a data classification policy and it is used to determine application access/authentication, network and end user device security requirements.	42.90%
We do not have a data classification policy.	33.00%
We have a data classification policy and it is used to determine application security requirements.	13.20%
We have a data classification policy, but it is not used nor enforced.	11.00%

The data in **Table 32** represents a classic good news/bad news situation. The good news is that the majority of IT organizations have a data classification policy that they use to determine requirements. The bad news is that 44% of IT organizations either don’t have a data classification policy or they have one that isn’t used or enforced.

In order to understand how IT organizations are responding to the BYOD movement, The Survey Respondents were asked, "If your organization does allow employee owned devices to connect to your network, please indicate which of the following alternatives are used to register employee owned devices and load authentication (e.g. certificate/private key) data onto those devices before they are allowed to connect to your company's network." The Survey Respondents were asked to check all of the alternatives that applied in their environment. Their responses are shown in **Table 33**.

Table 33: Alternatives to Support Employee Owned Devices

Alternative	Percentage
Employees must install a VPN client on their devices for network access	53.90%
IT Administrator and/or Service Desk must register employee owned device for network access	47.40%
Employees can self-register their devices for network access	28.90%
Employees must generate and/or load X.509 certificates & private keys network access	13.20%
Employees must install a token authentication app on their devices for network access	10.50%

The data in **Table 33** indicates that while using a VPN is the most common technique that a wide range of techniques are used. VPN's popularity comes in part from the fact that remote access VPN solutions implemented on new generation mobile devices have various capabilities to enforce security policies when connecting to the corporate network. Popular security checks include ensuring that a screen password is present, that anti-virus software is present and is up to date, that there is not rogue software on the device and that the operating system has not been modified.

Two different approaches have emerged to protect against lost devices. For the traditional PC, full disk encryption is typically used to protect data if the PC is lost or stolen. However, on new generation mobile devices, remote erase solutions are typically used to protect data. New generation mobile devices with smaller displays are often used more for content reading rather than content creation. As screen sizes and resolution improves, this situation may change. In order to understand how IT organizations have implemented full disk encryption, The Survey Respondents were asked to indicate which alternatives their organization implements relative to using full disk encryption on laptops and desktop PCs. Their responses are shown in **Table 34**.

Table 34: Techniques for Implementing Full Disk Encryption

Alternative	Percentage
We do not use full disk encryption on PCs.	52.5%
We use software based disk encryption on PCs.	49.5%
We use hardware based self-encrypting rotating drives on PCs.	6.1%
We use hardware based self-encrypting Solid State Drives on PCs.	6.1%

The data in **Table 34** indicates that just over half of all IT organizations don't use full disk encryption on PCs. The data also indicates that those IT organizations that do use full disk encryption do so by using a software solution and that a small percentage of IT organizations use multiple techniques.

The Survey Respondents were asked to indicate the approach that best describes their company's approach to Identity and Access Management (IAM). Their responses are shown in **Table 35**.

Table 35: How IAM is Implemented

Approach	Percentage
We do not have a formal IAM program.	36.6%
We have an IAM program, but it only partially manages identities, entitlements and policies/rules for internal users.	25.8%
We have an IAM program and it manages identities, entitlements and policies/rules for all internal users.	20.4%
We have an IAM program and it manages identities, entitlements and policies/rules for end users for internal, supplier, business partner and customers.	17.2%

The data in **Table 35** indicates that only a minority of IT organizations has a IAM program that has broad applicability.

The Survey Respondents were asked to indicate how their company approaches the governance of network and application security. Their responses are shown in **Table 36**.

Table 36: Governance Models in Use

Approach	Percentage
Network Security and Application Security are funded, architected, designed and operated together.	46.9%
Network Security and Application Security are funded, architected, designed and operated separately.	30.2%
Network Security and Application Security are funded jointly, but architected, designed and operated separately.	22.9%

The data in **Table 36** indicates that in the majority of instances, network security and application security are architected, designed and operated separately.

Cloud-Based Security

The section of this handbook that focused on the emerging application and service delivery challenges presented the results of a survey in which The Survey Respondents were asked how likely it was over the next year that their company would acquire a traditional IT service from an IaaS provider. Their responses are shown in **Table 37**.

Table 37: Interest in Obtaining IT Services as a Cloud-Based Service

	Will Not Happen	Might Happen	50/50 Chance	Will Likely Happen	Will Happen
VoIP	32.6%	18.6%	15.3%	13.5%	20.0%
Unified Communications	30.2%	22.8%	20.5%	14.9%	11.6%
Security	42.6%	17.1%	14.4%	11.6%	14.4%
Network and Application Optimization	32.1%	28.8%	16.0%	14.6%	8.5%
Network Management	41.4%	22.3%	13.5%	13.5%	9.3%
Application Performance Management	37.9%	26.5%	15.6%	11.4%	8.5%
Virtual Desktops	38.8%	28.0%	15.9%	12.1%	5.1%

As shown in **Table 37**, the interest shown by The Survey Respondents in obtaining security as a Cloud-based service is bimodal. When looking just at the percentage of The Survey Respondents that indicated that it either will happen or will likely happen, security is one of the most likely services that IT organizations will acquire from a CCSP. However, a higher percentage (42.6%) of The Survey Respondents indicated that they will not acquire security from a CCSP than made that indication for any other form of IT service listed in the survey.

One way that a Cloud-based Security Service (CBSS) could provide value is if it provides protection against the growing number of malware attacks. To effectively protect against malware attacks, a CBSS should be able to identify suspicious content or sites that are either suspicious or are known to distribute malware. In order to be effective, a CBSS that provides Web content filtering or malware protection needs a source of intellectual capital that identifies known and suspected vulnerabilities. This source needs to be both dynamic and as extensive as possible.

One part of the value proposition of a CBSS that provides security functionality is the previously discussed value proposition of any cloud based service. For example, a security focused CBSS reduces the investment in security that an organization would have to make. In addition, a security focused CBSS reduces the amount of time it takes to deploy new functionality. The speed at which changes can be made to a CBSS adds value in a variety of situations, including providing better protection against zero-day attacks[38]. Another part of the value proposition of a security focused CBSS is that unlike a traditional security solution that relies on the

[38] http://en.wikipedia.org/wiki/Zero-day_attack

implementation of a hardware based proxy, a CBSS can also protect mobile workers. The CBSS does this by leveraging functionality that it provides at its POPs as well as functionality in a software agent that is deployed on each mobile device. The use of a Cloud-based solution to provide mobile device management and security was discussed previously in this section.

In many instances, the best security solution is a hybrid solution that combines traditional on-premise functionality with one or more Cloud-based solutions. For example, in many cases IT organizations already have functionality such as web filtering or malware protection deployed in CPE at some of their sites. In this case, the IT organization may choose to implement a CBSS just to protect the sites that don't have security functionality already implemented and/or to protect the organization's mobile workers. Alternatively, an organization may choose to implement security functionality in CPE at all of their sites and to also utilize a CBSS as part of a defense in depth strategy.

Other situations in which a CBSS can serve to either be the only source of security functionality, or to compliment CPE based implementations include cloud-based firewall and cloud-based IPS services. Such a service should support equipment from the leading vendors. Given the previously mentioned importance of hybrid solutions, the service should allow for flexibility in terms of whether the security functionality is provided in the cloud or from CPE as well as for flexibility in terms of who manages the functionality – a CCSP or the enterprise IT organization.

In addition to the specific security functionality provided by the CBSS, the CBSS should also:

- Provide predictive analytics whereby the CBSS can diagnose the vast majority of potential enterprise network and security issues before they can impact network health.

- Incorporate expertise, tools, and processes to ensure that the service that is provided can meet auditing standards such as SAS-70 as well as industry standards such as ITIL.

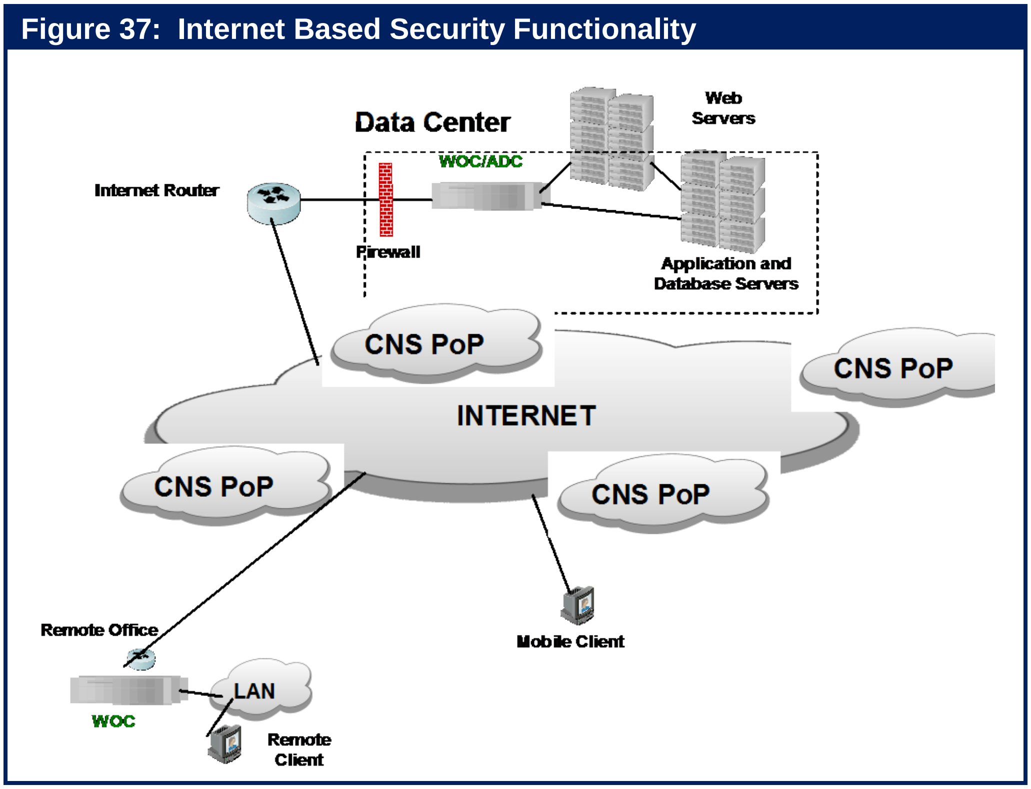
- Integrate audit and compliance tools that provide the necessary event-correlation capabilities and reporting to ensure that the service meets compliance requirements such as Sarbanes-Oxley, HIPAA, GLB and PCI.

- Provide the real-time notification of security events.

Web Application Firewall Services

The section of this report entitled *Network and Application Optimization*, discussed how a Cloud-based service, such as the one shown in **Figure 37**, can be used to optimize the performance of the Internet. As will be discussed in this sub-section of the handbook, that same type of service can also provide security functionality.

Figure 37: Internet Based Security Functionality



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. What is new in the current environment is the use of a CBSS 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.

In the current environment, high-end DDoS attacks can generate 100 Gbps of traffic or more[39]. Attacks of this magnitude cannot be prevented by onsite solutions. They can, however, be prevented by utilizing a CBSS 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

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

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 CBSS that provides Web application firewall functionality 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.
- 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[40], 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.

A CBSS that provides Web application firewall functionality is complimentary to a premise-based Web application firewall. That follows because while the Cloud-based Web application firewall service can perform many security functions that cannot be performed by an on premise Web application firewall, there are some security functions that are best performed by an on premise 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, it can prevent the data from leaving the site.

Evaluating the Security of Cloud Based Services

The primary concern that limits IT organization's use of public cloud services is concerns over security. Realizing that, the following is a set of security focused criteria that IT organizations can use to evaluate CCSP provided services.

- 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?

[40] Wikipedia Tarpit(networking)

- 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?

Conclusions

The following is a summary of the conclusions that were reached in the preceding sections of the handbook.

- IT organizations need to plan for optimization, security and management in an integrated fashion.

- The goal of the 2012 Application and Service Delivery Handbook is to help IT organizations ensure acceptable application and/or service delivery when faced with both the first generation, as well as the emerging second generation of application and service delivery challenges.

- If the applications and networks that support an organization's business processes are not running well, neither are those business processes.

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

- If a business critical application is performing poorly, it has a very significant business impact and it also has a very significant impact on the IT organization.

- Over the next year, the most important optimization task facing IT organizations is optimizing the performance of 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.

- 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.

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

- 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.

- As the complexity increases, so does the vulnerability to security attacks.

- IT organizations are beginning to face a set of new challenges which is expected to significantly complicate the task of ensuring acceptable application and service delivery.
- The BYOD movement has resulted in a loss of control and policy enforcement.

- Adopting BYOD increases a company's vulnerability to security breaches.

- IT organizations are under more pressure for agility than they ever have been in the past.

- In the current environment, public cloud providers sometimes play the role of a shadow IT organization.

- IT organizations can provide a lot of value by acting as a broker of services provided both internally and externally.

- Within most organizations the number of business critical applications is increasing dramatically.

- Over a third of large companies have more than 100 business critical applications.

- The interest on the part of IT organizations to manage services that they acquire from an IaaS vendor has increased over the last year.

- Getting better at managing a business service that is supported by multiple, inter-related applications is an important task for the vast majority of IT organizations

- 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, the number of IT organizations who have implemented at least some desktop virtualization will increase dramatically.

- 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.

- Improving the performance of virtualized desktops is becoming increasingly important to IT organizations.

- 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.

- 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.
- 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.

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

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

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

- 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 currently approach implementing optimization functionality is on a case-by-case basis.

- The deployment of WAN optimization is evolving from being narrowly focused to being broadly focused.

- Getting better at optimizing the transfer of storage data between different data centers is one of the most important optimization tasks facing IT organizations.

- IT organizations have a variety of options for how they acquire WOC functionality.

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

- There is a significant and growing interest on the part of IT organizations to implement integrated WOCs.

- There is broad interest in deploying a wide range of virtual functionality in branch offices.

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

- Effective BW = BW Efficiency x BW Multiplication Factor x Physical BW

- Average Effective BW = 0.75 x 5 x 1 Gbps = 3.75 Gbps

- Although the vast majority of IT organizations currently have a centralized approach to Internet access, IT organizations are continually adopting a more decentralized approach.

- An ADC provides more sophisticated functionality than a SLB does.

- Network appliances such as ADCs are evolving along two paths. One path is comprised of general-purpose hardware, a general-purpose hypervisor and a specialized O/S. The other path is comprised of specialized network hardware, specialized network hypervisors and a specialized O/S.

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

- The primary 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.

- IT organizations need to adopt an approach to management that is based on the assumption that the components of a service, and the location of those components, can and will change frequently.

- 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 roughly half of all 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.

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.

A10 Customer Driven Innovation

AX Series Application Delivery and Advanced Server Load Balancing

Flexibility to Solve Critical Business Challenges

A10 Networks was founded with a mission to be the leader in Application Networking. With the rapid speed of innovation allowed by advances in communication, customers choose A10 Networks to help their applications keep pace.

It is predicted that by 2020, there will be 31 billion devices and four billion people connected to the Internet (source: Intel). This massive and accelerating growth in network traffic is driving Application Networking momentum. As business critical applications continue to grow in number and complexity, intelligent tools are required for efficient performance.

We are only touching the surface for what is possible today, and it is certain that the need for intelligent Application Networking tools will only increase. Predicting this trend, A10 developed a new generation platform with the flexibility to solve critical business challenges for three key initiatives: Any App, Any Cloud and Any Size.

Any App

Web Scalability and Availability

Today's web servers are conduits for complex applications that require intelligence at every layer. If an application is slow or unavailable, or an Internet connection or server goes down, business productivity and profits are lost. A10's flexible Application Networking platforms give customers full control of their web, and any application environment, enabling scalability and availability for all mission-critical applications. In addition, partnerships and certifications with major vendors such as Microsoft, Oracle and VMware, enable rapid and predictable deployments.

IPv4 Exhaustion and IPv6 Migration

Amid rapid network growth, a key challenge is to ensure that expansion can continue unabated for brand protection and uninterrupted business, avoiding costly IT fire drills. A10 delivers powerful, enterprise and carrier class IPv4/IPv6 solutions at attractive price points that will enable organizations to extend and preserve existing IPv4 investments and provide a clear path to IPv6 while enabling communication and connectivity between the two protocols, with many of the largest deployments worldwide.

Any Cloud

Enterprises, Web Giants, Service Providers

With over 2,000 customers across all verticals, including companies such as GE Healthcare, LinkedIn and Microsoft, A10 has focused expertise to service constantly evolving network requirements with a rapid return on investment (ROI). Customer benefit examples include the ability to deploy differentiated customer services, reduce costs through data center consolidation, increase efficiency with large traffic volumes, accelerate web speed to drive customer satisfaction and many more. A10's flexible platform addresses needs for any cloud today, and in the future.

Multi-tenancy and Virtual Clustering

A10 delivers multi-tenancy through advanced high-performance Application Delivery Partitions, allowing customers to provide many services and applications to different groups on a single platform, with full network separation and without any hidden license costs. Any organization sharing the same infrastructure can greatly reduce Total Cost of Ownership (TCO) for Application Networking. Unique clustering technology extends unmatched scaling from millions to billions of connections as required.

Any Size

On-demand Virtual Appliances

A10 offers virtual appliances via hypervisor solutions as alternatives to its hardware platforms. With scale-as-you-grow options in numerous different sizes, A10's virtual machines can be rapidly deployed on commodity hardware, scaling up and down on-demand for changing traffic volumes and use cases.

Scalable and Faster Appliances

At A10, performance is a path to data center efficiency, and not the end itself. With the industry's fastest Application Networking platforms in the most compact form factors, A10's performance delivers overall optimization, ensuring non-stop commerce and applications with lower operational costs. All features are included without licenses so that additional budgets are not needed for new features, allowing for rapid deployments without any license complexity, streamlining internal operations.

Contact us

Contact us today to discuss how A10's

AX Series Application Networking platforms can solve critical business challenges within your mission-critical IT infrastructure: for any app, any cloud or any size.

Aryaka's WAN Optimization as-a-Service Brings a Bold New Direction to the Modern Distributed Enterprise

THE CLOUD has become the next logical step in the evolution of optimizing the enterprise wide area network (WAN) for today's global workforce.

WAN optimization is about improving the performance of business applications over WAN connections. This means matching the allocation of WAN resources to business needs and deploying the optimization techniques that deliver measurable business benefits. Since the WAN is the foundation of the globally connected enterprise, the performance of the WAN is critical to business success.

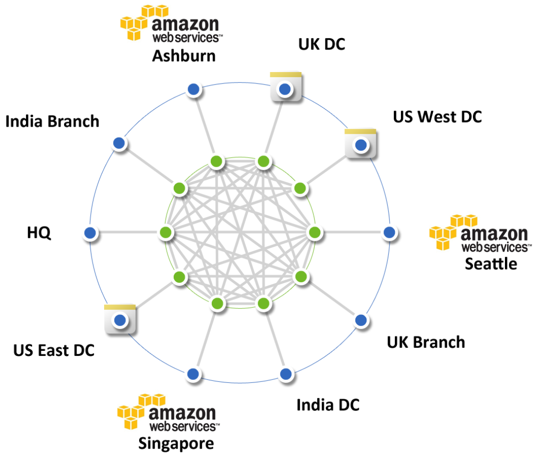
> *"Simplicity is the ultimate sophistication."*
> -Leonardo da Vinci

In the last decade, enterprises seeking to improve application performance across the WAN had little choice but to symmetrically deploy hardware-heavy WAN optimization controllers in data centers and remote locations, invest further in bandwidth, provision MPLS links or a combination of these. These dated solutions do not scale, create other problems and are beyond the affordable reach of 90 percent of the world's businesses. Enterprises suffer inasmuch as underperforming applications have a significant impact on a company's operational performance, including slower access to critical information and higher IT costs.

New cloud-based WAN optimization as-a-Service technology changes all that. This technology better addresses application performance problems caused by bandwidth constraints, latency or protocol limitations. WAN optimization as-a-Service dramatically improves response time of business-critical applications over WAN links and maximizes the return on investment in WAN bandwidth. Enterprises can ensure collaboration and avoid the need for costly, complicated hardware appliances or dedicated MPLS links.

The "Cloud" Defined,
WAN Architecture Redefined

The term "cloud" is intriguing and varied in its description. Vendors within the WAN optimization space and other service providers are trying to find a way to optimize access to the cloud. The only way they can achieve this is by installing another appliance where possible – a virtual appliance – in limited situations within the cloud provider's infrastructure. The cloud for any enterprise can mean public, private or hybrid; it can be data or applications hosted within a private data center or offered as a global on-demand (SaaS) application. Every enterprise requiring optimized access to the cloud will have to install a virtual appliance for each cloud service they need to access, and another few at locations or users that want to access this cloud service.



An Aryaka customer's locations, data centers and Amazon instances are meshed to Aryaka's closest POPs to leverage transport of all traffic across one optimized network.

There is a simpler way to achieve optimized access to cloud services worldwide, irrespective of their purpose and infrastructure location. Aryaka has created multiple Points of Presence (PoPs) across the world connected by a dedicated, secure and highly redundant network. This optimized network connects the enterprise WAN to any cloud service and remote locations worldwide in a simple, CAPEX-free, seamless way without any appliances or dedicated access links.

The cloud has redefined the architecture to optimize the enterprise WAN as the third and most important part needed for the success of compute and storage. Aryaka's purpose-built network drastically increases throughput to reduce the time required and data transmitted between enterprise locations and cloud services. Using compression, deduplication, Quality of Service (QoS) and TCP optimization technologies that are the cornerstones of these optimization solutions, enterprises can experience significant application performance gains 2-100X faster.

Global enterprises leveraging WAN optimization as-a-Service are improving productivity, enhancing collaboration and increasing network and application performance.

Aryaka's WAN optimization as-a-Service solution is sophisticated simplicity. The solution eliminates the need for expensive and complex appliances as well as long-haul connectivity worldwide. With Aryaka's WAN optimization as-a-Service solution, globally distributed teams can communicate and collaborate with the security, reliability, end-to-end visibility and control required by the enterprise.

By SONAL PURI

Optimize and Secure Cloud, SaaS, BYOD, and Social Media

How to Re-architect to Lower Networking Costs and Safely Improve Performance

So many of the dominant trends in applications and networking are driven from outside the organization, including cloud and Software-as-a-Service (SaaS), Bring Your Own Device (BYOD), Internet streaming video, and social networking. These technologies of an Internet connected world are fundamentally changing how we live and work every day. Yet, today's network and security architectures struggle to adapt.

A design that concentrates Internet access at a few data centers and backhauls branch Internet access over the Wide Area Network (WAN) is expensive; it creates overburdened networks and slows the response of both cloud-based and internally delivered applications. The reason this architecture persists is fear. Today's threat landscape has migrated to the web causing many security professionals to prevent direct Internet access at the branch.

But with new cloud-based security solutions from Blue Coat you can re-architect your network to embrace the Internet – safely – and optimize application performance.

First: Re-Architect Branch Connectivity with Cloud-based Security to Lower Costs

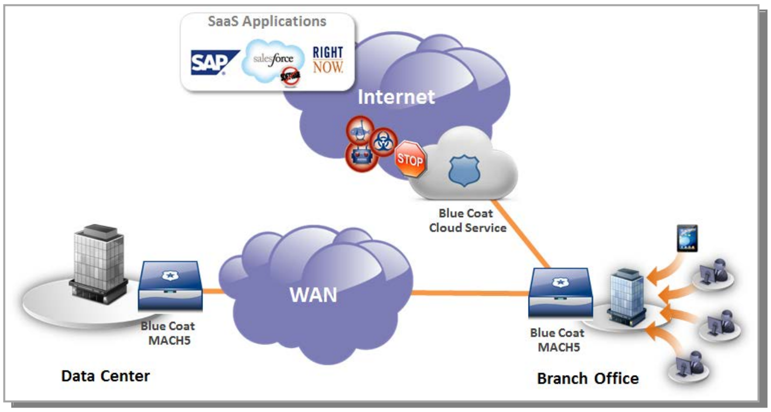
Blue Coat Cloud Service allows you to provide the same enterprise policies and technology to branch and mobile users. By leveraging Blue Coat WebPulse™, a collaborative defense powered by a global community of 75 million users, the Cloud Service is able to deliver real-time protection against the latest web threats from wherever users access the Internet.

WebPulse is based on sound analysis-system design principles:

- Massive input: WebPulse analyzes up to 1 billion web requests per day.
- In-depth analysis: 16 layers of analysis support over 80 categories in 55 languages.
- Granular policy: Up to 4 categories can be applied to each web request for multi-dimensional ratings.
- Speed: Automated systems process inputs – in most cases, in real time.
- Results: This collective intelligence allows WebPulse to block 3.3 million threats per day.

The Cloud Service extends WebPulse protection beyond the WAN, providing secure access to cloud and SaaS for all users at any location. The benefits are clear:

- Lower costs, better performance. By enabling branch Internet, you reduce Internet traffic on the WAN by 60-70%; and directly connected cloud users enjoy better performance.
- The Industry's best analysis and threat detection technology powered by WebPulse provide immediate, continuous protection against known and unknown web threats.
- Universal policy and reporting provides you a single pane of glass to configure policies and report on usage across your entire user base.



Second: Optimize Performance

SaaS, BYOD, Video and Social Media present challenges to network capacity and user patience. Blue Coat WAN Optimization helps overcome these challenges.

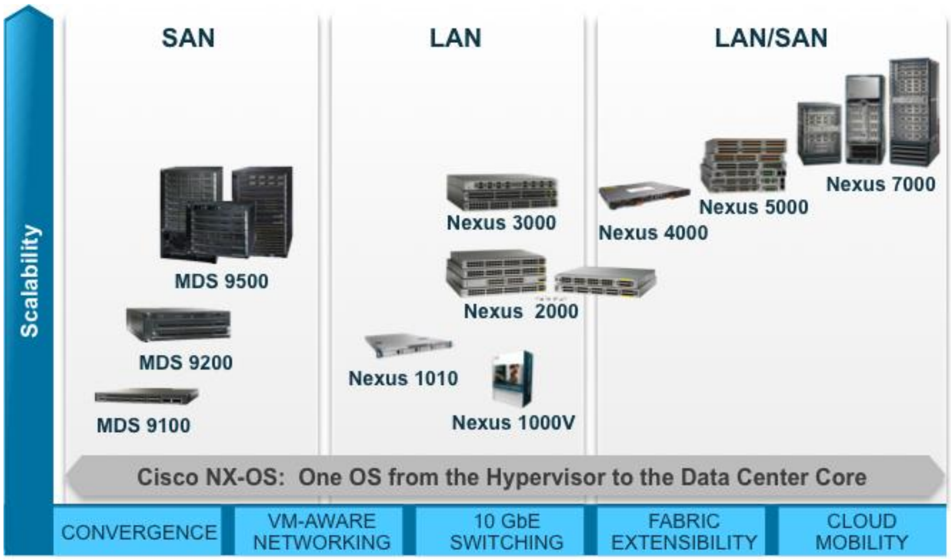
Chatty protocols and multi-megabyte files can hurt SaaS performance. Video requirements destroy capacity plans. Blue Coat's asymmetric, on-demand video caching and live stream splitting boost video capacity up to 500x – whether it's corporate or recreational video. For SaaS, our CloudCaching Engine improves performance by 3-93x, dramatically raising productivity for SaaS users at branch locations.

And now Blue Coat MACH5 technology secures SaaS applications as it accelerates their performance. MACH5 connects directly to the Blue Coat Cloud Service, enforcing SaaS user policies and leveraging WebPulse to scan and filter cloud traffic. Branch users can access applications like SAP, Salesforce, and RightNow without the burden of bandwidth slowdowns or risk of malware threats.

Cisco Unified Fabric

Converged. Scalable. Intelligent.

Cisco Unified Fabric is a flexible, innovative, and proven platform for physical, virtual or cloud deployments. It provides the foundational connectivity within and across data centers so resources are highly available wherever and whenever they are needed.



A key building block for cloud-based environments and virtualized data centers, the Cisco Unified Fabric brings unmatched architectural flexibility and scale to meet the diverse requirements of massively scalable data centers, bare-metal infrastructures, high performance and big data applications.

- Revolutionary fabric scale with over twelve thousand 10 GbE server connectivity with Cisco Nexus
- Highest 10Gb Ethernet density in the industry with Cisco Nexus 7000
- High performance and ultra-low latency networking at scale with Cisco Nexus
- Network services delivered in virtual and physical form factors with Cisco ASA, ASA 1000v, WAAS, vWAAS, VSG and more
- Virtual networking from the hypervisor layer on up with Cisco Nexus 1000v, VSS, VDC, and more
- High availability within and across devices with ISSU, VSS, vPC, and more.
- Flattened and scalable networking at Layer 2 and Layer 3 with Cisco FabricPath, TRILL, L3 ECMP, and more
- Overcome the challenges of expanding networks across locations and the limitations of network segmentation at scale with Cisco OTV, LISP, VXLAN, and more
- Unified operational, control, and management paradigms across the entire fabric with Cisco NX-OS, DCNM and open APIs
- Converged networking to carry every kind of traffic on a single fabric with DCB and FCoE with Cisco Nexus and MDS

Cisco Unified Fabric is a flexible, innovative, and proven platform for physical, virtual or cloud deployments with a non-disruptive, evolutionary approach to create future-proofed, service- and cloud-ready data centers and prevent ‘rip and replace’ for existing data centers. For more info: http://www.cisco.com/go/unifiedfabric

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.**

Beyond the Network…

Application Performance for Business Efficiency

The unique way to guarantee business application performance over the WAN, increase IT productivity and save on IT costs.

Ipanema Technologies – Fact Sheet 2012

www.ipanematech.com

Business Overview

IT departments are witnessing change at a pace never seen before. Transformation is occurring as CIOs seek to access the benefits offered by unified communications, cloud computing, internet-based applications and consolidation, amongst many other strategic projects.

These initiatives are aimed at increasing an enterprise's business efficiency. While they simplify the way IT is delivered to users, they increase the complexity of corporate networking as applications and users rely on the continuous, reliable and consistent flow of data traffic.

Today many organizations are being held back from achieving the true value of their strategic IT programs due to overloaded and poorly understood networks, which were not designed for the symmetric, data-heavy, internet-driven environments that proliferate today. Application usage habits are changing rapidly too. Just a few years ago the extensive use of social media, video and unified communications applications was the exception. For many large enterprises it's now the norm. These new usages and applications have serious implications for the network. The change outlined above can have a dramatic impact, not least on the critical applications that support core functions of the business. Application performance problems including slowness and non- responsiveness impact the user experience and overall productivity of the organization.

In order to protect the business and the significant investments made in transformative applications such as unified communications and SaaS the network must be more intelligent, more responsive and more transparent.

Ipanema at a Glance

- Corporate Headquarters: Paris (France)
- NA Headquarters: Waltham (MA)
- Used by worldwide market leaders across all industry sectors
- Over 150,000 managed sites with many 1,000+ site networks
- Leader for Application-Aware Network services
  (BT, Colt, C&WW, KDDI, KPN, OBS, Telecom Italia, Telefonica, Swisscom, etc.)
- Recognized as "Visionary" by Gartner
- A unique technology (Autonomic Networking) for automatic operations
- A system that tightly integrates all the necessary features
- A management platform that scales to over 400,000 sites

Ipanema automatically drives application performance over the enterprise's WAN from the priority of the business. With Ipanema, enterprises understand which applications run over their network, guarantee the performance they deliver to each user, succeed in their strategic IT transformations - like cloud computing, Unified Communications and hybrid networking - and control Internet traffic growth while reducing their IT expenses.

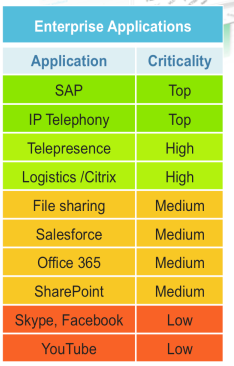
You can get Ipanema products through our distributor and reseller channels. You can also use them "as a Service" through numerous Managed Service Providers and Telecom Operators' offerings. SMBs/SMEs have access to Ipanema through AppsWork, a streamlined cloud service offering.

Solution Overview

Set your objectives and let Ipanema works for you – automatically!

Ipanema's revolutionary self-learning, self-managing and self-optimizing Autonomic Networking System™ (ANS) automatically manages all its tightly integrated features to guarantee the application performance your business requires over the global network:

- Global Application Visibility
- Per connection QoS and Control
- WAN Optimization
- Dynamic WAN Selection
- SLA-based Network Rightsizing

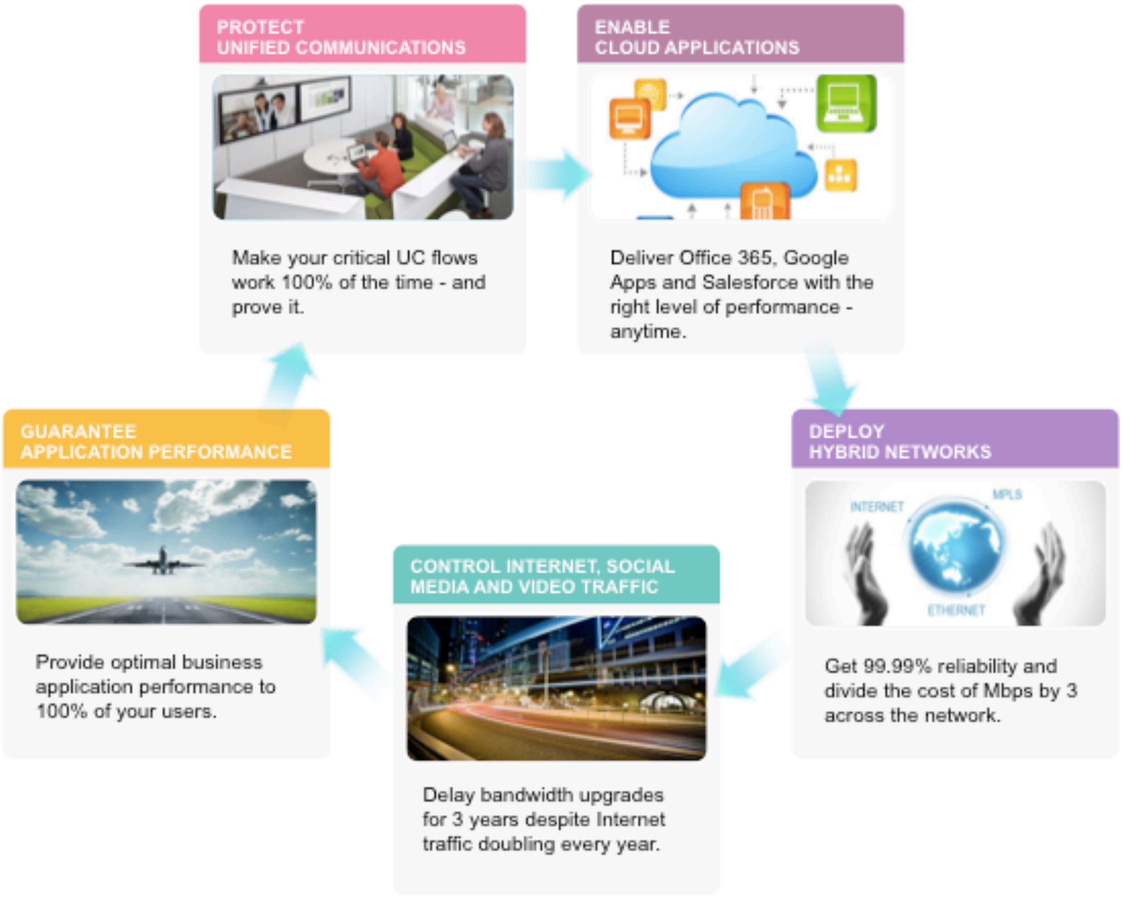


Enterprise Applications	
Application	Criticality
SAP	Top
IP Telephony	Top
Telepresence	High
Logistics /Citrix	High
File sharing	Medium
Salesforce	Medium
Office 365	Medium
SharePoint	Medium
Skype, Facebook	Low
YouTube	Low

Business efficiency requires guaranteed application performance

- Know which applications make use of your network…
- Guarantee the application performance you deliver to users…
- Manage cloud applications, Unified Communications and Internet growth at the same time…
- Do more with a smaller budget in a changing business environment, and to prove it…

With Ipanema, control all your IT transformations



PROTECT UNIFIED COMMUNICATIONS

Make your critical UC flows work 100% of the time - and prove it.

ENABLE CLOUD APPLICATIONS

Deliver Office 365, Google Apps and Salesforce with the right level of performance - anytime.

DEPLOY HYBRID NETWORKS

Get 99.99% reliability and divide the cost of Mbps by 3 across the network.

CONTROL INTERNET, SOCIAL MEDIA AND VIDEO TRAFFIC

Delay bandwidth upgrades for 3 years despite Internet traffic doubling every year.

GUARANTEE APPLICATION PERFORMANCE

Provide optimal business application performance to 100% of your users.

For $3/user/month or less, you guarantee the performance of your business applications… and can save 10 times more!

Ipanema's global and integrated approach allows enterprises to align the application performance to their business requirements. With an average TCO of $3/employee/month, Ipanema directly saves x10 times more and protects investments that cost x100 times more:

- **Application performance assurance:** Companies invest an average of $300/employee/month to implement the applications that support their business. At a mere 1% of this cost, Ipanema can ensure they perform according their application SLAs in every circumstance, maximizing the users' productivity and customers' satisfaction. While they can be seen as "soft money", business efficiency and investment protection are real value to the enterprise.

- **Optimized IT efficiency:** Ipanema proactively prevents most of the application delivery performances problems that load the service desk. It automates change management and shortens the analysis of the remaining performance issues. Global KPIs simplify the implementation of WAN Governance and allow better decision making. This provides a very conservative direct saving of $15/employee/month.

- **Maximized network efficiency:** Ipanema's QoS & Control allows to at least doubling the actual capacity (goodput) of networks, deferring upgrades for several years and saving an average of $15/employee/month. Moreover, Ipanema enables hybrid networks to get access to large and inexpensive Internet resources without compromising the business, typically reducing the cost per Mbps by a factor of 3 to 5.

What our customer say about us

Do more with less

"Whilst data volume across the Global WAN has increased by 53%, network bandwidth upgrades have only grown by 6.3%. With Ipanema in place we have saved $987k this year alone."

Guarantee Unified Communications and increase network capacity

"Ipanema is protecting the performance our Unified Communication and Digital Signage applications, improving our efficiency as well as our customers' satisfaction. Moreover, we have been able to multiply our available capacity by 8 while preserving our budget at the same time."

Reduce costs in a cloud environment

"With Ipanema, we guaranteed the success of our cloud messaging and collaboration deployment in a hybrid network environment, while dividing per 3 the transfer cost of each gigabyte over our global network."

ABOUT IPANEMA TECHNOLOGIES

The Ipanema System enables any large enterprise to have full control and optimization of their global networks; private cloud, public cloud or both. It unifies performance across hybrid networks. It dynamically adapts to whatever is happening in the traffic and guarantees constant control of critical applications. It is the only system with a central management and reporting platform that scales to the levels required by Service Providers and large enterprises. With solutions used extensively by many of the world’s largest telecom providers and enterprises across business and public sectors, Ipanema controls and optimizes over 100,000 sites among 1,000+ customers.

For more information www.ipanematech.com

www.ipanematech.com

Beyond the Network…

Do You Have the Best Choice in Application Delivery?

Overview

The data center has some well known challenges - including application availability, performance and security – problems that can be addressed using Application Delivery Controllers (ADC). However, taking a closer look at businesses whose operations depend on agile and efficient data centers reveals additional challenges. Enterprise data centers need to scale flexibly in a cost-effective manner, ensure connectivity to current and next generation switching infrastructure, provide guaranteed reliability, be able to handle rapid growth and spikes in network traffic, and be capable of harnessing the benefits of virtualized resources and ecosystems. And of course, it goes without saying that all of these requirements must be satisfied while reducing both capital and operational expense.

Radware **Alteon® 5224** is an advanced ADC specifically targeted to address all of these challenges. Offering the very latest in next generation application delivery technology with benchmark affordability, it's simply the best application delivery choice.

Here are four reasons why, we know you'll appreciate:

Reason 1: Unmatched OnDemand Scalability

The Alteon 5224 delivers unmatched on-demand scalability up to 16Gbps based on a simple software license-based mechanism. The platform supports the scaling of throughput capacity, additional advanced features and services (such as global server load balancing, bandwidth management, DoS protection and link optimization), as well as virtual ADC instances without device replacement or restart.

The result is that you pay only for the capacity you need. When you need more you upgrade the device you have and thereby eliminate costly capacity planning exercises and forklift upgrades projects. In contrast, if you were to scale from 1 to 16Gps with an ADC from a different vendor you may need to deploy up to 6 different platforms.

Reason 2: Highest Performance in Class

Alteon 5224 offers the best all round performance metrics – compared to any other competing ADC platform in its class. It is simply the best solution for supporting traffic growth, can process more secured SSL transactions (for both 1024 and 2048 bit keys), and deliver more Connections per Second (CPS). All at the lowest price point available with:

- **3-8x more layer 4 CPS vs. F5** – delivering 500,000 layer 4 CPS
- **4-20x more layer 7 TPS vs. F5** – delivering 200,000 layer 7 TPS
- **1.5-3x more concurrent connections vs. F5** – delivering 12M concurrent connections
- **2.5-7x more SSL CPS (1024 bit keys) vs. F5** – delivering 35,000 SSL CPS
- **4-11x more SSL CPS (2048 bit keys) vs. F5** - delivering 11,200 SSL CPS

Reason 3: The Only Enterprise Grade ADC with 10GE ports

Alteon 5224 is equipped with a total of 26 ports - the highest port density in the industry. This guarantees versatile connectivity options, enabling each Alteon 5224 to connect directly to more server farms or to ensure the physical separation of different networks without the need for intermediate switches. The result is simplified network architectures with fewer devices, reduced electrical and cooling costs, less rack space = greater savings.

In addition, Alteon 5224 offers a unique feature not found on any other 4Gbps ADC on the market: 10GE SFP+ ports. Connection to existing 1GE-interface switches as well as to next-generation 10GE-interface switches is straightforward. So as core switching fabric is refreshed over the next few years, the Alteon 5224 will continue to play well with its neighbors while your investment is protected.

Reason 4: Virtualization Ready for Any Enterprise Size

Looking to virtualize your environment or already there? Alteon 5224 is capable of supporting multiple virtual ADCs on each physical device – each effectively equivalent in capabilities to a physical device.

How does it work? Similar to the concept of sever virtualization, each of the physical devices supplied as part of the Alteon 5224 can host a single ADC service or two ADC services or "instances" (at no additional charge) and can be expanded on-demand to support up to ten fully-independent vADC instances.

In addition, Alteon 5224 enables use of a separate vADC instance per application to ensure high application SLA compliance. The provisioning of additional vADC instances is easy and is achieved once again via on-demand software license updates with no service interruption. And all at a fraction of the cost of deploying additional hardware appliances.

Simply Your Best Application Delivery Choice

The combination of these advantages – along with an industry unique 5-yeaar longevity guarantee – makes Alteon 5224 simply your best application delivery choice. Want to see for yourself? We invite you to download the competitive brief here or contact us at: info@radware.com.

KICK
YOUR NETWORK INTO EARTH-SHATTERING, MIND-BOGGLING HIGH GEAR.