The 2011
Cloud Networking Report

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

Produced by:

Webtorials

Table of Contents

Introduction and Forward to the 2011 Edition

The majority of IT organizations have either already adopted, or are in the process of evaluating the adoption of one or more classes of cloud computing. Gartner, for example, estimates that between 2010 and 2015 enterprises will spend $112 billion cumulatively on Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS), combined[1].

The broad interest in cloud computing is understandable given that the goal of cloud computing is to enable IT organizations to become dramatically more agile and cost effective and that evidence exists that that goal is achievable. The primary goal of this report is to describe the challenges and solutions that are associated with cloud networking.

The phrase cloud networking refers to the LAN, WAN and management functionality that must be in place to enable cloud computing.

As will be discussed in this report, a traditional network will not be able to successfully support cloud computing.

In order to support cloud computing, a cloud network must be dramatically more agile and cost effective than a traditional network.

In order to describe the networking challenges that are associated with enabling cloud computing, the rest of this section of the report will identify what cloud computing is today and will also describe how cloud computing is likely to evolve in the near term. Subsequent sections focus on the key components of a cloud network: Data Center LANs, WANs, and Network Management. Given the breadth of fundamental technology changes that are impacting the data center LAN, the data center LAN section is very technical. The sections on WANs and Network Management are moderately technical. This year's edition of the cloud networking report leverages last year's edition of the report[2]. However, every section of The 2010 Cloud Networking Report has been significantly updated to reflect the changes that have occurred in the last year.

As noted, the primary goal of this report is to describe the challenges and solutions that are associated with cloud networking. A secondary goal of this report is to identify how IT organizations are currently approaching cloud networking and where possible, indicate how that approach is changing. To accomplish that goal, this report includes the results of surveys that were recently given to the subscribers of Webtorials.com and to the attendees of the Interop conferences. Throughout this report, those two groups of respondents will be respectively referred to as The Webtorials Respondents and The Interop Respondents. In some cases, the results of the surveys given to The Webtorials Respondents and The Interop Respondents will be compared to the results of surveys given to these two groups in 2010. The purpose of these comparisons is to quantify the ongoing changes that are occurring.

The results of surveys such as the ones described in the preceding paragraph that ask IT organizations about their plans are always helpful because they enable IT organizations to see

[1] http://www.gartner.com/it/page.jsp?id=1389313
[2] http://www.webtorials.com/content/2010/12/2010-cloud.html

how their own plans fit with broad industry trends. Such surveys are particularly beneficial in the current environment when so much change is occurring.

Executive Summary

Background

The majority of IT organizations have either already adopted, or are in the process of evaluating the adoption of one or more classes of cloud computing. The broad interest in cloud computing is understandable given that the goal of cloud computing is to enable IT organizations to become dramatically more agile and cost effective and that evidence exists that that goal is achievable.

Throughout this report, the phrase cloud networking refers to the LAN, WAN and management functionality that must be in place to enable cloud computing. As is discussed in this report, in order to support cloud computing, a cloud network must be dramatically more agile and cost effective than a traditional network is. To help IT organizations deploy a network that can enable cloud computing, the primary goal of this report is to describe the challenges and solutions that are associated with cloud networking.

The first section of this report will identify what cloud computing is today and will also describe how cloud computing is likely to evolve in the near term. Subsequent sections focus on the key components of a cloud network: Data Center LANs, WANs, and Network Management. This year's edition of the cloud networking report leverages last year's edition of the report[3]. However, every section of The 2010 Cloud Networking Report has been significantly updated to reflect the changes that have occurred in the last year.

As noted, the primary goal of this report is to describe the challenges and solutions that are associated with cloud networking. A secondary goal of this report is to identify how IT organizations are currently approaching cloud networking and where possible, indicate how that approach is changing. To accomplish that goal, this report includes the results of surveys that were recently given to the subscribers of Webtorials.com and to the attendees of the Interop conferences. The results of surveys such as these that ask IT organizations about their plans are always helpful because they enable IT organizations to see how their own plans fit with broad industry trends. Such surveys are particularly beneficial in the current environment when so much change is occurring.

The Emergence of Cloud Computing and Cloud Networking

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. In order to demonstrate the concept behind the phrase *good enough,* consider just the availability of an IT service. In those cases in which the IT service is business critical, *good enough* could mean five or six 9's of availability. However, in many other cases *good enough* has the same meaning as *best effort* and in these cases *good enough* could mean two or three 9's of availability if that approach results in a notably less expensive solution.

In most instances the SLAs that are associated with public cloud computing services such as Salesforce.com are weak and as such, it is reasonable to say that these services are delivered on a best effort basis. For example, most of the SLAs that are associated with public cloud

[3] http://www.webtorials.com/content/2010/11/2010-cloud-networking-report.html

computing services don't contain a goal for the end-to-end performance of the service in part because these services are typically delivered over the Internet and no provider will give an end-to-end performance guarantee for the Internet. While this situation will not change in the near term, as discussed in the WAN section of this report, there are technologies and services that can improve the performance of the Internet.

While there is not a litmus test that determines whether or not a service is a cloud service, cloud services are usually involve the centralization and virtualization of IT resources along with significant levels of automation and orchestration. The primary factors that are driving the use of public cloud computing solutions are the same factors that drive any form of out-tasking; e.g., lowering cost and reducing the time it takes to deploy new functionality. As discussed in this report, there are significant differences amongst the solutions offered by Infrastructure-as-a-Service (IaaS) providers, especially when it comes to the SLAs they offer. And, as is also discussed in the report, the availability of IaaS solutions can vary widely.

A form of hybrid cloud computing that is discussed in detail in this report is cloud balancing. The phrase *cloud balancing* refers to routing service requests across multiple data centers based on myriad criteria. Cloud balancing can be thought of as the logical extension of global server load balancing (GSLB). The advantages of cloud balancing are that it enables IT organizations to maximize performance, minimize cost and manage risk. The challenges that are associated with cloud balancing are discussed in the WAN section of this report.

This report discusses a new class of solutions that has begun to be offered by cloud computing service providers (CCSPs). These are solutions that have historically been provided by the IT infrastructure group itself and include network and application optimization, VoIP, Unified Communications (UC), security, network management and virtualized desktops. This new class of solutions is referred to in this report as Cloud Networking Services (CNS). As discussed in this report, IT organizations have shown significant interest in CNSs and because of that interest, the adoption of CNSs represents the beginning of what could be a fundamental shift in terms of how IT services are provided.

As much as cloud computing is about technologies, it is also about changing the culture of the IT organization. One of the cultural shifts that is associated with the adoption of cloud computing is that IT organizations become less of a provider of IT services and more of a broker of IT services. In their role as a broker of IT services, IT organizations can facilitate contract negotiations with CCSPs. IT organizations can also ensure that the acquired application or service doesn't create any compliance issues, can be integrated with other applications as needed, can scale, is cost effective and can be managed.

Another cultural change that is associated with the adoption of cloud computing is the implementation of more usage sensitive chargeback. Usage sensitive chargeback is not new. Many IT organizations, for example, allocate the cost of the organization's network to the company's business unit managers based on the consumption of that network by the business units. Most of the increased use of usage sensitive chargeback will come from having the business unit managers pay the relevant cloud computing service providers for the services that their organization consumes. However, the movement to implement more usage sensitive chargeback over the next two years will not be dramatic in part because the culture of an IT organization changes very slowly.

The Emerging Data Center LAN

One of the key characteristics of the current generation of data center LANs is that they are usually designed around a three-tier switched architecture comprised of access, distribution and core switches. They are also characterized by the use of the spanning tree protocol to eliminate loops, the use of Ethernet on a best effort basis and the separation of the data network from the storage network. Today, a number of factors are causing IT organizations to rethink their approach to data center LAN design. One of the primary factors driving change in the data center LAN is the ongoing virtualization of servers. Server virtualization creates a number of challenges including the requirement to manually configure parameters such as QoS settings and ACLs in order to support the dynamic movement of VMs.

There are many on-going IT initiatives that are aimed at improving the cost-efficiency of the enterprise data center; e.g., server virtualization and SOA. In many cases these initiatives place a premium on IT organizations being able to provide highly reliable, low latency, high bandwidth communications among both physical and virtual servers. Whereas the hub and spoke topology of the traditional data center LAN was optimized for client-to-server communications, it is decidedly sub-optimal for server-to-server communications. As discussed in this report, one approach for improving server-to-server communications is to flatten the network from three tiers to two tiers consisting of access layer and aggregation/core layer switches.

The survey data contained in this report indicates that there is significant desire on the part of IT organizations to flatten their data center LANs, but that there is also significant uncertainty relative to how flat those LANs will become in the next two years. This survey data is consistent with other survey data in this report that indicates that IT organizations have significant interest in changing their approach to data center LAN design, but that the majority of IT organizations haven't made up their mind about how they will design their next generation of data center LAN.

One of the key design considerations relative to the next generation data center LAN is what technologies, if any, will IT organizations use to replace the spanning tree protocol (STP). One of the reasons why IT organizations are interested in replacing STP is because this protocol only allows for a single active path between any two network nodes and this feature of STP artificially limits the throughput of a LAN. One way to build a loop-free LAN without using STP is to use switch virtualization and multi-chassis Link Aggregation Group (MC LAG) technologies. With switch virtualization, two or more physical switches are made to appear to other network elements as a single logical switch or virtual switch. MC LAG allows the links of the LAG to span the multiple physical switches that comprise a virtual switch.

While some vendors are actively advocating switch virtualization and MC LAG as the best approach to data center LAN design for at least the next few years, this is not the only approach being discussed in the industry. Standards bodies have been working on technologies that will eliminate loops and allow active-active traffic flows. One of those technologies, TRILL (Transparent Interconnection of Lots of Links), is an Internet Engineering Task Force project to develop a Layer 2 shortest-path first routing protocol for Ethernet. The TRILL RFC is currently on the standards track and is being used as the basis for some pre-standard implementations. A similar competing effort is being pursued by the IEEE 802.1aq working group which is defining a standard for shortest path bridging (SPB) of unicast and multicast frames and which supports multiple active topologies. The SPB standard is expected to be ratified by the IEEE by early 2012.

A couple of emerging concepts that have the potential to impact data center LAN design are software defined networking and OpenFlow. One of the key underlying concepts of a software defined network is that the switch control plane is decoupled from the data plane and placed in a separate centralized server or controller. This centralization of the control plane makes it relatively easy to programmatically control the entire network. The programmatic control of the network is enabled by an abstraction layer or *network hypervisor* between the network operating system (NOS) control software and the packet forwarding data plane hardware. OpenFlow is an open API/protocol that is used between a network controller and a controlled physical switch that provides the forwarding hardware. The protocol is used to set flow table entries within the physical switch. The abstraction layer allows OpenFlow-enabled switches from different vendors to be mixed and matched without impacting the NOS.

As mentioned, one of the characteristics of the current generation of data center LANs is the separation of the data and storage networks. However, a possible characteristic of the next generation of data center LANs will be the convergence of block-level storage and data traffic over a common high-speed Ethernet data center switching fabric.
Traditional Ethernet, however, only provides a best effort service that relies on upper level protocols such as TCP to manage congestion and to recover lost packets through re-transmissions. In order to emulate the lossless behavior of a Fibre Channel (FC) SAN, Ethernet needs enhanced flow control mechanisms that eliminate buffer overflows for high priority traffic flows, such as storage access flows. Lossless Ethernet is based on a set of emerging standards, which are commonly referred to as IEEE Data Center bridging (DCB).

As was also mentioned, one of the primary factors that is driving IT organizations to redesign their data center LANs is the requirement to support server virtualization. In particular, when virtual machines (VMs) are migrated between servers, the network has to accommodate the constraints imposed by the VM migration utility; e.g., VMotion. Typically the VM needs to be on the same VLAN when migrated from source to destination server. This allows the VM to retain its IP address which helps to preserve user connectivity after the migration. An emerging approach that addresses some of the major limitations of live migration of VMs across a data center network is the Virtual eXtensible LAN (VXLAN). In addition to allowing VMs to migrate transparently across Layer 3 boundaries, VXLAN provides support for virtual networking at Layer 3, circumventing the traditional limitation of 4,094 VLANs, which is proving to be inadequate for VM-intensive enterprise data centers and for multi-tenant cloud data centers.

Many of the benefits of cloud computing depend on the ability to dynamically provision VMs and to migrate them at will among physical servers located in the same data center or in geographically separated data centers. The task of creating or moving a VM is a relatively simple function of the virtual server's management system. There can, however, be significant challenges in assuring that the VM's network configuration state, including VLAN memberships, QoS settings, and ACLs, is established or transferred in a timely fashion. In the current environment, the most common approach to automating the manual processes involved in VM provisioning and migration is based on communication between the Hypervisor Management system and the switch element management system (EMS) via APIs supported by both vendors. A somewhat different approach to automating data center configuration, including the provisioning and migration of VMs is based on orchestration engines, which are discussed in more detail in the management section of this report.

The Wide Area Network

As previously explained, after a lengthy period in which there was little or no fundamental innovation, the data center LAN is experiencing broad fundamental change. In contrast, after a lengthy period in which the WAN underwent repeated fundamental change, there are currently no fundamental changes in store for the WAN. In addition, the LAN follows Moore's Law. In contrast, the WAN doesn't follow Moore's Law and as a result, the price/performance of WAN services such as MPLS tends to improve by only a couple of percentage points per year.

One of the characteristics of cloud computing is increased reliance on the network. The increased reliance on the WAN in particular stems from the fact that the resources that support cloud computing solutions are centralized in a small number of data centers and the vast majority of users access these solutions over the WAN. Specific factors that are also driving more WAN traffic include the requirement to support virtual machine migrations, virtual desktops and collaboration.

As demonstrated by the survey data contained in this report, the increased WAN traffic is having an impact on WAN budgets. That survey data shows that over the next year, roughly forty percent of IT organizations will increase their WAN budget and in many cases, the increase will be significant. Given that IT organizations are expected to make increased use of cloud computing, IT organizations must either make changes to how they use WAN services, or else accept ongoing increases in their WAN budget.

The survey data also indicates that the primary WAN services used by IT organizations are MPLS and the Internet and that while IT organizations will increase their reliance on both MPLS and the Internet, they will make a relatively greater increase in their reliance on the Internet. The primary concerns that IT organizations have with the use of MPLS are cost, the lead-time to implement new circuits and uptime. The primary concerns that IT organizations have with the use of the Internet are uptime, latency and cost.

As previously discussed, 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. In a growing number of instances, Internet-based VPNs that use DSL for access are *good enough* to be a cloud network in part because the majority of IT organizations don't regard the SLAs that they receive from their network service providers for services such as MPLS as being effective. Another shift relative to the use of the Internet is that in a growing number of instances, IT organizations will avoid backhauling Internet traffic and will instead implement distributed access to the Internet from their branch offices. However, as important as the traditional Internet is to cloud computing, in many instances IT organizations will not utilize the traditional Internet to support cloud computing.

One of the key trends in network and application optimization is the deployment of virtual appliances; e.g., virtual WAN Optimization Controllers (vWOCs) and virtual Application Delivery Controllers (vADCs). 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. Another benefit of virtualized appliances is that in many instances the benefits of the dynamic movement of a VM from one server to another are maximized if the supporting infrastructure, including the WOCs and ADCs, is virtualized and can also be dynamically moved. In addition, there is significant interest in placing a WOC on premise at an IaaS provider's data centers. IT organization will have a notably easier time

placing an optimization device, whether that is a WOC or an ADC, at an IaaS provider's data center if the device is virtualized. That follows because if the device is virtualized, the IT organization can control the deployment of the functionality. If the device is physical, then the IT organization needs to get the IaaS provider to offer space for the device and to install it.

One of the ways that an IT organization can get better performance out of the Internet is by using an Internet overlay. An Internet overlay leverages service provider resources that are distributed throughout the Internet in order to optimize the performance, security, reliability, and visibility of the enterprise's Internet traffic. Another approach is to combine multiple ISP connections and to share traffic over the connections using policy based routing (PBR). Assuming that each ISP connection is a diversely routed DSL or cable access line and that one of the access lines has an availability of 99% and the other has an availability of 98%, then the system has an availability of 99.98%. Alternatively, if both access lines have an availability of 99%, then the system is available 99.99% of the time.

Unfortunately PBR can be difficult to administer and manage. The biggest limitation of the PBR approach, however, is that it creates a static allocation of traffic to multiple links and it doesn't have the ability to reallocate the traffic when the quality of one of the links degrades. The static nature of the policies means that unless there is an outage of one of the links that a given class of traffic will always be allocated to the same network connection.

Another way that an IT organization can better leverage the Internet is by implementing an aggregated virtual WAN (avWAN). This technology enables IT organizations to implement WANs based on multiple WAN services (e.g., MPLS, Frame Relay and the Internet) and/or WANs based on just multiple Internet VPN connections. An aggregated virtual WAN transcends simple PBR by dynamically recognizing application traffic and allocating traffic across multiple paths through the WAN based on real-time traffic analytics. An avWAN allows IT organizations to add significant amounts of additional bandwidth to an existing MPLS-based WAN at a relatively low incremental cost. In addition to enabling the augmentation of an MPLS WAN with inexpensive Internet connectivity, aggregated virtual WANs also give IT organizations the option to reduce its monthly ongoing expense by either eliminating or reducing its MPLS connections while simultaneously providing more bandwidth than the original network design provided.

As previously discussed, cloud balancing provides a lot of benefits. There are, however, a number of challenges associated with cloud balancing. For example, the VLANs within which VMs are migrated must be extended over the WAN between and amongst 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. In addition, application performance must meet user expectations regardless of the location of the users or the IT resources that the users are accessing. This means that the public cloud data centers need to offer the same WAN optimization and application acceleration capabilities that are deployed within the enterprise. 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.

There are a number of services and technologies that IT organizations can use to optimize the performance of applications and services that they get from a CCSP. One such service was previously mentioned – an Internet overlay. Similar services are beginning to be offered by CCSPs who provide a core network of their own, and then bundle in WOC functionality in their cloud data centers or POPs.

Somewhat of a new class of product is cloud optimized WOCs. These are purpose-built virtual WOC appliances for deployment in public cloud environments. Cloud optimized features include compatibility with cloud virtualization environments, SSL encryption and acceleration, and automated migration or reconfiguration of virtual WOCs in conjunction with VM provisioning or migration. A somewhat related new class of products is data mobility controllers (DMCs). The purpose of DMCs is to facilitate the transfer of high volume data between enterprise data centers or private cloud data centers.

Another emerging class of product is hypervisor–based multi-tenant ADC Appliances. Partitioned ADC hardware appliances have for some time allowed service providers to support a multi-tenant server infrastructure by dedicating a single partition to each tenant. Enhanced tenant isolation in cloud environments can be achieved by adding hypervisor functionality to the ADC appliance and by dedicating an ADC instance to each tenant. Each ADC instance is then afforded the same type of isolation as a virtualized server instance, with protected system resources and address space. A combination of hardware appliances, virtualized hardware appliances and virtual appliances provides the flexibility for a cloud service provider to offer highly customized ADC services that are a seamless extension of an enterprise customer's IT environment.

The Management of Cloud Computing

In spite of all of the hype around cloud computing, cloud computing has not been aggressively adopted by the majority of Global 2000 companies. However, in spite of that, the majority of IT organizations believe that getting better at managing private cloud computing solutions is either very or extremely important.

Just as IT organizations are getting somewhat comfortable with managing the performance of applications, they are being tasked with managing the performance of services. In part because the ongoing adoption of virtualization and cloud computing has created the concept of everything as a service (XaaS), the term service as used in this report will sometimes refer to services that IT organizations acquired from a public cloud computing provider; e.g., compute, storage, applications. The survey data indicates that managing this class of service is not a top priority for the majority of IT organizations.

The term service as used in this report will sometimes refer to business services that involve multiple inter-related applications. The majority of IT organizations believe that getting better at managing inter-related applications that comprise a business service is either very or extremely important. Unfortunately, the majority of IT organizations don't do a good job of managing the performance of traditional applications. Adding to that challenge is the fact that the adoption of cloud computing will further complicate the task of managing both traditional applications as well as the inter-related applications that comprise a service. That follows in part because in a cloud computing environment, the applications that comprise the service will increasingly be supported by an infrastructure that is virtual and managing a virtual infrastructure is very challenging.

As recently as two or three years ago, few IT organizations offered an SLA to the company's business and functional managers; a.k.a., an internal SLA. However, in the current environment, the vast majority of IT organizations provide an internal SLA for at least some applications. While IT organizations have made significant progress over the last few years relative to offering an internal SLA, two thirds of IT organizations believe that over the next year that it is either very or extremely important to get better at effectively managing internal SLAs. As previously noted, most services and applications that are provided by CCSPs are provided on a best effort basis. As such, the lack of meaningful SLAs for public cloud services is a deterrent to the Global 2000 adopting these services for delay-sensitive, business-critical applications.

A fact that has been true for years and which remains true today is that getting better at doing root cause analysis is the most important management task facing the vast majority of IT organizations. It is not surprising that rapidly identifying the root cause of degraded application performance is so important to IT organizations in part because on an ever increasing basis a company's key business processes rely on a handful of applications. That means that if those applications are not running well, neither are those key business processes.

As previously noted, the adoption of server virtualization is one of the factors that is driving IT organizations to re-think their approach to designing data center LANs. Server virtualization also causes management challenges. One of the primary management challenges associated with server virtualization is the requirements to perform management tasks on a per-VM basis. In fact, 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.

Unfortunately, the adoption of cloud computing makes troubleshooting application performance an order of magnitude more difficult than it is in a traditional environment.
One of the challenges associated with managing in any environment is that it is difficult to know the end-to-end path that packets take across a network. This management complexity comes in part from the distributed nature of IP which results in the lack of a single repository of routing information in the network. This lack of knowledge complicates tasks such as troubleshooting. The difficulty of knowing the path from origin to destination is greatly increased in a cloud computing environment because applications and services can be dynamically moved between servers both within and between data centers. Another fundamental issue relative to managing either a public or hybrid cloud computing service is that the service has at least three separate management domains: the enterprise, the WAN service provider(s) and the various cloud computing service providers.

There are a number of services and technologies that IT organizations can use to manage the applications and services that they get from a CCSP. One such class of service was previously mentioned – a cloud networking service. Over the next year, more than a quarter of IT organizations will either likely acquire or will acquire security and/or management functionality from a CCSP. A technology that can help manage CCSP provided applications and services is route analytics. Route analytics provides visibility, analysis, and diagnosis of the issues that occur at the routing layer in complex, meshed networks. As such, route analytics can help IT organizations identify changes in the Layer 3 network, such as the movement of a VM.

Another technology that can help IT organizations to manage the applications and services that they get from a CCSP is a highly scalable and integrated DNS/DHCP/IPAM solution, which is also well integrated with the virtual server management system.
Where DNS/DHCP/IPAM share a common database, the integration obviates the need to coordinate records in different locations and allows these core services to accommodate any different addressing and naming requirements of physical and virtual servers. Potential advantages of this approach include the automated generation of IP addresses for newly created VMs, the automated allocation of subnets for new VLANs, and the population of an IP address database with detailed information about the current location and security profiles of VMs.

An increasingly popular approach to building cloud data centers is based on pre-integrated and certified infrastructure packages from either a broadly-based IT equipment vendor, a group of partners or a joint venture formed by a group of complementary vendors. These packages typically are offered as turn-key solutions and include compute, server virtualization, storage, network, and management capabilities. Other data center functions such as WOCs, ADCs, APM and security functionality may also be included. In order to realize the full potential of the converged IT infrastructure, the management system must provide a unified, cross-domain approach to automated element management, provisioning, change management and operations management. Some of the most critical aspects of managing a cloud data center include integrated and automated infrastructure and service management; secure multi-tenancy and support for enterprise co-management.

Service orchestration is another technique that helps IT organizations automate many of the manual tasks that are involved in provisioning and controlling the capacity of dynamic virtualized services. Orchestration engines are available as standalone management products or as part of complete suites of management tools that are focused on the data center. In addition, the management systems that are integrated with converged infrastructure solutions typically include some orchestration capabilities.

The Emergence of Cloud Computing and Cloud Networking

The Goal of Cloud Computing

Within the IT industry there isn't a universally accepted definition of what is meant by cloud computing. This report 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.

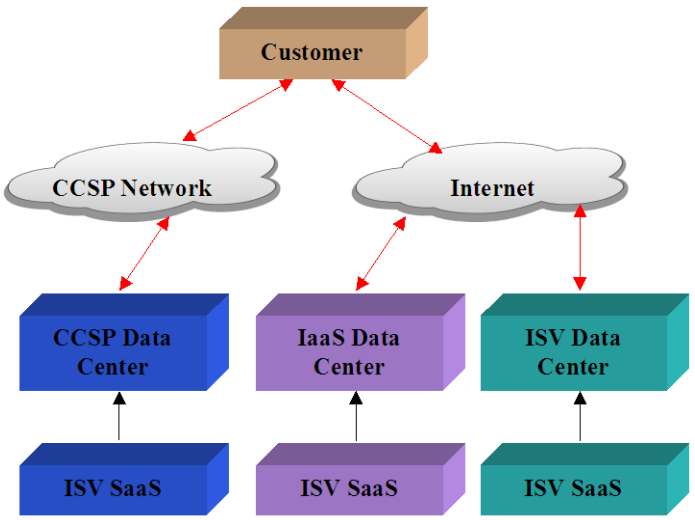
In order to demonstrate the concept behind the phrase *good enough,* consider just the availability of an IT service. In those cases in which the IT service is business critical, *good enough* could mean five or six 9's of availability. However, in many other cases *good enough* has the same meaning as *best effort* and in these cases *good enough* could mean two or three 9's of availability. The instances in which an approach that provides two or three 9's of availability is acceptable are those instances in which the IT service isn't business critical and that approach is notably less expensive than an alternative approach that offers higher availability.

On a going forward basis, IT organizations will continue to need to provide the highest levels of availability and performance for a small number of key services. However, an ever-increasing number of services will be provided on a best effort basis.

In most instances the SLAs that are associated with public cloud computing services such as Salesforce.com or Amazon's Simple Storage System are weak and as such, it is reasonable to say that these services are delivered on a best effort basis. For example, most of the SLAs that are associated with public cloud computing services don't contain a goal for the end-to-end performance of the service. The reason for the lack of performance guarantees stems from the way that most public cloud computing services are delivered. As shown in Figure 1, one approach to providing public cloud computing services is based on the service 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. Both of these approaches 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.

The fact that cloud computing service providers (CCSPs) don't provide an end-to-end performance SLA for applications delivered over the Internet will not change in the foreseeable future. However, as will be described in a subsequent section of this report, there are things that can be done to improve the performance of applications delivered over the Internet.

Figure 1: Distribution Models for Cloud-Based Solutions



An approach to providing public cloud computing services that does lend itself to offering more meaningful SLAs is based on a CCSP providing these solutions to customers from the CCSP's data center and over a network that is provided by the CCSP and based on a technology such as MPLS.

Organizations that utilize best effort cloud computing services do so with the implicit understanding that if the level of service they experience is not sufficient; their primary recourse is to change providers. It may seem counter-intuitive that a company would utilize public cloud computing services for which end-to-end performance SLAs are essentially non-existent. However, as described in a subsequent section of this report, two thirds of The Webtorials Respondents indicated that the SLAs that they receive from their network service providers for services such as MPLS are either not worth the paper they are written on, or that the SLAs they receive are not much better than nothing.

SLAs from both traditional network service providers as well as public cloud computing providers are a work in progress.

Characteristics of Cloud Computing Solutions

The following set of bullets identifies the primary characteristics of cloud computing solutions. There is not, however, a litmus test to determine if a particular service is or is not a cloud computing service.

- Centralization of applications, servers, data and storage resources.
- Extensive virtualization of every component of IT, including servers, desktops, applications, storage, switches, routers and appliances such as WAN optimization controllers, application delivery controllers and firewalls.
- Automation and Orchestration of as many tasks as possible; e.g., provisioning, troubleshooting, change and configuration management.
- The dynamic creation and movement of resources such as virtual machines and the associated storage.
- Heavy reliance on the network.
- Self-service to allow end users to select and modify their use of IT resources without the IT organization being an intermediary.
- Usage sensitive chargeback that is often referred to as pay-as-you-go. An alternative is for IT organizations to show the consumption of IT resources by certain individuals or organizations; a.k.a., showback.
- Simplification of the applications and services provided by IT.
- Standardization of the IT infrastructure.
- Technology convergence such as the convergence of LAN and SAN and of switch and server.
- The development of standards that enable, among other things, the federation of disparate cloud computing infrastructures with one another (see below).
- The federation of disparate cloud computing infrastructures with one another.

Classes of Cloud Computing Solutions

There are three classes of cloud computing solutions that will be described in this section of the report. Those classes are private, public and hybrid.

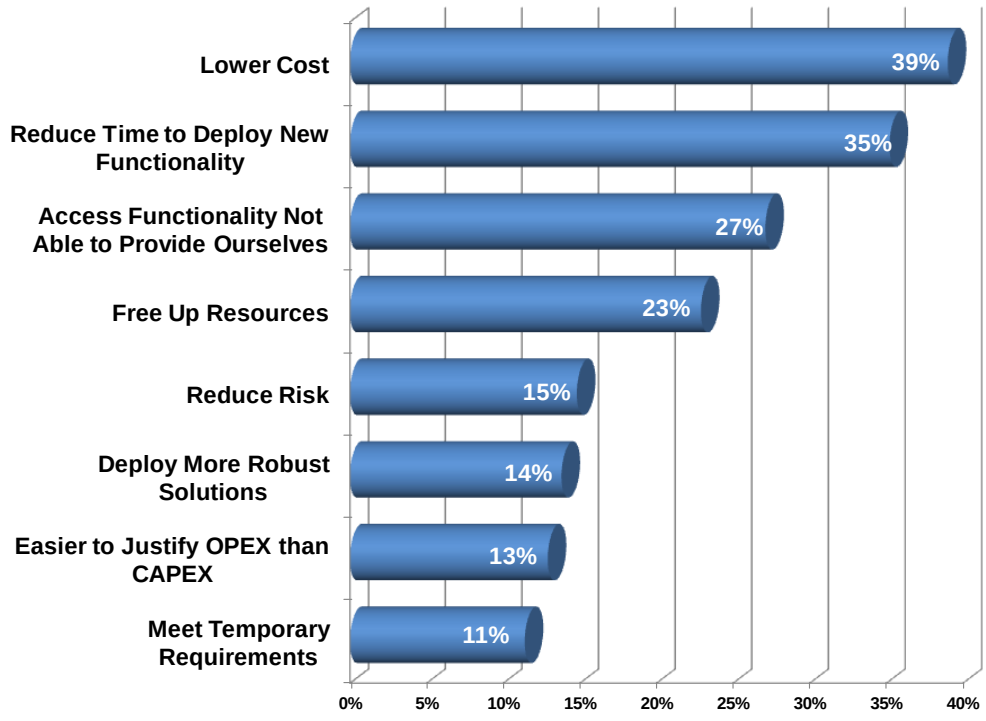
Private Cloud Computing

Many IT organizations have decided to implement some of the characteristics of cloud computing solutions described in the preceding subsection within their internal IT environment. This approach is usually referred to as a *Private Cloud*. As previously noted there is not a litmus test to determine which characteristics have to be in a solution for the solution to be deemed to be a cloud computing solution. As a result, an IT organization that has centralized some of all of its servers into their data centers or into a collocation site, virtualized some of all of those servers, implemented some additional automation and that also moves virtual machines (VMs) between servers can reasonable claim that they have implemented a private cloud.

Public Cloud Computing

CCSPs that provide their services either over the public Internet or over other WAN services are offering a class of solution that is often referred to as the *public cloud* or *public cloud computing.* The research report entitled *Cloud Computing: A Reality Check and Guide to Risk Mitigation* presented the results of a survey in which 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 2.

Figure 2: The Drivers of Public Cloud Computing



One of the observations that can be drawn from Figure 2 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.

That research report also pointed out that the primary factor that inhibits IT organizations from acquiring public cloud computing solutions is the concern over the security and confidentiality of data. Hence, it appears to be counter intuitive that almost 15% of the survey respondents indicated that reducing risk was a factor that would cause them to use a public cloud computing solution. In most cases the survey respondent's 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.

As described in the report entitled A Guide for Understanding Cloud Computing[4], the two primary types of services provided by CCSPs are Software-as-a-Service (SaaS) and Infrastructure-as-a-Service (IaaS)[5].

Software-as-a-Service

According to IDC[6], 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. One of the key characteristics of the SaaS marketplace is that:

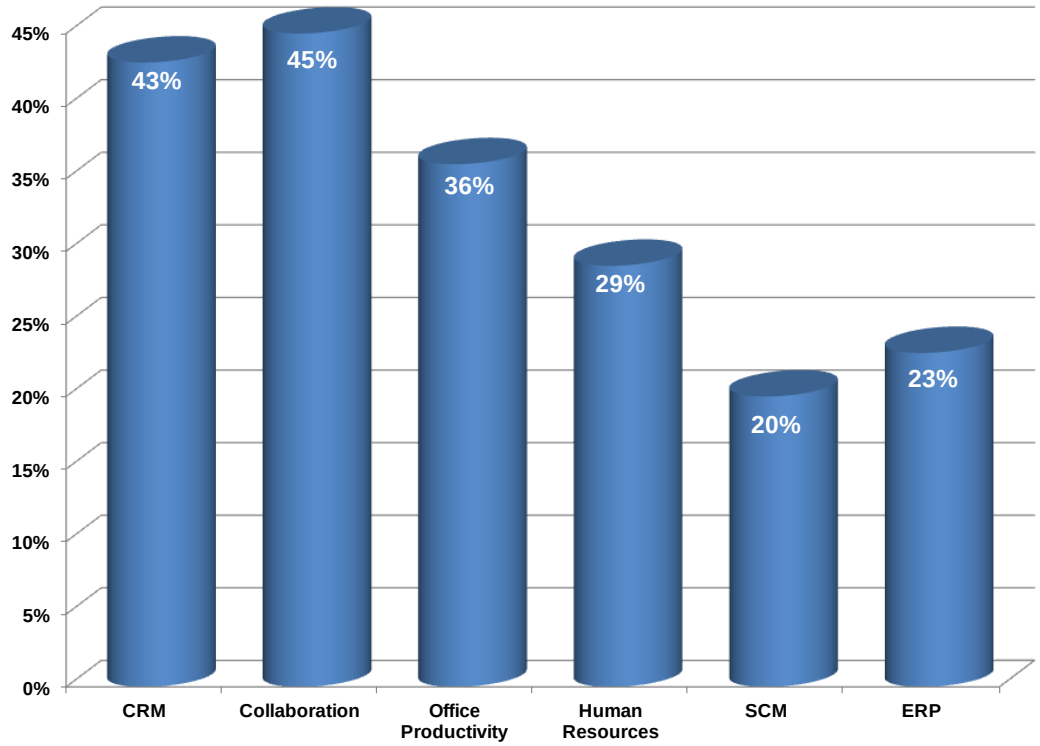
The SaaS marketplace is comprised of a small number of large players such as Salesforce.com, WebEx and Google Docs as well as thousands of smaller players.

One of the reasons why there are so many players in the SaaS market is that the barrier to entry is relatively low.

The research report entitled Cloud Computing: A Reality Check and Guide to Risk Mitigation[7] reported on the results of a survey in which the survey respondents were asked about their company's use of SaaS-based applications. Figure 3 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.

[4] http://www.webtorials.com/content/2009/11/a-guide-for-understanding-cloud-computing.html
[5] A third form of service provided by a CCSP is Platform-as-a-Service (PaaS). Because it hasn't been widely adopted, it will not be included in this report.
[6] http://www.businesswire.com/news/home/20100726005135/en/SaaS-Revenue-Grow-Times-Faster-Traditional-Packaged
[7] http://www.webtorials.com/content/2009/12/cloud-computing-a-reality-check-guide-to-risk-mitigation.html

Figure 3: Popular Categories of SaaS-Based Applications



The functionality provided by each of the six categories of applications listed in Figure 3 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 3, there are tens, and sometimes hundreds, of SaaS-based solutions currently available[8]. Table 1 contains a listing of some representative SaaS providers for each category.

Table 1: Representative SaaS Providers

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

[8] http://www.saas-showplace.com/saasproviderdirectory/saasapplicationcategory.html

One of the key challenges facing IT organizations that use SaaS-based applications is improving the performance, management and security of those applications.

Infrastructure as a Service (IaaS)

Over the last few years, IaaS solutions have been comprised primarily of the basic compute and storage resources that are required to run applications. The barrier to enter the IaaS marketplace is notably higher than is the barrier to enter the SaaS marketplace. That is one of the primary reasons why there are fewer vendors in the IaaS market than there are in the SaaS market. Representative IaaS vendors include Amazon, AT&T, CSC, GoGrid, IBM, Joyent, NTT Communications, Orange Business Services, Rackspace, NaviSite (recently acquired by Time Warner), Savvis (recently acquired by Century Link), Terremark (recently acquired by Verizon) and Verizon. As the preceding sentence indicates, the IaaS market is going through a period that is characterized by mergers and acquisitions. The IaaS market is also expected to exhibit significant growth in the next few years. For example, Gartner[9] estimates that the IaaS market will grow from $3.7 billion in 2011 to $10.5 billion in 2014.

Table 2 provides a high level overview of some of the services offered by IaaS vendors. The data in Table 2 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[10] is off line for 5 minutes and a replacement instance cannot be launched from another Availability Zone[11] within Amazon's geographical region. Not all IaaS providers have a similar definition of availability.

Table 2: 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	320 GB included	Included with

9 http://www.qas.com/company/data-quality-news/iaas_market_to_record_strong_growth__7178.htm
10 http://aws.amazon.com/ec2/instance-types/
11 http://docs.amazonwebservices.com/AWSEC2/latest/UserGuide/index.html?Welcome.html

Table 2: Representative IaaS Providers			
	Amazon AWS	**RackSpace**	**GoGrid**
	Store) 10₵/GB/month 10₵/million I/O requests/month	with server	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

Table 2 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[12]. Cedexis presented data that represented roughly 17 billion measurements that were taken between March 15, 2011 and April 15 2011. As shown in Figure 4, none of the IaaS providers that were monitored delivered availability that was greater than 95%.

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

Figure 4: Availability of Server Instances at Various IaaS Providers
(source: Cedexis)

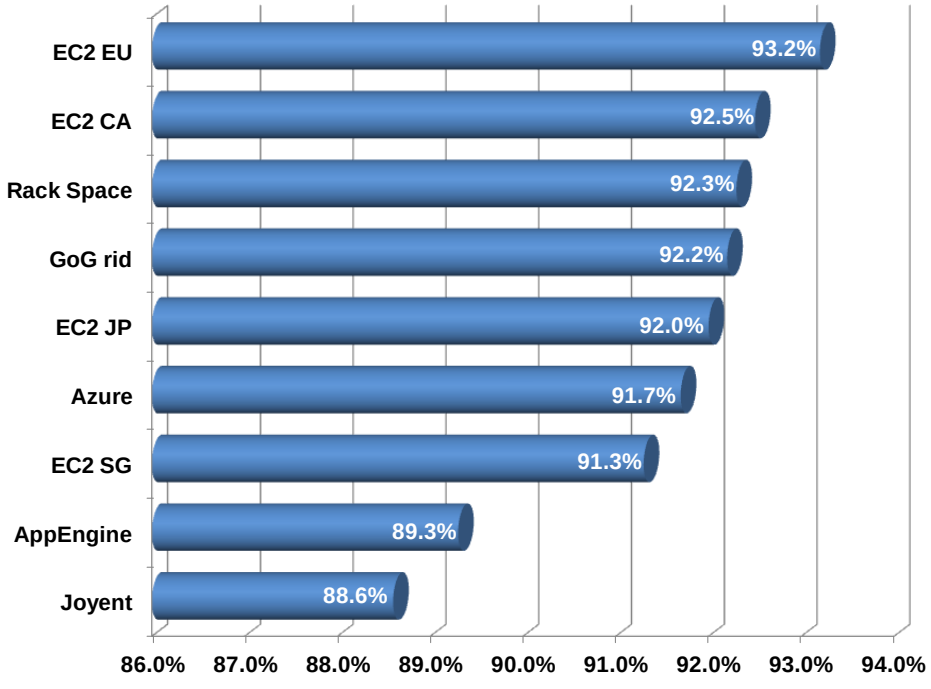


Figure 4 illustrates that:

The availability of IaaS solutions can vary widely.

In addition, similar to the situation with SaaS-based applications,

One of the key challenges facing IT organizations that use IaaS-based solutions is improving the performance, management and security of those solutions.

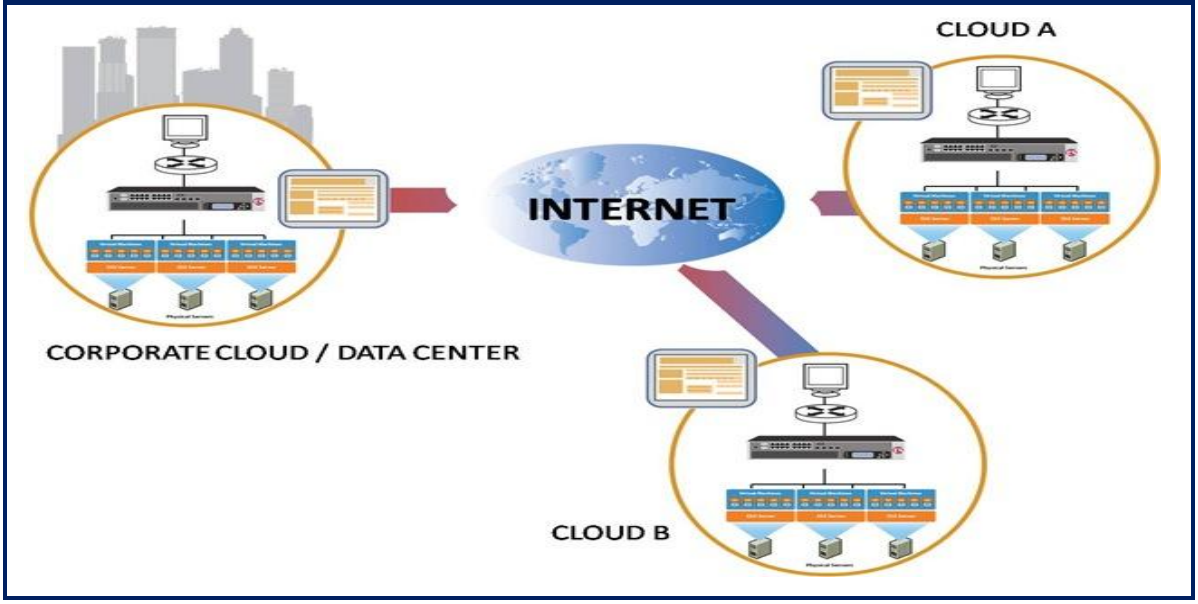
Hybrid Cloud Computing

Like so much of the terminology of cloud computing, there is not a uniformly agreed to definition of the phrase *hybrid cloud computing*. According to Wikipedia[13], "Hybrid cloud is a composition of two or more clouds (private, community, or public) that remain unique entities but are bound together, offering the benefits of multiple deployment models. Briefly it can also be defined as a multiple cloud systems which are connected in a way that allows programs and data to be moved easily from one deployment system to another."

Based on this definition, one form of a hybrid cloud is an n-tier application in which the web tier is implemented within one or more public clouds while the application and database tiers are implemented within a private cloud. Another form of hybrid cloud that receives a lot of attention is cloud balancing. The phrase *cloud balancing* refers to routing service requests across multiple data centers based on myriad criteria. As shown in **Figure 5**, cloud balancing involves one or more corporate data centers and one or more public cloud data centers.

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

Figure 5: 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:

[13] http://en.wikipedia.org/wiki/Cloud_computing#Hybrid_cloud

- 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.

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.

Manage 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 it can be performed across multiple independent locations of a single cloud service provider.

Emerging Public Cloud Computing Services

Data Center Services

Most of the IaaS providers do not want to compete entirely based on providing commodity services such as basic compute and storage. As such, many IaaS providers are implementing higher value-added data center services such as the ones described below.

Private Cloud Data Center Services

These services are based on outsourcing the enterprise's multi-tier private data center to a service provider. The data center could be located at either a site controlled by the enterprise or at a service provider's site. In most cases service providers will structure these services so that the customers receive the highest levels of support, as well as assurances written into the corresponding SLA for high levels of availability, performance and security. A private WAN service would typically be used to provide access to these services.

Virtual Private Data Center (VPDC)

These services provide an instance of an entire data center hosted on a service provider's infrastructure that is optimized to provide a high level of security and availability for multiple tenants. From the service provider's perspective, the data center architecture for the VPDC would be similar to the architecture used for a private cloud data center except that the resources would be shared among a number of customers rather than being dedicated to a single customer or tenant. The service provider's architecture needs to effectively leverage virtualization in order to maximize the efficient usage of a shared pool of resources. The architecture also needs to allow for a high degree of flexibility in providing a broad range of required network capabilities. This includes WAN optimization, load balancing and firewall services. Service management software should be in place to enable the co-management of the VPDC by customers and providers.

The hybrid cloud computing model works best in those instances in which the VPDC and the private cloud data center are based on the same hypervisors, hypervisor management systems and cloud controllers. This maximizes the enterprise's control over the hybrid cloud and allows application and server management to remain the responsibility of the enterprise. Access to a VPDC could be provided either over the Internet or a private WAN service.

Cloud Networking Services

As shown in Figure 3, with the exception of collaboration, the applications that organizations have acquired from CCSPs have typically been enterprise applications such as CRM. As was previously mentioned, over the last few years IaaS solutions have been comprised primarily of the basic compute and storage resources that are required to run applications. Recently, a new class of solutions has begun to be offered by CCSPs. These are solutions that have historically been provided by the IT infrastructure group itself and include network and application optimization, VoIP, Unified Communications (UC), security, network management and virtualized desktops. This new class of solutions will be referred to in this report as Cloud Networking Services (CNS).

A recent research report entitled Cloud Networking Services[14] presented the results of a survey in which 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 3.

Table 3: Interest in Cloud Networking Services

	Will Not Happen	Might Happen	50/50 Chance	Will Likely Happen	Will Happen
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%
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 3 shows that the interest in CNS 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 six rows of Table 3 would either likely be acquired or would be acquired.

Cloud Networking Services represents the beginning of what could be a fundamental shift in terms of how IT services are provided.

As noted, the two primary forms of public cloud computing are SaaS and IaaS. It would be possible to make a technical argument that at least some CNS solutions are SaaS solutions and that some others are IaaS solutions. While technology is one way to classify CNS solutions, a more compelling way is to look at how the typical IT organization is structured. Most IT organizations have an applications organization whose primary role is to develop, acquire and maintain enterprise applications such as CRM, ERP and SCM. Most IT organizations also have an infrastructure organization whose primary role is to provide, manage, secure and optimize the networks and servers that support the applications that enable the company's business processes. In most cases, services such as voice, collaboration, disaster recovery, management, security, optimization and virtual desktops are provided by the infrastructure organization – not the applications organization. Because of the way that IT organizations are

[14] http://www.webtorials.com/content/2011/09/2011-cloud-networking-services.html

typically structured, throughout this report CNS solutions will be considered to be the next wave of IaaS solutions.

Since CNS solutions are just one more form of public cloud computing, when evaluating these solutions IT organizations also need to understand the degree to which these solutions overcome the factors that impede the use of any public cloud computing solution. As previously mentioned, concerns about security is the primary impediment to the adoption of public cloud computing solutions and hence evaluating the security of the CNS provider's facilities is a critical component of evaluating a CNS solution.

However, just as important as whether or not the CNS solution provides adequate security is whether or not the solution actually provides the benefits (Figure 2) that drive IT organizations to use public cloud computing solutions. The primary benefit of using a public cloud computing solution is lower cost. While it can be tricky to compare the usage sensitive pricing of the typical CNS solution with the fully loaded cost of a premise based solution, the cost information provided by the CCSP should give the IT organization all the information it needs to do that analysis. The second most important benefit of using a public cloud computing solution is being able to reduce the time it takes to deploy new functionality. Evaluating the agility of a CCSP is notably more difficult than evaluating their cost structure.

One way for an IT organization to evaluate the agility of a CCSP is to identify the degree to which the CCSP has virtualized their infrastructure.

This follows because a virtual infrastructure is notably easier to initialize, scale and migrate than a physical infrastructure is. Since the vast majority of CCSPs implement virtualized servers, server virtualization is unlikely to distinguish one CCSP from another. What can distinguish one CCSP from another is the degree to which they have virtualized other components of their infrastructure. One such component is networking. By implementing routing software that runs on top of the most common hypervisors, CCSPs increase their ability to quickly provision and configure capacity. This approach to providing routing functionality also maps more closely to the usage sensitive pricing that most CCSPs offer.

The Culture of Cloud Computing

The rest of this report will discuss the networking technologies that enable cloud computing. However, as much as cloud computing is about technologies it is also about changing the culture of the IT organization. One such cultural shift was described in the preceding subsection entitled "The Goal of Cloud Computing".

To put this cultural shift into perspective, it is important to realize that it is implicit in the traditional IT culture to implement ongoing enhancements to make the network and the IT services that are delivered over the network, increasingly resilient. The adoption of cloud computing changes that and as previously described, in some instances it is becoming acceptable for IT services to be delivered on a best effort basis. A clear indication of that change is the success of Salesforce.com. Salesforce.com has three million customers who use their solutions to support critical sales processes. Yet in spite of the importance of the application, in virtually all cases Salesforce.com will not give a customer an availability guarantee and since the application is typically accessed over the Internet, it doesn't come with an end-to-end performance guarantee.

One of the other cultural shifts that is associated with the adoption of cloud computing is that IT organizations become less of a provider of IT services and more of a broker of IT services. In the traditional IT environment, the IT organization is the primary provider of IT services. Part of the challenge that is associated with the IT organization being the primary provider of IT services is that sometimes the IT organization can't meet the needs of the business units in a timely fashion. In the past the way that business unit managers have dealt with this lack of support is by having their own shadow IT organization whereby the business unit managers have some people on their staff whose role is to provide the IT services that the business unit manager can't get from the IT organization. In the current environment, public cloud providers often play the role of a shadow IT organization by providing a company's business unit managers services or functionality that they either can't get from their IT organization or they can't get in a timely manner. In some instances the IT function is in a position to stop the non-sanctioned use of public cloud computing once they find out about it. However, in many other instances they aren't.

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

IT organizations provide considerable value by being the broker between the company's business unit managers and cloud computing service providers.

Another cultural change that is associated with the adoption of cloud computing is the implementation of more usage sensitive chargeback. Usage sensitive chargeback is not new. Many IT organizations, for example, allocate the cost of the organization's network to the company's business unit managers based on the consumption of that network by the business units. Since there has traditionally been a lot of overhead associated with usage sensitive chargeback, usage sensitive chargeback has only made sense in those situations in which the

IT organization is in a position both to explain to the business unit managers in easily understood language, what they are paying for and to provide suggestions as to how the business unit managers can reduce their cost. In the current environment, roughly fifty percent of all IT organizations implement usage sensitive chargeback for at least some components of IT. However, relatively few implement it broadly. Input from The Webtorials Respondents indicates that over the next two years IT organizations will make increased use of usage sensitive chargeback. Most of this increased use will come from having the business unit managers pay the relevant cloud computing service providers for the services that their organization consumes. The movement to implement more usage sensitive chargeback over the next two years will not be dramatic because:

The culture of an IT organization changes very slowly.

The Emerging Data Center LAN

First and Second Generation Data Center LANs

As recently as the mid 1990s Local Area Networks (LANs) were based on shared media. Throughout this report these shared media LANs will be referred to as First Generation LANs. In the mid 1990s, companies such as Grand Junction introduced Ethernet LAN switches to the marketplace. The two primary factors that drove the deployment of Second Generation LANs based on switched Ethernet were performance and cost. For example, performance drove the deployment of switched Ethernet LANs in data centers because FDDI, which was the only viable, high-speed First Generation LAN technology, was limited to 100 Mbps whereas there was a clear path for Ethernet to evolve to continually higher speeds. Cost was also a factor that drove the deployment of Ethernet LANs in data centers because FDDI was fundamentally a very expensive technology.

A key characteristic of Second Generation data center LANs is that they are usually designed around a three-tier switched architecture comprised of access, distribution and core switches. The deployment of Second Generation LANs is also characterized by:

- The use of the spanning tree protocol at the link layer to ensure a loop-free topology.
- Relatively unintelligent access switches that did not support tight centralized control.
- The use of Ethernet on a best-effort basis by which packets may be dropped when the network is busy.
- Support for applications that are neither bandwidth intensive nor sensitive to latency.
- Switches with relatively low port densities.
- High over-subscription rate on uplinks.
- The separation of the data network from the storage network.
- VLANs to control broadcast domains and to implement policy.
- The need to primarily support client server traffic; a.k.a., north-south traffic.
- Redundant links to increase availability.
- Access Control Lists (ACLs) for rudimentary security.
- The application of policy (QoS settings, ACLs) based on physical ports.

Drivers of Change

The Webtorials Respondents were asked "Has your IT organization already redesigned, or within the next year will it redesign, its data center LAN in order to support cloud computing in general, and virtualized servers in particular?" Their responses are shown in Table 4.

Table 4: Redesign of the Data Center LAN			
	Already Have	**Will Within the Next Year**	**No Plans**
Cloud Computing in General	21.8%	51.1%	27.1%
Virtualized Servers in Particular	53.7%	34.0%	12.2%

One conclusion that can be drawn from the data in Table 4 is that the majority of IT organizations have already begun the process of redesigning their data center LANs. Another conclusion is that:

One of the key factors driving IT organizations to redesign their data center LANs is the deployment of virtual servers.

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

Table 5: Deployment of Virtualized Servers					
	None	**1% to 25%**	**26% to 50%**	**51% to 75%**	**76% to 100%**
Have already been virtualized	15%	33%	21%	18%	14%
Expect to be virtualized within a year	6%	25%	28%	20%	20%

In early 2010, the Webtorials survey base was asked to indicate the percentage of their data center servers that had already been virtualized. Their responses are shown in Table 6.

Table 6: Deployment of Virtualized Servers as of Early 2010					
	None	**1% to 25%**	**26% to 50%**	**51% to 75%**	**76% to 100%**
Have already been virtualized	30%	34%	17%	11%	9%

The data in Table 5 and Table 6 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 Webtorials 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 5 and Table 6 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 Webtorials Respondents predict that within a year, that only 6% of IT organizations will not have virtualized any of their data center servers.

As pointed out in Virtualization: Benefits, Challenges and Solutions[15], server virtualization creates a number of challenges for the data center LAN. One of these challenges is the requirement to manually configure parameters such as QoS settings and ACLs in order to support the dynamic movement of VMs. In order to quantify the extent to which IT organizations move VMs between physical servers, The Webtorials Respondents were asked to indicate whether they agreed or disagreed with the statements in the left hand column of Table 4.

Table 7: Movement of VMs

	Agree	Disagree
We currently manually migrate VMs between servers in the same data center	67.4%	32.6%
We currently automatically migrate VMs between servers in the same data center	55.9%	44.1%
We currently manually migrate VMs between servers in disparate data centers	42.6%	57.4%
We currently automatically migrate VMs between servers in disparate data centers	26.6%	73.4%

The data in Table 7 indicates the great interest that IT organizations have in moving VMs between physical servers. However, as will be described throughout this section of the report, moving VMs between physical servers can be very complex.

Manually configuring parameters such as QoS settings and ACLs in order to support the dynamic movement of VMs is not the only challenge that is associated with server virtualization. Other challenges include:

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

- **Limited VM-to-VM Traffic Visibility**
  Traditional vSwitches don't have the same traffic monitoring features as do physical access switches. This limits the IT organization's ability to do security filtering, performance monitoring and troubleshooting within virtualized server domains in both private, public and hybrid clouds.

- **Inconsistent Network Policy Enforcement**

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

Traditional vSwitches can lack some of the advanced features that are required to provide the degree of traffic control and isolation required in the data center. This includes features such as private VLANs, quality of service (QoS) and sophisticated ACLs.

- **Layer 2 Network Support for VM Migration**
  When VMs are migrated, the network has to accommodate the constraints imposed by the VM migration utility; e.g., VMotion. 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.

  Server virtualization, however, is not the only factor that is causing IT organizations to redesign their data center LANs. The left hand column in Table 8 contains a list of the factors that are driving data center redesign. The center column shows the percentage of The Interop Respondents who in the fall of 2010 indicated that the corresponding factor was the primary factor that is driving their organization to redesign their data center LAN. The right hand column shows the percentage of The Webtorials Respondents who recently indicated that the corresponding factor was the primary factor that is driving their organization to redesign their data center LAN.

Table 8: Factors Driving Data Center LAN Redesign

Factor	Percentage of The Interop Respondents in 2010	Percentage of The Webtorials Respondents in 2011
To reduce the overall cost	22.4%	24.6%
To support more scalability	11.6%	20.8%
To create a more dynamic data center	11.6%	12.6%
To support server virtualization	11.2%	12.1%
To reduce complexity	9.9%	5.3%
To make it easier to manage and orchestrate the data center	9.2%	13.0%
To support our storage strategy	7.5%	3.4%
To reduce the energy requirements	6.5%	1.0%
Other (please specify)	6.1%	3.4%
To make the data center more secure	4.1%	3.9%

The data in Table 8 indicates that a broad range of factors are driving IT organizations to re-design their data center LANs. For example, making it easier to manage and orchestrate the

data center is becoming a key driver in how IT organizations design their data center LANs. However, as was the case with the adoption of the second generation of data center LANs:

The primary factors driving IT organizations to re-design their data center LAN is the desire to reduce cost and support scalability.

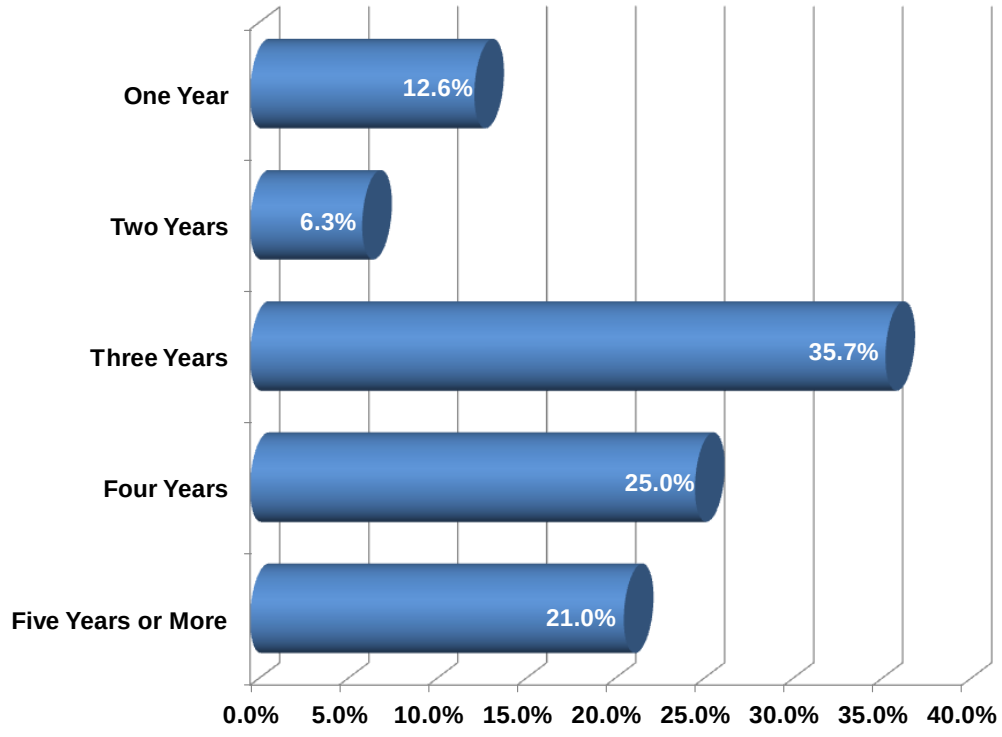
The conventional wisdom in the IT industry is that the cost of the power consumed by data center LAN switches is not significant because it is a small percentage of the total amount of power that is consumed in the typical data center. There is the potential for that situation to change going forward as 10 Gbps, 40 Gbps and 100 Gbps LAN interfaces will potentially consume considerably more power than 1 Gbps LAN interfaces currently do. As such, a requirement of third generation data center LAN switches is that the amount of power that they consume is only marginally more than what is consumed by second generation data center LAN switches and that these switches provide functionality to intelligently manage the power consumption during off peak hours.

Third Generation Data Center LAN Architecture and Technology Options

During the transition from First Generation LANs to Second Generation LANs there was considerable debate over the underlying physical and data link technologies. Alternative technologies included Ethernet, Token Ring, FDDI/CDDI, 100VG-AnyLAN and ATM. One of the few aspects of Third Generation Data Center LANs that is not up for debate is that they will be based on Ethernet. In fact, the Third Generation LAN will provide the possibility of leveraging Ethernet to be the single data center switching fabric, eventually displacing special purpose fabrics such as Fibre Channel for storage networking and InfiniBand for ultra low latency HPC cluster interconnect.

Many of the technologies that are discussed in this chapter are still under development and will not be standardized for another year or two. In order to understand whether or not IT organizations account for emerging technologies in their planning, The Webtorials Respondents were asked to indicate their company's planning horizon for the evolution of their data center LANs. To avoid ambiguity, the survey question stated "A planning horizon of three years means that you are making decisions today based on the technology and business changes that you foresee happening over the next three years." Their answers are shown in Figure 6.

Figure 6: Planning Horizon for Data Center LANs



The data in Figure 6 indicates that almost 75% of IT organizations have a planning horizon of three years or longer. Since most of the technologies discussed in this chapter will be standardized and ready for production use in three years, that means that the vast majority of IT

organizations can incorporate most of the technologies discussed in this chapter into their plans for data center LAN design and architecture.

Below is a discussion of some of the primary objectives of a Third Generation Data Center LAN and an analysis of the various alternatives that IT organizations have relative to achieving those objectives.

Two Tier Data Center LAN Design

There are many on-going IT initiatives that are aimed at improving the cost-efficiency of the enterprise data center. This includes server virtualization, SOA, Web 2.0, access to shared network storage as well as the implementation of HPC and cluster computing. In many cases these initiatives are placing a premium on IT organizations being able to provide highly reliable, low latency, high bandwidth communications among both physical and virtual servers. Whereas the hub and spoke topology of the traditional three-tier Second Generation LAN was optimized for client-to-server communications that is sometimes referred to as *north-south* traffic, it is decidedly sub-optimal for server-to-server communications, which is sometimes referred to as *east-west* traffic.

One approach for improving server-to-server communications is to flatten the network from three tiers to two tiers consisting of access layer and aggregation/core layer switches.

A two-tier network reduces the number of hops between servers, reducing latency and potentially improving reliability. The typical two-tier network is also better aligned with server virtualization topologies where VLANs may be extended throughout the data center in order to support dynamic VM migration at Layer 2.

The Interop Respondents were asked, “Two years from now, what is the fewest number of layers that you expect will be in any of your company’s data center LANs.” The answers of one hundred and ninety respondents are summarized in Table 9.

Table 9: Anticipated Number of Layers in a Data Center LAN

Number of Layers	Percentage of Respondents
4	8%
3	37%
2	38%
1	17%

The data in Table 9 indicates that while just over a third of IT organizations expect to still be running traditional, three-tier data center LANs in two years, the majority of the IT professionals who answered the question expect to be running a flatter data center LAN in that time frame. However, what is even more interesting is that two hundred and sixty five members of the pool of survey respondents answered the question with “don’t know”. That means that the number of survey respondents who don’t know how many layers will be in their data center LANs in two years is notably greater than the number of survey respondents that do know.

There is significant desire on the part of IT organizations to flatten their data center LANs, but there is also significant uncertainty relative to how flat they will become in the next two years.

As discussed below, two tier networks require switches that have very high densities of high-speed ports and a higher level of reliability to protect the soaring volumes of traffic flowing through each switch. As is also discussed below, the requirement for increased reliability and availability creates a requirement for redundant switch configurations in both tiers of the network.

High Port Density and Port Speed

The network I/O requirements of multi-core physical servers that have been virtualized are beginning to transcend the capacity of GbE and multi-GbE aggregated links. As the number of cores per server increases, the number of VMs per physical server can increase well beyond the 10-20 VMs per server that is typical today. With more VMs per server, I/O requirements increase proportionally. Thankfully, the traditional economics of Ethernet performance improvement[16] is falling into place for 10 Gigabit Ethernet (10 GbE). As a result, Third Generation data center LAN switches will need to support high densities of 10 GbE ports to provide connectivity for high performance virtualized servers, as well as an adequate number of 10 GbE ports and 40 GbE, plus 100 GbE ports when these are available. These high-speed ports will be used for multiple purposes, including connecting the access switches to the core tier.

As noted, second generation LAN switches had fairly low port density. In contrast:

The current generation of switches has exploited advances in switch fabric technology and merchant silicon switch-on-a-chip integrated circuits (ICs) to dramatically increase port densities.

Modular data center switches are currently available with up to 768 non-blocking 10 GbE ports or 192 40 GbE ports. The typical maximum port density for TOR switches which are generally based on merchant silicon, is 64 10 GbE ports. Today, high-speed uplinks are often comprised of multiple 10 GbE links that leverage Link Aggregation (LAG)[17]. However, a 40 GbE uplink typically offers superior performance compared to a 4 link 10 GbE LAG. This is because the hashing algorithms that load balance traffic across the LAG links can easily yield sub-optimal load distribution whereby a majority of traffic is concentrated in a small number of flows. Most high performance modular switches already have a switch fabric that provide 100 Gbps of bandwidth to each line card, which means that as 40 GbE and 100 GbE line cards become available, these can be installed on existing modular switches, preserving the investment in these devices. Most vendors of modular switches are currently shipping 40 GbE line cards, while 100 GbE line cards will not be widely deployed until 2012 or 2013.

In the case of stackable Top of Rack (ToR) switches, adding 40 or 100 GbE uplinks often requires new switch silicon, which means that the previous generation of ToR switches will probably need to be swapped out in order to support 40 GbE and, at some future date, 100 GbE uplink speeds.

[16] Ethernet typically provides a 10x higher performance for a 3-4x increase in cost. This is an example of how Moore's Law impacts the LAN.
[17] www.ieee802.org/3/hssg/public/apr07/frazier_01_0407.pdf

High Availability

As previously noted, IT organizations will be implementing a growing number of VMs on high performance multi-core servers.

The combination of server consolidation and virtualization creates an "all in one basket" phenomenon that drives the need for highly available server configurations and highly available data center LANs.

One approach to increasing the availability of a data center LAN is to use a combination of redundant subsystems within network devices such as LAN switches in conjunction with redundant network designs. A high availability modular switch can provide redundancy in the switching fabric modules, the route processor modules, as well as the cooling fans and power supplies. In contrast, ToR switches are generally limited to redundant power supplies and fans. Extensive hardware redundancy is complemented by a variety of switch software features, such as non-stop forwarding, that ensure minimal disruption of traffic flow during failovers among redundant elements or during software upgrades. Modular switch operating systems also improve availability by preventing faults in one software module from affecting the operation of other modules. Multi-chassis Link Aggregation Group is described below. Implementing this technology also tends to increase availability because it enables IT organizations to dual home servers to separate physical switches.

Alternatives to the Spanning Tree Protocol

The bandwidth efficiency of Layer 2 networks with redundant links can be greatly improved by assuring that the parallel links from the servers to the access layer and from the access layer to the core layer are always in an active-active forwarding state. This can be accomplished by eliminating loops in the logical topology without resorting to the Spanning Tree Protocol (STP). In the current state of evolution toward a Third Generation data center LAN, loops can be eliminated using switch virtualization and multi-chassis LAG (MC LAG) technologies, which are described below. Implementing one of the two emerging shortest path first bridging protocols, TRILL and SPB, that support equal cost multi-path bridging can also eliminate loops. TRILL and SPB are also described below.

Switch Virtualization and Multi-Chassis Link Aggregation Group

With switch virtualization, two or more physical switches are made to appear to other network elements as a single logical switch or virtual switch, with a single control plane.

In order for multiple physical switches to form a virtual switch, they need a virtual switch link (VSL) or interconnect (VSI) that supports a common control plane and data flows between the members of the virtual switch. In redundant configurations, connections between end systems and virtual access switches and between virtual access switches and virtual aggregation switches are based on multi-chassis (MC) link aggregation group (LAG) technology[18], as shown in Figure 7. MC LAG allows the links of the LAG to span the multiple physical switches that comprise a virtual switch. The re-convergence time associated with MC LAG is typically under 50 ms., which means that real time applications such as voice are not impacted by the re-convergence of the LAN. From the server perspective, links to each of the physical members of a virtual access switch appear as a conventional LAG or teamed links, which means that switches can be virtualized without requiring any changes in the server domain.

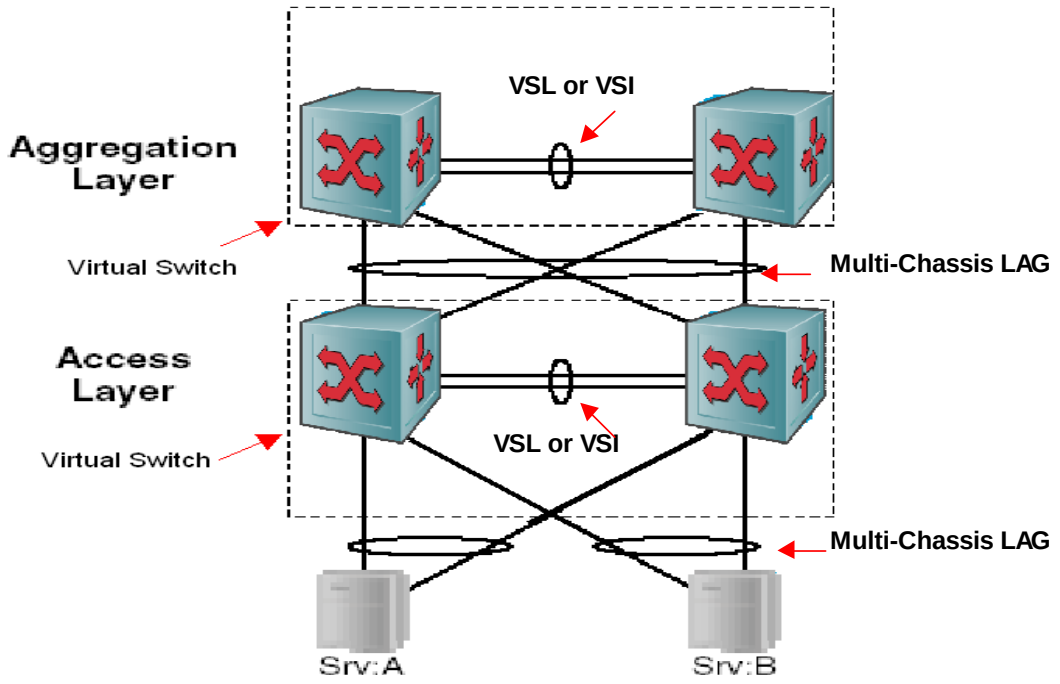
[18] http://en.wikipedia.org/wiki/Link_aggregation

The combination of switch virtualization and multi-chassis LAG can be used to create a logically loop-free topology

This means that data center LANs can be built without using the spanning tree protocol (STP) and first hop router redundancy protocols (e.g., VRRP). This is important because these protocols prevent all available forwarding resources in a redundant network design from being simultaneously utilized.

In Figure 7, loops are eliminated because from a logical perspective, there are only two switches with a single LAG from the server to the access switch and a single LAG from the access switch to the aggregation switch. The traffic load to and from each server is load balanced across the two links participating in the multi-chassis LAG connecting each server to the virtual access switch. Therefore, both server connections are actively carrying traffic in both directions rather than being in an active state for some VLANs and in an inactive state for others. In the same fashion, traffic between the access virtual switch and the aggregation virtual switch is load balanced across all four physical links connecting these devices. Both physical switches participating in the aggregation layer virtual switch are actively forwarding traffic to the network core that is not shown in Figure 7. The traffic is load balanced via the LAG hashing algorithms rather than being based on VLAN membership, as is the case with more traditional redundant LAN designs. The virtual switch not only improves resource utilization but also enhances availability because the relatively long convergence times of STP topology calculations are circumvented. Virtual switch technology also simplifies management because multiple physical switches can be managed as a single entity.

Figure 7: Switch Virtualization and Multi-Chassis LAG



Most vendors of data center switches support switch virtualization and MC LAG in their ToR and modular switches, and these technologies are fully utilized in the two-tier LAN designs that they

are currently recommending to enterprise customers. As a result, most two tier LAN designs being proposed by vendors will not be based on STP for loop control. There are some differences among vendors in the VSL/VSI technology and in the LAG hashing algorithms. For example, some vendors of stackable ToR switches take advantage of the stacking interconnect as the VSL/VSI link, while other vendors will use 10 GbE or 40 GbE ports when available for VSL/VSI. Most LAG implementations conform to the IEEE 802.3ad standard. However, LAG hashing algorithms are outside the 802.3ad standard and more sophisticated hashing algorithms can provide for some differentiation between LAN switches by improving load balancing across the LAG links. In addition, there are some differences in the number of ports or links that can participate in a LAG. Some vendors support up to 32 links per LAG, while 8 links per LAG is the most common implementation.

SPB and TRILL

It must be noted that two-tier LANs and switch virtualization are far from the final word in the design of data center networks. Standards bodies have been working on technologies that will allow active-active traffic flows and load balancing of Layer 2 traffic in networks of arbitrary switch topologies. TRILL (Transparent Interconnection of Lots of Links) is an Internet Engineering Task Force (IETF) project to develop a Layer 2 shortest-path first (SPF) routing protocol for Ethernet. The TRILL RFC (RFC 6325) is currently on the standards track and is being used as the basis for some pre-standard implementations. A similar competing effort is being pursued by the IEEE 802.1aq working group which is defining a standard for shortest path bridging (SPB) of unicast and multicast frames and which supports multiple active topologies. The SPB standard is expected to be ratified by the IEEE by early 2012.

With either TRILL or 802.1aq SPB, it would be possible to achieve load-balanced, active-active link redundancy without having to resort entirely to switch virtualization, MC LAG, and VSL/VSI interconnects. For example, dual homing of servers can be based on MC LAG to a virtual access switch comprised of two physical access switches, while the rest of the data center LAN is based on TRILL or SPB.

There is currently considerable debate in the industry about which is the best technology – TRILL or SPB. While that is an important debate:

> ***In many cases, the best technology doesn't end up being the dominant technology in the marketplace.***

TRILL and SPB have some points of similarity but they also have some significant differences that preclude interoperability. Both approaches use IS-to-IS as the Layer 2 routing protocol and both support equal cost multi-path bridging, which eliminates the blocked links that are a characteristic of STP. Both approaches also support edge compatibility with STP LANs. Some of the major differences include:

- TRILL involves a new header for encapsulation of Ethernet packets, while SPB uses MAC-in-MAC Ethernet encapsulation. Therefore, TRILL requires new data plane hardware, while SPB doesn't for Ethernet switches that support 802.1ah (MAC-in-MAC), 802.1ad (Q-in-Q) and 802.1ag (OAM).

- SPB's use of MAC-in-MAC Ethernet encapsulation eliminates the potential for a significant increase in the size of MAC address tables that are required in network switches.

- SPB forwards unicast and multicast/broadcast packets symmetrically over the same shortest path, while TRILL may not forward multicast/broadcast packets over the shortest path.

- SPB eliminates loops using Reverse Path Forwarding (RPF) checking for both unicast and multicast traffic, while TRILL uses Time to Live (TTL) for unicast and RPF for multicast.

- TRILL can support multi-pathing for an arbitrary number of links, while SPB is currently limited to 16 links.

- With TRILL, network virtualization is limited to 4K VLANs, while SPB supports a 16 million service instances via Q-in-Q.

- SPB is compatible with IEEE 802.1ag and ITU Y.1731 OAM which means that existing management tools will work for SPB, while TRILL has yet to address OAM capability.

- SPB is compatible with Provider Backbone Bridging (PBB), the protocol used by many service providers to provide MPLS WAN services. This means that SPB traffic can be directly mapped to PBB. Also, virtual data centers defined with SPB can be mapped to separate traffic streams in PBB and given different QoS and security treatment.

SPF bridging should have major implications for data center LAN designs and most of the larger switch vendors are well along in developing switches that can support either TRILL or SPB and network designs based on these technologies. A number of vendors are already shipping pre-standard versions of these protocols, in some cases with proprietary enhancements. It may well turn out that two-tier networks based on switch virtualization and MC LAG are just a mid-way point in the evolution of the Third Generation LAN.

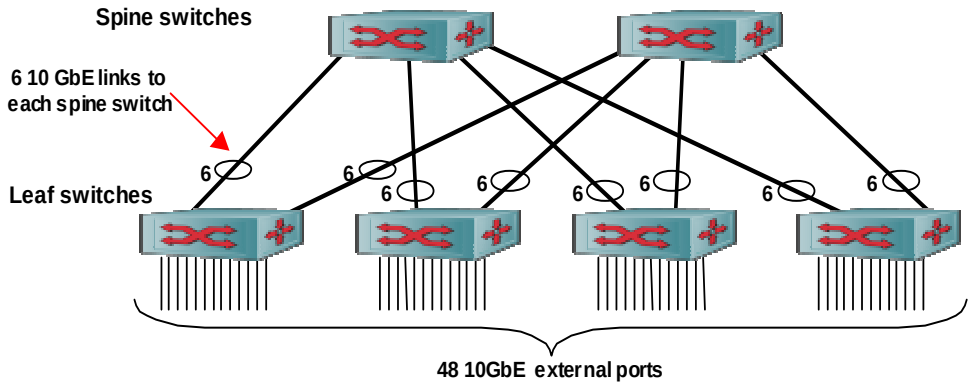
With technologies like TRILL and SPB, the difference between access switches and core switches may shrink significantly.

As a result of TRILL or SPB, the switch topology may shift from a two-tier hub and spoke, such as the one in Figure 7, to a highly meshed or even fully meshed array of switches that appears to the attached devices as a single switch. SPF bridging can support a variety of other topologies, including the fat tree switch topologies[19] that are popular in cluster computing approaches to HPC. Fat trees are also used by Ethernet switch vendors to build high density, non-blocking 10 GbE switches using merchant silicon switch chips. This trend may eventually lead to the commoditization of the data plane aspect of Ethernet switch design. Figure 8 shows how a 48 port 10 GbE TOR switch can be constructed using six 24-port 10 GbE switch chips. By increasing the number of leaf and spine switches, larger switches can be constructed[20]. A number of high density 10 GbE switches currently on the market use this design approach.

[19] http://www.mellanox.com/pdf/whitepapers/IB_vs_Ethernet_Clustering_WP_100.pdf

[20] The maximum density switch that can be built with a two-tier fat tree architecture based on 24 port switch chips has 288 ports.

Figure 8: TOR Switch Fat Tree Internal Architecture



The Interop Respondents were asked, "Two years from now, which of the following is likely to be the most commonly used L2 Ethernet protocol in your company's data center LANs?" The answers of one hundred and eighty one respondents are summarized in Table 10.

Table 10: Most Common L2 Ethernet Protocol

L2 Ethernet Protocol	Percentage of Respondents
Spanning Tree Protocol (STP)	43%
A Vendor's Proprietary Protocol	18%
Multi-Switch Link Aggregation (M-LAG)	16%
Transparent Interconnect of Lots of Links (TRILL)	12%
Shortest Path Bridging (SPB)	10%
Other	1%

The data in Table 10 indicates that just under a half of IT organizations expect to still be running STP in their data center LANs in two years. For those IT professionals who indicated that STP would not be the most commonly used L2 Ethernet protocol in two years, there was no consensus as to what protocol would be the most common. However, similar to the situation with flattening the data center LAN, what is even more interesting is that two hundred and seventy four members of the pool of survey respondents answered the question with "don't know". That means that the number of survey respondents that don't know which L2 Ethernet protocol will be the most commonly used in their data center LANs in two years is notably greater than the number that do know.

There is significant desire on the part of IT organizations to move away from using STP in their data center LANs, but there isn't a consensus as to what the most common replacement technology will be.

A discussion of the alternatives to STP amongst six of the primary data center LAN switch vendors can be found at Webtorials[21].

[21] http://www.webtorials.com/content/tls.html

Controlling and Managing Inter-VM Traffic

With server virtualization, each physical server is equipped with a hypervisor-based virtual switching capability that allows connectivity among VMs on the same physical platform. Traffic to external destinations also traverses this software switch. From the network perspective, the hypervisor vSwitch poses a number of potential problems:

1. The vSwitch represents another tier of switching that needs to be configured and managed, possibly requiring an additional management interface. This can partially defeat an effort to flatten the network to two–tiers.

2. The vSwitch adds considerable complexity, because there is an additional vSwitch for every virtualized server.

3. vSwitch control plane functionality is typically quite limited compared to network switches, preventing a consistent level of control over all data center traffic

4. As more VMs per server are deployed, the software switch can place high loads on the CPU, possibly starving VMs for compute cycles and becoming an I/O bottleneck.

5. VM-VM traffic on the same physical server is isolated from the rest of the network, making these flows difficult to monitor and control in the same fashion as external flows.

6. The vSwitch functionality and management capabilities will vary by hypervisor vendor and IT organizations are increasingly deploying hypervisors from multiple vendors.

The vSwitch presents a number of concerns related to management, security, functionality and organizational responsibilities.

There are two approaches to the problems posed by the early generation vSwitch: Distributed Virtual Switching (DVS) and Edge Virtual Bridging (EVB). With DVS, the control and data planes of the embedded vSwitch are decoupled. This allows the data planes of multiple vSwitches to be controlled by an external centralized management system that implements the control plane functionality. Decoupling the data plane from the control plane makes it easier to tightly integrate the vSwitch control plane with the control planes of physical access and/or aggregation switches and/or the virtual server management system. Therefore, DVS can simplify the task of managing a large number of vSwitches, and improve control plane consistency, but it doesn't address the other issues listed above.

With EVB, all the traffic from VMs is sent to the network access switch. If the traffic is destined for a VM on the same physical server, the access switch returns the packets to the server over the same port on which it was received. The shipping of traffic from a VM inside of a physical server to an external access switch and then back to a VM inside the same physical server is often referred to as a hair pin turn. With Edge Virtual Bridging, the hypervisor is relieved from all switching functions, which are now performed by the physical access network. With EVB, the vSwitch now performs the simpler function of aggregating hypervisor virtual NICs to a physical NIC. Basic EVB can be supported by most existing access switches via a relatively simple firmware upgrade. The IEEE 802.1Qbg Working Group is creating an EVB standard based on a technology known as Virtual Ethernet Port Aggregator (VEPA) that deals with hair-pin turns and a definition of a multi-channel service for remote ports to access local VMs. A companion effort, the IEEE's 802.1Qbh Port

Extension is defining a technique for a single physical port to support a number of logical ports and a tagged approach to deal with frame replication issues in the EVB. EVB/VEPA standards supported in switches and hypervisors will address all of the six potential problems listed above.

Essentially all vendors of data center switches support the IEEE's EVB standards efforts. Some vendors are waiting until the standard is finalized and are supporting hypervisor vSwitches in the interim. Other vendors have pre-standard implementations of basic EVB/VEPA already available or under development.

Software Defined Networks and Network Virtualization

With DVS, the switch control plane is decoupled from the data plane and placed in a separate server or controller. This concept can also be applied to the entire data center or campus LAN by removing the control plane from every physical and virtual switch and centralizing it in a control plane server. This centralization would make it relatively easy to programmatically control the entire network. Programmatic control is a key aspect of the concept of a Software Defined Network (SDN) that uses an abstraction layer or *network hypervsior* between the network operating system (NOS) control software and the packet forwarding data plane hardware.

OpenFlow[22] is an open API/protocol that is used between a network controller and a controlled physical switch that provides the forwarding hardware. The protocol is used to set flow table entries within the physical switch. The abstraction layer allows OpenFlow-enabled switches from different vendors to be mixed and matched without impacting the NOS. The Open Networking Foundation (ONF)[23] established in 2012 is now responsible for maintaining the OpenFlow specification, which is currently at Version 1.1.

Building an SDN with OpenFlow requires two components:

- A NOS supporting OpenFlow that is also capable of presenting a logical map of the entire network to the network administrators. This NOS could a modification of an existing proprietary NOS or possibly an open source NOS. The NOS should also be extensible by providing a northbound API to allow new functions to be added.

- Packet forwarding hardware that also supports OpenFlow. In principle, the SDN could be based on a physical network built with OpenFlow switches from a number of different vendors.

Some the potential benefits of an SDN with OpenFlow include:

- Network virtualization where multiple independent virtual networks can share a common physical infrastructure. Virtual networks are based on segmenting flows. Within OpenFlow, flows are defined using a ten-tuple of header fields including Ethernet SA/DA, IP SA/DA, TCP/UDP ports, and VLAN ID. This also provides enhanced security via firewall-style granular control of traffic flows within virtual networks. Network virtualization beyond VLANs is of particular interest in public cloud data centers that provide services to multiple tenants. This concept is elaborated upon in the subsection below that discusses the network support that is required to support the dynamic creation and migration of VMs.

- Network operations are streamlined via the global nature of the network-wide NOS, which results in lower OPEX. In the public cloud, OpenFlow allows the network to be programmatically controlled in conjunction with server and storage resources in order to provision and modify services to tenants. An OpenFlow-enabled NOS with an open API to server virtualization and cloud management systems can potentially be exploited to achieve higher levels of management integration across the data center or the cloud.

[22] http://www.openflow.org/
[23] http://www.opennetworking.org

- SDNs are well suited for highly meshed data center switching fabrics based on the fat tree topologies common in HPC and in web data centers that are dealing with the challenges that are associated with Big Data. Because the control plane has a global view of the network topology, loops can be avoided without resorting to bridging protocols such as STP, TRILL, or SPB.

- Where both the NOS and the packet forwarding hardware support open APIs, network designers would be free to independently optimize each level of the network; e.g., NOS, switches, and applications that extend the functionality of the network. OpenFlow proponents believe this would make the networking industry more innovative and competitive, lowering the overall CAPEX and OPEX cost of network infrastructure.

At Interop 2011, twelve vendors demonstrated prototype switches supporting OpenFlow. Three of these vendors are already shipping switches that support OpenFlow. There are also a number of vendors working on open source OpenFlow-enabled NOS packages and applications that extend NOS functionality. The Open Networking Summit in October 2011 had several demonstrations of OpenFlow implementations on physical hardware combined with network controllers from various vendors. A discussion of OpenFlow amongst six of the primary data center LAN switch vendors can be found at Webtorials[24].

Network Convergence and Fabric Unification

In contrast to Second Generation Data Center LANs:

> ***A possible characteristic of Third Generation Data Center LANs will be the convergence of block-level storage and data traffic over a common high-speed Ethernet data center switching fabric.***

This unified fabric offers significant cost savings in multiple areas including converged network adapters on servers and a reduction in rack space, power and cooling capacity, cabling, and network management overhead.

Traditional Ethernet, however, only provides a best effort service that allows buffers to overflow during periods of congestion and which relies on upper level protocols such as TCP to manage congestion and to recover lost packets through re-transmissions. In order to emulate the lossless behavior of a Fibre Channel (FC) SAN, Ethernet needs enhanced flow control mechanisms that eliminate buffer overflows for high priority traffic flows, such as storage access flows. Lossless Ethernet is based on the following standards, which are commonly referred to as IEEE Data Center bridging (DCB):

- **IEEE 802.1Qbb Priority-based Flow Control** (**PFC**) allows the creation of eight distinct virtual link types on a physical link, with each virtual link mapped to an 802.1p traffic class. Each virtual link can be allocated a minimum percentage of the physical link's bandwidth. Flow is controlled on each virtual link via the pause mechanism which can be applied on a per priority basis to prevent buffer overflow, eliminating packet loss due to congestion at the link level. In particular, block-level or file-level storage traffic on one of the virtual lanes can be protected from loss by pausing traffic on one or more of the remaining lanes.

[24] http://www.webtorials.com/content/tls.html

- **IEEE 802.1Qau Congestion Notification (CN)** is a traffic management technique that eliminates congestion by applying rate limiting or back pressure at the edge of the network in order to protect the upper network layers from buffer overflow. CN is intended to provide lossless operation in end-to-end networks that consist of multiple tiers of cascaded Layer 2 switches, such as those typically found in larger data centers for server interconnect, cluster interconnect and to support extensive SAN fabrics.

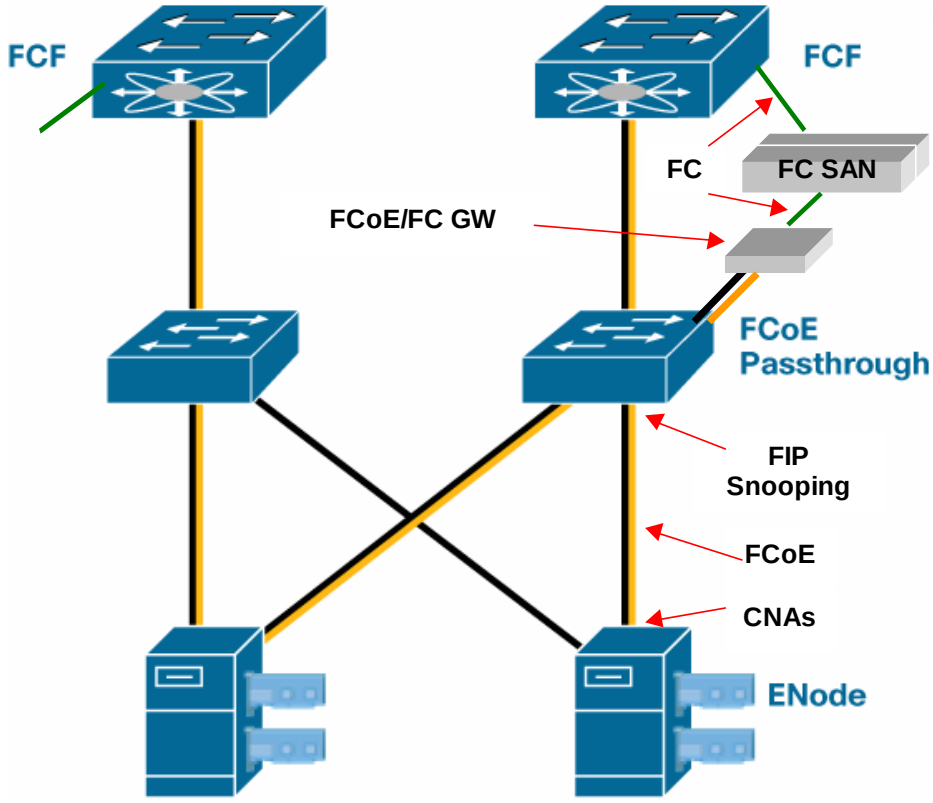
- **IEEE 802.1Qaz Enhanced Transmission Selection (ETS)** specifies advanced algorithms for allocation of bandwidth among traffic classes including the priority classes supported by 802.1Qbb and 802.1Qau. While the queue scheduling algorithm for 802.1p is based on strict priority, ETS will extend this by specifying more flexible drop-free scheduling algorithms. ETS will therefore provide uniform management for the sharing of bandwidth between congestion managed classes and traditional classes on a single bridged network. Priorities using ETS will coexist with priorities using 802.1Qav queuing for time-sensitive streams. **The Data Center Bridging Exchange (DCBX)** protocol is also defined in the 802.1Qaz standard. The DCBX protocol is an extension of the Link Layer Discovery Protocol (LLDP) that allows neighboring network elements to exchange request and acknowledgment messages to ensure consistent DCB configurations. DCBX is also used to negotiate capabilities between the access switch and the adapter and to send configuration values to the adapter.

  DCB Lossless Ethernet will play a key role in supporting Fibre Channel over Ethernet (FCoE) technology that will allow the installed base of Fibre Channel storage devices and SANs to be accessed by Ethernet-attached servers with converged FCoE network adapters over the unified data center switching fabric. DCB will benefit not only block-level storage, but also all other types of loss and delay sensitive traffic. In the storage arena, DCB will improve NAS performance and will make iSCSI SANs based on 10/40/100 GbE a more competitive alternative to Fibre Channel SANs at 2/4/8 Gbps. In order to take full advantage of 10 GbE and higher Ethernet bandwidth, servers accessing iSCSI storage resources may also need intelligent converged NICs that offload iSCSI and TCP/IP processing from the host.

Fibre Channel over Ethernet (FCoE) is an industry standard that is being developed by the International Committee for Information Technology Standards (INCITS) T11 committee.

The FCoE protocol specification maps Fibre Channel upper layer protocols directly over a bridged Ethernet network. FCoE provides an evolutionary approach to the migration of FC SANs to an Ethernet switching fabric while preserving Fibre Channel constructs and providing reliability, latency, security, and traffic management attributes similar to those of native FC. FCoE also preserves investments in FC tools, training, and SAN devices; e.g., FC switches and FC attached storage. Implementing FCoE over a lossless Ethernet fabric requires converged server network adapters (e.g., CNAs with support for both FCoE and IP) and some form of FC Forwarding Function (FCF) to provide attachment to native FC devices (FC SAN switches or FC disk arrays). FCF functionality can be provided by a FCoE switch with both Ethernet and FC ports or by a stand alone gateway device attached to a FCoE passthrough switch, as shown in Figure 9.

Figure 9: FCoE Converged LAN *Source: Cisco Systems*



As shown in Figure 9, End Nodes (servers) don't need to connect directly to a FCF capable switch. Instead the FCoE traffic can pass through one or more intermediate FCoE passthrough switches. The minimal requirements for a simple FCoE passthrough switch is support for lossless Ethernet or DCB. The FCoE Initialization Protocol (FIP) supports handshaking between a FCoE End Node and an FCF in order to establish and maintain a secure virtual FC link between these devices, even if the end-to-end path traverses FCoE passthrough switches. For DCB passthrough switches that support FIP Snooping, the passthrough switches can inspect the FIP frames and apply policies based on frame content. FIP Snooping can be used to enhance FCoE security by preventing FCoE MAC spoofing and allowing auto-configuration of ACLs.

As this discussion illustrates:

There are several levels of support that data center switch vendors can provide for FCoE.

For example:

1. The lowest level of support is FCoE passthrough via lossless Ethernet or DCB alone.

2. The next step up is to add FIP Snooping to FCoE passthrough switches

3. A third level of support is to add standalone FCF bridges/gateways to front end FC SAN switches or disk arrays.

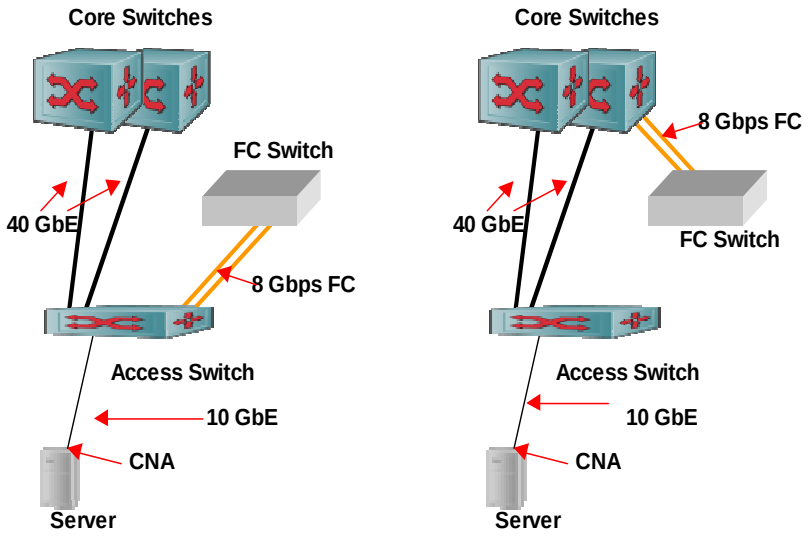
4. The highest level of support is to provide DCB and FIP Snooping for FCoE passthrough switches and also to provide FCoE switches that incorporate FCF ports, creating hybrid switches with both DCB Ethernet and native FC ports.

Most vendors of Ethernet data center switches that don't also have FC SAN switches among their products are planning FCoE support at levels 1, 2, or 3 described above. In fact, most of these Ethernet-only vendors are considerably more enthusiastic about iSCSI SANs over 10/40/100 GbE than they are about FCoE.

The primary drivers of FCoE are the vendors that offer both Ethernet and FC products.

These are the vendors that are already shipping lossless 10 GbE Ethernet switches and hybrid lossless 10 GbE/FCF switches. Even among the vendors providing early support for FCF there are some significant differences, as shown in Figure 10.

Figure 10: FCF Support Options



The left side of the figure shows single hop FCoE with the FCF function integrated into the access switch. It would also be possible to use intervening FCoE/FCF gateways, either standalone or incorporated in the FC switch, which would be connected to the access switch via 10 GbE, making the access switch an FCoE passthrough switch, as shown in the previous figure. The advantage of single hop FCoE is that the storage traffic doesn't compete for bandwidth in the uplinks or the core switches and the core switches aren't required to support DCB or FIP Snooping. The right side of the figure shows multihop FCoE with the FCF function integrated into the core switch, and the access switch in FCoE passthrough mode. Again it would be possible to use FCoE/FCF gateways, either standalone or incorporated in the FC switch, connected to the core switch via 10 GbE. FC SANs and disk arrays connected at the

core offer the advantage of a more centralized pool of storage resources that can be shared across the data center LAN.

The Interop Respondents were asked about their company's current approach to converging the LAN and SAN in their data centers as well as what they thought their company's approach will be two years from now. Fifty nine percent of the Interop Respondents indicated that their company has not currently made any implementation of a converged LAN and SAN. Almost half of the Interop Respondents didn't indicate what they thought their company's approach would be two years from now. Of The Interop Respondents who did indicate what they thought their company's approach would be two years from now, thirty percent indicated that over the next two years that they would either make a significant effort to converge their data center LANs and SANs or they would converge all of their data center LANs and SANs. A discussion of converging the data center LAN and SAN amongst six of the primary data center LAN switch vendors can be found at Webtorials[25].

The ambiguity expressed by The Interop Respondents about their company's direction relative to converging their LANs and SAN, combined with their previously discussed ambiguity about how they will replace the spanning tree protocol and how many layers of switches they will have in their data center LANs leads to the observation that:

The majority of IT organizations have not developed concrete, broad-based plans for the evolution of their data center LANs.

Network Support for the Dynamic Creation and Movement of VMs

When VMs are migrated between servers, the network has to accommodate the constraints imposed by the VM migration utility; e.g., VMotion. Typically the VM needs to be on the same VLAN when migrated from source to destination server. This allows the VM to retain its IP address which helps to preserve user connectivity after the migration. When migrating VMs between disparate data centers, these constraints generally require that the data center LAN be extended across the physical locations or data centers without compromising the availability, resilience and security of the VM in its new location. VM migration also requires the LAN extension service have considerable bandwidth and low latency. VMware's VMotion, for example, requires at least 622 Mbps of bandwidth and less than 5 ms of round trip latency between source and destination servers over the extended LAN[26].

The data storage location, including the boot device used by the virtual machine, 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, the two data sets must be identical. One approach is to extend the SAN to the two sites and maintain a single data source. Another option is to migrate the data space associated with a virtual machine to the secondary storage location. In either case, there is a significant impact on the WAN.

MPLS/VPLS offers one approach to bridging remote data center LANs together. Another alternative is to tunnel Layer 2 traffic through a public or private IP network using Generic Router Encapsulation (GRE). A more general approach that addresses some of the major limitations of live migration of VMs across a data center network is the Virtual eXtensible LAN

[25] http://www.webtorials.com/content/tls.html
[26] www.vce.com/pdf/solutions/vce-application-mobility-whitepaper.pdf

(VXLAN)[27]. VXLAN is the subject of a recently submitted IETF draft supported by VMware, Cisco, Arista Networks, Broadcom, Red Hat and Citrix. In addition to allowing VMs to migrate transparently across Layer 3 boundaries, VXLAN provides support for virtual networking at Layer 3, circumventing the 802.1Q limitation of 4,094 VLANs, which is proving to be inadequate for VM-intensive enterprise data centers and for multi-tenant cloud data centers. VXLAN also addresses the requirement for multi-tenancy where multiple tenants within a cloud data center could have overlapping MAC and IP addresses.

VXLAN creates a Layer 2 overlay on a Layer 3 network via encapsulation. The VXLAN segment is a Layer 3 construct that replaces the VLAN as the mechanism that segments the network for VMs. Therefore, a VM can only communicate or migrate within a VXLAN segment. The VXLAN segment has a 24 bit VXLAN Network identifier, which supports up to 16 million VXLAN segments within an administrative domain. VXLAN is transparent to the VM, which still communicates using MAC addresses. The VXLAN encapsulation is performed through a function known as the VXLAN Tunnel End Point (VTEP), which is typically present in a hypervisor or a physical switch. The encapsulation allows Layer 2 communications with any end points that are within the same VXLAN segment even if these end points are in a different IP subnet. This allows live migrations to transcend Layer 3 boundaries. Since MAC frames are encapsulated within IP packets, there is no need for the individual Layer 2 switches to learn MAC addresses. This alleviates MAC table hardware capacity issues on these switches. Overlapping IP and MAC addresses are handled by the VXLAN ID, which acts as a qualifier/identifier for the specific VXLAN segment within which those addresses are valid.

The VXLAN draft was submitted to the IETF in August 2011, so ratification of a standard is not imminent. However, VMware and Cisco are likely to include pre-standard implementations in their hypervisor switches in the relatively near future. The IETF draft also discusses VXLAN gateways that connect VXLAN environments to the current VLAN based environments. These gateways are likely to be implemented in hardware switches within the data center.

As noted earlier, the requirement to support the dynamic creation and movement of VMs is one of the primary factors driving IT organizations to redesign their data center LANs. As was also noted earlier, the requirements for VM migration with VLAN boundaries has provided a major impetus for flattening the LAN with two-tier designs featuring Layer 2 connectivity end-to-end.

Many of the benefits of cloud computing depend on the ability to dynamically provision VMs and to migrate them at will among physical servers located in the same data center or in geographically separated data centers. The task of creating or moving a VM is a relatively simple function of the virtual server's management system. There can, however, be significant challenges in assuring that the VM's network configuration state, including VLAN memberships, QoS settings, and ACLs, is established or transferred in a timely fashion. In many instances today, these network configuration or reconfigurations involves the time-consuming manual process involving multiple devices.

Regulatory compliance requirements can further complicate this task. For example, assume that the VM to be transferred is supporting an application that is subject to PCI compliance. Further assume that because the application is subject to PCI compliance that the IT organization has implemented logging and auditing functionality. In addition to the VM's network configuration state, this logging and auditing capability also has to be transferred to the new physical server.

[27] http://searchservervirtualization.techtarget.com/news/2240074318/VMware-Cisco-propose-VXLAN-for-VM-mobility

The most common approach to automating the manual processes involved in VM provisioning and migration is based on communication between the Hypervisor Management system and the switch element management system (EMS) via APIs supported by both vendors[28].

When a Virtual Machine is created or when the movement of a VM is initiated, the Hypervisor manager signals to the EMS that the event is about to occur and provides a partial VM network profile including a virtual MAC, VLAN memberships and the target hypervisor. Based on existing policies, the EMS extends the VM network profile to include appropriate QoS and security parameters such as ACLs. The EMS can then determine the target hypervisor's access switch and can configure or reconfigure it accordingly. Where VLANs need to be created, the EMS can also create these on the uplinks and neighboring switches as appropriate. In a similar manner, when a VM is deleted from a hypervisor, the EMS can remove the profile and then prune the VLAN as required. All of these processes can be triggered from the hypervisor.

An interesting benefit of the VXLAN overlay over Layer 3 networks is that IT organizations no longer need to plumb the VLAN for a VM on the link connecting to the ToR switch when the VM migrates. This requirement has been addressed to date through static configuration at the destination hypervisor/ToR switch. As indicated above, dynamic configuration is available via protocols being defined for Edge Virtual Bridging within the IEEE 802.1Qbg working group. With VXLAN, all VM traffic from the hypervisor to the ToR switch is encapsulated within an IP packet so there is no need to plumb the VM's VLAN information on the link between the hypervisor and the ToR switch.

Most data center switch vendors have already implemented some form of VM network profile software, including linking their switches to at least one brand of hypervisor. Some differences exist between the range of hypervisors supported and the APIs that are used. Distribution of VM network profiles is only one of many management processes that can benefit greatly from automation, so it would benefit IT departments to develop expertise in open APIs and powerful scripting languages that can be exploited to streamline time-consuming manual processes and thereby reduce operational expense while improving the ability of the data center to dynamically reallocate its resources in response to changes in user demand for services.

A somewhat different approach to automating data center configuration, including the provisioning and migration of VMs is based on orchestration engines, which are discussed in more detail in the management section of this report. Service orchestration is a centralized server function that can automate many of the manual tasks involved in provisioning and controlling the capacity of dynamic virtualized services across myriad technology domains; e.g., networking, servers, storage and security. In the case of VM provisioning and migration, the orchestration engine would function as the point of integration between the network device EMS and the hypervisor's management system. This capability requires that third generation data center LANs provide APIs that enable integration with third party orchestrations solutions. Orchestration solutions are available from a number of network management vendors and hypervisor vendors.

[28] While this approach is the most common, some vendors have alternative approaches.

Summary of Third Generation Data Center LAN Technologies

The data center LAN is on the cusp of a number of quite dramatic technology developments, as summarized in Table 11. As shown in the table, most the items on this list are still in flux and require additional development, and/or additional work from the standards bodies[29].

Table 11: Status of Data Center Technology Evolution

Technology Development	Status
Two-tier networks with Layer 2 connectivity extending VLANs across the data center.	On-going deployment
Reduced role for blade switches to eliminate switch tier proliferation.	On-going
Changing role for the hypervisor vSwitch as a port aggregator (VEPA) for EVB, essentially eliminating the vSwitch tier.	A standard is in progress and pre-standard implementations are available.
STP issues are being addressed by switch virtualization and multi-chassis LAG technology, as well as by newer protocols such as TRILL/SPB.	On-going deployment
Multi-core servers with notably more VMs per server and 10 GbE connectivity to the LAN.	Early adoption stage.
40 GbE and 100 GbE uplinks and core switches.	A standard is in place: 40 GbE is available 100 GbE due in 2012
DCB delivering lossless Ethernet for 10 GbE and higher speed Ethernet	Standards are in place. Implementations are being announced.
SDN and OpenFlow	Specifications have bees released There is some prototype switch support and some NOS support from startups.
VXLAN extended virtual networks address VLAN scalability, multi-tenancy and switch hardware Layer 2 table capacity issues that are caused by the proliferation of virtualization.	A draft was recently submitted to the IETF. Pre-standard implementations are expected.
FCoE approach to fabric unification	FCoE standard is in place. Early implementations are based on pre-standard DCB.
10 GbE iSCSI approach to fabric unification	Early implementations were over pre-standard DCB.
TRILL/SPB enabling new data center LAN topologies; e.g., fully meshed, fat tree with equal cost multi-path forwarding	Standards are in progress. Pre-standard implementations of both SPB and TRILL are available. SPB is expected to be finalized by early 2012
Management tools that integrate, coordinate, and automate provisioning and configuration of server, storage and network resource pools	These are proprietary and have varying levels of maturity.

[29] Exceptions to this statement are entries number 1, 2, 4 and to some extent 11.

The Wide Area Network (WAN)

Introduction

Background

The modern WAN got its start in 1969 with the deployment of ARPANET which was the precursor to today's Internet. The technology used to build the Internet began to be commercialized in the early 1970s with the development of X.25 based packet switched networks.

In addition to the continued evolution of the Internet, the twenty-year period that began in 1985 saw the deployment of four distinct generations of enterprise WAN technologies[30]. For example, in the mid to late 1980s, it became common for enterprise IT organizations to deploy integrated TDM-based WANs to carry both voice and data traffic. In the early 1990s, IT organizations began to deploy Frame Relay-based WANs. In the mid to late 1990s, some IT organizations replaced their Frame Relay-based WANs with WANs based on ATM (Asynchronous Transfer Mode) technology. In the 2000s, many IT organizations replaced their Frame Relay or ATM-based WANs with WANs based on MPLS. Cost savings was the primary factor that drove the adoption of each of the four generations of WAN technologies. The cost savings, however, were very modest when compared to the price performance improvements that are associated with local area networking.

However, in contrast to the volatility of this twenty-five year period:

Today there is not a fundamentally new generation of WAN technology in development.

Relative to the deployment of new WAN services, what sometimes happens in the current environment is that variations are made to existing WAN technologies and services. An example of that phenomenon is Virtual Private LAN Service (VPLS)[31]. As described later in this section of the report, within VPLS an Ethernet frame is encapsulated inside of MPLS. While creating variations on existing services can result in significant benefits, it does not produce fundamentally new WAN services.

Contrasting the LAN and the WAN

As noted, the WAN is notably different than the data center LAN. These differences include the fact that:

After a lengthy period in which there was little or no fundamental innovation, the LAN is experiencing broad fundamental change. In contrast, after a lengthy period in which the WAN underwent repeated fundamental change, there are currently no fundamental changes in store for the WAN.

[30] An enterprise WAN is designed to provide for connectivity primarily within the enterprise and between the enterprise and key contacts such as partners. This is in contrast to the Internet that is designed to provide universal connectivity.
[31] http://vlt.me/vpls-0810

In the vast majority of instances, the latency, jitter and packet loss that the LAN exhibits doesn't have an appreciable impact on application performance. In many instances, the latency, jitter and packet loss that the WAN exhibits has an appreciable impact on application performance. This is particularly true of 3G/4G networks.

One of the primary design criteria for designing a data center LAN is scalability. A manifestation of the ongoing improvements in LAN scalability is that over the last fifteen years the speed of a data center LAN has increased from 10 Mbps to 10 Gbps – which is a factor of a thousand. In contrast, in many cases the primary design criterion for designing a WAN is to minimize cost. For example, in many parts of the world it is possible to get high-speed WAN links such as an OC-192 link. These links, however, are usually not affordable.

- The LAN follows Moore's Law. In contrast, the price/performance of WAN services such as MPLS tends to improve by only a couple of percentage points per year.

The WAN doesn't follow Moore's Law.

WAN Budgets

Both in 2010 and again in 2011, The Webtorials Respondents were asked how their budget for the forthcoming year for all WAN services compares to what it is in the current year year. Their responses are contained in Table 12.

Table 12: WAN Budget Increases

	Responses in 2010	Responses in 2011
REDUCED BY MORE THAN 10%	5.9%	3.2%
Reduced by 1% to 10%	17.0%	11.1%
Basically static	36.3%	34.9%
Increased by 1% to 10%	30.4%	32.8%
Increased by more than 10%	10.4%	18.0%

The change in the budget for WAN services in 2011 bears a lot of similarities to what the change in the budget for WAN services was in 2010. The biggest differences are at the extremes. For example, in 2011 the percentage of The Webtorials Respondents that expect that their WAN budgets will decrease has been cut almost in half when compared to what it was in 2010. In addition, in 2011 the percentage of The Webtorials Respondents that expect that their WAN budgets will increase by more than 10% is almost double what it was in 2010.

Over the next year, roughly forty percent of IT organizations will increase their WAN budget and in many cases, the increase will be significant.

As is explained in the next subsection, the adoption of cloud computing will increase the rate of growth in the amount of traffic that transits the WAN. As such,

IT organizations must either make changes to how they use WAN services, or else accept ongoing increases in their WAN budget due to the increased traffic generated by the use of cloud computing.

Drivers of Change

As explained in the section of this report entitled *The Emergence of Cloud Computing and Cloud Networking*, one of the characteristics of cloud computing is increased reliance on the network. The increased reliance on the WAN in particular stems from the fact that the resources that support cloud computing solutions are centralized in a small number of data centers and the vast majority of users access these solutions over the WAN. Hence, the more use that organizations make of cloud computing, the more traffic transits the WAN.

Below are some of the specific factors that are putting more traffic onto the WAN and hence, driving the need for IT organizations to change their approach to wide area networking.

Virtual Machine Migration

The section of this report entitled *The Emerging Data Center LAN* quantified the great interest that IT organizations have in server virtualization in general and in moving virtual machines (VMs) between data centers in particular. That section of the report also discussed the fact that one of the requirements associated with moving VMs between data centers is that the data storage location, including the boot device used by the VM being migrated, 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, the two data sets must be identical. One approach to enabling data access is to extend the SAN to the two sites and to maintain a single data source. Another option is to migrate the data along with the VM to the secondary site. In either case, it is necessary to coordinate VM and storage migrations and to be able to move large data sets efficiently between data centers, which will have a significant impact on the WAN.

Virtual Desktops

Another form of virtualization that will drive a further increase in WAN traffic is desktop virtualization. In order to quantify the interest that IT organizations have in desktop virtualization, The Webtorials 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 13.

Table 13: 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%

The data in Table 13 indicates the growing interest that IT organizations have in desktop virtualization. For example,

> ***Over the next year, the percentage of IT organizations that have not implemented any desktop virtualization will be cut roughly in half.***

Part of the challenge in supporting virtualized desktops is that the implementation of virtualized desktops puts more traffic on the WAN, which typically leads to the need for more bandwidth. In addition to the bandwidth challenges, as explained in The 2011 Application and Service Delivery Handbook[32], there are performance challenges associated with each of the two primary form of desktop virtualization; e.g., client side (a.k.a., streamed desktops) and server side (a.k.a., hosted desktops).

In the case of client side desktop virtualization, 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[33]. Server side protocols such as ICA and RDP, tend to work relatively well when supporting traditional applications[34]. However, these protocols can behave badly when supporting graphics and video. Some newer protocols, such as Teradici's PC-over-IP (PCoIP)[35], consume considerable WAN bandwidth and are latency sensitive.

Collaboration

As was described in the section of this report that is entitled *The Emergence of Cloud Computing and Cloud Networking*, many organizations are beginning to acquire services such as collaboration from a cloud computing service provider (CCSP). Independent of whether the collaboration service is provided by a CCSP or by the IT organization, it stresses the WAN. This stress comes in part from the fact that the performance of applications such as video and telepresence is very sensitive to delay, jitter and packet loss. The stress also comes in part because video and telepresence consume considerable WAN bandwidth. It is common, for example, to allocate several megabits per second of WAN bandwidth to a single telepresence session.

The current conventional wisdom in the IT industry is that organizations are increasing their use of video. In order to evaluate that assertion, The Webtorials Respondents were asked to indicate how much change in the use of all forms of video they anticipated that their organization would make over the next year. Their responses are shown in Table 14.

Table 14: Anticipated Change in the Use of Video

Down by More than 25%	Down by 1% to 25%	No Change	Up 1% to 25%	Up 26% to 50%	Up 51% to 75%	Up 76% to 100%	Up more than 100%
0.0%	0.5%	19.9%	34.6%	24.6%	9.9%	4.7%	5.8%

[32] http://www.webtorials.com/content/2011/07/2011-application-service-delivery-handbook.html
[33] A chatty protocol requires hundreds, if not thousands of round trips to complete a transaction.
[34] Even though they work relatively well in native mode, many IT organizations choose to implement WAN optimization in order to improve the performance of these protocols.
[35] http://en.wikipedia.org/wiki/PCoIP

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

> ***Over the next year almost 80% of IT organizations will increase their use of video, and in many cases the increased use of video will be substantial.***

Mobile Workers

In the last few years there has been an explosive growth in the number of mobile workers. There are a number of key concerns relative to supporting mobile workers. One such concern is the number and types of devices that mobile workers use. As recently as a couple of years ago, many IT organizations tried to control the types of devices that their users could utilize. In the current environment the majority of IT organizations are in a position where they have to support a large and growing set of mobile devices from a range of vendors. It most cases mobile workers have two mobile devices (i.e., a laptop and a smartphone) and in a growing number of cases, mobile workers have three mobile devices; i.e., a laptop, a smartphone and a tablet.

Another key concern relative to supporting mobile workers is how the applications that these workers access has changed. At one time, mobile workers tended to primarily access either recreational applications or applications that are not delay sensitive; e.g., email. However, in the current environment mobile workers also need to access a wide range of business critical applications, many of which are delay sensitive. This shift in the applications accessed by mobile workers was highlighted by SAP's recent announcement[36] that it will leverage its Sybase acquisition to offer access to its business applications to mobile workers.

One of the technical issues associated with supporting mobile workers' access to delay sensitive, business critical applications is that because of the way that TCP functions, even the small amount of packet loss that is often associated with wireless networks results in a dramatic reduction in throughput. A related issue is that typically there is a large amount of delay associated with 3G and 4G networks.

Traditional WAN Services

Background

The Webtorials Respondents were given a set of eleven WAN services and asked to indicate the extent to which they currently utilize each WAN service. The survey question included Frame Relay and ATM among the set of WAN services. In the not too distant past, these services were widely deployed. However, over half of The Webtorials Respondents don't have any Frame Relay in their networks and almost two thirds of The Webtorials Respondents don't have any ATM in their networks. In addition, few IT organizations are increasing their use of these technologies[37], while many IT organizations are decreasing their use of these technologies[38].

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

[37] Roughly 2% of IT organizations are increasing their use of Frame Relay and 6% of IT organizations are increasing their use of ATM.

[38] Roughly 34% of IT organizations are decreasing their use of Frame Relay and 22% of IT organizations are decreasing their use of ATM.

One of the observations that can be drawn from the response to this survey question is that:

The primary WAN services used by IT organizations are MPLS and the Internet.

The Webtorials Respondents were also asked to indicate the change that they anticipated that their organization would make over the next year relative to their usage of MPLS, VPLS and the Internet. Table 15 shows the percentage of The Webtorials Respondents that indicated that their organization would increase their use of those services.

Table 15: Increase in the use of key WAN Services

WAN Service	Percentage of Organizations Increasing their Use of this Service
MPLS	50.0%
VPLS	37.5%
Internet traffic to external sites	83.5%
Internet traffic to internal sites	74.3%

One of the conclusions that can be drawn from the data in Table 15 is that:

While IT organizations will increase their reliance on both MPLS and the Internet, they will make a relatively greater increase in their reliance on the Internet.

WAN Design Criteria and Challenges

The Webtorials Respondents were given a list of possible concerns and were asked to indicate which two were their company's primary concerns relative to its use of MPLS and the Internet. The set of concerns that were presented to The Webtorials Respondents is shown in the left hand column of Table 16. The second and third columns from the left in Table 16 show the percentage of The Webtorials Respondents who indicated that the concern is one of their company's two primary concerns with MPLS and the Internet respectively. The right hand column is the difference between the second and third columns from the left. This column will be referred to as the delta column.

The delta column contains positive and negative numbers. A positive number means that that concern was mentioned more often relative to MPLS than it was mentioned relative to the Internet. For example, The Webtorials Respondents mentioned cost as one of their primary concerns about the use of MPLS 22.1% more often than they mentioned cost as one of their primary concerns about the use of the Internet. Analogously, a negative number means that that concern was mentioned more often relative to the Internet than it was relative to MPLS. For example, The Webtorials Respondents mentioned latency as one of their primary concerns about the use of the Internet 19.3% more often than they mentioned latency as one of their primary concerns about use of MPLS.

Table 16: Concerns about MPLS

Concern	MPLS	Internet	Delta
Cost	60.1%	38.0%	22.1%
Lead time to implement new circuits	32.2%	11.4%	20.8%
Uptime	30.1%	46.3%	-16.2%
Latency	27.0%	46.3%	-19.3%
Lead time to increase capacity on existing circuits	23.5%	13.1%	10.4%
Jitter	14.8%	18.8%	-4.0%
Packet Loss	12.2%	26.2%	-14.0%

The primary concerns that IT organizations have with the use of MPLS are cost, the lead time to implement new circuits and uptime. The primary concerns that IT organizations have with the use of the Internet are uptime, latency and cost.

IT organizations typically design their WAN based on the following criteria:

1. Minimize cost
2. Maximize availability
3. Ensure appropriate performance

As shown in Table 16, MPLS is regarded by The Webtorials Respondents as doing a good job at ensuring appropriate performance because it exhibits relatively small amounts of delay, jitter and packet loss. Unfortunately, MPLS is regarded poorly relative to the goal of minimizing cost. In contrast, the Internet is regarded relatively well on the goal of minimizing cost but is regarded relatively poorly on the goal of ensuring appropriate performance. In addition, The Webtorials Respondents expressed concerns about both MPLS and the Internet relative to the goal of maximizing availability.

As was pointed out in the section of this report entitled *The Emergence of Cloud Computing and Cloud Networking*, 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*. As that section also pointed out, in order to support a small number of business critical services and applications, a cloud network that is *good enough* will have to provide the highest possible levels of availability and performance. However, in a growing number of instances, a cloud network is *good enough* if it provides a best effort level of service at a reduced price. Hence, independent of the concerns that IT organizations have about the Internet:

> ***In a growing number of instances, Internet-based VPNs that use DSL for access are 'good enough' to be a cloud network.***

Some of the concerns that IT organizations have with the use of the Internet such as uptime, stem from the fact that in many cases IT organizations access the Internet over a single DSL link. The availability of DSL is somewhat lower than the availability of access technologies such as T1/E1 links. One impact of this reduced availability is that Internet VPNs based on DSL access are often used only as a backup connection to a primary private WAN circuit. This is unfortunate because the shortfall in quality is fairly small when compared to the dramatic cost savings and additional bandwidth that can be realized by using broadband connections such as DSL and cable. One technology that addresses this issue is referred to as an *aggregated virtual WAN*.

> ***The key concept behind an aggregated virtual WAN is that it simultaneously utilizes multiple enterprise WAN services and/or Internet connections in order to optimize reliability and minimize packet loss, latency and jitter.***

Aggregated virtual WANs and other types of alternate WAN services are discussed later in this section of the report. As that discussion highlights, aggregated virtual WANs have the potential to maximize the benefits of the Internet and possibly MPLS while minimizing the negative aspects of both.

Local Access to the Internet

The traditional approach to providing Internet access to branch office employees has been to carry their Internet traffic on the organization's enterprise network (e.g., their MPLS network) to a central site where the traffic was handed off to the Internet. The advantage of this approach is that it enables IT organizations to exert more control over their Internet traffic and it simplifies management in part because it centralizes the complexity of implementing and managing security policy. One disadvantage of this approach is that it results in extra traffic transiting the enterprise's WAN, which adds to the cost of the WAN. Another disadvantage of this approach is that it usually adds additional delay to the Internet traffic. The fact that centralized Internet access exhibits these disadvantages is significant because as highlighted in Table 16, cost and delay are two of the primary concerns that IT organizations have relative to the use of the Internet.

Some of the concerns that IT organizations have about the use of the Internet are exacerbated by backhauling Internet traffic to a central site.

The Webtorials Respondents were asked to indicate how they currently route their Internet traffic and how that is likely to change over the next year. Their responses are contained in Table 17.

Table 17: Routing of Internet Traffic

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

Driven in part to save money and in part to improve application performance:

Over the next year, IT organizations will make an increased use of distributed access to the Internet from their branch offices.

Cloud Networking Without the Internet

There is a temptation to associate the WAN component of *cloud networking* either exclusively or primarily with the traditional Internet[39]. However, due to a variety of well-known issues, such as packet loss at peering points, BGP's inability to choose the path with the lowest delay, the TCP Slow start algorithm, the Internet often exhibits performance problems. As such, the Internet is not always the most appropriate WAN service to use to access cloud computing solutions. To

[39] Throughout this report, the phrase "traditional Internet" will refer to the use of the Internet with one access link and not optimization functionality.

put the use of the Internet into context, The Webtorials Respondents were asked to indicate which WAN service their users would most likely use when accessing public and private cloud computing services over the next year. Their responses are shown in Table 18.

Table 18: WAN Services to Access Cloud Computing Services

	The Internet	An Internet overlay from a company such as Akamai	A traditional WAN service such as MPLS	WAN Optimization combined with a traditional WAN service; e.g. MPLS
Public Cloud Computing Services	58.8%	10.3%	17.7%	13.2%
Private Cloud Computing Services	27.2%	4.1%	37.8%	30.9%

The Webtorials survey base was asked the same question in 2010 and the answers they provided in 2010 are remarkably similar to what they provided in 2011. This implies that few IT organizations are making a significant change to their WAN in order to support cloud computing.

The data in Table 18 indicates that IT organizations understand the limitations of the traditional Internet relative to supporting cloud computing. In particular:

> ***In somewhat less than half of the instances that business users are accessing public cloud computing services, the primary WAN service that they intend to use is not the traditional Internet.***
>
> ***In almost three quarters of the instances that business users are accessing private cloud computing services, the primary WAN service that they intend to use is not the traditional Internet.***

However, techniques that IT organizations can use to mitigate their concerns about the use of the Internet are discussed later in this section of the report.

Service Level Agreements

As previously stated, the majority of IT organizations utilize MPLS and the usage of MPLS is expected to increase significantly. One of the reasons for the popularity of MPLS is that the major suppliers of MPLS services offer a number of different classes of service (CoS) designed to meet the QoS requirements of the varying types of applications that transit a WAN. For example, real-time applications are typically placed in what is often referred to as a Differentiated Services Code Point (DSCP) Expedited Forwarding class that offers minimal latency, jitter, and packet loss. Mission critical business applications are typically relegated to what is often referred to as a DSCP Assured Forwarding Class.

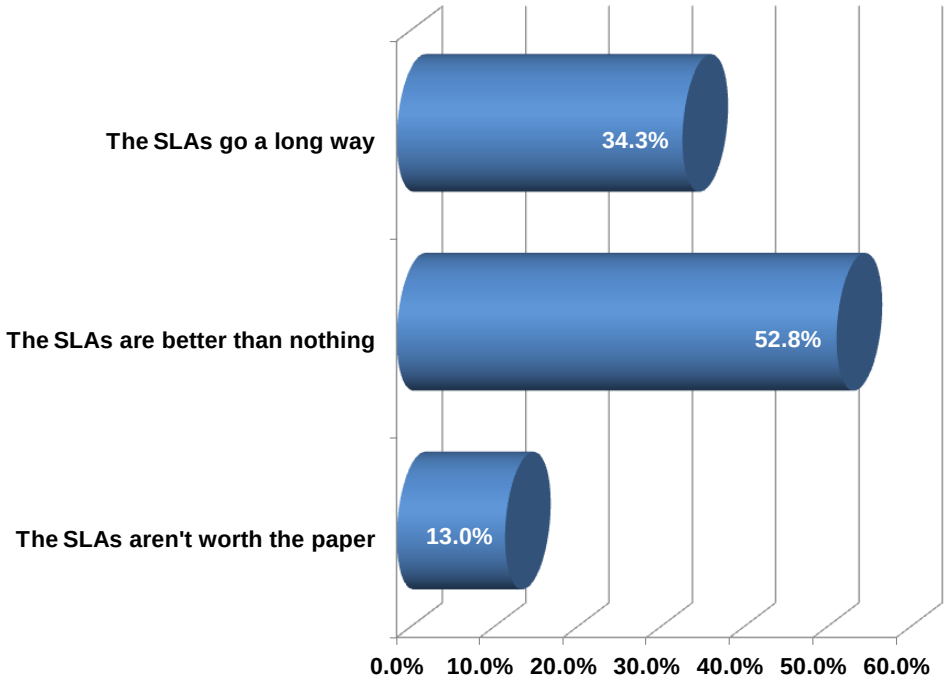
Each class of MPLS service is typically associated with a service level agreement (SLA) that specifies contracted ranges of availability, latency, packet loss and possibly jitter. Unfortunately, in many cases the SLAs are weak. In particular, it is customary to have the SLAs be reactive in focus; i.e., the computation of an outage begins when the customer opens a trouble ticket. In most cases, the carrier's SLA metrics are calculated as network-wide averages rather than for a specific customer site. As a result, it is possible for a company's data center to receive notably poor service in spite of the fact that the network-wide SLA metrics remain within agreed bounds. In addition, the typical level of compensation for violation of service level agreements is quite modest.

To gauge the effectiveness of SLAs that IT organizations receive from their network service providers (NSPs), The Webtorials Respondents were asked to indicate which of the following best describes the SLAs that they get from their NSPs for services such as MPLS.

- The SLAs go a long way towards ensuring that we get a quality service from the network service provider.
- The SLAs are better than nothing, but not by much.
- The SLAs are not worth the paper they are written on.

Their responses are shown in Figure 11.

Figure 11: The Effectiveness of SLAs



The fact that two thirds of The Webtorials Respondents indicated that the SLAs that they receive from network service providers are either not worth the paper they are written on, or that the SLAs they receive are not much better than nothing, demonstrates the weak nature of most SLAs.

The majority of IT organizations don't regard the SLAs that they receive from their network service providers as being effective.

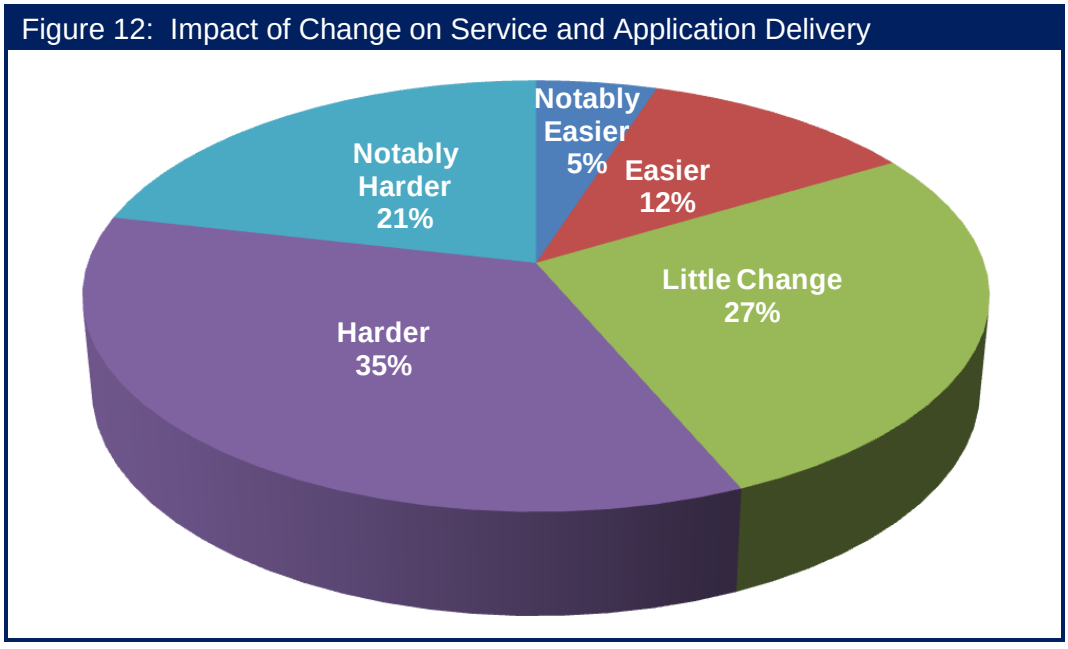
Optimizing the Performance of IT Resources

Background

This subsection of the report will discuss techniques that IT organizations can implement to overcome the limitations of protocols and applications and to optimize the use of their servers. The focus of this subsection is on how these techniques enable IT organizations to ensure acceptable application and service delivery over a WAN. The discussion in this subsection will focus on two classes of products: WAN Optimization Controllers (WOCs) and Application Delivery Controllers (ADCs).

The introduction to this section of this report discussed how the adoption of cloud computing in general is impacting the WAN and also discussed some of the specific factors that are driving change in the WAN. These factors included both the increasing number of mobile workers and the impact of multiple forms of virtualization. In order to gauge the effect that these factors have on the ability of an IT organizations to ensure acceptable application and service delivery, The Webtorials Respondents were asked "How will the ongoing adoption of mobile workers, virtualization and cloud computing impact the difficulty that your organization has with ensuring acceptable application performance?" Their responses are shown in Figure 12.

Figure 12: Impact of Change on Service and Application Delivery



One conclusion that can be drawn from Figure 12 is that:

The majority of IT organizations believe that factors such as the growth in the number of mobile workers and the increase in the use of virtualization and cloud computing will make ensuring acceptable service and application delivery either harder or notably harder.

WAN Optimization Controllers (WOCs)

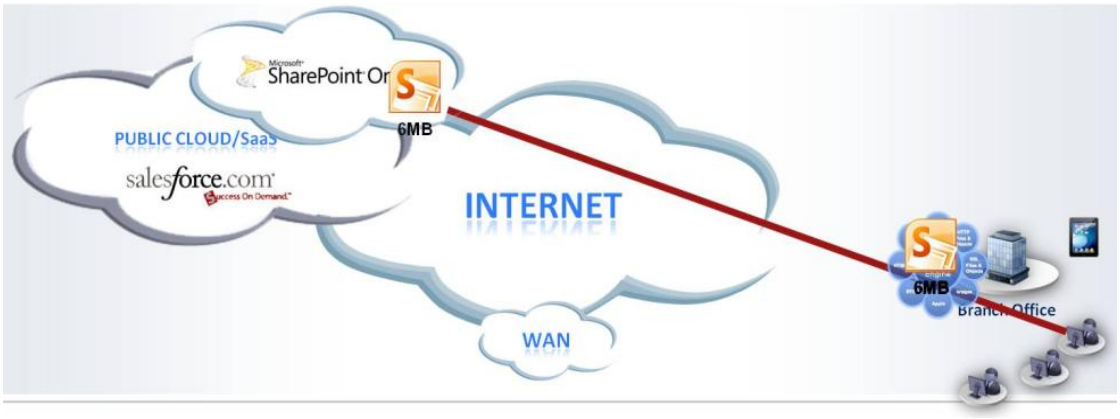
Goals of a WOC

The goal of a WOC is to improve the performance of applications and services that are delivered across a WAN from the data center either to a branch office, a home office or directly to a mobile user. In some cases the data center is owned and managed by the enterprise IT organization and in other cases it is owned and managed by a cloud computing service provider (CCSP). The WOC accomplishes this goal by implementing techniques to overcome the limitations of the WAN such as constrained bandwidth, delay and packet loss.

WOCs are often referred to as *symmetric solutions* because they typically require complementary functionality at both ends of the connection; i.e., a WOC in the data center and another WOC at the branch office. However, the requirement to improve the performance of applications and services acquired from a CCSP has been the impetuous for the deployment WOCs in an asymmetric fashion. As shown in Figure 13, in an asymmetric deployment of a WOC content is downloaded from a CCSP to a WOC in a branch office. Once the content is stored in the WOC's cache for a single user, subsequent users who want to access the same content will experience accelerated application delivery. Caching can be optimized for a range of cloud content, including Web applications, streaming video (e.g., delivered via Flash/RTMP or RTSP) and dynamic Web 2.0 content.

As previously described, IT organizations are moving away from a WAN design in which they backhaul their Internet traffic from their branch offices to a central site prior to handing it off to the Internet. Also, as is described in the next section of this report, there are a variety of techniques that enable IT organizations to improve both the price-performance and the availability of distributed Internet access. As a result of these factors, asymmetric WOC deployment as described in the preceding paragraph will increasingly be supported by a network design that features distributed Internet access. However, for this network design to be effective, IT organizations need to ensure that the design includes appropriate security functionality.

Figure 13: Asymmetric WOC Deployment



Modeling Application Response Time

A model is helpful to illustrate how the performance of a WAN can impact the performance of an application and it also serves to illustrate how a WOC can improve application performance. The following model (Figure 14) is a variation of the application response time model created by Sevcik and Wetzel[40]. Like all mathematical models, the following is only an approximation. For example, the model shown in Figure 14 doesn't account for the impact of packet loss.

As shown below, the application response time (R) is impacted by amount of data being transmitted (Payload), the WAN bandwidth, the network round trip time (RTT), the number of application turns (AppTurns), the number of simultaneous TCP sessions (concurrent requests), the server side delay (Cs) and the client side delay (Cc).

Figure 14: Application Response Time Model

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

In order to improve the performance of applications that are delivered over the WAN, WOCs implement a variety of techniques. For example, to mitigate the impact of a large payload, WOCs implement techniques such as compression and de-duplication. These techniques are explained in detail in The 2011 Application Delivery Handbook. The handbook also details criteria that IT organizations can use to evaluate WOCs as well as specific techniques that WOCs need to support in order to optimize:

- The rapidly growing amount of traffic that goes between data centers
- Desktop virtualization
- Delay sensitive applications such as voice, video and telepresence

The 2011 Application Delivery Handbook also describes techniques that can optimize the delivery of applications to mobile workers. Many IT organizations, however, resist putting any additional software on the user's device. In addition, many users resent having multiple clients (e.g., WOC, SSL VPN, IPSec VPN, wireless/cellular access) on their access device that are not integrated. One option for IT organizations on a going forward basis is to implement WOC software on mobile devices that is integrated with the other clients used by mobile workers. As is explained below, an alternative way that IT organizations can improve the performance of applications and services delivered to mobile users is to utilize an optimization service from a CCSP.

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

Application Delivery Controllers (ADCs)

ADCs provide some functionality, such as compression, that optimizes the delivery of bulk data over the Internet. However, the primary goal of an ADC is to improve the performance of servers.

The current generation of ADCs evolved from the earlier generations of Server Load Balancers (SLBs) that were deployed in front of server farms. While an ADC still functions as a SLB, the ADC has assumed, and will most likely continue to assume, a wider range of sophisticated roles that enhance server efficiency and provide asymmetrical functionality to accelerate the delivery of applications from the data center to individual remote users.

An ADC provides more sophisticated functionality than a SLB does.

Referring back to Figure 14, one of the factors that increase the application response time is server side delay. An ADC can reduce server side delay and hence can reduce the application response time. In particular, the ADC can allow a number of compute-intensive functions, such as SSL processing and TCP session processing, to be offloaded from the server. Server offload can increase the transaction capacity of each server, reducing the number of servers required for a given level of business activity.

The 2011 Application Delivery Handbook describes the primary techniques implemented by ADCs and identifies criteria that IT organizations can use to evaluate ADCs

Virtual Appliances

The section of this report entitled *The Emerging Data Center LAN* used the phrase *virtual switch* in two fundamentally different ways. One way referred to making two or more physical switches appear to be a single logical switch. The other way referred to the switching functionality that resides inside of a virtualized server.

In similar fashion, it is possible to look at a *virtual appliance* in a variety of fundamentally different ways. For example, two or more appliances, such as ADCs, can be combined to appear as a single logical ADC. Alternatively,
a single physical ADC can be partitioned into a number of logical ADCs or ADC contexts. Each logical ADC can be configured individually to meet the server-load balancing, acceleration and security requirements of a single application or a cluster of applications.

However, the most common use of the phrase *Virtual Appliance* refers to what is typically appliance-based software, together with its operating system, running in a VM. Virtual appliances can include WOCs, ADCs, firewalls, routers, IDS, IPS and performance monitoring solutions. As explained in the next subsection of this report, virtual appliances make it easier for an IT organization to deploy network and application optimization functionality at a CCSP's data center. That, however, is not the only advantage of a virtualized appliance.

One of the compelling advantages of a virtualized appliance is that the acquisition cost of a software-based appliance can be notably less than the cost of a hardware-based appliance with same functionality.

In many cases the acquisition cost of a software-based appliance can be a third less than the cost of a hardware-based appliance[41]. In addition, a software-based solution 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[42].

In addition to cost savings, another advantage of a virtual appliance is that it offers the potential to alleviate some of the management burdens because most of the provisioning, software updates, configuration, and other management tasks can be automated and centralized at the data center. An example of this is that if virtualized appliances have been deployed, then it is notably easier than it is in a more traditional environment for various networking functions (WOC, ADC, firewall, router, etc.) to be migrated along with VMs in order to replicate the VMs's networking environment in its new location.

> ***In many instances the benefits of the dynamic movement of a VM from one server to another are maximized if the supporting infrastructure is virtualized and can also be dynamically moved.***

A virtualized ADC also makes it easy for an IT organization to package and deploy a complete application. One example of this packaging is the situation in which an entire application resides on VMs inside a physical server. The virtualized ADC that supports the application resides in the same physical server and it has been tuned for the particular application. This makes it easy to replicate or migrate that application as needed. In this case, a virtualized ADC also provides some organizational flexibility. For example, the virtual ADC might be under the control of a central IT group or it might be under the control of the group that supports that particular application. The later is a viable option from an organizational perspective because any actions taken by the application group relative to their virtual ADC will only impact their application.

A virtual firewall appliance can also help IT organizations meet some of the challenges associated with server virtualization. That follows because virtual firewall appliances can be leveraged to provide isolation between VMs on separate physical servers as well as between VMs running on the same physical server. Through tight integration with the virtual server management system, virtual firewall appliances can also be dynamically migrated in conjunction with VM migration where this is necessary to extend a trust zone to a new physical location. In addition, hypervisor APIs, such as VMware's Vsafe, can allow physical/virtual firewall consoles to monitor servers for abnormal CPU, memory, or disk activity without the installation of special agent software.

The research report entitled Virtualization: Benefits, Challenges and Solutions, contains more detail on virtual appliances. Included in that report is a discussion of the challenges associated with virtual appliances, as well as suggested evaluation criteria.

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

Optimizing Access to Public Cloud Computing Solutions

As noted in the section of this report entitled *The Emergence of Cloud Computing and Cloud Networking*, one of the key challenges facing IT organizations that use either SaaS or IaaS solutions is improving the performance of those solutions. In order to quantify what IT organizations are doing to improve the performance of those solutions, The Webtorials Respondents were asked "If your company either currently acquires services from an Infrastructure-as-a-Service (IaaS) provider or you expect that they will within the next year, which of the following best describes the primary approach that your company will take to optimizing the performance of those solutions?" They were asked a similar question about their use of SaaS solutions.

The responses to these two questions are summarized in Table 19. The leftmost column in Table 19 lists the approaches that The Survey Respondents had to choose from. The middle column shows the percentage of The Survey Respondents that indicated that that would be their primary approach for optimizing services acquired from an IaaS provider. The rightmost column shows the percentage of The Survey Respondents that indicated that that would be their primary approach for optimizing services acquired from a SaaS provider.

Table 19: Optimizing CCSP Services

Approach	IaaS Services	SaaS Services
Leverage optimization functionality provided by the CCSP	29.3%	31.9%
Place a WOC on the service provider's site and on our premise	22.4%	13.9%
Use an optimization service from a company such as Akamai or Virtela	4.1%	3.6%
Do nothing	19.0%	17.5%
Don't know	25.2%	33.1%

A number of conclusions can be drawn from the data in Table 19, including:

There is significant interest in placing a WOC on premise at an IaaS provider's data centers.

Between a quarter and a third of IT organizations don't know how they will optimize the performance of services that they acquire from an IaaS or a SaaS provider.

In addition, referencing back to the discussion in the previous subsection, IT organization will have a notably easier time placing an optimization device, whether that is a WOC or an ADC, at an IaaS provider's data center if the device is virtualized. That follows because if the device is virtualized, the IT organization can control the deployment of the functionality. If the device is physical, then the IT organization needs to get the IaaS provider to offer space for the device and to install it.

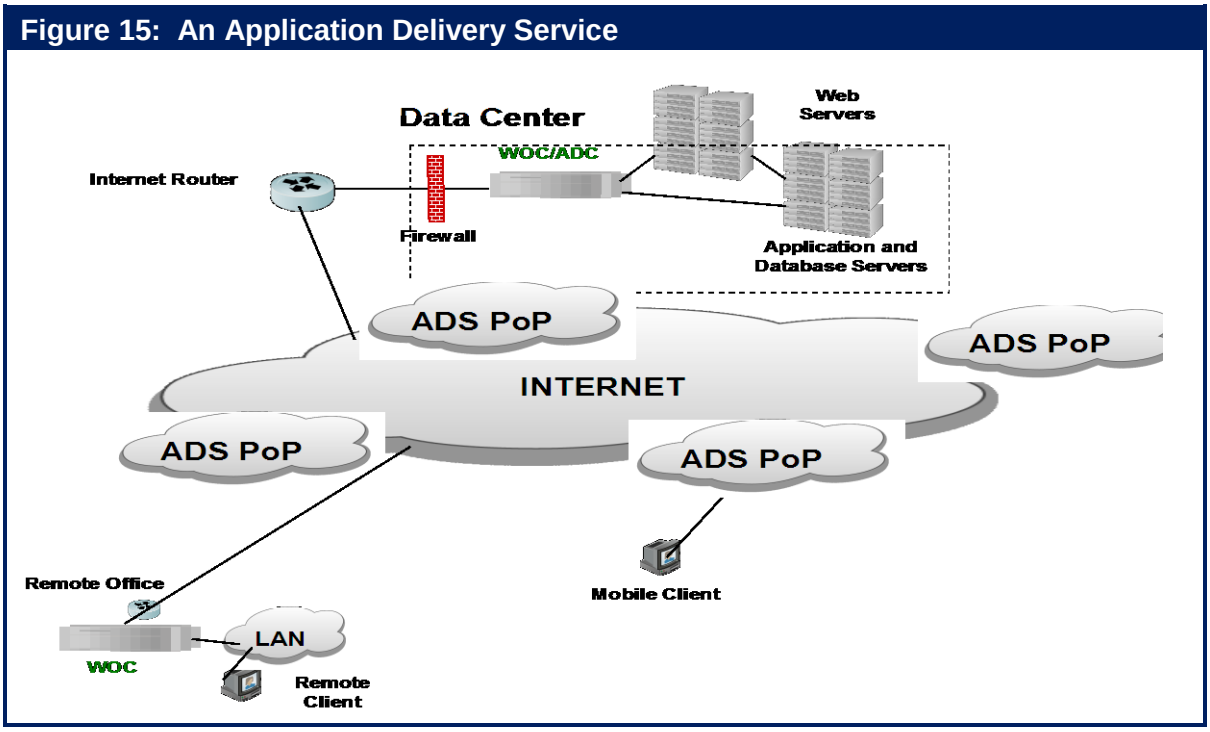
Alternative WAN Services

As noted, there is not a new generation of fundamentally new WAN technology currently under development. However, as is described below, there are a number of WAN service alternatives that are variations on existing WAN technologies and services that better enable IT organizations to meet their WAN design goals. A number of these alternatives are either complementary to the WAN optimization technologies previously discussed or they depend partially on WAN optimization technologies to deliver acceptable levels of service quality.

An Internet Overlay

As described in the preceding subsection, IT organizations often implement WOCs and ADCs in order to improve network and application performance. However, these solutions make the assumption that performance characteristics within the WAN itself can't be 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 WAN services such as MPLS. However, this assumption doesn't apply to enterprise application traffic that transits the Internet because there are significant opportunities to optimize performance within the Internet itself based on implementing an Internet overlay. An Internet overlay leverages service provider resources that are distributed throughout the Internet in order to optimize the performance, security, reliability, and visibility of the enterprise's Internet traffic. As shown in Figure 15, all client requests to the application's origin server in the data center are redirected via DNS to a server in a nearby point of presence (PoP) that is close to users of the application, typically within a single network hop. This edge server that is close to the users then optimizes the traffic flow to the server closest to the data center's origin server. Throughout this section, the Internet overlay that is depicted in Figure 15 will be referred to as an Application Delivery Service (ADS).

Figure 15: An Application Delivery Service



An ADS provides a variety of optimization functions that generally complements the functionality provided by WOCs and ADCs. One such function that is often provided by an ADS is content offload. This calls for taking static content out of a data-center and placing it in caches in servers and in replicated in-cloud storage facilities that are close to the users. Because the content is close to the users, IT organizations that offload content and storage improve response time and simultaneously reduce both their server utilization as well as the bandwidth utilization of their data center access links.

Some of the other functionality that is often associated with an ADS includes:

- Route optimization
- Transport optimization
- HTTP protocol optimization
- Visibility

In addition to the functionality listed above, some ADSs incorporate Web application firewall functionality.

One use case for an ADS that is growing in importance stems from that fact that not all CCSPs will support virtual WOC instances in their data centers. This is particularly true of SaaS providers. Access to services provided by a CCSP can be accelerated via an ADS. Even greater improvement in application delivery over the Internet can be achieved by migrating WOC functionality into ADS PoPs and by migrating ADS functionality into enterprise WOCs. The result places WOC functionality very close[43] to the CCSP's data center and moves the ADS PoP right to the on-premise edge of the enterprise network. This helps optimize application and service delivery over the Internet between the branch office and the CCSP's site.

Another approach to the asymmetrical acceleration of access to CCSP services over the Internet was mentioned in the preceding subsection: caching content at the branch office or other enterprise site. As mentioned, once the content is stored in the cache for a single user, subsequent users will experience accelerated application delivery.

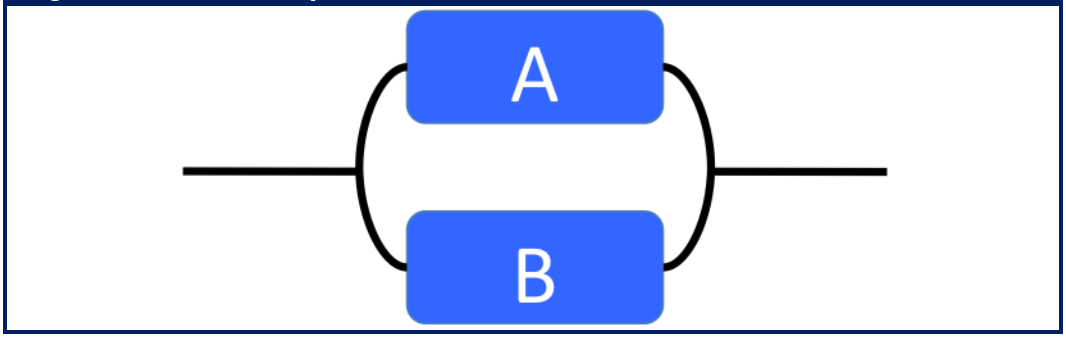
Dual ISP Internet VPN with Policy Based Routing

The preceding subsection of this report identified the concerns that IT organizations have with the use of the Internet. The two primary concerns are uptime and latency. Another approach to overcoming the limitations of the Internet is to connect each enterprise site to two ISPs. Having dual connections can enable IT organizations to add inexpensive WAN bandwidth and can dramatically improve the reliability and availability of the WAN.

For example, Figure 16 depicts a system that is composed of two components that are connected in parallel.

[43] While the closeness to the CCSP will vary based on the CCSP and the provider of the ADS, ideally the WOC functionality will be one hop away from the CCSP.

Figure 16: Two Components Connected in Parallel



The system depicted in Figure 16 is available unless both of the two components are unavailable. Assuming that each component is a diversely routed DSL or cable access line and that one of the access lines has an availability of 99% and the other has an availability of 98%, then the system has an availability of 99.98%. Alternatively, if both access lines have an availability of 99%, then the system is available 99.99% of the time[44]. This level of availability is equal to or exceeds the availability of most MPLS networks.

Traffic can be shared by the two connections by using Policy Based Routing (PBR). When a router receives a packet, it normally decides where to forward it based on the destination address in the packet, which is then used to look up an entry in a routing table. Instead of routing by the destination address, policy-based routing allows network administrators to create routing policies to select the path for each packet based on factors such as the identity of a particular end system, the protocol or the application.

Perhaps the biggest limitation of the PBR approach it that it creates a static allocation of traffic to multiple links and it doesn't have the ability to reallocate the traffic when the quality of one of the links degrades. The static nature of the policies means unless there is an outage of one of the links, that a given class of traffic will always be allocated to the same network connection.

Dual ISPs and PBR can be used in conjunction with WOCs to further alleviate the shortcomings of Internet VPNs, bringing the service quality more in line with MPLS at a much lower cost point. For example, a WOC can classify the full range of enterprise applications, apply application acceleration and protocol optimization techniques, and shape available bandwidth in order to manage application performance in accordance with enterprise policies. As a result,

In many situations, a dual ISP-based Internet VPN with PBR can deliver a level of CoS and reliability that is comparable to that of MPLS at a significantly reduced price.

Part of the cultural challenge that IT organizations have relative to migrating traffic away from their MPLS network and onto an Internet based network is that Internet based networks don't provide a performance based SLA. However, as previously described, the majority of IT organizations don't place much value in the SLAs that they receive from their network service providers.

[44] If, as described later, 4G is added as a third access technique and if each access technique has an availability of 99%, then the system as a whole has an availability of 99.9999%.

Hybrid WANs with Policy Based Routing

As noted, some IT organizations are reluctant to abandon traditional enterprise services such as MPLS. An alternative design that overcomes their concerns is a hybrid WAN that leverages multiple WAN services, such as traditional enterprise WAN services and the Internet, and which uses PBR for load sharing. The advantage of a hybrid WAN is that the CoS of MPLS can be leveraged for delay sensitive, business critical traffic with the Internet VPN used both for other traffic and as a backup for the MPLS network. As in the case of the dual ISP based Internet VPN, the major disadvantage of this approach is the static nature of the PBR forwarding policies. Since PBR cannot respond in real time to changing network conditions, it will consume more costly bandwidth than would a dynamic approach to traffic allocation. A second drawback of Hybrid WANs based on PBR is that they can prove to be overly complex for some IT departments. As with many other types of WAN services, hybrid WANs can also be used in conjunction with WOCs and ADCs.

Aggregated Virtual WANs

A relatively new class of device has emerged to address the shortcomings of PBR-based hybrid WANs. WAN path controller (WPC) is one phrase that is often used to describe devices that work in conjunction with WAN routers to simplify PBR and to make selections of the best WAN access link or the best end-to-end WAN path from a number of WAN service options.

Some members of this emerging class of products are single-ended solutions whereby a device at a site focuses on distributing traffic across the site's access links on a per-flow basis. Typical capabilities in single-ended solutions include traffic prioritization and bandwidth reservation for specific applications. These products, however, lack an end-to-end view of the available paths and are hence limited to relatively static path selections.

In contrast, symmetrical or dual-ended solutions are capable of establishing an end-to-end view of all paths throughout the network between originating and terminating devices and these solutions can distribute traffic across access links and specific network paths based on either a packet-by-packet basis or a flow basis. These capabilities make the multiple physical WAN services that comprise a hybrid WAN appear to be a single *aggregated virtual WAN*.

Aggregated virtual WANs (avWANs) represent another technique for implementing WANs based on multiple WAN services (e.g., MPLS, Frame Relay and the Internet) and/or WANs based on just multiple Internet VPN connections. An aggregated virtual WAN transcends simple PBR by dynamically recognizing application traffic and allocating traffic across multiple paths through the WAN based on real-time traffic analytics, including:

- The instantaneous end-to-end performance of each available network: This allows the solution to choose the optimal network path for differing traffic types. One differentiator among virtual WAN solutions is whether the optimal path is chosen on a per packet basis or on a per flow basis. Per packet optimization has the advantage of being able to respond instantaneously to short term changes in network conditions.

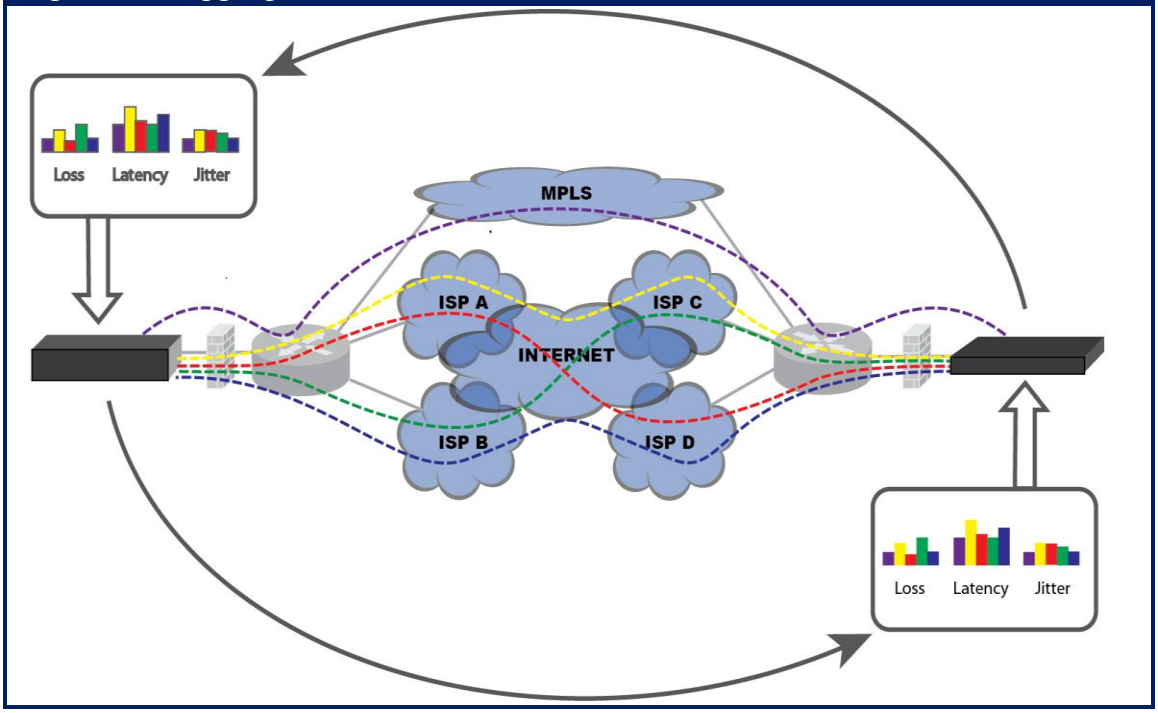
- The instantaneous load for each end-to-end path: The load is weighted based on the business criticality of the application flows. This enables the solution to maximize the business value of the information that is transmitted.

- The characteristics of each application: This includes the type of traffic (e.g., real time, file transfer); the performance objectives for delay, jitter and packet loss; as well as the business criticality and information sensitivity.

One of the primary reasons why IT organizations backhaul their Internet traffic to a central site over an enterprise WAN service is because of security concerns. In order to mitigate those concerns when using an avWAN for direct Internet access, the avWAN should support security functionality such as encryption.

Like other hybrid WANs, an avWAN (Figure 17) allows IT organizations to add significant amounts of additional bandwidth to an existing MPLS-based WAN at a relatively low incremental cost. In addition to enabling the augmentation of an MPLS WAN with inexpensive Internet connectivity, aggregated virtual WANs also give IT organizations the option to reduce its monthly ongoing expense by either eliminating or reducing its MPLS connections while simultaneously providing more bandwidth than the original network design provided.

Figure 17: Aggregated Virtual WANs



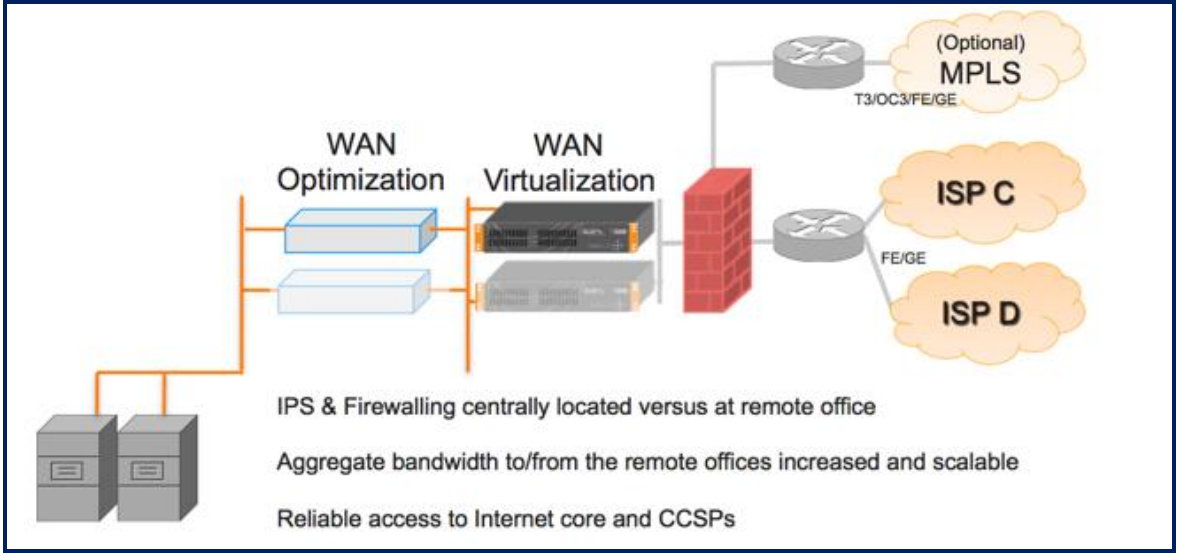
As shown in Figure 17, because the two avWAN appliances work together to continuously measure loss, latency, jitter and bandwidth utilization across all of the various paths between any 2 locations, an aggregated virtual WAN can rapidly switch traffic away from a path that is exhibiting an unacceptable level of performance. This capability, combined with the availability advantages of parallel systems as depicted in Figure 16, means that all of the bandwidth in each of the paths can be used most of the time, and that most of the bandwidth can be used virtually all of the time. This combination of capabilities also underscores the ability of aggregated virtual WANs to deliver performance predictability that equals, and in many cases exceeds, that of a single MPLS network.

Because of the high availability and performance predictability of aggregated virtual WANs, IT organizations can now leverage a number of WAN services that are dramatically lower in cost than traditional MPLS services. This includes DSL and cable Internet access from branch offices and fiber access to the Internet from data centers. It also positions IT organizations to take advantage of the huge volumes of very inexpensive Internet access bandwidth that are typically available at co-location facilities.

While the preceding discussion focused on DSL and cable access to the Internet it is important to realize that over the next year or two, there will be a broad scale deployment of 4G services on the part of most wireless service providers. There will be some variability in the effective bandwidth of 4G services based in part on the fact that the wireless service providers will not all implement the same technologies. It should generally be possible, however, for users of these services to realize throughput in the range of three to four megabits per second, which is roughly equivalent to two T1 or E1 access lines. This will make 4G services a viable access service for some branch offices. For example, a 4G service could be combined with Internet access via DSL as part of a virtual WAN. In addition to providing cost savings, due to the inherent diverse routing associated with 4G and DSL, this design would provide a very high level of reliability.

There are three scenarios in which an avWAN offers significant benefits for accessing cloud computing services. The first scenario is for enterprises that have implemented centralized Internet access and who want to access a wide range of public cloud network services on the Internet, such as those offered by SaaS and IaaS providers. The implementation of an avWAN makes Internet backhaul far less expensive, higher capacity and more scalable than if the backhaul was done entirely using traditional enterprise WAN services. Between the remote sites and the central data center, all the backhauled traffic will benefit from the reliability, security, and QoS features of an avWAN, as shown in Figure 18. Between the central data center and public cloud services sites reliability will be improved because of the dual ISP connections (ISPs C and D), but the unique benefits of WAN virtualization will not generally be available on this portion of the end-to-end path because of the impracticality of the SaaS or IaaS CCSP provisioning a dedicated APN appliance for each service subscriber.

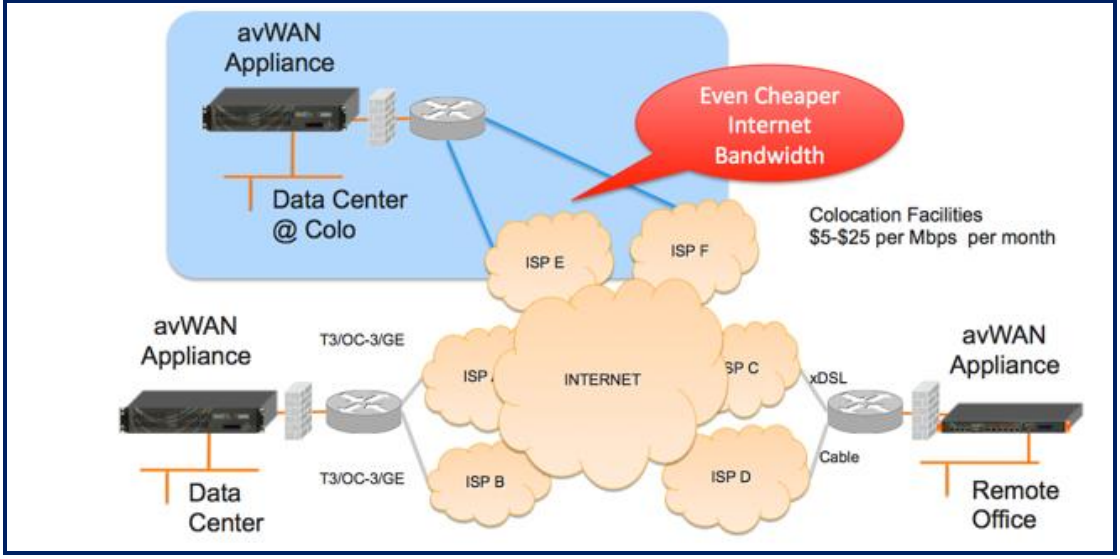
Figure 18: Scalable and Economical Internet Backhaul



The second scenario is the situation in which the enterprise has implemented a private cloud computing data center, either at a central site managed by the enterprise or at a co-location facility. Here the full benefits of an avWAN can be derived for all of the traffic, including the traffic that goes between enterprise users and the private cloud resources as well as the server-to-server communications between the data centers. As shown in Figure 19, the economic advantages of an avWAN are further enhanced when the private cloud data center is built at a co-location facility where Internet access costs are generally considerably lower than at the enterprise central sites.

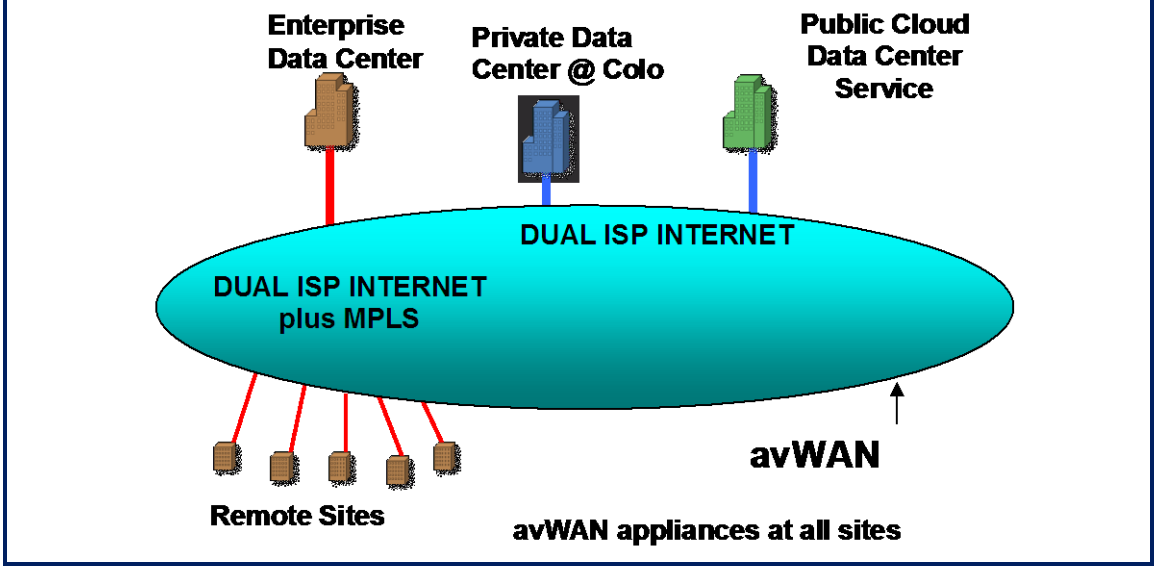
In this scenario, all Internet traffic can be backhauled to the enterprise's data center or to the co-location facility. If there are a number of geographically dispersed co-location sites, directing backhauled Internet traffic to the nearest site can minimize the extra propagation latency associated with centralized Internet access.

Figure 19: Private Cloud Data center at Colo Facility



The third scenario is the situation in which the enterprise has subscribed to a public cloud based service, such as an outsourced private data center located on the CCSP premises or a Virtual Private Data Center hosted in a CCSP multi-tenant data center. With both types of data center services it should be possible to extend the avWAN to include the CCSP's data center by having avWAN appliances provisioned at these data center sites. Figure 20 shows one way that a hybrid cloud computing environment could be supported by an avWAN that is comprised of dual ISP connections at all sites plus an MPLS connection at some or all of the remote sites and at the enterprise's data center. As before, all general purpose Internet traffic could be economically backhauled to the enterprise data center, while all intra-enterprise cloud traffic can flow directly via the virtual WAN.

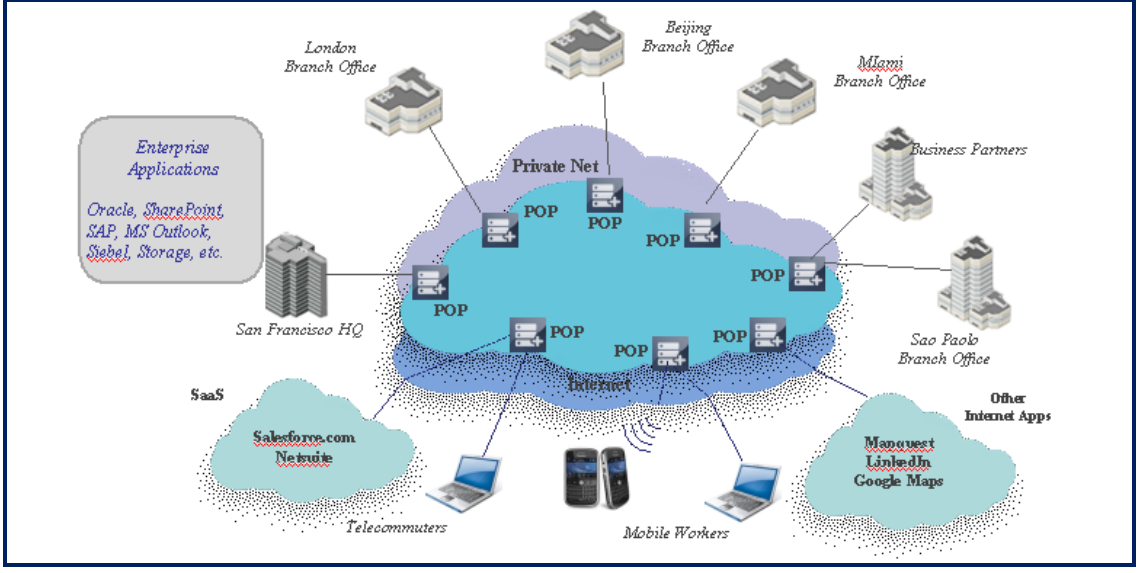
Figure 20: Hybrid Cloud Computing based on Aggregated Virtual WAN



Cloud-Based Network and Application Optimization

As mentioned in the section of this report entitled *The Emergence of Cloud Computing and Cloud Networking*, network and application optimization has become available from CCSPs as a Cloud Networking Service (CNS). In this situation, instead of a physical or virtual WOC at each site, the WOC functionality is provided at the CCSP's cloud data centers or POPs, which ideally are in close proximity to the enterprise users, the data centers and the providers of other cloud services. As shown in Figure 21, the PoPs are interconnected by the CCSP's core network with customer access to each PoP provided via the Internet or via an enterprise WAN service. The CNS core network could be an Internet overlay, a private IP network or possibly a multi-carrier MPLS/IP network that uses intelligent routing capabilities similar to an aggregated virtual WAN or ADS in order to provide high levels of performance and reliability.

Figure 21: Network and Application Optimization CNS



Therefore, a network and application optimization CNS can be considered to be another form of alternative WAN service that is layered on top of the Internet or existing private WAN services from one or more carriers. The key differentiation is that this class of WAN service has WAN Optimization built into the PoPs. The form of a CNS can be used for the traditional challenge of optimizing communications between users in a branch office and IT services that are provided at the enterprise's data centers. This form of a CNS can also be used in those situations in which installing WOC functionality is either not economical or it is problematical; e.g., at home offices, small branch offices, IaaS data centers, SaaS data centers or on mobile devices.

VPLS

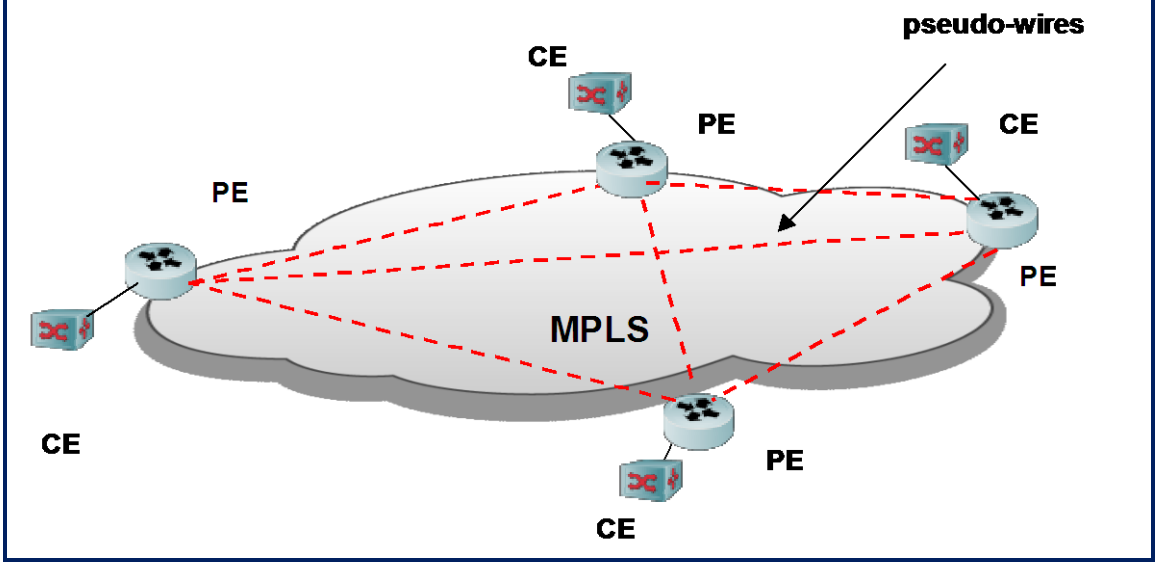
As previously mentioned:

VPLS represents the combination of Ethernet and MPLS.

While VPLS is not widely implemented today, the data in Table 15 indicates that more than a third of IT organizations will increase their use of VPLS over the next year.

VPLS is a class of VPN that supports the connection of customer edge (CE) Layer 2 switches at multiple sites into a single bridged, multipoint-to-multipoint domain over a service provider's IP/MPLS network, as shown in Figure 22. VPLS presents an Ethernet interface to customers that simplifies the LAN/WAN boundary for Service Providers and customers, and enables rapid and flexible service provisioning. All sites in a VPLS appear to be on the same LAN, regardless of location. A companion technology, Virtual Private Wire Services (VPWS), provides point-to point services.

Figure 22: A VPLS Service Linking Four Customer Sites



With VPLS, either the Border Gateway Protocol (BGP) or the Label Distribution Protocol (LDP) is used to create the required pseudo-wires to fully mesh the provider edge (PE) devices serving the customer sites. Meshed pseudo-wires support the multipoint-to-multipoint nature of the virtual LAN and improve reliability. Reliability is enhanced because in case of failure in the

MPLS network, traffic will automatically be routed along available backup paths, providing very short failover times.

VPLS MPLS packets have a two-label stack. The outer label is used for normal MPLS forwarding in the service provider's network. If BGP is used to establish the VPLS, the inner label is allocated by a PE as part of a label block. If LDP is used, the inner label is a virtual circuit ID assigned by LDP when it first establishes a mesh between the participating PEs. Every PE keeps track of assigned inner label, and associates these labels with the VPLS instance.

Table 20 provides a high level comparison of the different types of Ethernet WAN services available for LAN extension between data centers . It should be noted that there are other options for LAN extension, such as Ethernet over leased dark fiber and Ethernet over GRE tunneling through a private IP network. As described previously, VXLAN is VM-specific overlay solution for LAN extension.

Table 20: Ethernet WAN Service Types

Service Topology	Access Link	Provider Core	Service Type	Tunneling
Ethernet end-end	Ethernet	Ethernet	Pt-Pt or Mpt-Mpt	802.1Q or Q in Q
Ethernet/IP	Ethernet	IP	Pt-Pt or Mpt-Mpt	L2TPv3
VPLS/VPWS	Ethernet	MPLS	Pt-Pt or Mpt-Mpt	EoMPLS

Emerging Cloud Networking Specific Solutions

The preceding discussion of WAN services provided some insight into the interplay between the general requirements of cloud computing and the capabilities of WAN services to meet those requirements. One of the goals of this section of the report is to describe the functionality that is required to support a particular form of hybrid cloud computing – cloud balancing. Another goal of this section of the report is to describe some of the optimization functionality that is being developed specifically to support cloud computing.

Cloud Balancing

As previously described, a hybrid cloud relies on a WAN to provide the connectivity between the enterprise's locations, including the enterprise's data center(s) and its remote sites, and the public cloud data center(s) that is providing the IaaS or other cloud service. 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 creates a number of requirements. This includes:

- **VLAN Extension**
  As is the case for private clouds, hybrid clouds depend heavily on VM migration among geographically dispersed servers connected by a WAN in order to ensure high availability

and dynamic response to changes in user demand for services. The VLANs within which VMs are migrated must be extended over the WAN between and amongst 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 enhances security as well as transparency by allowing these application services to be provisioned from the private enterprise data center and eliminating manual intervention to modify server configurations as the application and its 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 users or the IT resources that the users are accessing. This means that the public cloud data centers need to offer the same WAN optimization and application acceleration capabilities that are deployed within the enterprise. In addition, WOCs may well be needed between the enterprise's private cloud data center(s) and the public cloud data center(s) in order to accelerate VM migration, system backups, and other bulk data transfers between these data centers.

- **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.

- **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.

WAN Optimization and Application Delivery for Cloud Sites

One of the most significant trends in the WAN optimization market is the development of new products and new product features that are designed to enable IT organizations to leverage public and hybrid clouds as extensions of their enterprise data centers. Some recent and anticipated developments include:

Cloud Optimized WOCs: These are purpose-built virtual WOC appliances for deployment in public cloud environments. Cloud Optimized features include compatibility with cloud virtualization environments, SSL encryption and acceleration, and automated migration or

reconfiguration of virtual WOCs in conjunction with VM provisioning or migration. As previously mentioned, WOCs can either be deployed in a symmetric fashion, with a WOC on each end of the WAN link; or in an asymmetric fashion, with a WOC deployed just in a branch office.

Cloud Storage Optimized WOCs: These are purpose-built virtual or physical WOC appliances for deployment in the enterprise's data center(s) and also at public cloud Storage as a Service environments that are used for backup and archival storage. Cloud optimized features can include support for major backup and archiving tools, de-duplication to minimize the required data transfer bandwidth and the storage capacity that is required, and support for SSL and AES encryption.

Data Mobility Controller Enhancements: Data Mobility Controllers (DMCs) facilitate the transfer of high volume data between enterprise data centers or private cloud data centers. DMC products are still in a early stage of evolution and a number of developments can be expected in this space, including enhanced hardware support for various functions including encryption and higher speed WAN and LAN interfaces at 10 GbE and higher in order to support a combination of highly efficient data reduction and high bandwidth WAN services.

Cloud Optimized Application Delivery Controllers: One trend in the evolution of ADCs is increasing functional integration with more data center service delivery functions. As organizations embrace cloud computing models, service levels need to be assured irrespective of where the applications are hosted. As is the situaion with WOCs, ADC vendors are in the process of adding enhancements that support the various forms of cloud computing, including:

- **Hypervisor–based Multi-tenant ADC Appliances:** Partitioned ADC hardware appliances have for some time allowed service providers to support a multi-tenant server infrastructure by dedicating a single partition to each tenant. Enhanced tenant isolation in cloud environments can be achieved by adding hypervisor functionality to the ADC appliance and by dedicating an ADC instance to each tenant. Each ADC instance is then afforded the same type of isolation as a virtualized server instance, with protected system resources and address space. A combination of hardware appliances, virtualized hardware appliances and virtual appliances provides the flexibility for a cloud service provider to offer highly customized ADC services that are a seamless extension of an enterprise customer's IT environment.

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

cloud balancing routing decisions. When architectural consistency extends to the hypervisors across the cloud, integration of cloud balancing/bursting ADCs with the hypervisors management systems can help the routing of application traffic synchronized with private and public cloud resource availability and performance. Access control systems integrated within the GSLB and ADC make it possible to maintain control of applications wherever they reside in the hybrid cloud.

The following table summarizes the applicability of the various WAN services and optimization solutions to different types of cloud computing.

Table 21: Applicability of WAN Technologies in the Cloud

	Private Cloud	Hybrid Cloud	IaaS	ISV SaaS
MPLS	Yes	Yes	Depends on the Service Provider	Depends on the Service Provider
Internet	Yes	Yes	Yes	Yes
Internet Overlay	Yes	Yes	Yes	Yes
Hybrid WAN	Yes	Yes	Depends on the Service Provider	Depends on the Service Provider
WAN Virtualization	Yes	Yes	No	No
VPLS	Yes	Yes	Depends on the Service Provider	Depends on the Service Provider
Cloud Bridging	No	Yes	No	No
WOC/ADC	Yes	Yes	No	No
VA WOC/ADC	Yes	Yes	Yes	No
Cloud WOC	Yes	Yes	Yes	Yes
Storage Service WOC	No	No	Yes	No
DMC	Yes	Yes	Yes	No
Hypervisor ADC	In multi-tenant data centers	Yes	Yes	No
Cloud Bursting/ Balancing ADC	Yes	Yes	No	No
Cloud Balancing	Yes	Yes	No	No

The Management of Cloud Computing

Importance of Managing Cloud Computing

One of the questions that was administered to The 2011 Webtorials Respondents was "Please indicate how important it is to your organization to get better at each of the following tasks over the next year." The question included twenty wide-ranging management tasks, many of which were included in a similar question that was administered to The 2010 Webtorials Respondents. The possible answers were to the question were:

- Extremely important
- Very important
- Moderately important
- Slightly important
- Not at all important

In order to avoid restating that question each time it is referenced in this section of the report, it will be referred to as The Question. Three of the twenty tasks that were included in The Question were managing private, managing hybrid and managing public cloud computing solutions. The responses of The 2011 Webtorials Respondents are summarized in Table 22.

Table 22: 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 22 is that

The majority of IT organizations believe that getting better at managing private cloud computing solutions is either very or extremely important.

Another observation that can be drawn from the data in Table 22 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 reasons for this phenomenon is that enterprise IT organizations are making more use of private cloud solutions than they are of either public or hybrid cloud solutions. Another reason 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.

The Evolving Management Environment

The Increased Focus on Services

Just as IT organizations are getting somewhat comfortable with managing the performance of applications, they are being tasked with managing the performance of services. IT professionals use the term *service* in a variety of ways. For example, the ITIL definition of service[45] states that a service:

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

In part because the ongoing adoption of virtualization and cloud computing has created the concept of everything as a service (XaaS), the term service as used in this section of the report will sometimes refer to services that IT organizations acquired from a public cloud computing provider; e.g., compute, storage, applications.

In order to quantify the interest that IT organizations have in managing this type of service, three of the twenty tasks that were included in The Question were:

- Effectively monitoring and managing compute services acquired from a third party such as Rackspace.

- Effectively monitoring and managing storage services acquired from a third party such as Rackspace.

- Effectively monitoring and managing applications acquired from a software-as-a-service provider such as Salesforce.com.

The responses of The 2011 Webtorials Respondents are summarized in Table 23.

Table 23: Importance of Effectively Monitoring and Managing Cloud Solutions

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

As shown in Table 23, 28.6% of The Webtorials Respondents responded with “not at all important” when asked about the importance of getting better at monitoring and managing

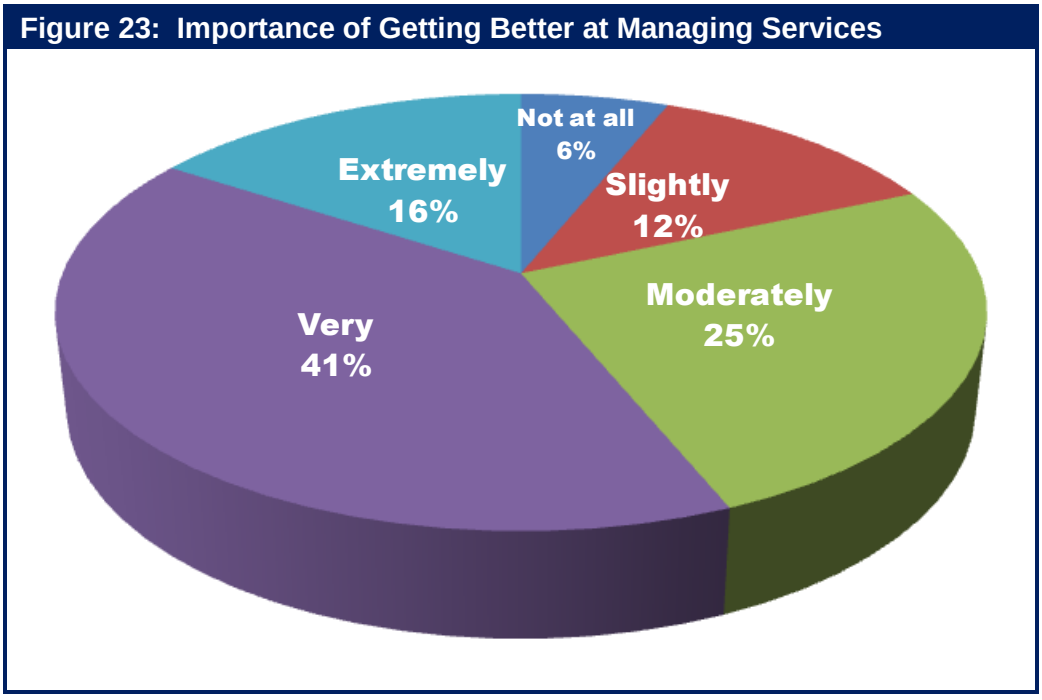
[45] ITIL definition of service

storage services that they acquire from a public cloud computing vendor; a.k.a., an Infrastructure as a Service (IaaS) vendor.

The 28.6% was the largest percentage to respond with "not at all important" for any of the twenty management tasks that were presented to The Webtorials Respondents. Given that, it is possible to conclude that monitoring and managing the services obtained from an IaaS vendor is not an important task. However, that conclusion is contradicted by the fact that almost a quarter of The Webtorials Respondents indicated that getting better at monitoring and managing storage services acquired from an IaaS vendor was either very or extremely important. A more reasonable conclusion is based on the observation that many companies don't make any use of storage and compute services from an IaaS vendor and the ones that do often make only minor use of such services. Based on that observation, the data in Table 23 suggests that if a company makes significant use of the services provided by an IaaS vendor, then monitoring and managing those services is indeed an important task.

The term service as used in this section of the report will sometimes refer to business services that involve multiple inter-related applications. One of the management tasks that was included in The Question was "Manage a business service, such as CRM, that is supported by multiple, inter-related applications." The answers of The 2011 Webtorials Respondents are summarized in Figure 23.

Figure 23: Importance of Getting Better at Managing Services



One observation that can be drawn from Figure 23 is that:

The majority of IT organizations believe that getting better at managing inter-related applications that comprise a business service is either very or extremely important.

Unfortunately, the adoption of cloud computing will further complicate the task of managing the inter-related applications that comprise a service. That follows because in a cloud computing

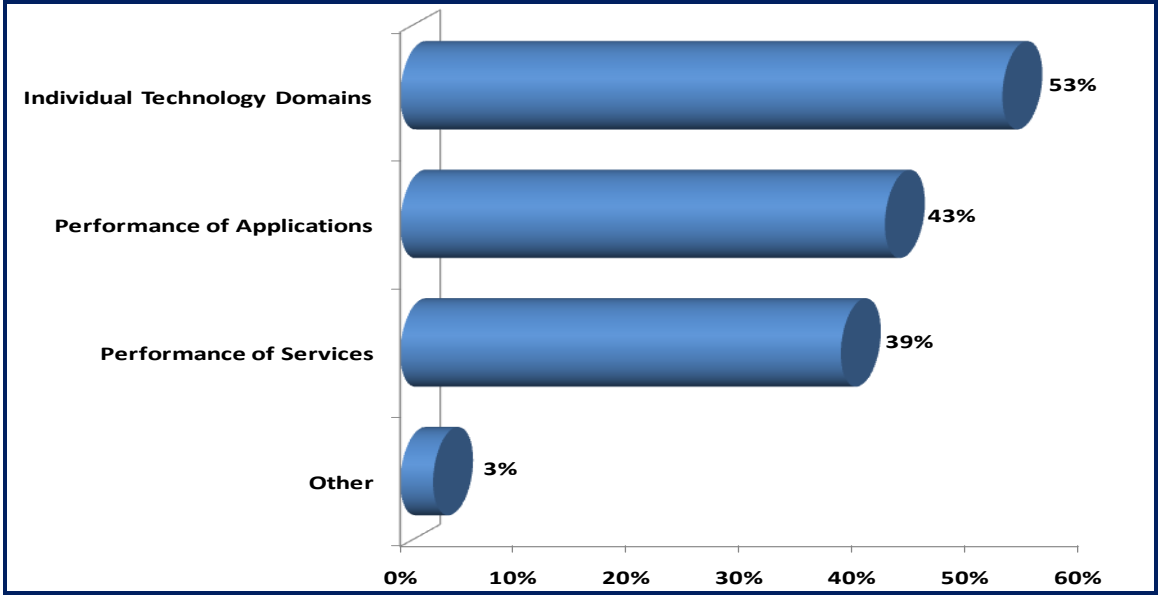
environment, the applications that comprise the service will increasingly be supported by an infrastructure that is virtual. The challenges that are associated with managing server virtualization are discussed below. In addition, as is also discussed below, managing application performance in a cloud computing environment is extremely complex.

The 2010 Webtorials Respondents were asked to indicate the approach their organization takes to management. They were given the following choices and allowed to choose all that applied to their environment.

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

Their responses are summarized in Figure 24.

Figure 24: Focus of Management



The data in Figure 24 indicates that the most frequent approach that IT organizations take to management is to focus on individual technology domains. However:

A significant percentage of IT organizations focus their management activities on the performance of applications and/or services.

The Growing Importance of Application Performance Management

In order to quantify how successful IT organizations are with their growing focus on managing the performance of applications and services, The 2011 Webtorials Respondents were given a set of statements and were asked to indicate which of the statements described their

organization's approach to application performance management (APM). They were allowed to indicate all that applied.

Only about fifteen percent of The 2011 Survey Respondents indicated that their organization currently does a good job of APM. In addition, The 2011 Survey Respondents indicated by a significant margin that the approach that their organization takes to APM is that each technical discipline does its own thing vs. their using an approach that is top down and pretty tightly coordinated.

There is growing discussion in the industry about the best technical approach to implement APM. One approach is to be able to infer application performance based on management data, such as NetFlow, that is routinely collected by the network elements. An alternative approach is to use specialized agents to gather more sophisticated management data. Approximately twelve percent of The 2011 Survey Respondents indicated that their approach to APM makes heavy use of specialized agents to monitor the status of the various components of the application delivery chain.

One conclusion that can be drawn from the data discussed in the preceding two paragraphs is that

> ***APM is a work in progress. By that is meant that in spite of its importance, the vast majority of IT organizations don't do a good job of it.***

The statement that APM is both important and a work in progress is supported by the fact that a third of The 2011 Survey Respondents indicated that it was important to their organization to get better at APM over the next year.

Communications Based Applications

Communications based applications are an important class of application in part because these applications tend to be highly visible and their performance can degrade quickly if they experience impairments such as undo delay, jitter or packet loss. These applications are also important because as explained in the section of this report entitled *The Wide Area Network*, over the next year almost 80% of IT organizations will increase their use of video, and in many cases the increased use of video will be substantial.

Another reason why communications based applications are an important class of applications in general and important relative to cloud computing in particular is that as discussed in the section of this report entitled *The Emergence of Cloud Computing and Cloud Networking*, services such as VoIP and unified communications are now available from a cloud computing service provider (CCSP). As that section of the report also discussed, there is significant interest on the part of IT organizations to acquire both VoIP and unified communications from a CCSP.

To quantify the challenges associated with supporting a range of communications traffic, The 2011 Webtorials Respondents were asked to indicate how important it was over the next year for their IT organization to get better at managing the use of VoIP, traditional video traffic and telepresence. Their answers are summarized in Table 24**.**

Table 24: Importance of Managing the Use of Communications Based Traffic			
	VoIP	**Traditional Video Traffic**	**Telepresence**
Extremely Important	13.4%	6.8%	4.8%
Very Important	33.9%	20.3%	25.6%
Moderately Important	29.9%	29.7%	25.6%
Slightly Important	14.2%	28.0%	24.8%
Not at all Important	8.7%	15.3%	19.2%

The data in Table 24 shows that almost 50% of The Survey respondents indicated that getting better at managing the use of VoIP traffic is either very or extremely important to their IT organization. This is a significant percentage, particularly given that VoIP is not a new application. The challenge of managing VoIP will increase in those situations in which VoIP is acquired from a CCSP. In those instances, the IT organization will have to be able to gather and correlate management data from the CCSP, the network or networks that carry the VoIP traffic and the users' devices.

Internal SLAs

As recently as two or three years ago, few IT organizations offered an SLA to the company's business and functional managers; a.k.a., an internal SLA. However, that situation has changed and now it is common for IT organizations to offer internal SLAs. To understand the prevalence and effectiveness of internal SLAs, The 2010 Webtorials Respondents were asked to indicate their agreement or disagreement with three statements. The three statements and the percentage of The Webtorials Respondents that agreed with the statement are shown in Table 25.

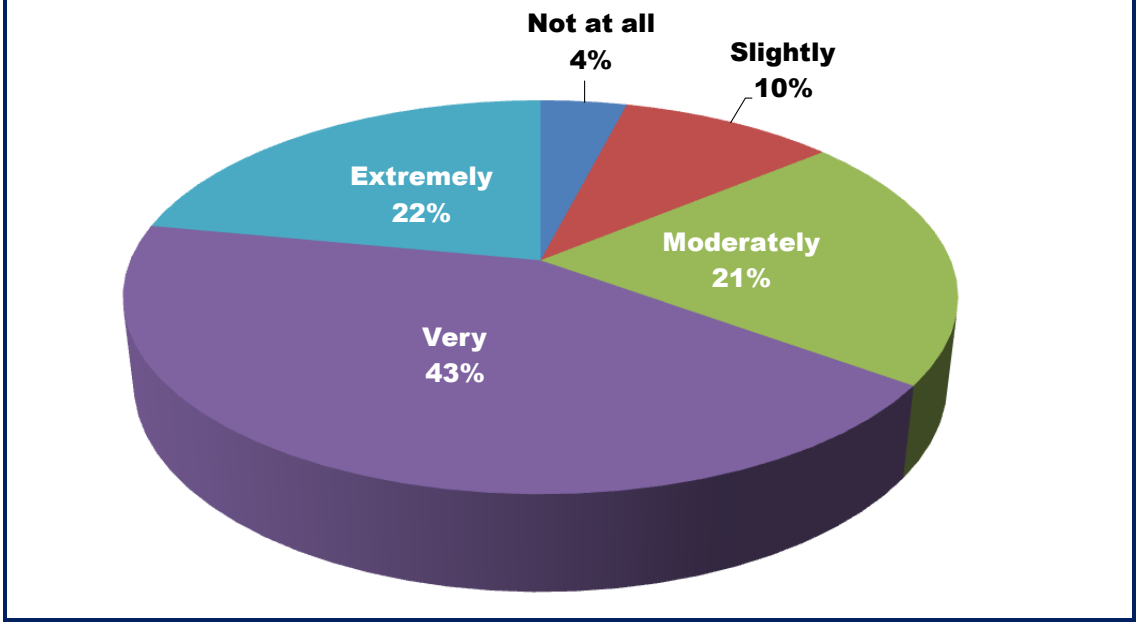
Table 25: Status of Internal SLAs	
Statement	**Percentage**
We provide an SLA internally for every application that we support	30.0%
We provide an SLA internally for at least some applications	69.9%
We do a good job of managing our internal SLAs	55.8%

The data in Table 25 highlights the growing interest that IT organizations have in providing internal SLAs for at least some applications.

The vast majority of IT organizations provide an internal SLA for at least some applications.

One of the answers to The Question was managing internal SLAs for one or more business-critical applications. The responses of The 2011 Webtorials Respondents are summarized in Figure 25.

Figure 25: The Importance of Getting Better at Managing Internal SLAs



The data in Figure 25 leads to two related conclusions. One conclusion is that

Two thirds of IT organizations believe that it is either very or extremely important to get better at effectively managing internal SLAs.

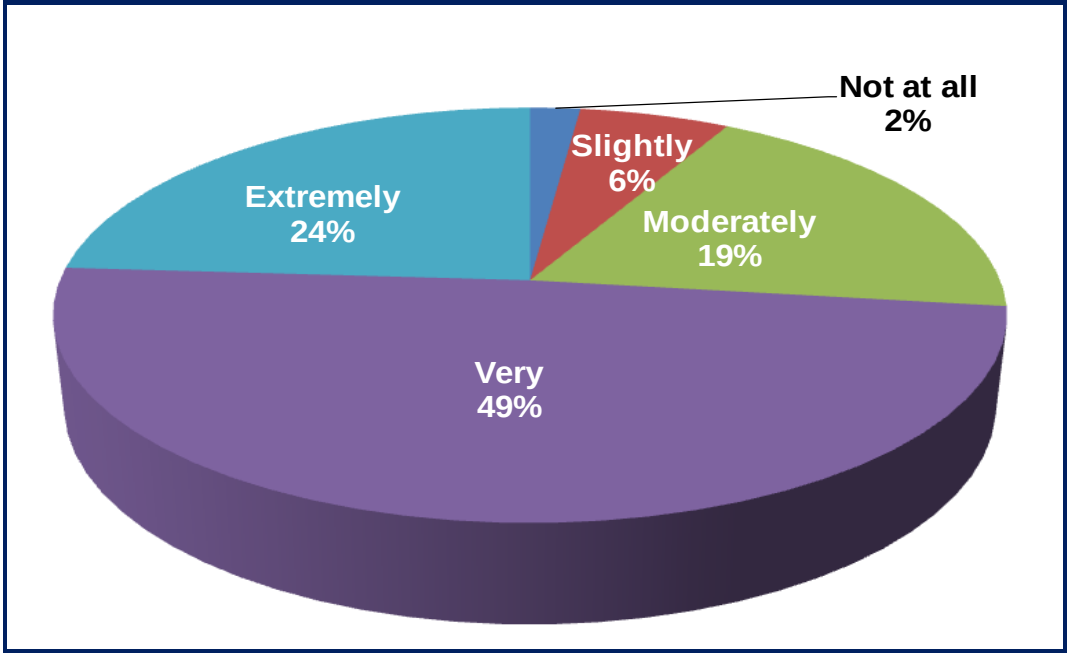
A somewhat more subtle conclusion is that managing internal SLAs is difficult or else the majority of IT organizations would already be doing a good job of managing these SLAs and hence would not be striving to get better at the task. Unfortunately, the movement to utilize public cloud computing services greatly increases the difficulty associated with managing an internal SLA. That follows in part because of the difficulty of gathering all of the management data on an end-to-end basis that is necessary to effectively monitor an SLA. It also follows because as pointed out in the section of this report entitled *The Emergence of Cloud Computing and Cloud Networking*, it is common for CCSPs to deliver their services over the Internet and no vendor will provide an end-to-end performance guarantee for services and applications that are delivered over the Internet.

The lack of meaningful SLAs for public cloud services is a deterrent to the Global 2000 adopting these services for delay-sensitive, business-critical applications.

Root Cause Analysis

The 2011 Webtorials Respondents were asked how important it was over the next year for their organization to get better at rapidly identifying the causes of application degradation. Their responses are shown in Figure 26.

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



Comparing the answers that The 2011 Webtorials Respondents gave to the this management task to the other nineteen management tasks shows that:

Getting better at doing root cause analysis is the most important management task facing the vast majority of IT organizations.

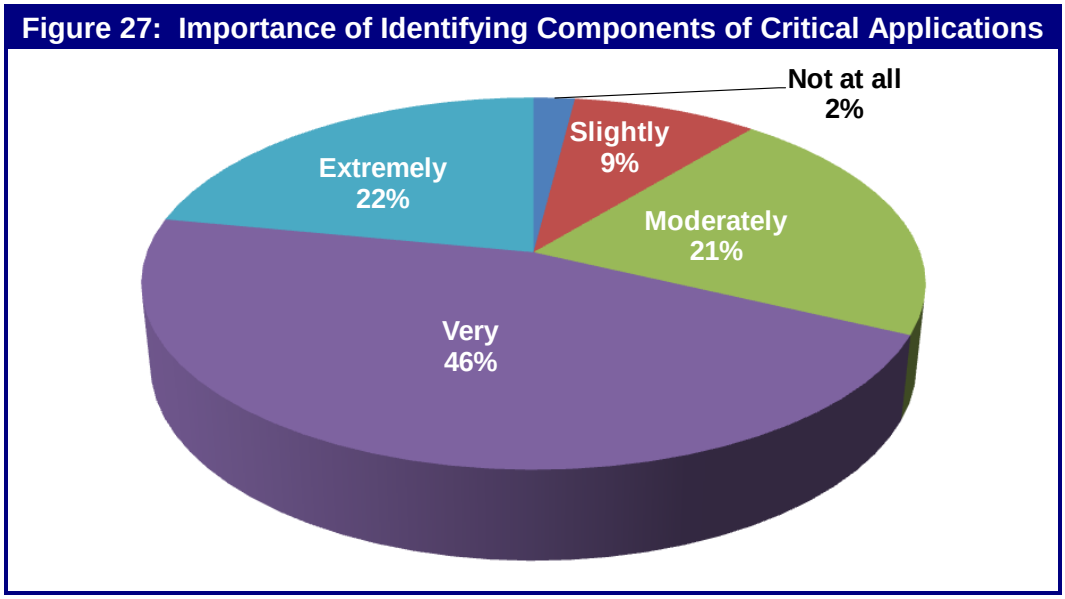
It is not surprising that rapidly identifying the root cause of degraded application performance is so important to IT organizations in part because on an ever increasing basis a company's key business processes rely on a handful of applications. That means that if those applications are not running well, neither are those key business processes.

A prerequisite to being able to perform effective root cause analysis is the automatic discovery of all the elements in the IT infrastructure that support each service or application. If IT organizations can effectively identify which components of the infrastructure support a particular application or service, monitoring can much more easily identify when services are about to degrade due to problems in the infrastructure. As part of this approach, predictive techniques such as heuristic-based trending of software issues and infrastructure key performance indicators can be employed to identify and alert management of problems before they impact end users. In addition, outages and other incidents that generate alerts can be prioritized based on their potential business impact. Prioritization can be based on a number of factors including the affected business process and its value to the enterprise, the identity and number of users affected and the severity of the issue.

Once the components of the infrastructure that support a given application or service has been identified, triage and root cause analysis can be applied at both the application and the infrastructure levels. When applied directly to applications, triage and root cause analysis can identify application issues such as the depletion of threads and pooled resources, memory leaks or internal failures within a Java server or .NET server. At the infrastructure level, root cause analysis can determine the subsystem within the component that is causing the problem.

The 2011 Webtorials Respondents were asked how important it was over the next year for their organization to get better at identifying the components of the IT infrastructure that support the company's critical business applications. Their responses are shown in Figure 27.

Figure 27: Importance of Identifying Components of Critical Applications



A clear observation that can be drawn from Figure 27 is that

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

Server Virtualization

As discussed in the section of this report entitled *The Emergence of Cloud Computing and Cloud Networking*, there isn't a universally accepted definition of what is meant by cloud computing. That section of the report included a number of characteristics of a cloud computing solution, but also pointed out that there is not a litmus test to determine if a particular service is indeed a cloud computing service based on how many of the characteristics it supports. That said, the vast majority of private, public and hybrid cloud computing solutions are based at least in part on server virtualization. Hence, the management challenges that are associated with server virtualization can reasonably be regarded as management challenges for cloud computing.

As pointed out in Virtualization: Benefits, Challenges and Solutions, server virtualization creates a number of management challenges. For example, the need to manually reconfigure the network to support VM migration that was discussed in the section of the report entitled ***The Emerging Data Center LAN*** can be regarded as either a LAN challenge or a management challenge. Additional management challenges that are associated with server virtualization include:

Breakdown of Network Design and Management Tools

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

Limited VM-to-VM Traffic Visibility

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

Poor Management Scalability

Many IT organizations have experienced VM proliferation sometimes called VM sprawl. In addition, the normal best practices for virtual server configuration call for creating separate VLANs for the different types of traffic to and from the VMs. The combined proliferation of VMs and VLANs places a significant strain on the manual processes that are traditionally used to manage servers and the supporting infrastructure.

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.

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 that are offered by physical access switches. This situation leads to the implementation of inconsistent end-to-end network policies.

Multiple Hypervisors

It is becoming common to find IT organizations using multiple hypervisors, each of which comes with their own management system and their own management interface. In addition, the management functionality provided by each hypervisor varies as does the degree to which each hypervisor management system is integrated with other management systems.

Management on a per-VM Basis

IT organizations typically perform management tasks such as discovery, capacity planning and troubleshooting on a per server basis. While that is still required, IT organizations must also perform those tasks on a per-VM basis.

In order to quantify the interest that IT organizations have in responding to the management challenges that are created by server virtualization, three of the twenty tasks that were included in The Question were:

- Manage the traffic that goes between virtual machines (VMs) on a single physical server.
- Support the movement of VMs between servers in different data centers.
- Perform traditional management tasks such as troubleshooting and performance management on a per VM basis.

The responses of The 2011 Webtorials Respondents are summarized in Table 26.

Table 26: Importance of Managing Server Virtualization

	Traffic Between VMs	Move VMs Between Servers	Manage on a per VM Basis
Extremely	7.3%	15.4%	12.9%
Very	29.0%	32.5%	37.9%
Moderately	29.8%	20.5%	29.8%
Slightly	17.7%	18.8%	16.1%
Not at All	16.1%	12.8%	3.2%

One conclusion that can be drawn from the data in Table 26 is that managing the traffic that goes between VMs on a single physical server is not a very important task for the majority of IT organizations. Another conclusion is that

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.

Management Challenges Associated with Cloud Computing

Even in the traditional IT environment[46] when the performance of an application is degrading the degradation is typically noticed first by the end user and not by the IT organization. In addition, when IT is made aware of the fact that application performance has degraded, the process to identify the source of the degradation can be lengthy.

Unfortunately:

The adoption of cloud computing makes troubleshooting application performance an order of magnitude more difficult than it is in a traditional environment.

One of the challenges associated with managing in any environment is that it is difficult to know the end-to-end path that packets take across a network. This management complexity comes in part from the distributed nature of IP. In particular, routers exchange reachability information with each other via a routing protocol such as OSPF (Open Shortest Path First). Based on this information, each router makes its own decision about how to forward a packet. There is, however, no single repository of routing information in the network. This lack of knowledge complicates tasks such as troubleshooting. The difficulty of knowing the path from origin to destination is greatly increased in a cloud computing environment because applications and services can be dynamically moved between servers both within and between data centers.

One of the fundamental issues relative to managing in a cloud computing environment is that the network topology becomes even more complex and hence understanding the end-to-end path becomes even more difficult.

In order to illustrate some of the other challenges of managing a cloud computing environment, assume that a hypothetical company called SmartCompany has started down the path of implementing private cloud computing by virtualizing their data center servers. Further assume that one of SmartCompany's most important applications is called BusApp and that the users of the application complain of sporadic poor performance and that BusApp is implemented in a manner such that the web server, the application server and the database server are each running on VMs on separate physical servers which have been virtualized using different hypervisors.

In order to manage BusApp in the type of virtualized environment described above, an IT organization needs detailed information on each of the three VMs that support the application and the communications amongst 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 BusApp 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.

Part of the challenge associated with troubleshooting this scenario is that as previously noted, in most cases once an IT organization has virtualized its servers it looses insight into the inter-VM

[46] This refers to an IT environment prior to the current wave of virtualization and cloud computing.

traffic that occurs within a physical server. Another part of the challenge is that as was also previously noted, each of the hypervisors comes with their own management system.

Staying with this example, now assume that SmartCompany has decided to evaluate the viability of deploying BusApp using either a public or hybrid cloud computing solution. For the sake of this example, consider two alternative approaches that SmartCompany might implement. Those approaches are:

1. Public Cloud Computing

   SmartCompany acquires BusApp functionality from a SaaS provider. The employees of SmartCompany that work in branch and regional offices use an MPLS service from a network service provider (NSP) to access the application, while home office workers and mobile workers use the Internet.

2. Hybrid Cloud Computing

   SmartCompany hosts the application and data base servers in one of their data centers and the web servers are provided by a cloud computing service provider. All of the users access the web servers over the Internet and the connectivity between the web server layer and the application server layer is provided by an MPLS service.

In order to monitor and manage either deployment, consistent and extensive management data needs to be gathered from the cloud computing service provider(s), the MPLS provider(s) and the provider(s) of Internet access. In the case of the first option (public cloud computing) similar management data also needs to be gathered on the components of the on-site infrastructure that are used by SmartCompany's employees and supported by the IT organization. In the case of the second option (hybrid cloud computing) similar management data also needs to also be gathered on both the on-site infrastructure as well as the web and application servers that are supported by the IT organization. In either case, effective tools are also necessary in order to process all of this data so that IT organizations can identify when the performance of the application is degrading before end users are impacted and can also identify the root cause of that degradation.

Another fundamental issue relative to managing either a public or hybrid cloud computing service is that the service has at least three separate management domains: the enterprise, the WAN service provider(s) and the various cloud computing service providers.

Cloud Management Solutions

The Growing Use of Cloud Networking Services

As pointed out in the section of this report entitled *The Emergence of Cloud Computing and Cloud Networking*, a new class of solutions has begun to be offered by CCSPs. These are solutions that have historically been provided by the IT infrastructure group itself and include management, security, network and application optimization, VoIP, Unified Communications (UC) and virtualized desktops. This new class of solutions is referred to as Cloud Networking Services (CNS). That section of this report also presented the results of a survey in which The 2011 Webtorials Respondents were asked to indicate how likely it was over the next year that their company would acquire specific CNSs. The survey respondents were given nine types of services. Table 27 below highlights the interest that The 2011 Webtorials Respondents have in acquiring three specific CNSs.

Table 27: Interest in Cloud Networking Services

	Will Not Happen	Might Happen	50/50 Chance	Will Likely Happen	Will Happen
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%

One observation that can be drawn from the data in Table 27 is that:

Over the next year, more than a quarter of IT organizations will either likely acquire or will acquire security and/or management functionality from a CCSP.

Security as a Cloud Networking Service

Security is a very broad topic. That said, one of the largest, if not the largest sources of security vulnerabilities is Web based applications. Part of the growing security challenge associated with Web based applications is the continually increasing business use of social media sites such as Facebook and of major Webmail services such as Yahoo. A company could implement a simple acceptable use policy that either allows or denies access to these sites. However, such a policy ignores the fact that these sites typically provide a variety of functions, some of which fall into the acceptable use policies of a growing number of organizations. To deal with the evolving use of multi-faceted social media sites, a security based CNS needs to be able to allow access to a social media site such as Facebook, but block specific activities within the site, such as gaming or posting. Analogously, the CNS needs to have the granular controls to be able to allow users to send and receive mail using Yahoo, but block email attachments.

Another one of the security challenges associated with the use of Web based applications that is rapidly increasing in importance is the growth of malware. To protect against malware, a security based CNS should be able to identify sites that are either suspicious or are known to distribute malware. In order to be effective, a CNS that provides Web content filtering or

malware protection needs a source of intellectual capital that identifies known and suspected vulnerabilities. To be effective, this source needs to be both dynamic and as extensive as possible.

One component of the value proposition of a CNS that provides web filtering and/or malware protection is the standard value proposition of any cloud based service. That value proposition is that a cloud based service has the potential to lower the cost of providing the service, reduce the time it takes to implement the service and give the company that is using the service access to functionality that they couldn't otherwise acquire. Another component of the value proposition of a CNS that provides web filtering and/or malware protection is that

Unlike a traditional security solution that relies on the implementation of a hardware based proxy, a security based CNS can also protect mobile workers.

The security based CNS does this by leveraging functionality that it provides at its cloud data centers as well as functionality in a software agent that is deployed on each mobile device.

In many cases, the best use of a CNS is as part of a hybrid solution. For example, in some cases, the IT organization already has functionality such web filtering or malware protection deployed in CPE at some of their sites. In this case, the IT organization may choose to implement a CNS just to protect the sites that don't have security functionality already implemented and/or to protect the organization's mobile workers. Alternatively, an organization may choose to implement security functionality in CPE at all of their sites and to also utilize a CNS as part of a defense in depth strategy.

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

Management as a Cloud Networking Service

As is the case with security, management is a very broad topic and hence it is possible to find a CNS that provides almost any possible form of management capability. For example, the preceding subsection discussed how a security based CNS could support mobile employees. In a similar fashion, a management based CNS can add value by helping IT organization to manage the burgeoning deployment of mobile devices.

One class of management based CNS is focused on managing specific types of devices, such as branch office routers, WiFi access points, mobile devices or security devices. In some cases, the CNS supports customer-owned CPE from a wide range of vendors. In other cases, the CNS could be bundled with CCSP-owned devices located at the customer's premise. A variation on the latter approach involves a CNS vendor that provides devices, such as branch office routers, that have been specifically designed to be centrally managed from the cloud via a web portal. In this case, the vendor can move the device's control plane into the cloud in a manner analogous to the separation of control plane and data plane provided by OpenFlow, as discussed in the section of this report entitled *The Emerging Data Center LAN*.

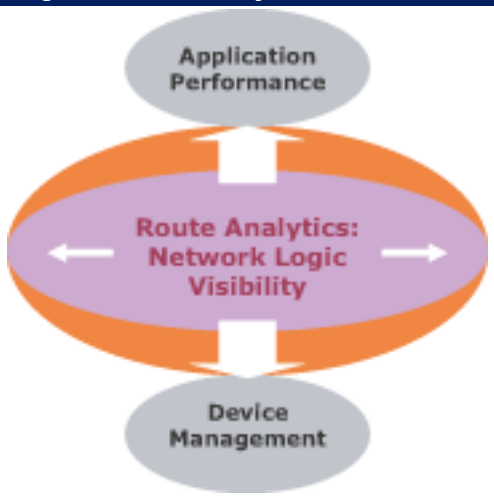
A second class of management based CNS is focused on managing other CNS services provided by a CCSP. These services typically are aimed at addressing the weaknesses in management capability generally associated with early CCSP provided services. For example, the initial wave of CCSP services came with little if any commitment on the part of the service provider relative to an SLA. One example of this class of management based service is a CNS that provides an enhanced level of management for a VoIP service that an IT organization acquires from a CCSP.

Route Analytics

As was previously mentioned, due to the distributed nature of IP it is sometimes difficult to know the end-to-end path that packets take across a network. While that is a challenge in any IT environment, it is a particularly difficult challenge in a cloud computing environment due to the dynamic nature of creating and moving virtual machines.

As shown in Figure 28, route analytics provides IT organizations and service providers with insight into the routing layer.

Figure 28: The Positioning of Route Analytics



The value proposition of route analytics is that

Route analytics provides visibility, analysis, and diagnosis of the issues that occur at the routing layer in complex, meshed networks.

A route analytics appliance draws its primary data directly from the network in real time by participating in the IP routing protocol exchanges. This allows the route analytics device to compute a real-time Layer 3 topology of the end-to-end network, detect routing events in real time and correlate routing events or topology changes with other information, including application performance metrics. As a result, route analytics can help both IT organizations and service providers determine the impact on performance of both planned and actual changes in the Layer 3 network.

Dynamic Infrastructure Management

A traditional environment can benefit from implementing dynamic infrastructure management. However, due to the challenges that are associated with cloud computing:

A dynamic virtualized environment can benefit greatly from a highly scalable and integrated DNS/DHCP/IPAM solution, which is also well integrated with the virtual server management system.

Where DNS/DHCP/IPAM share a common database, the integration obviates the need to coordinate records in different locations and allows these core services to accommodate any different addressing and naming requirements of physical and virtual servers. Potential advantages of this approach include the automated generation of IP addresses for newly created VMs, the automated allocation of subnets for new VLANs, and the population of an IP address database with detailed information about the current location and security profiles of VMs. The integration of infrastructure utilities with the virtual server management system can also facilitate automated changes to the DHCP and DNS databases.

Virtualized Performance and Fault Management

In a traditional IT environment it is common to implement adaptive performance thresholding solutions that can identify systemic deviations from normal patterns of behaviour as well as time over threshold violations and can also automatically update thresholds based on changes to historic levels of utilization. That same capability is needed in a virtualized environment so that IT organizations can monitor the performance of individual VMs.

Virtual switches currently being introduced into the market can export traffic flow data to external collectors in order to provide some visibility into the network flows between and among the VMs in the same physical machine. Performance management products are currently beginning to leverage this capability by collecting and analysing intra-VM traffic data. Another approach to monitoring and troubleshooting intra-VM traffic is to deploy a virtual performance management appliance or probe within the virtualized server. This approach has the advantage of potentially extending the fault and performance management solution from the physical network into the virtual network by capturing VM traffic at the packet level, as well as at the flow level.

While changes in the virtual topology can be gleaned from flow analysis, a third approach to managing a virtualized server is to access the data in the server's management system. Gathering data from this source can also provide IT organizations with access to additional performance information for specific VMs, such as CPU utilization and memory utilization.

Management Solutions Packaged with Converged Infrastructure

An increasingly popular approach to building cloud data centers is based on pre-integrated and certified infrastructure packages from either a broadly-based IT equipment vendor, a group of partners or a joint venture formed by a group of complementary vendors. These packages typically are offered as turn-key solutions and include compute, server virtualization, storage, network, and management capabilities. Other data center functions such as WOCs, ADCs, APM and security functionality may also be included.

One of the primary reasons why IT organizations implement a converged IT infrastructure is to reduce the overall complexity of a pervasively virtualized infrastructure. The reduction in complexity makes it feasible for IT organizations to fully capitalize on the virtualized infrastructure's inherent potential to serve as an agile, demand-driven platform that can deliver dynamic IT services with unprecedented levels of control, security and compliance, reliability, and efficiency. In order to realize the full potential of the converged IT infrastructure, the management system must provide a unified, cross-domain approach to automated element management, provisioning, change management and operations management. Some of the most critical aspects of managing a cloud data center include:

- **Integrated and Automated Infrastructure and Service Management:** Integrated management reduces the number of management interfaces that are involved in implementing administrative workflows. Automation allows services to be dynamically provisioned, modified or scaled without requiring time-consuming manual configuration across the various technology domains of the data center; e.g., compute, network, storage and security. The management suite should also include application and service level management capabilities that will support end-to-end SLAs. From an operational management perspective, the management system should provide additional capabilities, such as cross-domain root cause analysis and service impact analysis, in order to support the highest levels of service reliability.

- **Secure Multi-tenancy:** A robust multi-layer security architecture is required to ensure confidentiality and integrity of the services and the subscriber's data, particularly in a multi-tenant environment.

- **Support for Enterprise Co-Management:** The service management system should provide a web portal supporting the self-service provisioning of new services or the scaling of existing services. The portal should also include dashboards that provide real-time visibility of application and service performance as well as the consumption of on-demand services. The service management system should also facilitate turning off resources such as VMs that are acquired from a CCSP when they are not needed so that the company using the resources does not incur unnecessary expenses.

- **Compatibility with Enterprise Cloud Implementations:** The efficiency of hybrid clouds is optimized where there is a high degree of consistency across the private and public portions of the solution in terms of the cloud management systems, the hypervisors and the hypervisors' management systems. This consistency facilitates the movement of VMs between enterprise data centers and service provider data centers, and this movement also enables the dynamic reallocation of cloud resources.

Management systems for a converged infrastructure typically support APIs for integration with other management systems that may be currently deployed in order to manage the end-to-end data center. These APIs can provide integration with enterprise management systems, automated service provisioning systems, fault and performance management systems and orchestration engines.

While IT departments or CCSPs can themselves achieve some degree of cross-domain management integration by leveraging available element manager plug-ins and APIs, ad hoc automation and integration across the end-to-end infrastructure is quite time-consuming and involves considerable specialized programming expertise. Therefore, the completeness and

effectiveness of pre-integrated management functionality are likely to be two of the key differentiators among converged infrastructure solutions.

Cross-domain integrated management of the converged infrastructure will bring added benefits in those situations in which a single administrator has the authority to initiate and complete cross-domain tasks, such as provisioning and modifying infrastructure services. The use of a single administrator can eliminate the considerable delays that are typical in a traditional management environment in which the originating administrator must request other administrators in the other domains to synchronize the configuration of elements within their domains of responsibility. However, a well-known cliché describes the difficulty of realizing these benefits.

Culture eats strategy for breakfast.

That cliché refers to the fact that in many cases the culture of an IT organization resists any changes that involve changing the roles of the members of the organization. Exacerbating the challenge of the IT organization's resistance to change is the fact that, as was pointed out in the section of this report entitled *The Emergence of Cloud Computing and Cloud Networking*, the culture of an IT organization typically changes very slowly.

Orchestration and Provisioning

Service orchestration is an operational technique that helps IT organizations automate many of the manual tasks that are involved in provisioning and controlling the capacity of dynamic virtualized services. Orchestration engines are available as standalone management products or as part of complete suites of management tools that are focused on the data center. In addition, the management systems that are integrated with converged infrastructure solutions typically include some orchestration capabilities.

By automatically coordinating provisioning and resource reuse across servers, storage, and networks, service orchestration can help IT organizations streamline operational workloads and overcome technology and organizational silos and boundaries. The value proposition of an orchestration engine is that

Orchestration engines use business policies to define a virtual service and to translate that service into the required physical and virtual resources that are needed for deployment.

The orchestration engine then disseminates the needed configuration commands to the appropriate devices across the network in order to initiate the requested service. The orchestration engine can automatically initiate the creation of the required virtual machines while simultaneously deploying the network access and security models across all of the required infrastructure components. This includes routers, switches, security devices and core infrastructure services. The entire process can allow for the setup and deployment of network routes, VPNs, VLANs, ACLs, security certificates, firewall rules and DNS entries without any time consuming manual entries via device-specific management systems or CLIs.

Orchestration engines are available that are pre-configured to interface with certain families of infrastructure devices. Therefore, it is possible to think of the orchestration engine as providing some degree of management integration for non-converged infrastructure. As such, orchestration engines might be a highly desirable approach in those instances in which an existing heterogeneous (i.e., non-converged) data center infrastructure is being transitioned to perform as a cloud data center.

Orchestration solutions would benefit greatly from the emergence of an open standard for the exchange of information among the full range of devices that may be used to construct a dynamic virtual data center. In the Cloud Computing arena there are a number of standards under development, including the Open Cloud Computing Interface (OCCI) from the Open Grid Forum[47]. These standards activities may also provide value within the enterprise virtual data center, since the stated scope of the specification is to encompass “all high level functionality required for the life-cycle management of virtual machines (or workloads) running on virtualization technologies (or containers) supporting service elasticity”.

IF-MAP is another emerging standard proposed by the Trusted Computing Group[48] and implemented by a number of companies in the security and network industries. It is a publish/subscribe protocol that allows hosts to lookup meta-data and to subscribe to service or host-specific event notifications. IF-MAP can enable auto-discovery and self-assembly (or re-assembly) of the network architecture. As such, IF-MAP has the potential to support the

[47] http://www.gridforum.org/
[48] http://www.trustedcomputinggroup.org/

automation and dynamic orchestration of not only security systems, but also other elements of the virtual data center. For example, IF-MAP could facilitate the automation of the processes associated with virtual machine provisioning and deployment by publishing all of the necessary policy and state information to an IF-MAP database that is accessible by all other elements of the extended data center.

Conclusions and Observations

Throughout the 2011 Cloud Networking Report, the following conclusions were drawn and observations were made.

- The phrase cloud networking refers to the LAN, WAN and management functionality that must be in place to enable cloud computing.
- In order to support cloud computing, a cloud network must be dramatically more agile and cost effective than a traditional network is.
- 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.
- On a going forward basis, IT organizations will continue to need to provide the highest levels of availability and performance for a small number of key services. However, an ever-increasing number of services will be provided on a best effort basis.
- SLAs from both traditional network service providers as well as public cloud computing providers are a work in progress.
- The primary factors that are driving the use of public cloud computing solutions are the same factors that drive any form of out-tasking.
- In some cases, the use of a public cloud computing solution reduces risk.
- The 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.
- One of the key challenges facing IT organizations that use SaaS-based applications is improving the performance, management and security of those applications.
- There are significant differences amongst the solutions offered by IaaS providers, especially when it comes to the SLAs they offer.
- The availability of IaaS solutions can vary widely.
- One of the key challenges facing IT organizations that use IaaS-based solutions is improving the performance, management and security of those solutions.
- Cloud balancing can be thought of as the logical extension of global server load balancing (GSLB).
- Cloud Networking Services represents the beginning of what could be a fundamental shift in terms of how IT services are provided.
- One way for an IT organization to evaluate the agility of a CCSP is to identify the degree to which the CCSP has virtualized their infrastructure.

- IT organizations provide considerable value by being the broker between the company's business unit managers and cloud computing service providers.
- The culture of an IT organization changes very slowly.
- One of the key factors driving IT organizations to redesign their data center LANs is the deployment of virtual servers.
- The primary factors driving IT organizations to re-design their data center LAN is the desire to reduce cost and support scalability.
- One approach for improving server-to-server communications is to flatten the network from three tiers to two tiers consisting of access layer and aggregation/core layer switches.
- There is significant desire on the part of IT organizations to flatten their data center LANs, but there is also significant uncertainty relative to how flat they will become in the next two years.
- The current generation of switches has exploited advances in switch fabric technology and merchant silicon switch-on-a-chip integrated circuits (ICs) to dramatically increase port densities.
- The combination of server consolidation and virtualization creates an "all in one basket" phenomenon that drives the need for highly available server configurations and highly available data center LANs.
- With switch virtualization, two or more physical switches are made to appear to other network elements as a single logical switch or virtual switch, with a single control plane.
- The combination of switch virtualization and multi-chassis LAG can be used to create a logically loop-free topology
- In many cases, the best technology doesn't end up being the dominant technology in the marketplace.
- With technologies like TRILL and SPB, the difference between access switches and core switches may shrink significantly.
- There is significant desire on the part of IT organizations to move away from using STP in their data center LANs, but there isn't a consensus as to what the most common replacement technology will be.
- The vSwitch presents a number of concerns related to management, security, functionality and organizational responsibilities.
- A possible characteristic of Third Generation Data Center LANs will be the convergence of block-level storage and data traffic over a common high-speed Ethernet data center switching fabric.
- Fibre Channel over Ethernet (FCoE) is an industry standard that is being developed by the International Committee for Information Technology Standards (INCITS) T11 committee.

- There are several levels of support that data center switch vendors can provide for FCoE.
- The primary drivers of FCoE are the vendors that offer both Ethernet and FC products.
- The majority of IT organizations have not developed concrete, broad-based plans for the evolution of their data center LANs.
- The most common approach to automating the manual processes involved in VM provisioning and migration is based on communication between the Hypervisor Management system and the switch element management system (EMS) via APIs supported by both vendors.
- Today there is not a fundamentally new generation of WAN technology in development.
- The WAN doesn't follow Moore's Law.
- Over the next year, roughly forty percent of IT organizations will increase their WAN budget and in many cases, the increase will be significant.
- IT organizations must either make changes to how they use WAN services, or else accept ongoing increases in their WAN budget due to the increased traffic generated by the use of cloud computing.
- Over the next year, the percentage of IT organizations that have not implemented any desktop virtualization will be cut roughly in half.
- Over the next year almost 80% of IT organizations will increase their use of video, and in many cases the increased use of video will be substantial.
- The primary WAN services used by IT organizations are MPLS and the Internet.
- While IT organizations will increase their reliance on both MPLS and the Internet, they will make a relatively greater increase in their reliance on the Internet.
- The primary concerns that IT organizations have with the use of MPLS are cost, the lead time to implement new circuits and uptime. The primary concerns that IT organizations have with the use of the Internet are uptime, latency and cost.
- In a growing number of instances, Internet-based VPNs that use DSL for access are 'good enough' to be a cloud network.
- The key concept behind an aggregated virtual WAN is that it simultaneously utilizes multiple enterprise WAN services and/or Internet connections in order to optimize reliability and minimize packet loss, latency and jitter.
- Some of the concerns that IT organizations have about the use of the Internet are exacerbated by backhauling Internet traffic to a central site.
- Over the next year, IT organizations will make an increased use of distributed access to the Internet from their branch offices.

- In somewhat less than half of the instances that business users are accessing public cloud computing services, the primary WAN service that they intend to use is not the traditional Internet.

- In almost three quarters of the instances that business users are accessing private cloud computing services, the primary WAN service that they intend to use is not the traditional Internet.

- The majority of IT organizations don't regard the SLAs that they receive from their network service providers as being effective.

- The majority of IT organizations believe that factors such as the growth in the number of mobile workers and the increase in the use of virtualization and cloud computing will make ensuring acceptable service and application delivery either harder or notably harder.

- One of the compelling advantages of a virtualized appliance is that the acquisition cost of a software-based appliance can be notably less than the cost of a hardware-based appliance with same functionality.

- In many instances the benefits of the dynamic movement of a VM from one server to another are maximized if the supporting infrastructure is virtualized and can also be dynamically moved.

- IT organizations have a significant interest in placing a WOC on premise at an IaaS provider's data centers.

- Between a quarter and a third of IT organizations don't know how they will optimize the performance of services that they acquire from an IaaS or a SaaS provider.

- In many situations, a dual ISP-based Internet VPN with PBR can deliver a level of CoS and reliability that is comparable to that of MPLS at a significantly reduced price.

- VPLS represents the combination of Ethernet and MPLS.

- The majority of IT organizations believe that getting better at managing private cloud computing solutions is either very or extremely important.

- The majority of IT organizations believe that getting better at managing inter-related applications that comprise a business service is either very or extremely important.

- A significant percentage of IT organizations focus their management activities on the performance of applications and/or services.

- APM is a work in progress. By that is meant that in spite of its importance, the vast majority of IT organizations don't do a good job of it.

- The vast majority of IT organizations provide an internal SLA for at least some applications.

- Two thirds of IT organizations believe that it is either very or extremely important to get better at effectively managing internal SLAs.

- The lack of meaningful SLAs for public cloud services is a deterrent to the Global 2000 adopting these services for delay-sensitive, business-critical applications.

- Getting better at doing root cause analysis is the most important management task facing the vast majority of IT organizations.

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

- 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.

- The adoption of cloud computing makes troubleshooting application performance an order of magnitude more difficult than it is in a traditional environment.

- One of the fundamental issues relative to managing in a cloud computing environment is that the network topology becomes even more complex and hence understanding the end-to-end path becomes even more difficult.

- Another fundamental issue relative to managing either a public or hybrid cloud computing service is that the service has at least three separate management domains: the enterprise, the WAN service provider(s) and the various cloud computing service providers.

- Over the next year, more than a quarter of IT organizations will either likely acquire or will acquire security and/or management functionality from a CCSP.

- Unlike a traditional security solution that relies on the implementation of a hardware based proxy, a security based CNS can also protect mobile workers.

- Route analytics provides visibility, analysis, and diagnosis of the issues that occur at the routing layer in complex, meshed networks.

- A dynamic virtualized environment can benefit greatly from a highly scalable and integrated DNS/DHCP/IPAM solution, which is also well integrated with the virtual server management system.

- Culture eats strategy for breakfast.

- Orchestration engines use business policies to define a virtual service and to translate that service into the required physical and virtual resources that are needed for deployment.

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.

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.

nano|engine

Full application control at 10% of the cost

www.ipanematech.com

A unique technology that breaks the price/performance barrier to guarantee business application performance in branch offices

- For the first time it is possible to guarantee application performance with a device compatible with branch office constraints;
- The nano|engines fully integrate with the other components of Ipanema's ANS solution;
- Plug-and-Play devices, nano|engines are managed under SALSA;
- Real-time changes in network performance and each user's behavior are taken into account in real-time.

Algorithms embedded in the nano|engine automatically adapt to real-time changes as they happen on the network:

- Traffic from private data centers mixed with traffic from external public clouds;
- Hybrid networks combining MPLS and Internet;
- Unified Communications branch-to-branch flows;
- Virtual desktops and rich media delivery…

The nano|engine's ability to guarantee application performance at the branch maximizes productivity, prevents brownouts and protects the business.

Ultra compact **nano|engine** appliances are tailored for providing full application control with unmatched performance/price ratio in broadband branch offices.

The **nano|engine** devices target broadband branch offices and provide:

- Application aware, **per connection Control and dynamic QoS** for public and private application flows to guarantee an excellent and stable Quality of Experience to each user;
- **End-to-end visibility** of application performance of each flow with comprehensive KPIs and application quality scores;
- **Dynamic WAN path selection** among up to 3 networks for optimized control of multi-attached branches, local Internet breakouts and hybrid networks.

Self-managed, nano|engines are installed at the edge locations of the WAN, typically between the CPE router and branch office LAN. Fully "Plug and Play," nano|engines require no on-site configuration. They operate under control of the central management software, SALSA. Customers simply need to plug the nano in, and configuration and provisioning are managed by SALSA.

The nano|engine family fits particularly well in B to C sectors like retail, finance and hospitality, where slow response times to access customer data or delays in processing an order lead to customer dissatisfaction and loss of productivity. Nano|engines' ability to guarantee application performance prevents any brownouts and protects the business.

The nano|2 addresses branch offices with up to 20 users and 4 Mbps while the nano|5 targets branch offices with up to 50 users and 20 Mbps.