The 2011
Application & Service Delivery Handbook

Part 2: Virtualization, Cloud Computing and Optimizing & Securing the Internet

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

Produced by: Webtorials

Executive Summary

The ***2011 Application and Service Delivery Handbook*** will be published both in its entirety and in a serial fashion. This is the second of the serial publications and it consists of three sections – one on virtualization, one on cloud computing and one on optimizing and securing the Internet. The section on virtualization focuses on three forms of virtualization: server virtualization, desktop virtualization and virtualized appliances. One of the goals of this section is to identify both the current interest and the challenges associated with these three forms of virtualization. Another goal of this section is to discuss the technologies, both existing and emerging, that enable IT organizations to respond to those challenges. The latter discussion will be continued in subsequent sections of this handbook that are devoted to optimization and management.

One goal of the section on cloud computing is to identify the characteristics that are most commonly associated with cloud computing solutions. Another goal of this section is to characterize public and hybrid cloud computing solutions with a focus on topics such as the SLAs that are offered, the availability that is actually provided by these solutions and the interest that IT organization have in better managing and optimizing these solutions. Also included in this section is a discussion of the advantages and challenges that are associated with cloud balancing.

The section on optimizing and securing the Internet introduces a new and rapidly growing category of services that are available from a cloud computing service provider (CCSP). That class of services is traditional network and infrastructure services such as VoIP, unified communications, management, optimization and security. Throughout this handbook, if such a service is provided by a CCSP it will be referred to as a Cloud Networking Service (CNS). One goal of this section is to quantify the interest that IT organizations have in using CNSs. Other goals of this section are to describe some of the performance and security challenges associated with using the Internet and to also describe how CNSs can mitigate these challenges.

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

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 1.

Table 1: Deployment of Virtualized Servers

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

In early 2010, a similar group of IT professionals was asked to indicate the percentage of their data center servers that had already been virtualized. Their responses are shown in Table 2.

Table 2: Deployment of Virtualized Servers as of Early 2010

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

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

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

The Fractal Data Center

As noted in a previous section of the handbook, in the current environment almost every component of IT can be virtualized. One way to think about the current generation of virtualized data centers, and the related management challenges, draws on the concept of a fractal[1]. A fractal is a geometric object that is similar to itself on all scales. If you zoom in on a fractal object

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

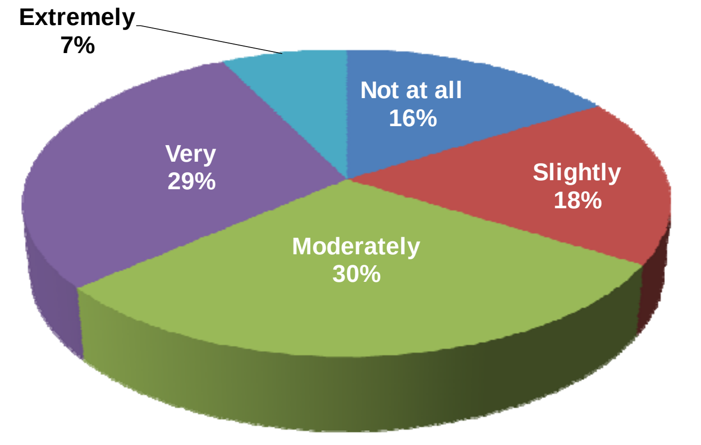
it will look similar or exactly like the original shape. This property is often referred to as self-similarity.

The relevance of fractals is that the traditional data center is comprised of myriad physical devices including servers, LAN switches, probes, WAN optimization controllers (WOCs) and firewalls. The virtualized data centers that most IT organizations are in the process of implementing are still comprised of physical servers, LAN switches and firewalls. However, as shown in Table 1, the vast majority of IT organizations have virtualized at least some of their data center servers. These virtualized data center servers are typically comprised of a wide range of functionality including virtual machines (VMs), a virtual LAN switch (vSwitch) that switches the traffic between the VMs and in many cases devices such as virtual probes, WOCs and firewalls. Hence, if you take a broad overview of the data center you see certain key pieces of functionality. If you were to then zoom inside of a virtualized data center server you would see most, if not all of that same functionality. Hence:

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

One of the challenges that is introduced by the deployment of virtualized servers is that due to the limitations of vSwitches, once a server has been virtualized IT organizations loose visibility into the inter-VM traffic. This limits the IT organization's ability to perform functions such as security filtering or performance monitoring and troubleshooting. To quantify the impact of loosing visibility into the inter-VM traffic, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at managing the traffic that goes between virtual machines on a single physical server. Their responses are shown in Figure 1.

Figure 1: Manage the Inter-VM Traffic

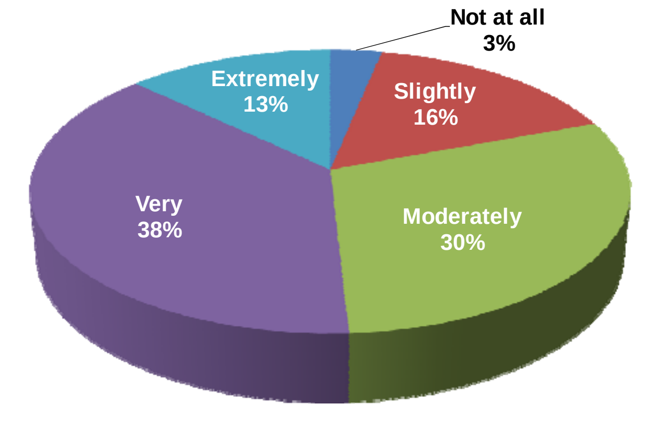


The data in Figure 1 indicates that while there is significant interest in getting better at managing inter-VM traffic, the level of interest is less than the level of interest that The Survey Respondents indicated for many other management tasks

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

To quantify the impact that managing on a per-VM basis is having on IT organizations, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at performing traditional management tasks such as troubleshooting and performance management on a per-VM basis. Their responses are shown in Figure 2.

Figure 2: Managing on a per-VM Basis



One observation that can be drawn from the data in Figure 2 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.

To put the challenge of troubleshooting on a per-VM basis into perspective, consider a hypothetical 4-tier application that will be referred to as BizApp. For the sake of this example, assume that BizApp is implemented in a manner such that the web server, the application server and the database server are each running on VMs on separate servers, each of which have been virtualized using different hypervisors. One challenge that is associated with troubleshooting performance problems with BizApp is that each server has a different hypervisor management system and a different degree of integration with other management systems.

In order to manage BizApp in the type of virtualized environment described in the preceding paragraph, an IT organization needs to gather detailed information on each of the three VMs and the communications between them. For the sake of example, assume that the IT organization has deployed the tools and processes to gather this information and has been able to determine that the reason that BizApp sporadically exhibits poor performance is that the application server occasionally exhibits poor performance.   However, just determining that it is the application server that is causing the application to perform badly is not enough. The IT organization also needs to understand why the application server is experiencing sporadic performance problems. The answer to that question might be that other VMs on the same physical server as the application server are sporadically consuming resources needed by the application server and that as a result, the application server occasionally performs poorly. A way to prevent one VM from interfering with the performance of another VM on the same physical server is to implement functionality such as VMotion[2] that would move a VM to another physical server if performance degrades. However, as discussed in the next sub-section, the dynamic movement of VMs creates a whole new set of challenges.

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

[2] VMotion

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

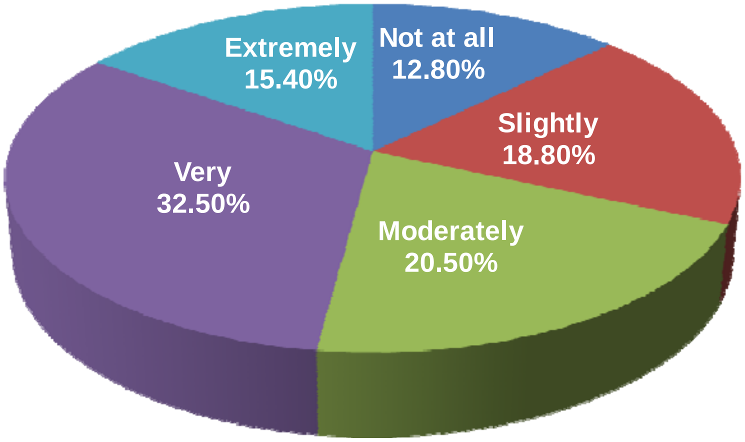
Challenges of Server Virtualization

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

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

In order to quantify the concern that IT organization have with the mobility of VMs, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at supporting the movement of VMs between servers in different data centers. Their responses are shown in Figure 3.

Figure 3: Supporting the Mobility of VMs



Given that the data in Table 1 indicates that IT organizations plan to increase their deployment of virtualized servers, one observation that can be drawn from the data in Figure 3 is that:

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

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

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

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

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

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

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

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

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

- ***Over-subscription of Server Resources***
  With a desire to cut cost, there is the tendency for IT organizations to combine too many VMs onto a single physical server. The over subscription of VMs onto a physical server can result in performance problems due to factors such as limited CPU cycles or I/O bottlenecks. This challenge is potentially alleviated by functionality such as VMotion.

- ***Layer 2 Network Support for VM Migration***
  When VMs are migrated, the network has to accommodate the constraints imposed by the VM migration utility. Typically the source and destination servers have to be on the same VM migration VLAN, the same VM management VLAN, and the same data VLAN.

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

Meeting the Challenges of Server Virtualization

Listed below are some the key developments that can help IT departments meet the challenges of virtualization:

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

- ***Virtualized Performance and Fault Management***
  Virtual switches currently being introduced into the market can export traffic flow data to external collectors. Another approach to monitoring and troubleshooting intra-VM traffic is to deploy a virtual performance management appliance or probe[3] within the virtualized server. A third approach is to access the data in the virtual server management system.

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

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

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

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

Desktop[4] 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 3.

Table 3: Deployment of Virtualized Desktops

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

Comparing the data in Table 3 with the data in Table 1 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 3 with the data in the second row of Table 3 yields the following conclusion:

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

The two fundamental forms of desktop virtualization are:

- Server-side virtualization
- Client-side virtualization

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

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

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

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

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

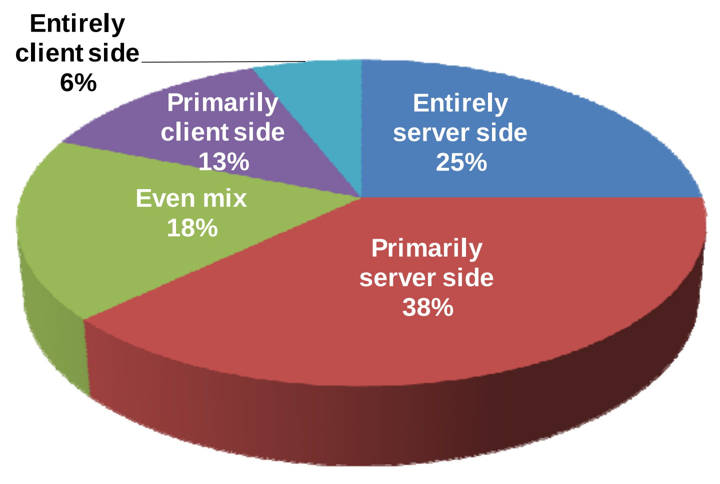
Client-side application virtualization is based on a model in which applications are streamed on-demand from central servers to client devices over a LAN or a WAN. On the client-side, streamed applications are isolated from the rest of the client system by an abstraction layer inserted between the application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

The Survey Respondents whose company will have implemented desktop virtualization by the end of 2011 were asked to indicate which form(s) of desktop virtualization they will have implemented. Their answers are shown in Figure 4.

One conclusion that can be drawn from the data in Figure 4 is:

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

Figure 4: Implementation of Desktop Virtualization



Challenges of Desktop Virtualization

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

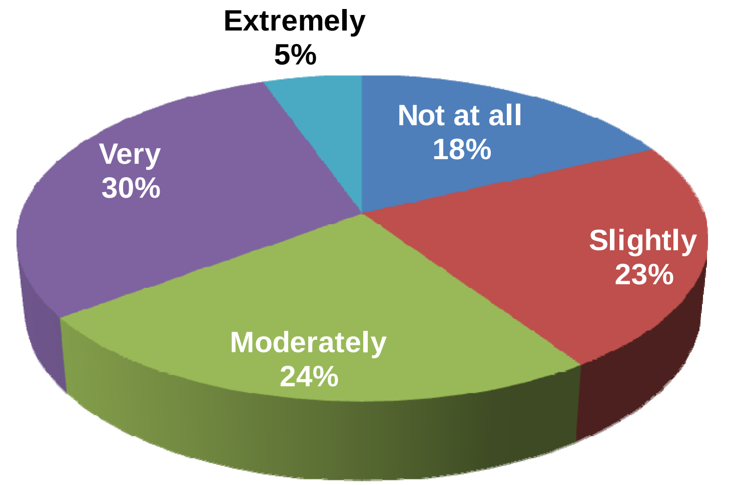
From a networking perspective, the primary challenge in implementing desktop virtualization is achieving adequate performance and an acceptable user experience for client-to-server connections over a WAN.

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 5.

One conclusion that could be drawn from the data in Figure 5 is that getting better at optimizing the performance of virtualized desktops is not that important to IT organizations. However, given that in the current environment there is a very limited deployment of virtualized desktops, and that the forecast is for only a modest increase in deployment, a more viable conclusion is that IT organizations who are implementing virtualized desktops realize the importance of optimizing performance.

Ensuring acceptable performance for VDI presents some significant challenges. One such challenge is that, as is the case in with any TCP based application, packet loss causes the network to retransmit packets. This can dramatically increase the time it takes to refresh a user's screen. While this is a problem in any deployment, it is particularly troublesome in those situations in which there is a significant amount of packet loss.

Figure 5: Optimizing the Performance of Virtualized Desktops



The ICA and RDP protocols employed by many hosted application virtualization solutions are somewhat efficient in their use of the WAN because they incorporate a number of compression techniques including bitmap image compression, screen refresh compression and general data compression. While these protocols can often provide adequate performance for traditional data applications, they have limitations with graphics-intensive applications, 3D applications, and applications that require audio-video synchronization.

Before implementing desktop virtualization, IT organizations need to understand the network implications of that implementation. One of those implications is that other WAN traffic such as large file transfers, can negatively impact the user's experience with desktop virtualization. Another implication is that a large amount of WAN bandwidth may be required. For example, the first two columns of Table 4 show estimates for the amount of WAN bandwidth required by XenDesktop as documented in an entry in The Citrix Blog[5].

Table 4: Bandwidth Requirements from a Representative Branch Office

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

The two rightmost columns in Table 4 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.

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

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

Meeting the Challenges of Desktop Virtualization

As mentioned, protocols such as ICA and RDP have limitations with graphics-intensive applications, 3D applications, and applications that require audio-video synchronization. To respond to the challenges created by these types of applications, Teradici introduced the PC-over-IP (PCoIP) protocol. PCoIP is a proprietary protocol that renders the graphics images on the host computer and transfers compressed pixel level data to the client device. PCoIP is the display protocol used by VMware's View 4 VDI product, which also supports RDP.

While PCoIP resolves some challenges, it also creates others. For example, a document the Teradici published[6] stated that, "To support the lower bandwidth typically available over a WAN, the minimum peak bandwidth required for a PCoIP connection has been reduced to 1 Mbps." While the 1 Mbps required by PCoIP to support a single user represents a worst-case situation, it does underscore the fact that a significant amount of WAN bandwidth can be required to support desktop virtualization. Another challenge associated with PCoIP is that Teradici cannot turn off encryption which makes it difficult, if not impossible, to optimize PCoIP traffic.

As mentioned:

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

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

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

Some of the relevant optimization techniques include:

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

[6] Teradici.com: PCoIP WAN brief

Although virtually all WOCs on the market support the functions listed above, there are some significant differences in terms of how the functionality is implemented and how well it performs. For example, the ICA and RDP protocols can be difficult to optimize for a number of reasons. One of those reasons is that these protocols only send small request-reply packets and this form of communications is best optimized by byte-level caching that is not supported by all WOC vendors. In addition, before implementing any of the techniques listed above, an IT organization must determine which acceleration techniques are compatible with the relevant display protocols. For example, in order to be able to compress ICA traffic, a WOC must be able to decrypt the ICA workload, apply the optimization technique, and then re-encrypt the data stream.

In order to enable the growing population of mobile workers to access enterprise applications, the communications between the mobile worker and the data center has to be optimized. One way to optimize this communications is to deploy client software on the user's mobile device (e.g., laptop, smartphone) that provides WOC functionality. Until recently, the typical device that mobile workers used to access enterprise applications was a laptop. While that is still the most common scenario, today many mobile workers use their smartphones to access enterprise applications.

Over the next few years it is reasonable to expect that many IT organizations will support the use of smartphones as an access device by implementing server-side application virtualization for those devices.

This means that in a manner somewhat similar to remote workers, mobile workers will access corporate applications by running protocols such as ICA and RDP over a WAN.

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

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

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

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

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

Virtual Appliances

Interest in Virtual Appliances

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

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

In a branch office, a suitably placed virtualized server could potentially host a virtual WOC appliance as well as other virtual appliances forming what is sometimes referred to as a Branch Office Box (BOB). Alternatively, a router or a WOC that supports VMs could also serve as the infrastructure foundation of the branch office. Virtual appliances can therefore support branch office server consolidation strategies by enabling a single device (i.e., server, router, WOC) to perform multiple functions typically performed by multiple physical devices.

One of the compelling advantages of a virtualized appliance is that the acquisition cost of a software-based appliance can be notably less than the cost of a hardware-based appliance with same functionality[7]. In many cases the cost of a software-based appliance can be a third less than the cost of a hardware-based appliance. In addition, a software-based client can potentially leverage the functionality provided by the hypervisor management system to provide a highly available system without having to pay for a second appliance[8].

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

A virtual firewall can help IT organizations meet some of the challenges associated with server virtualization. That follows because virtual firewalls can be leveraged to provide isolation between VMs on separate physical servers as well as between VMs running on the same physical server. Ideally, the virtual firewall would use the same software as the physical firewalls already in use in the data center. In addition to firewall functionality, the virtual appliance may

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

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

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

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

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

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

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

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

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

Cloud Computing

Within the IT industry there is not an agreed to definition of exactly what is meant by the phrase ***cloud computing***. This handbook takes the position that it is notably less important to define exactly what is meant by the phrase ***cloud computing*** than it is to identify the goal of cloud computing.

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

The phrase ***good enough*** refers in part to the fact that as described in a following sub-section of the handbook:

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

As a result, the organizations that use these services do so with the implicit understanding that if the level of service they experience is not sufficient, their only recourse is to change providers.

There are several proof points that indicate that the goal of cloud computing as stated above is achievable. For example, an article in Network World identified some of the potential cost savings that are associated with cloud computing[9]. In that article, Geir Ramleth the CIO of Bechtel stated that he benchmarked his organization against some Internet-based companies. As a result of that activity, Ramleth determined that the price that Amazon charges for storage is roughly one fortieth of his internal cost for storage. Ramleth also estimated that YouTube spends between $10 and $15 per megabit/second for WAN bandwidth, while Bechtel is spending $500 per megabit/second for its Internet-based VPN.

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

Additional information on the topic of cloud computing can be found in two reports: *A Guide for Understanding Cloud Computing*[10] and *Cloud Computing: A Reality Check and Guide to Risk Mitigation*[11].

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

[10] Webtorials: A Guide for Understanding Cloud Computing

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

The Primary Characteristics of Cloud Computing

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

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

- Extensive ***virtualization*** of every component of IT. This includes servers, desktops, applications, storage, networks and appliances such as WAN optimization controllers, application delivery controllers and firewalls. The reason that virtualization is so often associated with cloud computing is that virtualization tends to reduce cost and increase the elasticity of service provisioning.

- ***Standardization*** of the IT infrastructure. Complexity drives up cost and reduces agility and elasticity. As such, complexity is the antithesis of cloud computing. One source of complexity is having multiple suppliers of equipment such as switches and routers, as well as having multiple operating systems (i.e., Linux, Windows, Solaris), or even multiple versions of the same network operating system such as IOS.

- ***Simplification*** of the applications and services provided by IT. In a simplified IT environment, the IT organization rarely develops a custom application or customizes a third party application, has just one system for functions such as ERP and SCM, and only supports one version of a given application.

- ***Technology convergence.*** Roughly a year and a half ago, Cisco announced its Unified Computing System[12] (UCS). UCS is intended to enable the convergence of technologies such as servers, networks, storage and virtualization. Cisco's stated rational for technology convergence is to lower the cost and improve the elasticity of the data center infrastructure. Several other vendors either already have, or soon will, announce similar solutions.

- ***Service orchestration*** is an operational technique that helps IT organizations to automate many of the manual tasks that are involved in provisioning and controlling the capacity of dynamic virtualized services. This enables IT to streamline operational workloads and overcome technology and organizational silos and boundaries,

- ***Automation*** of as many tasks as possible; e.g., provisioning, troubleshooting, change and configuration management. Automation can enable IT organizations to reduce cost, improve quality and reduce the time associated with management processes.

- ***Self-service*** allows end users to select and modify their use of IT resources without the IT organization being an intermediary. This concept is often linked with usage sensitive chargeback (see below) as well as the concept of providing IT services on-demand.

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

- ***Usage sensitive chargeback*** is often referred to as pay-as-you-go. One part of the rational for implementing usage sensitive chargeback is that it gives the users greater control over their IT spend because they determine how much of the IT services they consume. Another part of that the rationale is that it enables the IT organization providing the services to focus on what they can control - the unit cost of the services.

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

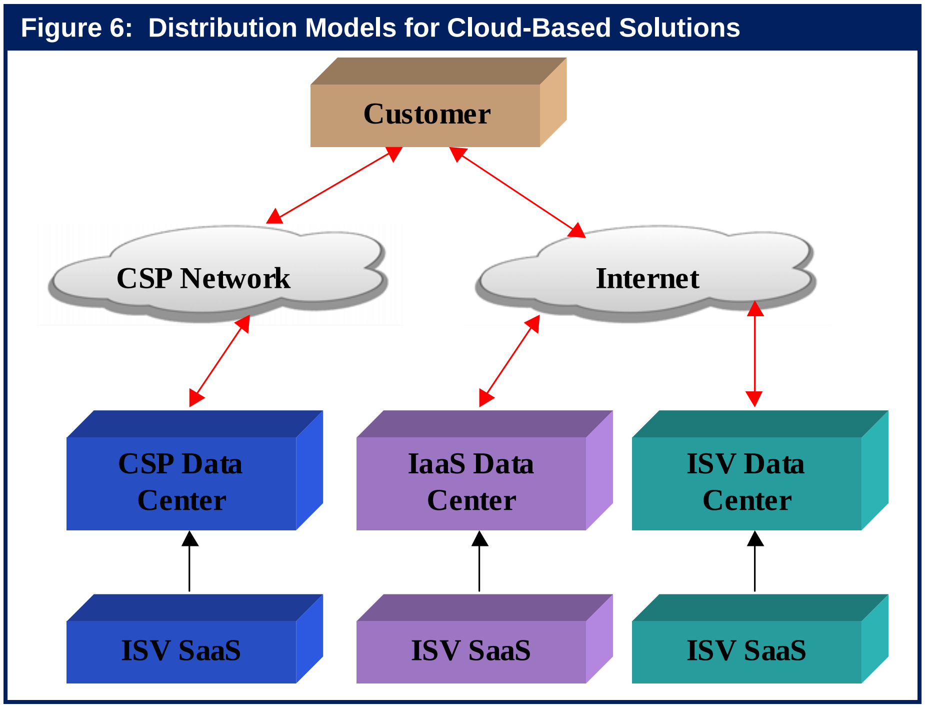
Public Cloud Computing

Background

CCSPs that provide their services either over the public Internet or over other WAN services such as MPLS are offering a class of solution that is often referred to as the *public cloud* or *public cloud computing.* One form of public cloud computing is referred to as Platform-as-a-Service (PaaS). Platform services provide software development environments, including application programming interfaces (APIs) and middleware that abstract the underlying infrastructure in order to support rapid application development and deployment. SalesForce.com provides one of the initial PaaS offerings: Force.com[13]. In April 2011, VMware announced its intention to provide a PaaS offering[14].

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 6 shows some of the common distribution models for SaaS and IaaS solutions. As shown in Figure 6, 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 6: Distribution Models for Cloud-Based Solutions



[13] Salesforce.com: Force.com
[14] CtoEdge: PaaS

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

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

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

SaaS and IaaS

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

SaaS

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

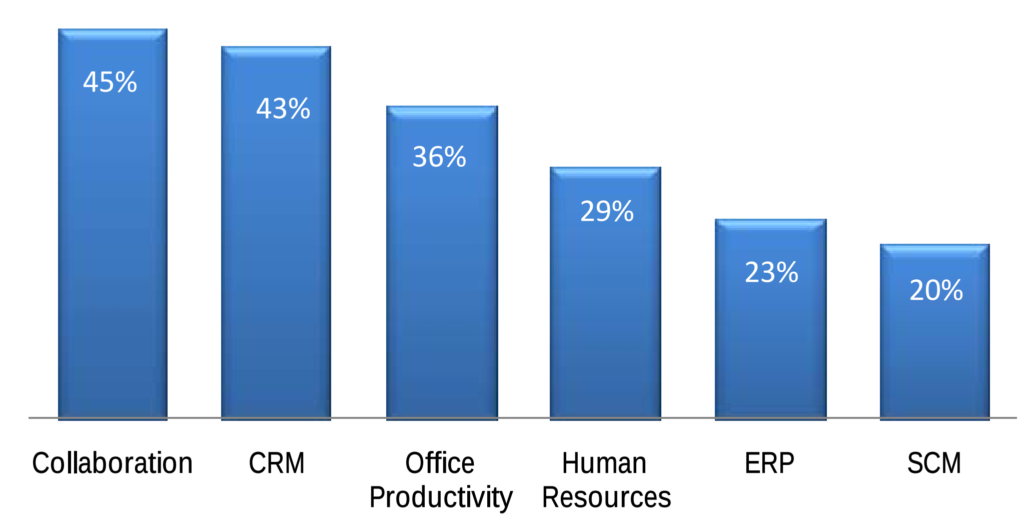
The SaaS marketplace is comprised of a small number of large players such as Salesforce.com, WebEx and Google Docs as well as thousands of smaller players.

According to IDC[15], the Software as a Service (SaaS) market had worldwide revenues of $13.1 billion in 2009 and is projected to reach $40.5 billion by 2014.

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

[15] BusinessWire.com

Figure 7: Popular Categories of SaaS-Based Applications



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

For each category of application shown in Figure 7, there are tens, and sometimes hundreds, of SaaS-based solutions currently available[16]. Table 5 contains a listing of some representative SaaS providers for each category.

Table 5: 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

[16] Saas-showplace.com

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, NTT Communications, Orange Business Services, Rackspace, Savvis, Terremark (recently acquired by Verizon) and Verizon. The IaaS market is expected to exhibit significant growth in the next few years. For example, Gartner[17] estimates that the IaaS market will grow from $3.7 billion in 2011 to $10.5 billion in 2014.

Table 6 provides a high level overview of some of the services offered by IaaS vendors. The data in Table 6 is for illustration purposes only. That follows because it is extremely difficult, if not impossible, to correctly summarize in a table the intricate details of an IaaS solution; e.g., how the solution is priced, the SLAs that are provided and the remedies that exist for when the SLAs are not met. For example, consider the availability of an IaaS solution. On the surface, availability appears to be a well-understood concept. In fact, vendors often have differing definitions of what constitutes an outage and hence, what constitutes availability. For example, within Amazon's EC2 offering an outage is considered to have occurred only when an instance[18] is off line for 5 minutes and a replacement instance cannot be launched from another Availability Zone[19] within Amazon's geographical region. Not all IaaS providers have a similar definition of availability.

Table 6: Representative IaaS Providers

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

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

[17] Qas.com
[18] Amazon.com EC2 Instance types
[19] AWSEC2 UserGuide

Table 6 illustrates that:

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

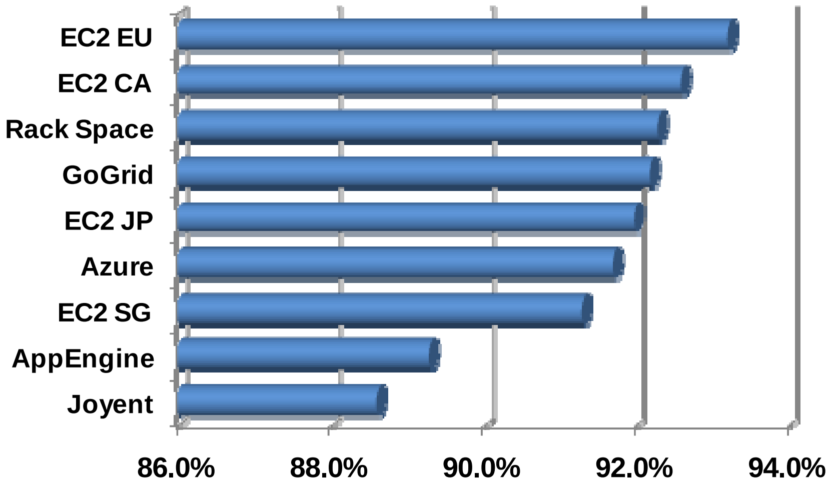
It is important to realize that the value of an availability SLA is only partially captured by the number of 9s it features. A number of factors can cause an SLA that promises four or more 9s of availability to become notably less meaningful. One such factor was previously mentioned – how the vendor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

Insight into the availability of a number of IaaS solutions was provided by Cedexis at the Interop conference in May, 2011[20]. Cedexis presented data that represented roughly 17 billion measurements that were taken between March 15, 2011 and April 15 2011. As shown in Figure 8, none of the IaaS providers that were monitored delivered availability that was greater than 95% (*Source: Cedexis*)

Figure 8 illustrates that:

The availability of IaaS solutions can vary widely.

Figure 8: Availability of Server Instances at Various IaaS Providers



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

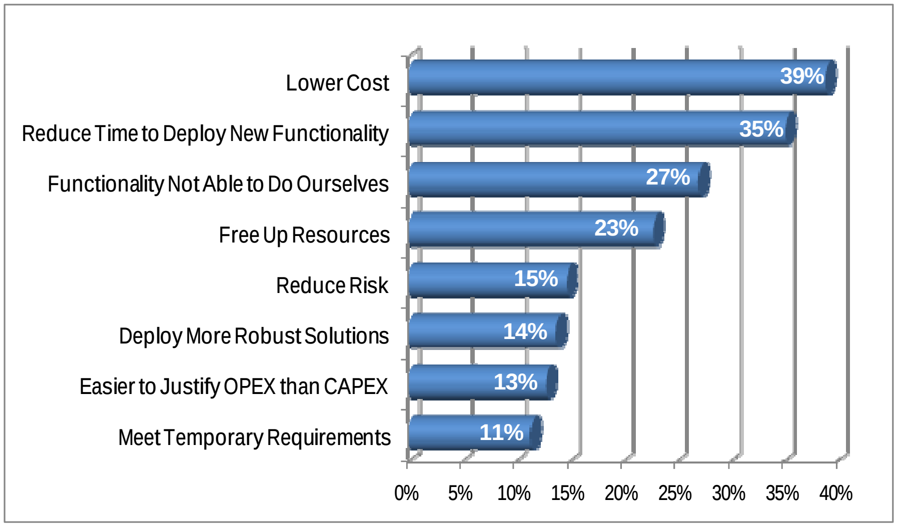
The Drivers of Public Cloud Computing

The Survey Respondents were asked to indicate the two primary factors that are driving, or would likely drive their company to use public cloud computing services. Their responses are shown in Figure 9.

One of the observations that can be drawn from Figure 9 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.

Figure 9: The Drivers of Public Cloud Computing



The next sub-section of this handbook analyzes a topic that is frequently discussed – the risks that are associated with public cloud computing. However, as shown in Figure 9, almost 15% of The Survey Respondents indicated that reducing risk was a factor that would cause them to use a public cloud computing solution. For the most part, their reasoning was that acquiring and implementing a large software application (e.g., ERP, CRM) presents considerable risk to an IT organization and one way to minimize this risk is to acquire the functionality from a SaaS provider.

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

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

The Risks Associated with Public Cloud Computing

One of the risks associated with public cloud computing is performance. One of the causes of those performance problems is that most cloud computing platforms (i.e., Amazon's EC2) are built on a small number of large data-centers that users access over the Internet. As a result of this design, the majority of users of the platform are a considerable distance removed from the data-center. As is always the case, the user's experience tends to degrade as the user gets further removed from the data-center. Even users who are close to the data-center can be subject to unacceptable performance as a result of sub-optimal routing and the inefficient protocols used within the Internet.

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

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

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

Private cloud computing has the advantage of not being burdened by many of the potential security vulnerabilities, data confidentiality and control issues that are associated with public cloud computing. Because of that fact, this section will focus on three categories of risk that are associated with public cloud computing and that IT organizations need to evaluate prior to using public cloud computing services.

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

Security

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

- What are the CCSP's policies and procedures relative to data recovery?

- What procedures does the CCSP have in place to avoid issues such as virus attacks, Cross-site scripting (XSS) and man in the middle intercepts?

- How well trained and certified is the CCSP's staff in security matters?

Management

- What is the ability of the CCSP to manage the challenges associated with virtualization that were discussed in the preceding section of this handbook?

- What management data will the CCSP make available to the IT organization?

- What is the ability of the CCSP to troubleshoot performance or availability issues?

- What are the CCSP's management methodologies for key tasks such as troubleshooting?

- Does the CCSP provide tools such as dashboards to allow the IT organization to understand how well the service they are acquiring is performing?

- Does the CCSP provide detailed information that enables the IT organization to report on their compliance with myriad regulations?

- What are the primary management tools that the CCSP utilizes?

- What is the level of training and certification of the CCSP's management personnel?

- What are the CCSP's backup and disaster recovery capabilities?

- What approach does the CCSP take to patch management?

- What are the specific mechanisms that the IT organization can use to retrieve its data back in general and in particular if there is a dispute, the contract has expired or the CCSP goes out of business?

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

- Will the CCSP allow the IT organization to test the data retrieval mechanisms on a regular basis?

- What is the escalation process to be followed when there are issues to be resolved?

- How can the service provided by the CCSP be integrated from a management perspective with other services provided by either another CCSP and/or by the IT organization?

- How can the management processes performed by the CCSP be integrated into the end-to-end management processes performed by the IT organization?

Performance

- What optimization techniques has the CCSP implemented?
- What ADCs and WOCs does the CCSP support?
- Does the CCSP allow a customer to incorporate their own WOC or ADC as part of the service provided by the CCSP?
- What is the ability of the CCSP to identify and eliminate performance issues?
- What are the procedures by which the IT organization and the CCSPs will work together to identify and resolve performance problems?
- What is the actual performance of the service and how does that vary by time of day, day of week and week of the quarter?
- Does the IT organization have any control over the performance of the service?
- What technologies does the CCSP have in place to ensure acceptable performance for the services it provides?
- Does the CCSP provide a meaningful SLA? Does that SLA have a goal for availability? Performance? Is there a significant penalty if these goals are not met? Is there a significant penalty if there is a data breach?
- To what degree is it possible to customize an SLA?
- What is the ability of the CCSP to support peak usage?
- Can the CCSP meet state and federal compliance regulations for data availability to which the IT organization is subject?

Managing and Optimizing Public Cloud Computing

As previously noted, in the current environment there are significant limitations to the steps that an IT organization can take to either manage or optimize a solution that involves one or more CCSPs. In spite of that limitation, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at monitoring and managing storage, compute and application services that they acquire from a CCSP. Their responses are shown in Table 7.

Table 7: The Importance of Managing Public Cloud Services

	Storage	Compute	Applications
Extremely	2.9%	8.1%	9.4%
Very	20.0%	20.7%	30.8%
Moderately	28.6%	25.2%	23.9%
Slightly	20.0%	22.5%	18.8%
Not at All	28.6%	23.4%	17.1%

The Survey Respondents were also asked how important it is for their IT organization over the next year to get better at optimizing the storage, compute and application services that they acquire from a CCSP. Their responses are shown in Table 8.

Table 8: The Importance of Optimizing Public Cloud Services

	Storage	Compute	Applications
Extremely	4.4%	3.1%	7.8%
Very	16.7%	17.7%	28.2%
Moderately	23.3%	26.0%	25.2%
Slightly	28.9%	26.0%	20.4%
Not at All	26.7%	27.1%	18.4%

There are many conclusions that can be drawn from the data in Table 7.and Table 8. One of which is that getting better at managing and optimizing SaaS solutions is more important to IT organizations than is getting better at managing and optimizing IaaS solutions. One reason for that situation is that IT organizations make more use of SaaS solutions than they do IaaS solutions. Another observation is that getting better at managing and optimizing SaaS and IaaS solutions is less important to IT organizations than is getting better at many other management and optimization tasks. One reason for that situation is that in the current environment it is often impossible to effectively manage and/or optimize a SaaS or an IaaS solution.

Private and Hybrid Cloud Computing

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

In those instances in which an enterprise IT organization uses a mixture of public and private cloud services, the result is often referred to as a *Hybrid Cloud*. The hybrid cloud approach can offer the scalability of the public cloud coupled with the higher degree of control offered by the private cloud. Hybrid clouds, however, do present significant management challenges. For example, the preceding section of the handbook discussed a hypothetical 4-tier application that was referred to as BizApp. As that section pointed out, it is notably more difficult to troubleshoot BizApp in a virtualized environment than it would be to troubleshoot the same application in a traditional environment. Now assume that BizApp is deployed in such a way that the web tier is supported by a CCSP and the application and database tiers are provided by the IT organization. This increases the difficulty of management yet again because all of the management challenges that were discussed previously still exist and added to them are the challenges associated with having multiple organizations involved in managing the application.

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

To quantify the concerns that IT organizations have in managing cloud computing environments, The Survey Respondents were asked to indicate how important it was over the next year for their organization to get better at managing private, hybrid and public cloud computing solutions. Their responses are shown in Table 9.

Table 9: Importance of Managing Cloud Solutions

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

One observation that can be drawn from the data in Table 9 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.

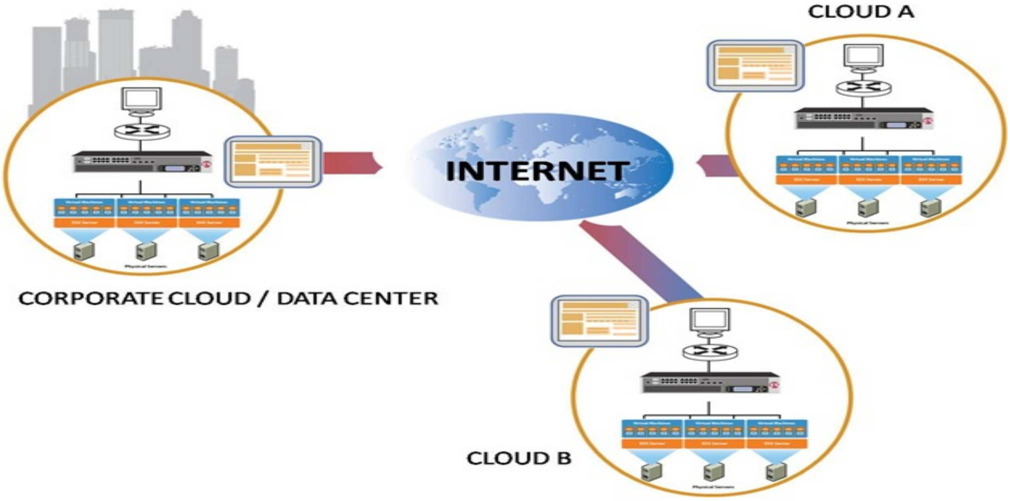
Cloud Balancing

Background

Cloud balancing refers to routing service requests across multiple data centers based on myriad criteria. As shown in Figure 10, 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 10: Cloud Balancing



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

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

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

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

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

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

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

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

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

  The global infrastructures of large cloud providers provide an opportunity for cloud balancing without the complexity of dealing with multiple providers. For example, Amazon EC2 locations are composed of Regions and Availability Zones. Availability Zones are distinct locations that are engineered to be insulated from failures in other Availability Zones and are provided with low latency network connectivity to other Availability Zones in the same Region. In theory, cloud balancing across Availability Zones or Regions can greatly reduce the probability of outages within the Amazon AWS global cloud. However, an outage that Amazon suffered in April 2011 gave the indication that the Availability Zones didn't provide the promised protection[21].

Enabling Cloud Balancing

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

[21] TheRegister.co.uk

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

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

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

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

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

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

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

Optimizing and Securing the Use of the Internet

Figure 6 in the preceding section of the handbook highlights some of the common distribution models for cloud based services. As was discussed in that section, it isn't possible to provide end-to-end quality of service over the Internet. The inability to ensure the end-to-end performance of the Internet can be a problem, however, whether or not it is a cloud based service that is being supported.

Figure 7 in the preceding section of the handbook identifies some of the popular categories of applications that can be acquired from a CCSP. The applications identified in that figure are well known enterprise applications including CRM, SCM and ERP. For the last few years, this is the class of applications that has been most closely associated with cloud computing. While those applications will continue to be closely associated with cloud computing, it is becoming increasingly common for organizations to acquire a different category of application from a CCSP. That class of applications is traditional network and infrastructure services such as VoIP, unified communications, management, optimization and security. Throughout this document, if such a service is provided by a CCSP it will be referred to as a Cloud Networking Service (CNS).

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

The Interest in CNSs

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

Table 10: Interest in Cloud Networking Services

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

Table 10: Interest in Cloud Networking Services

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

The data in Table 10 shows that there is strong interest in a number of CNSs – most notably application hosting and VoIP. The interest in CNSs, however, is quite broad as over twenty-five percent of The Survey Respondents indicated that over the next year that each of the services listed in the top seven rows of Table 10 would either likely be acquired or would be acquired. That represents the beginning of what could be a fundamental shift in terms of how IT services are provisioned.

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

The factors that are driving IT organizations to consider CNSs are the same factors that are driving companies to consider any cloud computing service. Those factors were identified in Figure 9 of the preceding section of the handbook and include:

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

The Use of the Internet

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

The growing attractiveness of the Internet is due in part to the fact that it is a lower cost alternative to WAN services such as Frame Relay and MPLS, and in part to the fact that for some of the enterprise's user constituencies (e.g., customers, suppliers, distributors) the Internet is the only viable WAN connectivity option. As the boundaries of the typical enterprise continue to be blurred due to an increasingly diverse user community, as well as the adoption of new distributed application architectures (e.g., Web-enabled applications and business processes, SOA/Web Services, Cloud Computing) that often traverse multiple enterprises, enterprise usage of the Internet will continue to increase at a significant rate.

Over the last few years that IT organizations have focused on ensuring acceptable application delivery, the vast majority of that focus has been on either making some improvements within the data center or on improving the performance of applications that are delivered to branch office employees over private WAN services[22].

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

Optimizing the delivery of applications that transit the Internet requires that flows be optimized within the Internet itself. This in turn requires subscription to a CNS that provides that functionality. These Internet optimization services are based primarily on proprietary application acceleration and WAN optimization servers located at points of presence (PoPs) distributed across the Internet and like most cloud based services, they don't require that remote sites accessing the services have any special hardware or software installed. The benefits of these services include complete transparency to both the application infrastructure and the end-users. This transparency ensures the compatibility of the Internet optimization service with complementary application acceleration technologies provided by WAN optimization controllers (WOCs) or application delivery controllers (ADCs) deployed in the data center or at remote sites.

The Limitations of the Internet

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

The primary reason for the limitation of the Internet is that as pointed out by Wikipedia[23], the Internet "Is a 'network of networks' that consists of millions of private and public, academic, business, and government networks of local to global scope." In the case of the Internet, the only service providers that get paid to carry Internet traffic are the providers of the first and last mile services. All of the service providers that carry traffic between the first and last mile do so without compensation. One of the affects of this business model is that there tends to be availability and performance bottlenecks at the peering points. Another affect is that since there is not a single, end-to-end provider, service level agreements (SLAs) for the availability and performance of the Internet are currently not available and are unlikely to ever be available.

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

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

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

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

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

TCP has a number of characteristics that can cause the protocol to perform poorly when run over a lossy, high latency network. The Survey Respondents recognized this fact. As previously mentioned, over 80% of The Survey Respondents indicated that over the next year it was at least moderately important to their organization to get better at optimizing the performance of TCP.

One of the characteristics of TCP that can lead to poor performance is TCP's retransmission timeout. This parameter controls how long the transmitting device waits for an acknowledgement from the receiving device before assuming that the packets were lost and need to be retransmitted. If this parameter is set too high, it introduces needless delay as the transmitting device sits idle waiting for the timeout to occur. Conversely, if the parameter is set too low, it can increase the congestion that was the likely cause of the timeout occurring.

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

The affect of packet loss on TCP has been widely analyzed[24]. 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 11: Factors that Impact Throughput

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

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

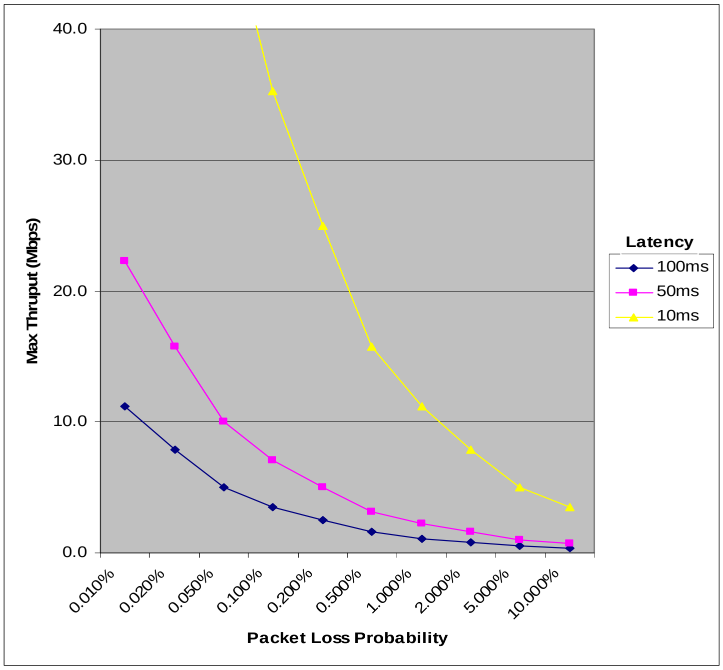
The preceding equation shows that throughput decreases as either the RTT or the packet loss rate increases. To illustrate the impact of packet loss, assume that MSS is 1,420 bytes, RTT is 100 ms. and p is 0.01%. Based on the formula, the maximum throughput is 1,420 Kbytes/second. If however, the loss were to increase to 0.1%, the maximum throughput drops to 449 Kbytes/second. Figure 12 depicts the impact that packet loss has on the throughput of a single TCP stream with a maximum segment size of 1,420 bytes and varying values of RTT.

One conclusion we can draw from Figure 12 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 12: Impact of Packet Loss on Throughput

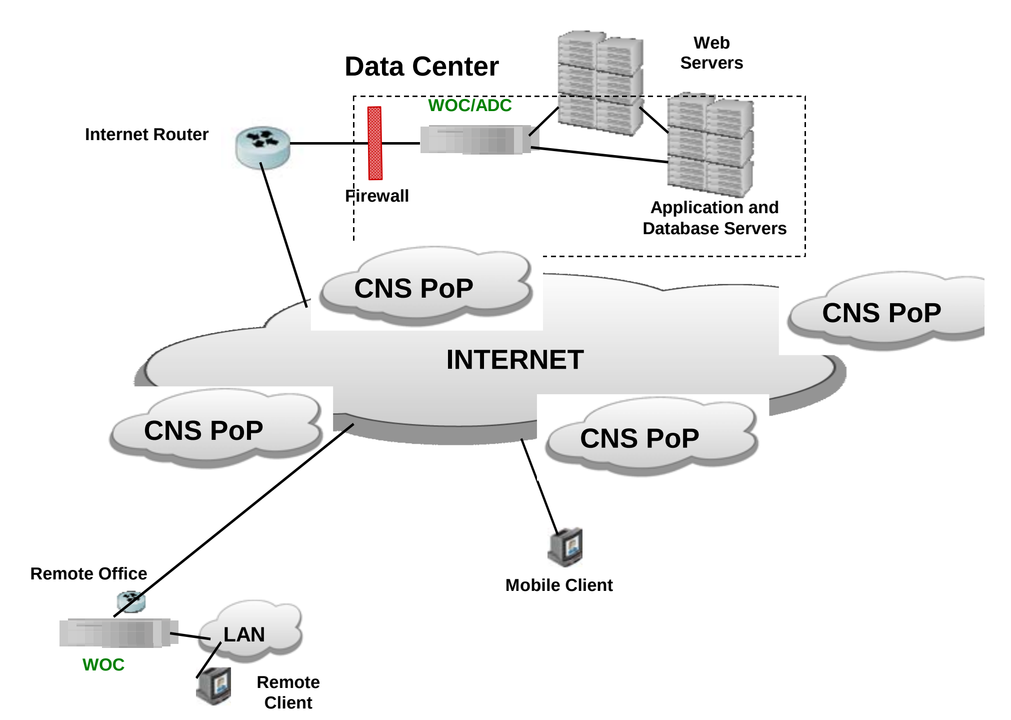


[24] The macroscopic behavior of the TCP congestion avoidance algorithm by Mathis, Semke, Mahdavi & Ott in Computer Communication Review, 27(3), July 1997

Internet-Based Application Delivery Optimization

The traditional classes of application delivery solutions (e.g., ADC, WOC) that are described in detail in a subsequent chapter of the handbook were designed to address application performance issues at both the client and server endpoints. These solutions make the assumption that performance characteristics within the WAN are not capable of being optimized because they are determined by the relatively static service parameters controlled by the WAN service provider. This assumption is reasonable in the case of private WAN services. However, this assumption does not apply to enterprise application traffic that transits the Internet because there are significant opportunities to optimize performance within the Internet itself based on the use of a CNS. Such a CNS would out of necessity leverage service provider resources that are distributed throughout the Internet in order to optimize the performance, security, reliability, and visibility of the enterprise's Internet traffic. As shown in Figure 13, all client requests to the application's origin server in the data center are redirected via DNS to a server in a nearby PoP that is part of the CNS. This edge server then optimizes the traffic flow to the CNS server closest to the data center's origin server.

Figure 13: The Internet Infrastructure for a CNS



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

- ***Route Optimization***
  Route optimization is a technique for circumventing the limitations of BGP by dynamically optimizing the round trip time between each end user and the application server. A route optimization solution leverages the intelligence of the CNS servers that are deployed in the service provider's PoPs to measure the performance of multiple paths through the Internet and to choose the optimum path from origin to destination. The selected route factors in the

degree of congestion, traffic load, and availability on each potential path to provide the lowest possible latency and packet loss for each user session.

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

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

  There is a strong synergy between route optimization and transport optimization because both an optimized version of TCP or a higher performance transport protocols will operate more efficiently over route-optimized paths that exhibit lower latency and packet loss.

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

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

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

Visibility

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

Web Application Firewall Services

As previously described, one of the characteristics of the current environment is the shifting emphasis and growing sophistication of cyber crime.

Role of a Traditional Firewall: Protect the Perimeter

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

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

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

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

There are fundamental flaws with an approach to security that focuses only on the perimeter of the organization. To overcome these flaws, most IT organizations have moved to an approach to security that is typically referred to as *defense in depth*. The concept of defense in depth is not new. This approach was widely used during the Application Delivery 1.0 era as IT organizations often deployed multiple layers of security functionality including virus scanning, authentication, firewalls, intrusion detection systems and intrusion protection systems.

In the Application Delivery 1.0 era, however, all of the layers of security functionality were typically deployed onsite. What is new in the Application Delivery 2.0 era is the use of a CNS to provide Web application firewall functionality that is distributed throughout the Internet. This means that Web application functionality is close to the source of security attacks and hence can prevent many security attacks from reaching the organization. The distribution of security functionality as part of a CNS is analogous to the distribution of optimization functionality as part of a CNS that was discussed in the preceding subsection.

In the current environment, high-end DDoS attacks can generate 100 Gbps of traffic or more[25]. Attacks of this magnitude cannot be prevented by onsite solutions. They can, however, be prevented by implementing a CNS that includes security functionality analogous to what is provided by a Web application firewall and that can identify and mitigate the DDoS-related traffic close to attack traffic origin.

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

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

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

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

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

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

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

- Block or redirect requests based on characteristics such as the originating geographic location and whether or not the originating IP addresses are on either a whitelist or a blacklist.

- Direct traffic away from specific servers or regions under attack.

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

- Issue slow responses to the machines conducting the attack. The goal of this technique, known as tarpits[26], is to shut down the attacking machines while minimizing the impact on legitimate users.

- Direct the attack traffic back to the requesting machine at the DNS or HTTP level.

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

[26] Wikipedia Tarpit(networking)

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.

Advertorial

A10 Networks

The Fastest Growing Application Networking Company

64-bit AX Series

Application Delivery

- Advanced Application Delivery Controller (ADC)
- New Generation Server Load Balancer (SLB)

IPv6 Migration

- Large Scale NAT
- Dual-Stack Lite
- NAT64 & DNS64
- IPv6 ↔ IPv4 (SLB-PT)

Cloud Computing & Virtualization

- SoftAX & AX-V
- AX Virtual Chassis
- AX Virtualization (Application Delivery Partitions)

Advanced Core Operating System (ACOS)

AX Series Advantage

- All inclusive pricing for hardware appliances, no performance or feature licenses
- Most scalable appliances in the market with unique modern 64-bit ACOS, solid-state drives (SSD) and multiple hardware acceleration ASICs
- Faster application inspection with aFleX TCL rules
- aXAPI for custom management

Application Solutions

The AX Series increases scalability, availability and security for enterprise applications. Visit A10's web site for deployment guides, customer usage scenarios and to participate in the Application Delivery Community.

Blackboard Microsoft Microsoft Exchange Microsoft Lync IMPERVA JUNIPER NETWORKS ORACLE vmware READY

www.a10networks.com

Transforming the Internet into a Business-Ready Application Delivery Platform

Akamai

Ensuring applications perform to support your business goals

As organizations expand globally, they need to make a variety of business-critical applications available to employees, partners and customers across the globe. Application delivery strategies are increasingly leveraging Cloud based options for hosting enterprise applications on Cloud infrastructure and outsourcing applications via SaaS vendors. Organizations must also be sensitive to the economic pressures driving IT consolidation and centralization initiatives.

Whether delivering applications from behind the firewall, hosting in the cloud, or using a hybrid model, the Internet remains an integral part of application delivery strategy. Though global delivery of enterprise applications over the Internet can provide remote users with essential business capabilities, poor application performance can quickly sour user experience. Business applications must perform quickly, securely, and reliably at all times, or adoption and intended benefits will suffer.

Key Challenges in Delivering Applications

IT organizations often use the public Internet to support globalization efforts because of its lower cost, quick time to deploy, and expansive reach. However, when delivering applications via the Internet to global users, business can face many challenges, including:

- Poor performance due to high latency and chatty protocols (like HTTP & XML)
- Spotty application availability caused by unplanned internet disruptions
- Inadequate application scalability and spiky peak usage
- Growing security threats, including distributed denial of service, cross-site scripting, and SQL injections

These problems can severely undermine application effectiveness and ROI and do not disappear by moving to the Cloud.

Akamai's Application Performance Solutions

Today, thousands of businesses trust Akamai to distribute and accelerate their content, applications, and business processes. Akamai Application Performance Solutions are a portfolio of fully managed services designed to accelerate performance and improve reliability of any application delivered over the Internet, hosted behind the firewall or in the Cloud, with no significant IT infrastructure investment.

Akamai leverages a highly distributed intelligent Internet platform, comprised of tens of thousands of servers, within a single network hop of 90% of the world's Internet users. The Akamai Protocol optimizes application delivery at the routing, transport, and application layers, not only caching content at the Internet's edge, close to end users, for fast delivery, but accelerating dynamic content from the origin to global users. This intelligent Internet platform also extends the security perimeter to the edge of the Internet with modules providing a cloud based Web Application Firewall and DDoS defense.

Application Performance Solutions drive greater adoption through improved performance, higher availability, and an enhanced user experience, ensuring consistent application performance, regardless of user location, and delivers capacity on demand, where and when it's needed. This helps reduce infrastructure costs and support data center consolidation. Examples of applications delivered by Application Performance Solutions include Web-based enterprise applications, Software as a Service (SaaS), applications deployed on IaaS and Paas, Web services, client/server or virtualized applications, live chat, productivity, and administration functions, such as secure file transfers.

To learn more about Akamai Application Performance Solutions, visit www.akamai.com/aps.

Next-Generation Application Acceleration

Blue Coat

Organizations everywhere face tough challenges in optimizing business application performance. For today's distributed enterprises, centralization and server consolidation can create user response and network capacity problems; business applications are often slow or unpredictable; and bandwidth costs are out of control. Now, IT is expected to deliver even more — including corporate communication videos and cloud delivered software-as-a-service (SaaS) applications — all while containing costs.

To solve these and other application delivery problems, you have to understand how application performance requirements have changed, know which technologies can meet your business demands today and prepare for capacity needs down the road.

The Foundation: Optimizing Traditional Applications

Rapid growth of files, email, storage and backup systems put an incredible burden on WAN connections and create significant end-user performance issues — unless you can accelerate traffic. Blue Coat's protocol optimization, byte caching, compression and QoS are the technologies required to accelerate remote and branch office access to centralized files, email and backup systems. These technologies offer significant performance benefits by mitigating the latency caused by chatty file protocols, caching data and expanding bandwidth for high-volume transfers. Besides data applications, however, you need specialized technologies to optimize performance of key emerging applications.

Next Generation WAN Optimization Requirements

Many of the latest applications are changing the way we collaborate, educate, and communicate. Video, for instance, is increasingly used for training and live communications, and Cloud delivered SaaS applications are enabling new business processes. However, the traditional acceleration technologies cannot address these newer types of applications.

Streaming video and rich media

Delivering high-quality, on-demand or live streaming video requires massive amounts of bandwidth on specialized protocols. For example, a single live stream can be 200KB to 1.5MB and large on-demand files can reach 25MB, 100MB and even 1GB in size. In addition, bandwidth-hungry rich media applications can dominate the entire network and still fail due to insufficient resources.

Cloud Delivered SaaS applications

SaaS applications, such as Salesforce.com, or SaaS-hosted SAP and SharePoint applications have unique management challenges due to their location and the encryption used to secure them. Because SaaS offerings are located outside of your network they are outside of your control, but still need to be accelerated. They are also encrypted with SSL and use certificates and keys controlled by the SaaS provider and the Web browser – not your organization.

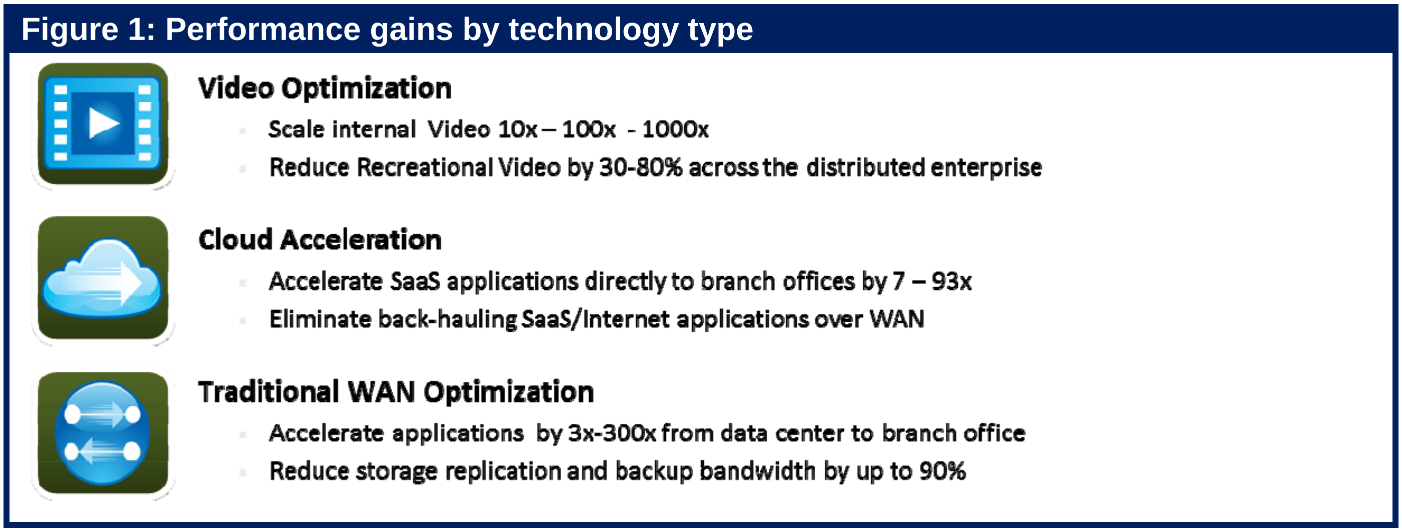
Traditional WAN Optimization technologies would require you to place an appliance on the SaaS provider's network, which is simply not possible. Because SaaS applications rely on HTTP and SSL delivery, you need optimization technologies that can asymmetrically accelerate HTTP and SSL, as well as secure client-side certificate handling so you can decrypt and accelerate the sessions.

Next-Generation Acceleration

The good news is next generation acceleration technologies available today can help you optimize your most critical applications and reclaim bandwidth from non-essential traffic. These new optimization technologies include:

- Video caching, stream splitting and Content Delivery Network (CDN) to enable optimized delivery of business video and minimize the impact of recreational video over the WAN.
- Asymmetric optimizations technologies and external SSL certificate handling that don't require changes to the SaaS infrastructure, like Blue Coat CloudCaching engine.
- URL classification and content filtering with usage and QoS policies to identify and contain recreational content and traffic.
- Integration with web security service to protect Internet-connected branch offices from malware and enable faster SaaS, 100% recreational offload and high availability networking.

Figure 1: Performance gains by technology type



Video Optimization

- Scale internal Video 10x – 100x - 1000x
- Reduce Recreational Video by 30-80% across the distributed enterprise

Cloud Acceleration

- Accelerate SaaS applications directly to branch offices by 7 – 93x
- Eliminate back-hauling SaaS/Internet applications over WAN

Traditional WAN Optimization

- Accelerate applications by 3x-300x from data center to branch office
- Reduce storage replication and backup bandwidth by up to 90%

Get the right acceleration strategy

Acceleration requirements have rapidly moved beyond CIFS and MAPI acceleration. Video and SaaS application delivery are IT's challenges today. With the right acceleration strategy, you can gain superior business value from internal and external infrastructure. Find out how Blue Coat can help at www.bluecoat.com

About Blue Coat

Blue Coat Systems secures and optimizes the flow of information to any user, on any network with leading web security and WAN Optimization solutions. Blue Coat enables the enterprise to tightly align network investments with business objectives, speed decision making and secure business applications for a long-term competitive advantage.

Application Performance: Your Window to Service Delivery

Virtually all organizations depend on online services to transact business. For online brokerage, retail companies and others, online services are their business. For insurance companies and manufacturers, online services enable their business. Regardless of your business, you want to deliver a positive customer experience. Satisfied customers come back and customer retention is the foundation of your bottom line. Pressure is mounting on IT departments to deliver on this requirement and, as a result, continually increasing amounts of IT budgets are spent on tools and processes to assure that services are performing.

Keeping customers happy is nothing new. For years, organizations focused on the domains – the network, the databases, the servers – assumed that if all the domains worked so would the service. But that strategy exposed an IT paradox: IT services are more than the sum of their parts. Managing each domain for peak performance is no guarantee of success. The information essential to assuring services include the service delivery pathway - the route through the infrastructure the service takes to reach the customer - and the components in that pathway - the network links, databases and servers that are essential to delivering the service. Tools and teams, dedicated to supporting individual domains, often have a hazy view of which components actually impact specific services. For example, a single server outage may have nothing to do with your critical business service…or it may have everything to do with it. Managing each domain for peak performance without a clear asset-to-service view is not a guarantee that you will stay ahead of calls to the help desk.

Applications as Bellwethers

Are applications another domain or a service? Actually, it depends. Some applications, such as online trading, are the end-user service. An email application is often an end-user service but can also be an enabling part of an online retail service, thereby putting the application into the role of a domain in a service delivery pathway. What can be said with certainty is that applications are an essential part of any service, and applications, like services, rely on the other domains to function. Consider how Web-based applications rely on the full range of IT infrastructure components to be operational. So, whether your application is the service itself or a service enabler, its performance is linked tightly with your business service delivery and that makes your application performance an open window into your service performance.

The CA Technologies Service Assurance portfolio is built around proactive performance management. On a foundation of CA eHealth® Performance Manager for client-server applications, CA Technologies added and integrated CA Introscope and CA Customer Experience Manager, the CA Application Performance Management (CA APM) solution, to detect, triage and diagnose performance problems in your complex, composite and Web application environments. CA APM supports both Java and .NET applications and provides end-to-end visibility to online transactions. To complete the picture, CA Technologies acquired NetQoS, bringing products like CA NetQoS SuperAgent® and CA NetQoS ReporterAnalyzer™ into the fold. CA NetQoS SuperAgent tracks every TCP application packet traversing the network between clients and servers, providing metrics such as network, server and application latency for all applications. CA NetQoS ReporterAnalyzer provides historical, real-time and predictive behavioral views through traffic composition metrics that show how applications tax and compete for network resources. With these detailed application performance metrics, application delivery bottlenecks are quickly pinpointed, root cause established and performance issues corrected, often before user impact. And the ripple effect on business services is all positive.

Service Assurance: Application Performance Plus

The CA Technologies Service Assurance portfolio provides a layer of intelligence that leverages data from your existing infrastructure and application performance management tools used to directly manage your IT assets, including Infrastructure Management products like CA Spectrum® Infrastructure Manager, CA eHealth Performance Manager and CA NetQoS Performance Center, and CA Application Performance Management products like CA Introscope® and CA Customer Experience Manager. Consolidating information from these performance managers, CA Service Operations Insight (formerly CA Spectrum® Service Assurance) provides the business service analytics, uniquely linking applications to infrastructure to calculate key performance indicators (KPIs) for service delivery and risk.

CA Spectrum Service Assurance creates a single service model, leveraging information from the domain managers, that is updated dynamically as things change, so you know what components – infrastructure or application – are in the pathway of your critical business service and you know if there is a problem that will impact service delivery, even as configurations and virtual machines change. With the CA Technologies Service Assurance portfolio, you can prioritize your efforts, have confidence in the information you have and fix the important things first to minimize customer and business impact. Even better, CA Technologies can show you where a potential problem is chipping away at performance, for example, telling you when a server farm is losing machine power even if it is not yet impacting service. This puts you where you want to be - two steps ahead of your customer.

Integration Works at Rooms To Go

Customers that have benefitted from the tight integration in the CA Technologies Service Assurance portfolio have compelling stories to tell. Putting it all together was the key for Rooms To Go. To enhance the customer experience at its 150 showrooms across the U.S., Rooms To Go added CA Technologies software for network, application and virtual system performance management to its existing Service Assurance products to maintain service availability and improve support of its retail and distribution outlets.

Rooms To Go is using the CA NetQoS Performance Center, a key component of the Service Assurance portfolio, and CA Virtual Assurance for Infrastructure Managers to improve the performance of its most business-critical, networked applications and their supporting infrastructure. For example, Rooms To Go uses the two CA Technologies solutions to monitor and manage its point-of-sale (POS) application that provides immediate purchase-related information and fast credit application processing and approvals.

“The CA NetQoS Performance Center and CA Virtual Assurance for Infrastructure Managers will help Rooms To Go be more proactive in ensuring a high level of service across our stores and improving the customer experience as a result,” said Jason Hall, Director of IT systems for Rooms To Go. “Combined with our other products from CA Technologies, the CA NetQoS and CA Virtualization Management solutions will give us a more complete understanding of what is happening across our network and virtualized infrastructure and where we need to direct our attention to solve problems faster, prepare for future capacity needs, and optimize application performance.”

In addition to monitoring how well the network delivers the POS application to the Rooms To Go showrooms, the CA NetQoS solution will help Rooms To Go understand how application traffic affects network performance, with views into the composition of traffic on every network link, and which applications and users consume bandwidth. Before installing NetQoS, Hall had no visibility into how end users were experiencing application and service performance across the WAN or LAN. "It was purely the end user," he said. "We waited for someone to call. Operationally, that gives the end user the perspective that the systems are slow … and that we're not doing anything about it. ” Hall said that adding NetQoS's performance management capabilities to his suite of tools has also helped him solve some service delivery mysteries, particularly with his company's intranet. You can read more on this story on SearchNetworking.com in their June 16, 2010 article by Shamus McGillicuddy titled, “Service delivery management: Integrating IT management tools.”

Jack Henry & Associates Put Service First

No one doubts the importance of accuracy and high performance when it comes to financial applications. Jack Henry & Associates processes transactions, automates business processes, and manages mission-critical information for more than 8,700 financial institutions and corporate entities, serving around six million end-users who depend on Jack Henry to run business-critical applications and financial processes. Initially, the company had no consistent means of monitoring end-to-end performance across its network and applications, which made it difficult to safeguard service levels and manage capacity.

“We have to prove every single day that our performance is meeting customer requirements, which, without end-to-end monitoring, was challenging,” said Josh Bovee, Senior Network Engineer, Jack Henry & Associates. “We needed to focus on application performance from the end-user perspective and create a baseline of how well we were serving those customers so we could understand when performance degraded and what impact things like infrastructure changes might have. We were reliant on getting all the IT groups in the same room, and then putting our heads together until we located the source of the issue. With limited insight into network and application performance metrics, this would often take days.”

Realizing they needed to take a more proactive approach to managing its business critical banking applications, Jack Henry looked for a solution that would address its performance management challenges. After struggling for several months with a competitive product, they arranged with CA Technologies for a Proof of Concept with the NetQoS Performance Center, starting with the CA NetQoS SuperAgent. “We started the POC at 8 a.m. and by 1 p.m. we were capturing more meaningful data with SuperAgent than after six months working with the competitive product. SuperAgent was also easier to implement. We didn’t need to install an agent on the server or re-architect our infrastructure, which was something we very much wanted to avoid,” notes Bovee. Having made the decision to deploy CA NetQoS SuperAgent, the company decided to implement additional modules of the CA NetQoS Performance Center.

Jack Henry now has a finger firmly on the pulse of its customers’ business-critical applications, furthering its commitment to industry-leading client satisfaction and retention rates. As a result of their investment in CA Service Assurance solutions, the company is already benefiting from improved service, more cost-effective support and greater business agility. “We now have a great foundation on which to continue to improve our service levels and customer satisfaction,” concludes Bovee.

CA Technologies Manages Risk to Assure Application and Service Delivery

Service Assurance and risk management is achieved through new, advanced technology that can model the IT assets that comprise services, track service quality (end-user experience), the status of each IT asset (network devices, systems, databases and applications) and calculate each asset’s risk to each service dynamically. With this information, you’ll know how to proactively fix problems before they impact users.

These capabilities also factor dimensions of risk beyond typical KPIs to include compliance, answering questions such as: “Are my business services at risk because configurations do not meet the gold standard? Do we have the latest security patches deployed on every device?”

Identifying and measuring risk to business services benefits both IT executives and the technical staff who manage the IT environment “hands-on.” By understanding risk, IT executives can make more informed decisions about capital and operational investments. Technical staff benefit because they can see the root cause of trends that will impact services in the future and can proactively prevent impact to quality.

CA Technologies Service Assurance is a mature, integrated portfolio that provides end-to-end visibility into business services, applications and transactions linked with top-to-bottom insight over the entire infrastructure. Providing great service in a consistent manner, meeting SLAs and having the agility in your infrastructure to roll out new services quickly and efficiently is just table stakes in today's complex IT environment. No matter what business you are in, service assurance is critical to your success, and CA Technologies can work with you to help you deliver the service your customers demand.

Advertorial

Software WAN Optimization

certeon®

Accelerate Your Business

“… application performance … one of the top three inhibitors of cloud adoption” [1]

aCelera™

Secure Automated Optimized

- Any deployment model:
  - Enterprise
  - Hosted
  - Cloud
  - Or - Any combination
- Any hypervisor & Windows Server 2008 R2
- Any number of instances
- Any throughput capacity
- Any security requirement
- Any routing mode
- Any Failover mode

- Automated Management
- Meet cost savings objectives
- Match footprint limitations
- Bundle best of breed applications:
  - Video streaming
  - Directory services
  - Security

Microsoft GOLD CERTIFIED Partner

vmware READY

Certeon Inc.
4 Van de Graaff Drive
Burlington, MA 01803
781 425 5200
http://www.certeon.com

APPLICATION DELIVERY PERFORMANCE From datacenter and cloud to any user anywhere

Maximum Value & Maximum Performance

The **business value** of any application must be measured by its ability to increase business agility, decrease cost through on-demand provisioning and teardown of infrastructure and services, accelerated development, and improved reliability. Solutions must be utility-based, self-service, secure and most importantly, have levels of application performance that improve productivity.

Maximizing the business value of any networked application requires full featured, secure, scalable, high performance WAN Optimization software that allows applications to perform as expected, and can be part of any on demand architecture. Tactical hardware or virtual appliances with limited performance don’t measure up.

aCelera: Built for Global Performance

aCelera software exceeds the scalability and performance of purpose-built hardware appliances. aCelera WAN Optimization software can support hundreds of thousands of connections and gigabits of throughput. It is built to support global enterprise scalability requirements and is ready for the Internet scale demands of managed services and cloud computing providers.

aCelera software and virtual appliances deliver these performance benefits and advantages without the costs or the friction of hardware appliances or limited scope virtualization. aCelera can easily be scaled on any existing hardware platform or migrated to more powerful platforms and processors when business conditions dictate, leveraging any industry standard management tool.

aCelera: Built for Global Deployment

Enterprise and cloud infrastructures are not uniform. aCelera software can be deployed in any heterogeneous mix of hardware, virtualization platforms, storage technologies, networking equipment and service providers supporting any custom or off the shelf application.

Hardware WAN optimization products require more planning and are more labor intensive to install. aCelera software packages are delivered over a network and installed in a data centers, remote sites, or end user PCs in less than 30 minutes. aCelera creates a high performance WAN infrastructure that can span the globe and scale to meet your application and user performance needs.

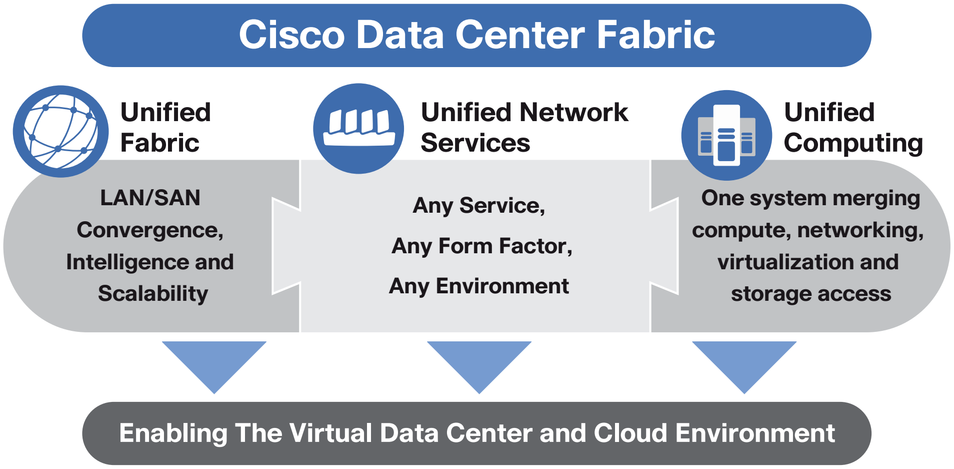
aCelera can be deployed in any private, public, and hybrid cloud computing environment and is poised to meet ANY future performance, scale and connection demands imposed by any enterprise IT environment, private network, private cloud, public cloud or a hybrid of them all.

aCelera software WAN optimization:
60% better 3 year TCO & 50% better scalability

[1], Clouds and Beyond: Positioning for the Next 20 Years of Enterprise IT, Frank Gens, IDC

Cisco Unified Network Services

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


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

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

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

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

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

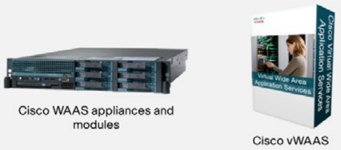
Application Delivery Controllers

Enhanced web application performance, availability, and server scalability

Cisco ACE module and appliance, Cisco GSS

WAN Optimization

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



Cisco WAAS appliances and modules

Cisco vWAAS

Network Analysis and Monitoring

Simplifies application performance monitoring

Cisco NAM appliances, modules, and virtual blades

Data Center Security

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

Cisco ASA 5585-x

Cisco VSG

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**.

Advertorial

exinda

UNIFIED PERFORMANCE MANAGEMENT

VISIBILITY | CONTROL | OPTIMIZATION

COMPLETE WAN OPTIMIZATION

Increase the speed and efficiency of your wide area network.

Exinda's Unified Performance Management (UPM) solution delivers everything you need to manage your application performance and ensure the highest quality user experience.

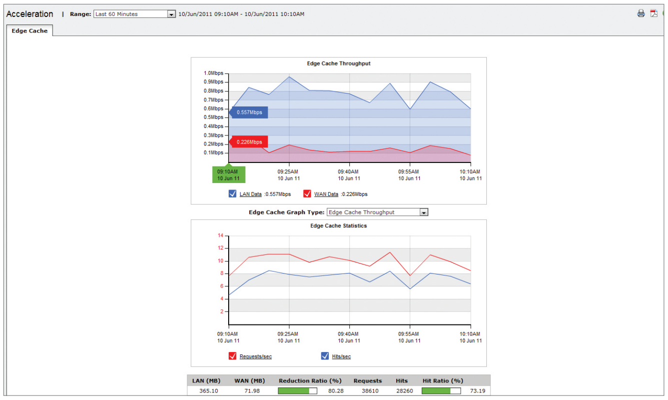
Point solutions lack inter-communication between the functions of visibility, control, and optimization. This creates contention between these independent solutions, as each function is unaware of the effect its actions has on the other.

Exinda's unique, holistic approach to WAN Optimization eliminates the communication barriers and contention of point solutions, by integrating visibility, control and optimization, into a single, unified solution.

LATEST ADVANCES IN UNIFIED PERFORMANCE MANAGEMENT

Exinda's development team is continually adding new features and functionality into our unified performance management solution. It is because of our agile development cycle and constant push to add innovation to our product line that Exinda has become the fastest growing WAN optimization vendor in the world. The following are some of the latest advances in our UPM solution.

EDGE CACHE



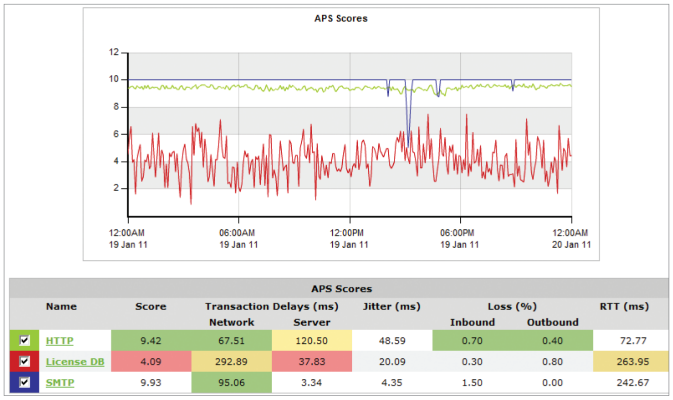
Acceleration | Range: Last 60 Minutes 10/Jun/2011 09:10AM - 10/Jun/2011 10:10AM

Edge Cache Graph Type: Edge Cache Throughput

LAN (MB)	WAN (MB)	Reduction Ratio (%)	Requests	Hits	Hit Ratio (%)
365.10	71.98	80.28	38610	28260	73.19

Exinda Edge Cache will allow you to reduce bandwidth usage, decrease network costs, and accelerate content delivery, improving user experience and productivity.

APPLICATION PERFORMANCE SCORE



APS Scores

	Name	Score	Transaction Delays (ms)		Jitter (ms)	Loss (%)		RTT (ms)
			Network	Server		Inbound	Outbound	
☑	HTTP	9.42	67.51	120.50	48.59	0.70	0.40	72.77
☑	License DB	4.09	292.89	37.83	20.09	0.30	0.80	263.95
☑	SMTP	9.93	95.06	3.34	4.35	1.50	0.00	242.67

Gain proactive reports on users perception of application performance & responsiveness.

Edge Cache

The Exinda Edge Cache^{TM} enables single-sided caching of Internet-based content at the network edge, including web objects, videos and software updates, delivering a superior user experience and reducing WAN resource utilization.

Web objects are cached at the network edge when they are first downloaded from the Internet or across WAN links. These objects can then be delivered to the users on subsequent requests over the corporate local area network much faster without needing to download the data over the WAN again, providing a better user experience and increased productivity to the workforce. By caching web objects in the local office, organizations can drive down the network traffic consumed by each office, which directly reduces network costs.

The Exinda Edge Cache enables caching of web objects, video, software update and other content on the WAN. It also offers cache statistics, which provide insight into the amount of repetitive data being off-loaded from the WAN link, how cacheable the network data is, how frequently the cache is being accessed, and by how many hosts, helping organizations to understand the nature of their network traffic over time.

The Exinda Edge Cache can also be aligned with an organization's optimization policies, allowing the administrator to only cache specific content for specific users or groups of users, and to maintain very precise controls over how much WAN bandwidth should be made available for each application traversing the network.

Application Performance Score

A significant feature of Exinda's WAN Optimization solutions is its ability to provide Application Performance Scores (APS). Exinda's APS provides a single data point to monitor and report on the overall health and performance of an application on your network. With APS, you can set performance thresholds for the applications on your network, and easily monitor if and when the thresholds are met or exceeded. When WAN application performance issues arise, the APS allows you to quickly troubleshoot the problems, by drilling down into individual metrics for the application, including network delay, server delay, jitter and loss, and round trip time, helping you to pinpoint and address the source of the performance issue.

Exinda also allows you to monitor and report on TCP efficiency and health. With Exinda, TCP efficiency reports let you examine how efficiently packets flow through the network, based on the number of dropped packets and retransmitted packets for the application. When combined with Exinda's TCP health monitoring, TCP efficiency reporting gives you a more in-depth view of network and application performance. TCP Health monitoring displays the health of TCP Connections by showing the total number of TCP connections, and how many were aborted, ignored, or refused by the server. With Exinda, you get a simple graphical view of the TCP health of the network, allowing rapid drill down for troubleshooting network and application performance issues.

exinda
performance you can see

Unified Performance Management

Network Visibility, Control and Optimization - All in a Single Appliance

> "Unified Performance Management is driven by improving the quality of a user's experience."
>
> - Ed Ryan, Exinda Vice President of Products

The Best Solution For You.

Identify and Improve Application Performance

- Application Performance Measurement technology measures user experience objectively.
- Identify the source of application performance issues - Network, Server or Application.
- Apply application performance scoring to more than 2,000 applications.

Offer a Superior User Experience

- Dramatically increases user download speeds for internet applications, videos, and software updates.
- Accelerate delivery of content to users at LAN speeds from a web cache with a single appliance.
- Optimize and accelerate mission critical applications.

Real-time and Historical Reporting

- Real time reporting showing all traffic on the network over the last 10-60 seconds.
- Up to 2 years of historical reporting on applications, hosts, conversations, URL's, and performance scores "on appliance".
- Microsoft Active Directory Integration allows you to report on users or groups regardless of IP Address.
- Netflow v9 export, providing in-depth layer 7 details of your network usage and application performance.

Conserve WAN Resources

- Guarantee bandwidth for critical applications while controlling recreational traffic.
- Byte and Object level caching with dual or single appliances reduces the footprint of traffic on the WAN serving files, software updates, and video to users at LAN speeds.
- Reclaim up to 90% of the bandwidth on your WAN circuits to deliver data more efficiently.

Leverage Your Investment

- Exinda is fully scalable supporting WAN circuits from 256k to 10Gbps, and includes mobile client support.
- Exinda auto-discovery limits the operational burden and cost of managing large scale multi-site deployments.
- Exinda's Service Delivery Platform (SDP) is available as an appliance or on a cloud-based management platform, offers a flexible and cost-saving option to manage your network.
- A single appliance delivering visibility, control, and optimization makes it easier and more cost-effective to manage and expand over time.

Features & Benefits

Visibility

Provides insight into network activity, usage and performance. Gives you the information you need to keep your network operating at peak performance

- Layer 7 Classification
- Heuristic Classification
- URL Classification
- Drill Down Capabilities
- Real Time Monitoring
- Top Talkers/Top Conversations
- Active Directory User ID
- Anonymous Proxy Detection
- Application Performance Score
- Service Level Agreements
- Network Health
- Citrix Published Applications
- Automated PDF Reporting

Control

Maximize network resources to the needs of your organization through comprehensive control over network traffic without placing heavy-handed restrictions on users.

- QoS / Dynamic per IP User
- Bandwidth Management
- Traffic-shaping
- Prioritization
- Active Directory Integration

Optimization

Rapidly, turn understanding into action that drives network performance, improves the user experience, and optimizes productivity.

- Layer 4 TCP Optimization
- Layer 7 Application Acceleration
- Universal Caching
- Compression
- Intelligent Acceleration
- Peer Auto-Discovery
- SSL Acceleration

Americas
+1 877 219 0603
info.americas@exinda.com

UK / Europe
+44 808 120 1996
info.emea@exinda.com

Asia Pacific
+61 3 9415 8332
info.apac@exinda.com

IMEA
+971 4 295 5049
info.imea@exinda.com

Advertorial

2011--May--UPM--Overview.pdf

Advertorial

EXPAND networks | VIRTUALLY EVERYWHERE™

EXPAND ENABLES SERVER CONSOLIDATION, THIN-CLIENT COMPUTING AND BANDWIDTH OPTIMIZATION AT RIDLEY INC – DELIVERS SAVINGS OF $250,000 PER ANNUM

Having initially deployed Expand Networks' Accelerators as part of a bandwidth consolidation project in 2006, Ridley Inc – the leading animal nutrition company - was already aware of the benefits that WAN optimization technology could bring; this initial $200,000 investment paid for itself through efficiency savings in just over six months.

However, with many of its 42 locations being extremely harsh and dusty environments, Ridley recently embarked on a thin-computing strategy, removing servers and computers from branches and delivering server based computing from a central location in Minnesota.

In order to meet renewed bandwidth requirements and ensure the company's new thin-computing IT initiatives were to succeed, Ridely Inc.re-assessed the company's WAN environment.

Chad Gillick, the IT Manager that led the project at Ridley Inc, explained, "Moving to a thin computing environment could help us streamline processes, increase productivity and reduce costs. However, I knew WAN optimization would be essential to the success of these projects, to ensure the user experience and productivity wouldn't suffer across our distributed network environment."

By investing further in new Expand technology, Ridley has been able to remove expensive desktop and laptop computers at the remote sites and replace with thin client terminals, without costly bandwidth upgrades.

The company chose Expand because of its superior capabilities in accelerating Citrix and web based traffic, and the Accelerators have been deployed in 31 key sites.

Combining compression, byte-level caching, layer 7 QoS and small packets mitigation techniques, Expand's technology enables available bandwidth and real-time interactive TCP traffic to be maximized, extending Ridley existing network infrastructure investments and providing 'virtual bandwidth' capacity to its users.

With substantially faster data transfer speeds over WAN links, Ridley is gaining an estimated 45 minutes of productivity per person, per day. Furthermore, Expand's Wide Area File Services (WAFS) capabilities with QoS have enabled the IT team to tailor traffic flows across the managed network and dynamically manage bandwidth requirements 'on the fly'.

"Without the Expand solution we would have needed a 45mbps connection at the central site that would have cost in the region of $26,000 per month. With Expand we were able to reduce this to a 9mbps link costing $4,500, an annual saving of over $250,000," said Gillick.

He concluded, "On top of this, using Expand as an enabler of server consolidation and thin client computing, we have managed to reduce our technical refresh costs which were running at $400,000 annually down to $220,000. We believe we will be reaping the benefits of the Expand solution for many years to come."

Enabling Strategic Initiatives

- **Virtualization** - The foundation infrastructure for delivering on all strategic IT initiatives, Expand's technology is unique in its combined ability to be deployed within a virtualized infrastructure and to accelerate and control virtualized traffic out of it. The software can be effectively integrated into virtual server environments, such as VMWare, Citrix XenServer and Microsoft HyperV, and as a truly virtualised solution Expand can also be deployed under extreme conditions such as on aircraft, mobile environments and remote and unattended locations..

- **VDI and Thin Computing** - - Expand accelerates within Virtual Desktop Infrastructure (VDI) and thin computing environments optimizing protocols including Microsoft Terminal Services (RDP), Citrix XenDesktop (ICA) and Sun Sunray (ALP). Unlike competitive offerings, Expand works on the IP layer, this enables Expand to accelerate all IP & uniquely UDP applications over the WAN, applying advanced compression, byte level caching, layer 7 QoS and small packet mitigation techniques.

- **Server Consolidation** - Expand's integrated 'virtual server' technology enables complete server consolidation by replacing the need for an additional branch office file server. Expand's unique "Virtual Branch Server" feature sets also enable to customer to replace features that used to be delivered by a remote server, such as DCHP, DNS and Printing, all within the AOS and not via third party plug-ins like other vendors.

- **Satellite** - With integrated Space Communication Protocol Specifications (SCPS) Standard technology, Expand helps distributed organizations overcome the traditional limited bandwidth, high latency obstacles that impede the speed and performance of applications and services over satellite links. Communication Protocol Standard technology, helps distributed organizations overcome the traditional low bandwidth, high latency obstacles that impede the speed and performance of applications and services over satellite links.

Advertorial

SOLUTION BRIEF

Software as a Service (SaaS)

A Cloud-Ready Network ensures rollout success

www.ipanematech.com

Cloud adoption adds complexity to network management. Cloud applications such as SaaS collaboration bring many of the same issues as licensed software, but each IT implementation project can have a larger impact because of its reliance on your WAN. By aligning your network with business and Application Performance Objectives, WAN Governance puts you in control of this complexity and network impact.

WAN Governance improves the IT Governance you already have in place by providing:

- A holistic approach to global visibility, control and optimization of application performance, as opposed to conventional solutions operating as independent agents
- Business continuity and control as SaaS applications are adopted
- Guaranteed application performance for any network architecture
- Network capabilities to absorb enterprise requirements for agility, flexibility and growth
- Next-generation solutions for implementing and managing a cloud-ready network

Using WAN Governance, your organization can:

- Understand the nature of application traffic
- Control and optimize this traffic
- Guarantee application performance
- Improve users' Quality of Experience
- Simplify network operations
- Control network costs and leverage savings

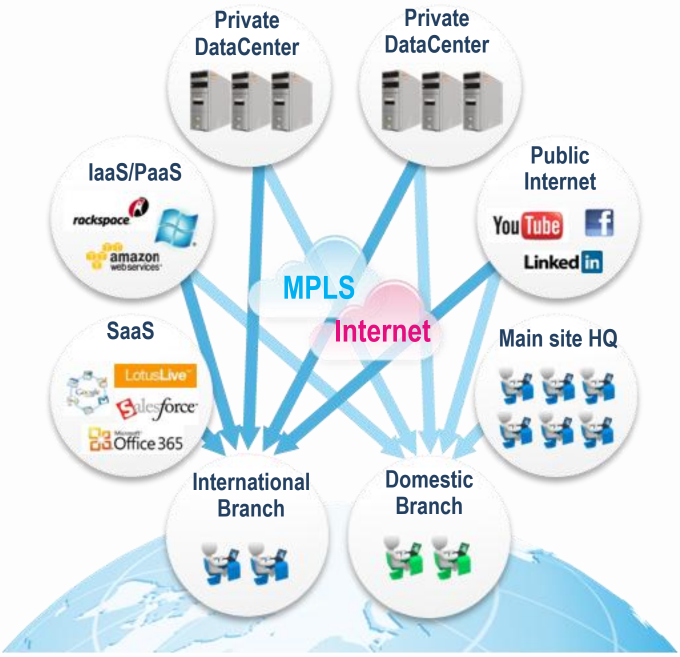
IT infrastructure directors today find themselves in one of two situations: the business side of their organization is planning for SaaS applications that the VPN will need to support, or existing SaaS applications are underperforming or impacting the performance of other business applications.

VPNs and the tools used to manage them are optimized for traditional private applications residing in data centers, not those stored in the cloud. For example, SaaS collaboration applications, such as Google Apps, Microsoft BPOS/Office 365 and IBM LotusLive, consume much more network bandwidth than many traditional applications. Moving from traditional on-premise collaboration to a SaaS counterpart dramatically changes the way traffic flows across the WAN.

In order to avoid application performance issues and ensure optimal end-user experience, infrastructure directors need to make their VPN "cloud ready." A cloud-ready network (CRN) is a network that provides full application performance visibility and total control of both SaaS and on-premise applications. Ideally, the best time to prepare is prior to your first SaaS implementation, so that the impact of SaaS on your VPN can be mastered from the pilot phase through full enterprise rollout.

With Ipanema for a fraction of the cost per user of your SaaS you can:

- Guarantee the performance of SaaS across the WAN
- Ensure peaceful co-existence of SaaS and existing applications (ERP, CRM…)
- Obtain a dashboard of application performance for all critical applications including SaaS
- Take full advantage of hybrid MPLS + Internet networks
- Shift to WAN governance, plan and grow your network according to business needs



Beyond the Network…

ipanema Technologies

www.ipanematech.com

Valeo

Valeo Embraces the Cloud and Maximizes Value

Valeo, one of the world's leading suppliers of components, integrated systems and mod-ules for automotive CO2 emissions reduction, rolled out a hybrid network with MPLS + Internet for its migration from conventional email and collaboration applications to Google Apps.

Valeo's network supports approximately 160 sites worldwide, 52,000 users, and the delivery of applications such as ERP and CATIA.

Using Ipanema's ANS to dynamically manage application performance over their hybrid network, Valeo successfully deployed Google Apps with full Applications Visibility, QoS & Control, and Dynamic WAN Selection.

"With Ipanema, we divided by three the transfer cost of each Gbyte of band-width over our global network," says **Alain Meurou, Infrastructure and Network Manager at Valeo**.

Return on investment

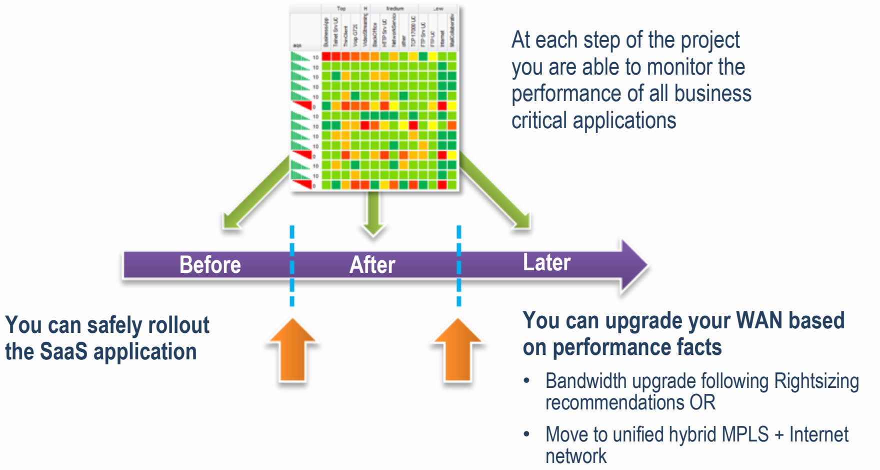
Enterprises that have chosen to move to a unified hybrid net-work controlled by ANS typically chose not to upgrade their MPLS bandwidth in favor of the less-expensive Internet bandwidth. Including the price of the deployed ANS solution, typically 1 to 2 € per user per month , most enter-prises were able to obtain a 20% decrease in overall network costs, upgrade available band-width by a factor of three, and adequately pre-pare for traffic increases over the next three to five years.

All-in-One Solution for Guaranteeing Application Performance

Ipanema's Autonomic Networking System (ANS) tightly couples into a single, all-in-one solution.

- QoS & Control
- Application Visibility
- WAN Optimization
- Dynamic WAN Selection (hybrid network unification)

With ANS, all application performance challenges can be managed with a holistic approach over the global network. The autonomic networking solution automates tasks that IT organizations cannot perform with traditional approaches. Orchestrating network traffic in real-time, ANS manages the complexity of the hybrid cloud and guarantees application performance for public and private applications. ANS not only helps to guarantee the performance of SaaS during and after implementation, but the end-user experience for all applications over your WAN, and much more cost effectively.



At each step of the project you are able to monitor the performance of all business critical applications

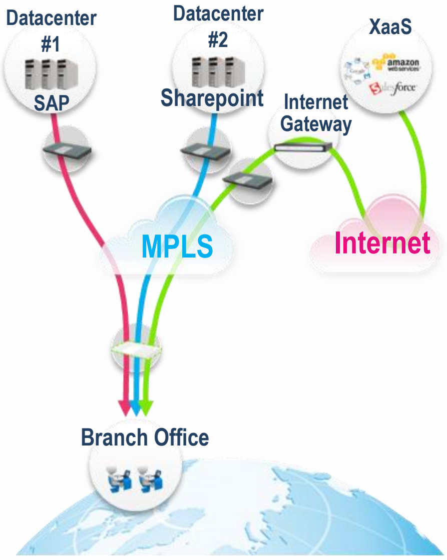
You can safely rollout the SaaS application

You can upgrade your WAN based on performance facts

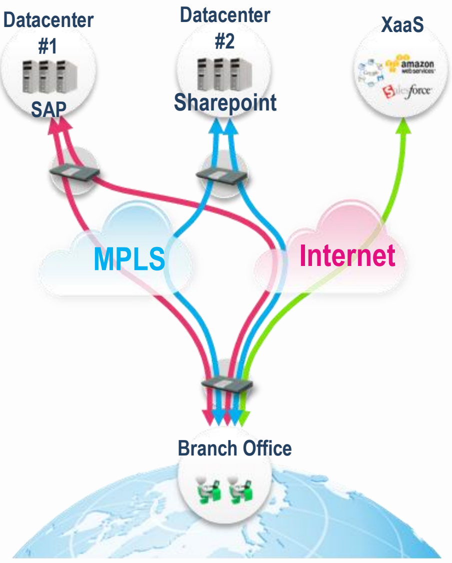
- Bandwidth upgrade following Rightsizing recommendations OR
- Move to unified hybrid MPLS + Internet network

Since every enterprise is different, IT strategy on whether or not to change network architecture for SaaS collaboration varies from one company to another. You do not necessarily need to change your architecture to make your network "cloud-ready".

All companies, however, must implement a minimum set of capabilities in order to avoid application performance issues during and after SaaS implementation, or to fix issues resulting from a prior SaaS deployment. Companies that use or plan to use a hybrid (MPLS + Internet) network architecture will also want to consider additional capabilities to further optimize their "cloud-ready network" (CRN).



An MPLS CRN



A unified hybrid MPLS + Internet CRN

Beyond the Network…

ipanema Technologies

Advertorial

Advertorial

Packet Design

Network-Wide IP Routing and Netflow Monitoring, History, Modeling & Planning

Optimize IP Networks with Traffic Explorer

- Monitor and analyze critical traffic dynamics across all IP network links and routes by Class of Service (CoS)
- Strengthen change management with operationally accurate network modeling based on real-time, network-wide routing and traffic state
- Reduce Internet transit costs with IGP/BGP-aware peering and transit routing and traffic analysis
- Analyze network-wide traffic usage, even per MPLS VPN
- Improve network continuity with easy traffic trending
- Perform network-wide traffic capacity planning

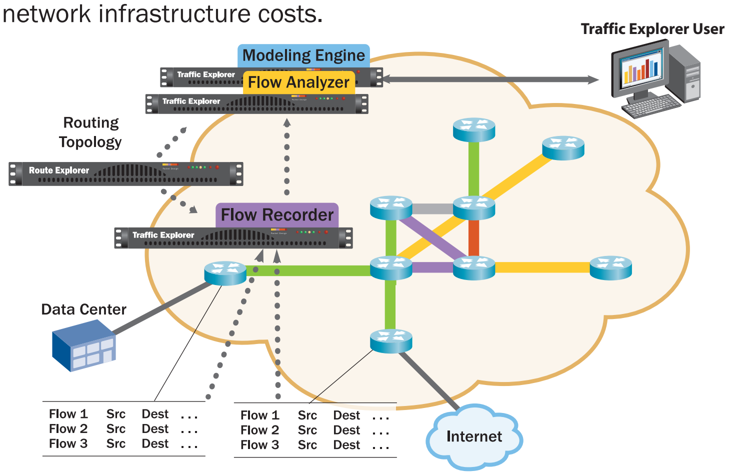
Packet Design Overview:

- Founded in 2003, Packet Design pioneered and is the market-leading provider of routing and traffic analysis solutions
- 500+ global enterprises, Service Providers, and government and military agencies utilize Packet Design solutions to manage their complex IP networks.
- Packet Design solutions offer IT departments significant operational cost savings by increasing the accuracy and efficiency of key IT business processes.

Overview of Traffic Explorer

Traffic Explorer is the first solution to combine real-time, integrated routing and Netflow traffic monitoring and analysis, with "what-if" modeliing and capacity planning capabilities. Unlike previous traffic analysis tools that only provide localized, link by link traffic visibility, Traffic Explorer's knowledge of IP routing enables visibility into network-wide routing and traffic behavior. Powerful "what-if" modeling capabilities empower network managers with new visibility to strengthen change management processes and optimize network infrastructure costs.

Traffic Explorer delivers the industry's only integrated analysis of network-wide routing and traffic dynamics. Standard reports and threshold-based alerts help engineers track significant routing and utilization changes in the network. An interactive topology map and deep, drill-down tabular views allow engineers to quickly perform root cause analysis of important network changes, including the routed path for any flow, network-wide traffic impact of any routing changes or failures, and the number of flows and hops affected. This information helps operators prioritize their response to those situations with the greatest impact on services or applications.



Traffic Explorer provides extensive "what-if" planning features to enhance ongoing network operations best practices. Traffic Explorer lets engineers model changes on the "as running" network, using the actual routed topology and traffic loads. Engineers can simulate a broad range of changes, such as adding or failing routers, interfaces and peerings; moving or changing prefixes, BGP policy configurations, link capacities or traffic loads; even adding new MPLS VPNs. Simulating the effect of these changes on the actual network results in faster, more accurate network operations and optimal use of existing assets, leading to reduced capital and operational costs and enhanced service delivery.

Proven, Market Leading Solutions: Based in Palo Alto, Packet Design Inc. is the pioneer and market leader in routing-aware network management solutions. Packet Design is a member of the Cisco Technology Developer Partner program. Find out more at www.packetdesign.com

CISCO PARTNER
Technology Developer

Packet Design Inc. • 2455 Augustine Drive, Santa Clara, CA 95054 • 408-490-1000 • info@packetdesign.com

Advertorial

Why an Application Delivery Fabric is Essential for Agile & Scalable Virtualization

Data Center Virtualization and Application Delivery

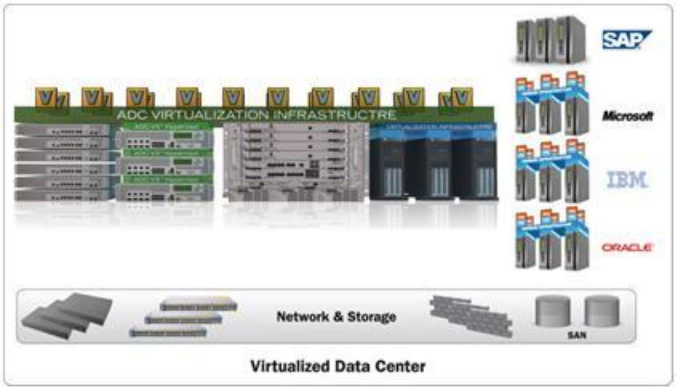
Network infrastructure virtualization/consolidation has had a major impact on the Application Delivery Controller (ADC) role, position and deployment models. For instance, ADCs which were previously tightly coupled with a single application must now be able to service a layer of virtualized applications sharing a common server infrastructure.

A New Paradigm: Virtual Application Delivery Infrastructure (VADI)

Radware's VADI strategy allows for the transformation of computing resources and ADC devices and software into an integrated, agile and scalable set of application delivery services that can be dynamically provisioned, decommissioned, and migrated.

Radware's VADI delivers the following business benefits:

- ✓ Significant cost reduction via ADC consolidation
- ✓ Simpler path to data center virtualization
- ✓ Improved business agility
- ✓ Greater IT efficiency via data center workflow automation
- ✓ Full application delivery resource elasticity
- ✓ On demand scalability in throughput, advanced services and virtual ADCs
- ✓ Full investment protection, increased asset ROI, and CAPEX savings



VADI Key Components

Virtual ADC Instances

A Virtual ADC (vADC) instance is a service providing a consistent and complete set of application delivery features such as load balancing, global server load balancing, application acceleration, integrated security, bandwidth management and more. A vADC runs on top of specialized and general purpose computing resources, thus transforming ADCs into services.

ADC Computing Resources via Three Form-Factors

- Dedicated ADC – a dedicated, physical ADC device running a single vADC, which is designed to provide application delivery services for siloed data center architectures, hybrid (virtualized and physical) data centers, and applications requiring high SLA and performance predictability.
- ADC-VX™ - the industry's first ADC hypervisor that runs multiple vADCs on a dedicated ADC hardware, Radware's OnDemand Switch platform.
- Alteon VA™ - Radware's Soft ADC is a vADC deployed on a general server virtualization infrastructure, running as a virtual appliance, providing the full functionality of a physical ADC.

Virtual Data Center Ecosystem Integration

Radware's vDirect™ is the industry's first ADC management orchestration plug-in, designed specifically for virtual data centers. It provides all the building blocks and management interfaces required for an orchestration system to provision, decommission, configure and monitor Radware's vADCs and computing resources within a virtual data center.

Advanced VADI Services

VADI services, such as ADC service provisioning, decommissioning, and migration of virtual ADC instances across form factors, enables business agility goals while delivering the matching resilience and SLA per application. Radware provides various VADI services such as:

- Provisioning and decommissioning - vADCs are instantly provisioned and/or decommissioned through the ADC management system or orchestration systems' API
- vADC migration - Easily move a vADC instance between different form factors, allowing scheduling ADC maintenance with zero downtime, thus reducing the potential loss of business and revenue
- Dynamic elasticity - Dynamically instruct the orchestration system to allocate additional resources for an application when the existing computing resources are completely utilized
- Cloud burst - Dynamically instruct the orchestration system to allocate additional resources in the cloud or in a second data center when the resources are completely utilized in the main data center

For additional information on ADC-VX, Alteon VA and vDirect please refer to http://www.radware.com/Solutions/Enterprise/Virtualization/DataCenterVirtualization.aspx or for customer case examples please visit our press release section: http://www.radware.com/NewsEvents/PressReleases.aspx.