The 2013
Application & Service Delivery Handbook

Part 1: Introduction and Challenges

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

Sponsored in part by:

Produced by:

Executive Summary

The ***2013 Application and Service Delivery Handbook*** will be published both in its entirety and in a serial fashion. This is the first of the serial publications. One goal of this publication is to describe how the ***2013 Application and Service Delivery Handbook*** differs from previous editions in this series. Another goal of this publication is to describe how a variety of factors, such the increasingly mobile work force and the continuing adoption of virtualization and cloud computing, are complicating the task of ensuring acceptable application and service delivery.

Subsequent publications of the ***2013 Application and Service Delivery Handbook*** will focus on describing the technologies, products and services that are available to improve:

- The performance of applications and services.
- The management and security of applications and services.

The fourth and final publication will include an executive summary as well as a copy of the complete document.

Introduction

Background and Goals of the *2013 Application and Service Delivery Handbook*

Throughout the ***2013 Application and Service Delivery Handbook***, the phrase ***ensuring acceptable application and service delivery*** will refer to ensuring that the applications and services that an enterprise uses:

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

There is a growing relationship between the requirements listed above. For example, in order to implement an appropriate level of security, an IT organization may implement encryption. However, the fact that the information flow is encrypted may preclude the IT organization from implementing the optimization techniques that are required to ensure acceptable performance.

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

At the same time that many IT organizations are still in the process of implementing solutions that respond to the first generation of application delivery challenges such as supporting chatty protocols or transmitting large files between a branch office and a data center, a second generation of challenges is emerging. These challenges are driven in large part by the:

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

The goal of the 2013 Application and Service Delivery Handbook is to help IT organizations ensure acceptable application and/or service delivery when faced with both the first generation, as well as the emerging second generation of application and service delivery challenges.

Foreword to the 2013 Edition

While this year's edition of the application delivery handbook builds on the previous edition of the handbook, every section of the 2012 edition of the handbook was modified before being included in this document. For example, on the assumption that a number of the concepts that were described in previous editions of the handbook are by now relatively well understood, the description of those concepts was made more succinct in this year's handbook. To compensate for those changes, the 2012 Application and Service Delivery Handbook is still accessible at Webtorials[1].

In early 2013, two surveys were given to the subscribers of Webtorials. Throughout this document, the IT professionals who responded to the surveys will be referred to as ***The Survey Respondents***. One of the surveys asked a broad set of questions relative to application delivery. The other survey focused on identifying the optimization and management tasks that are of most interest to IT organizations. With that later goal in mind, The Survey Respondents were given a set of twenty optimization tasks and twenty management tasks and asked to indicate how important it was to their IT organization to get better at these tasks over the next year. The Survey Respondents were given the following five-point scale:

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

The answers to all of surveys will be used throughout the ***2013 Application and Service Delivery Handbook*** to demonstrate both the challenges facing IT organizations as well as the relative importance that IT organizations place on a wide variety of optimization and management tasks.

[1] http://www.webtorials.com/content/2012/08/2012-application-service-delivery-handbook-2.html

The Importance of Ensuring Successful Application and Service Delivery

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

Table 1: Impact of Poor Application Performance	
Impact	**Percentage**
The Company Loses Revenue	**62.0%**
IT Teams Are Pulled Together	**59.8%**
Company Loses Customers	**45.1%**
CIO Gets Pressure from his/her Boss	**45.1%**
Harder for IT to get Funding	**44.6%**
CIO Gets Other Pressure	**42.9%**

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

In addition to the fact that the success of a company's key business processes depends on the performance of a wide variety of applications and the networks that support them, another reason why application and service delivery continues to be an important topic for IT organizations is the fact that approximately sixty five percent of The Survey Respondents indicated that when one of their company's key applications begins to degrade, that the degradation is typically noticed first by the end user and not by the IT organization.

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

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

To illustrate the importance that IT organizations place on improving application performance The Survey Respondents were asked how important it was over the next year for their IT organization to get better at optimizing the performance of a key set of applications that are critical to the success of the business. Their answers are shown **Table 2**.

Table 2: Importance of Optimizing Business Critical Applications	
	Percentage
Extremely Important	**21%**
Very Important	**51%**
Moderately Important	**18%**
Slightly Important	**7%**
Not at all Important	**3%**

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

An example of an application that is time sensitive and important to most businesses is VoIP. Since the first application delivery handbook was published in 2007, a growing percentage of the traffic on the typical enterprise data network is VoIP. To quantify the challenges associated with supporting a range of communications traffic, The Survey Respondents were asked to indicate how important it was over the next year for their IT organization to get better at managing the use of VoIP and they were also asked to indicate the importance of ensuring acceptable performance for VoIP traffic. Their answers are shown in **Table 3**.

Table 3: Importance of Managing and Optimizing VoIP		
	Managing	**Ensuring Acceptable Performance**
Extremely Important	**25%**	**25%**
Very Important	**30%**	**41%**
Moderately Important	**24%**	**20%**
Slightly Important	**13%**	**5%**
Not at all Important	**7%**	**9%**

The data in **Table 3** shows that over half of The Survey respondents indicated that getting better at managing VoIP traffic is either very or extremely important to their IT organization and that two thirds of The Survey Respondents indicated that ensuring acceptable performance for VoIP traffic is either very or extremely important to their IT organization.

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

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

Application delivery is more complex than merely accelerating the performance of all applications.

Successful application delivery requires that IT organizations are able to identify the applications running on the network and are also able to ensure the acceptable performance of the applications relevant to the business while controlling or eliminating applications that are not relevant.

First Generation Application & Service Delivery Challenges

There are a number of fairly well understood challenges that have over the years complicated the task of ensuring acceptable application and service delivery. Those challenges are listed below and are described in detail in two documents: 2012 Application and Service Delivery Handbook[2] and Traditional Application & Service Delivery Challenges[3].

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

[2] http://www.webtorials.com/content/2012/08/2012-application-service-delivery-handbook-2.html
[3] http://www.ashtonmetzler.com/Traditional%20App%20Delivery%20Challenges%20V2.0.pdf

Second Generation Application and Service Delivery Challenges

There are a number of emerging challenges that are beginning to complicate the task of ensuring acceptable application and service delivery. Some of these challenges are technical in nature and some are organizational. One of the emerging organizational challenges results from the fact that because of their awareness of the technology that is available to them in their homes, a growing number of business and functional managers have increased expectations of the IT organization. As a result, IT organizations are under more pressure for agility than they ever have been in the past. One of the emerging technical challenges results from the adoption of application architectures such as SOA, Web 2.0 and Rich Internet Applications. These application architectures tend to be more susceptible to performance problems due to WAN impairments than do traditional application architectures. In addition, the introduction of technologies such as AJAX creates significant security vulnerabilities.

Many of the second generation application and service delivery challenges, such as the ones described in the preceding paragraph, are described in 2012 Application and Service Delivery Handbook. ***The 2013 Application and Service Delivery Handbook*** will focus on three key second generation challenges:

- Mobility and BYOD
- Virtualization
- Cloud Computing

Mobility and BYOD

One of the traditional (a.k.a., first generation) application delivery challenges was the fact that many employees who had at one time worked in a headquarters facility now work someplace else; i.e., a regional, branch or home office. The logical extension of that challenge is that most IT organizations now have to support a work force that is increasingly mobile.

There are a number of concerns relative to supporting mobile workers. One such concern is that up through 2010, the most common device used by a mobile worker was a PC. In 2011, however, more tablets and smartphones shipped than PCs[4]. Related to the dramatic shift in the number and types of mobile devices that are being shipped, many companies have adopted the BYOD (Bring Your Own Device to work) concept whereby employees use their own devices to access applications.

In order to quantify the impact of mobility, The Survey Respondents were asked a couple of question. One question was: "In some cases employees of a company access business related data and applications by using a mobile device within a company facility and, in some cases, employees access business related data and applications by using a mobile device when they are at an external site. In the typical day, what percentage of your organization's employees use a mobile device at some time during the day to access business related data and applications, either from within a company facility or from an external site?" Their responses are show in **Table 4**.

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

Table 4: Amount of Mobile Access							
	0%	1% to 9.99%	10% to 24.995	25% to 49.99%	50% to 74.99%	75% to 99.99%	100%
Company Facility	6%	14%	26%	19%	22%	10%	4%
External Site	2%	23%	20%	20%	14%	15%	6%

The vast majority of employees require mobile access for at least part of their typical day.

The Survey Respondents were also asked to indicate the types of employee owned devices that their organization allows to connect to their branch office networks and which of these devices is actively supported, Their responses are shown in **Table 5.**

Table 5: Support for Employee Owned Devices			
	Not Allowed	Allowed but not Supported	Allowed and Supported
Company managed, employee owned laptop	22%	24%	54%
Employee owned and managed laptop	38%	38%	25%
Blackberry	17%	24%	58%
Apple iPhone	14%	30%	55%
Android phone	19%	33%	48%
Windows mobile phone	26%	40%	34%
Apple iPad	18%	40%	52%
Android based tablet	28%	37%	35%
Windows based tablet	28%	36%	37%

The data in **Table 5** indicates that there is wide acceptance BYOD. As a result, the typical branch office network now contains three types of end user devices that are all accessing business critical applications and services. This includes PCs as well as the new generation of mobile devices; i.e., smartphones and tablet computers. Because of their small size, this new generation of mobile devices doesn't typically have wired Ethernet ports and so they are typically connected via what is hopefully a secure WiFi network in the branch office.

This new generation of mobile devices, however, doesn't run the Windows O/S and the existing security and management services for PCs must be extended for mobile devices or alternatively, additional products and/or services added to perform these functions. Similar to PCs, smartphone and tablet computers are subject to malware and network intrusion attacks. On PCs, there are mature, robust products for malware protection (e.g. anti-virus software) and network intrusion protection (e.g., personal firewall), but these protections are just now

emerging for smartphones and tablet computers[5]. Similarly, inventorying and updating installed software on smartphone and tablet computers are emerging capabilities and a critical area for Mobile Device Management solutions.

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

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

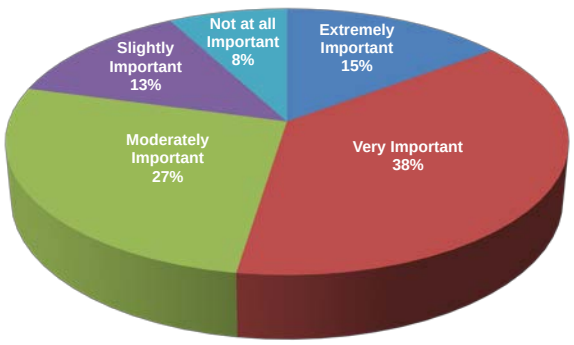
Adopting BYOD increases a company's vulnerability to security breaches.

Another key concern relative to supporting mobile workers is how the applications that these workers access have changed. At one time, mobile workers tended to primarily access either recreational applications or applications that are not delay sensitive; e.g., email. However, in the current environment mobile workers also need to access a wide range of business critical applications, many of which are delay sensitive. This shift in the applications accessed by mobile workers was highlighted by SAP's announcement[6] that it will leverage its Sybase acquisition to offer access to its business applications to mobile workers. One of the issues associated with supporting mobile workers' access to delay sensitive, business critical applications is that because of the way that TCP functions, even the small amount of packet loss that is often associated with wireless networks results in a dramatic reduction in throughput.

In order to quantify the concern amongst IT organizations about ensuring acceptable application and service delivery to mobile workers, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at improving the performance of applications used by mobile workers. Their responses are shown in **Figure 1**.

One conclusion that can be drawn from the data in **Figure 1** is that roughly half of all IT organizations consider it to be either extremely or very important to get better at improving the performance of applications used by mobile workers.

Figure 1: Importance of Optimizing Mobile Applications



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

6 Wall Street Journal, May 17, 2012, page B7

Virtualization

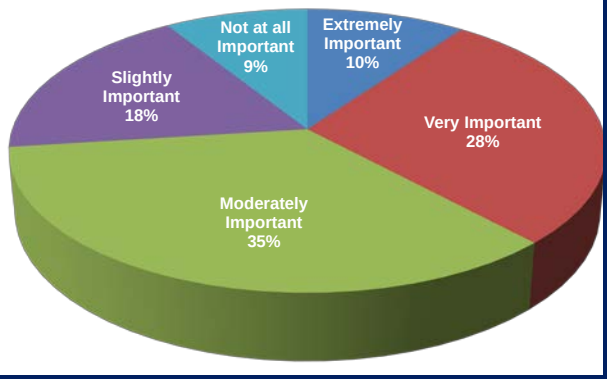
Server Virtualization

Interest in Server Virtualization

The vast majority of organizations have made at least some deployment of server virtualization and that the deployment of server virtualization will increase over the next year.

One of the challenges that is introduced by the deployment of virtualized servers is that, due to the limitations of vSwitches once a server has been virtualized, IT organizations often lose visibility into the inter-VM traffic. This limits the IT organization's ability to perform functions such as security filtering, performance monitoring and troubleshooting. To quantify the impact of losing visibility into the inter-VM traffic, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at managing the traffic that goes between virtual machines on a single physical server. Their responses are shown in **Figure 2**.

Figure 2: Manage the Inter-VM Traffic

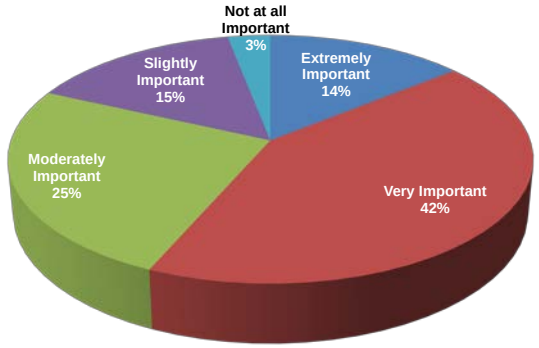


The data in **Figure 2** indicates that, while there is significant interest in getting better at managing inter-VM traffic, the level of interest is less than the level of interest that The Survey Respondents indicated for many other management tasks

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

To quantify the impact that managing on a per-VM basis is having on IT organizations, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at performing traditional management tasks such as troubleshooting and performance management on a per-VM basis. Their responses are shown in **Figure 3**.

Figure 3: Managing on a per-VM Basis



One observation that can be drawn from the data in **Figure 3** is that unlike the situation with managing inter-VM traffic:

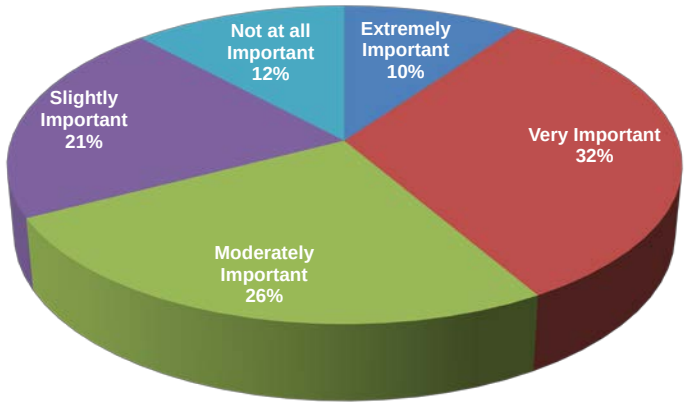
Over half of the IT organizations consider it to be either very or extremely important over the next year for them to get better performing management tasks such as troubleshooting on a per-VM basis.

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

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

In order to quantify the concern that IT organization have with the mobility of VMs, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at supporting the movement of VMs between servers in different data centers. Their responses are shown in **Figure 4**.

Figure 4: Supporting the Mobility of VMs



Combining the insight provided by the data in **Figure 4** with the fact that the use of server virtualization will increase:

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

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

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

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

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

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

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

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

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

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

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

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

Desktop Virtualization

Background

The two fundamental forms of desktop virtualization are:

- Server-side virtualization
- Client-side virtualization

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

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

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

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

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

While there are advantages to both forms of desktop virtualization:

The vast majority of virtualized desktops will utilize server side virtualization.

Challenges of Desktop Virtualization

Desktop virtualization can provide significant benefits. However:

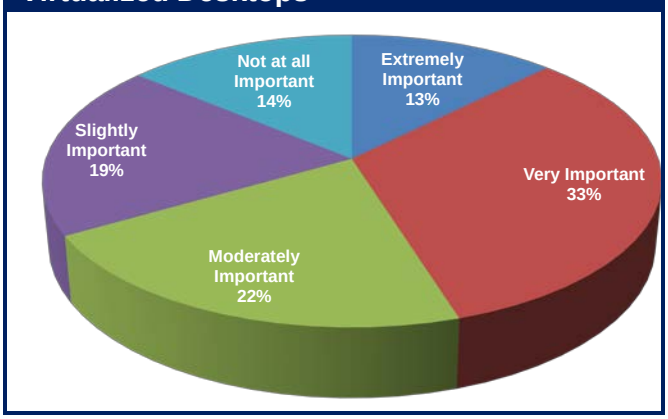
From a networking perspective, the primary challenge in implementing desktop virtualization is achieving adequate performance and an acceptable user experience for client-to-server connections over a WAN.

To quantify the concern that IT organizations have relative to supporting desktop virtualization, The Survey Respondents were asked how important it is for their IT organization over the next year to get better at optimizing the performance of virtualized desktops. Their responses are shown in **Figure 5**.

Half of The Survey Respondents indicated that getting better at optimizing the performance of virtualized desktops is either extremely or very important to their IT organization. That is in sharp contrast to the results of a survey given in 2012 when only a third of The Survey Respondents indicated that getting better at optimizing the performance of virtualized desktops was either extremely or very important to their IT organization.

Getting better at optimizing the performance of virtualized desktops is becoming significantly more important.

Figure 5: Optimizing the Performance of Virtualized Desktops



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

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

Cloud Computing

Over the last few years IT organizations have made a significant adoption of cloud computing in large part because:

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 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, the SLA[7] that Amazon offers for its Amazon Web Services (AWS) states that, "AWS will use commercially reasonable efforts to make Amazon EC2 available with an Annual Uptime Percentage of at least 99.95% during the Service Year." As part of the Amazon definition of Annual Uptime Percentage, Amazon excludes any outage of 5 minutes or less. The Amazon SLA also states that if their service doesn't meet the Annual Uptime Percentage commitment, the customer will receive 10% off its bill for the most recent month that the customer included in the SLA claim that it filed.

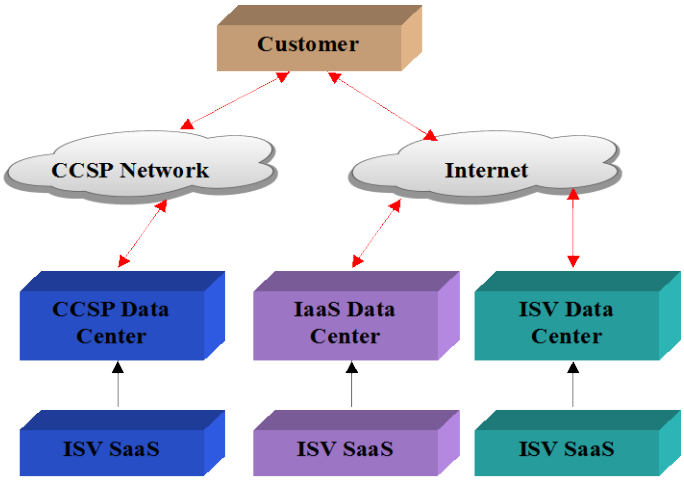
A key attribute of the vast majority of the SLAs that are associated with public cloud computing services is that they don't contain a goal for the end-to-end performance[8] of the service. The reason for the lack of performance guarantees stems from the way that most public cloud computing services are delivered. As shown in **Figure 6**, 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

[7] http://aws.amazon.com/ec2-sla/
[8] In this context, *performance* refers to metrics such as delay or response time.

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 handbook, there are things that can be done to improve the performance of applications delivered over the Internet.

Figure 6: 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.

The Primary Characteristics of Cloud Computing

The following set of characteristics are typically associated with cloud computing. More detail on these characteristics can be found in the 2012 Application and Service Delivery Handbook.

- ***Centralization*** of applications, servers and storage resources.
- Extensive ***virtualization*** of every component of IT.
- ***Standardization*** of the IT infrastructure.
- ***Simplification*** of the applications and services provided by IT.
- ***Technology convergence*** such as the integration of servers, networks and computing.
- ***Service orchestration*** to automate provisioning and controlling the IT infrastructure.
- ***Automation*** of as many tasks as possible.
- ***Self-service*** to enable end users to select and modify their use of IT resources.
- ***Usage sensitive chargeback*** on a user and/or departmental basis.

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

Classes of Cloud Computing Solutions

There are three classes of cloud computing solutions that will be described in this section of the handbook. 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*. One of the primary ways that IT organizations have adopted private cloud computing solutions is by implementing some or all of the previously mentioned characteristics of cloud computing solutions in order to be able to provide Infrastructure-as-a-Service (IaaS) solutions that are similar to the solutions offered by IaaS providers such as Rackspace.

The Survey Respondents were given a set of 7 possible approaches to IaaS and were asked to indicate which approach best described their company's approach to using IaaS solutions, either provided internally by their own IT organization, or provided externally by a CCSPs. The Survey Respondents were allowed to indicate as many approaches as were appropriate. Their responses are shown in **Table 6**.

Table 6: Approach to IaaS	*N=171*
Approach	**Percentage of Respondents**
We are in the process of developing a strategy	48.0%
We provide IaaS solutions internally for a wide range of applications	19.9%
We provide IaaS solutions internally for a small range of applications	19.9%
We have a well-defined and understood strategy	15.2%
We only use IaaS solutions from a CCSP for a small set of applications that are not business critical	14.6%
We use IaaS solutions from a CCSP for a wide range of applications	12.3%
Other	7.0%
We only outsource either a trial of the initial deployment of an application to a CCSP	6.4%
We have a policy against using any IaaS solutions provided by a CCSP	3.5%

One key conclusion that can be drawn from the data in**Table 6** is that:

Only a small percentage of IT organizations have a strategy for how they will acquire or implement IaaS solutions.

The Survey Respondents were asked to indicate the two primary factors that limit their company's interest in using internally provided IaaS solution. The five inhibitors to the adoption

of private IaaS solutions that were indicated the most times by the Survey Respondents and the percentage of times that they were mentioned were:

- Concerns about the security and confidentiality of data (36.3%)
- Their lack of an internal strategy about IaaS (28.7%)
- Their lack of personnel to design and implement the solutions (25.7%)
- The relative immaturity of the technologies that would have to installed and managed (19.9%)
- The lack of significant enough cost savings (19.3%)

While the conventional wisdom in our industry is that security and confidentiality of data is the major impediment to the adoption of public cloud based IaaS solutions, it is somewhat surprising that:

> ***Concern about the security and confidentiality of data is the primary impediment to the broader adoption of private IaaS solutions.***

Public Cloud Computing

This section of the handbook will focus on the two most popular types of public cloud computing solutions: Software-as-a-Service and Infrastructure-as-a-Service.

Software-as-a-Service

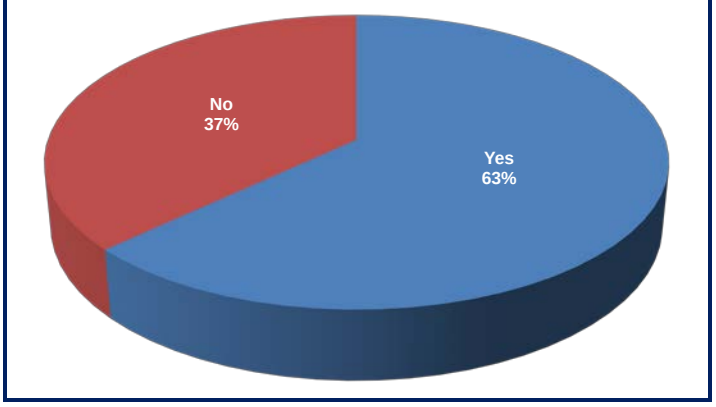
According to Gartner[9], the Software as a Service (SaaS) market will have worldwide revenues of $22.1 billion by 2015. One of the key characteristics of the SaaS marketplace is that:

> ***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 Survey Respondents were asked to indicate if their company currently acquires applications from a SaaS provider or if they are likely to within the next twelve months. Their responses are shown in **Figure 7**.

Figure 7: Adoption of SaaS Solutions *N=264*



The Survey Respondents were then given a set of 7 types of applications and were asked to indicate the types of applications that their company currently acquires from a SaaS provider and the types of applications that their organization would likely acquire from a SaaS provider over the next twelve months. Their responses are shown in **Table 7**.

[9] http://www.slideshare.net/rajeshdgr8/global-saa-s-2012

Table 7: Interest in SaaS		N=153
	Currently Acquire	**Will Acquire**
Collaboration	**55%**	**31%**
Customer Relationship Management CRM)	**53%**	**22%**
Human Resources	**45%**	**18%**
Office Productivity	**40%**	**33%**
Project and Portfolio Management	**27%**	**54%**
Enterprise Resource Planning (ERP)	**24%**	**16%**
Supply Chain Management (SCM)	**15%**	**27%**

The Survey Respondents were given a set of ten factors and were asked to indicate the two factors that were the primary drivers of their organization's interest in using SaaS solutions. The responses of the Survey Respondents are shown in **Table 8**. In **Table 8**, the column on the right is labeled *Percentage of Respondents*. That column contains the percentage of the Survey Respondents that indicated that the factor in the left hand column of **Table 8** was one of the two primary drivers of their organization's interest in using SaaS solutions.

Table 8: Factors Driving the Adoption of SaaS Solutions	N=153
Factor	**Percentage of Respondents**
Lower cost	39%
Reduce the amount of time it takes to implement an application	35%
Free up resources in the IT organization	29%
Deploy applications that are more robust; e.g., available and scalable	27%
Easier to justify OPEX than CAPEX	26%
Leverage the expertise of the SaaS provider	19%
Reduce risk	11%
Management mandate as our strategic direction	8%
Meet temporary requirements	3%
Other	2%

One conclusion that can be drawn from the data in **Table 8** is that:

The primary factors that are driving the adoption of SaaS are the same factors that drive the adoption of any form of out-tasking.

Given the concerns that IT organizations have relative to the security and confidentiality of their data, it appears to be counter intuitive that 11% 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 Respondents' 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.

Infrastructure as a Service (IaaS)

The initial set of IaaS solutions that were brought to market by IaaS providers were the basic compute and storage services that are necessary to run applications. However, the IaaS market is highly dynamic and IaaS providers are deploying myriad new services including:

- Disaster Recovery
- Virtual Private Data Centers
- High Performance Computing

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 (acquired by Time Warner), Savvis (acquired by Century Link), Terremark (acquired by Verizon) and Verizon.

The Survey Respondents were asked to indicate the IaaS services that their organization currently acquires from a CCSP and the services that their organization will likely acquire from a CCSP during the next year. Their responses are shown in **Table 9**.

Table 9: Current and Planned Adoption of IaaS Services *N = 142*

	Currently Acquire	Will Likely Acquire
Storage	26.8%	16.9%
Computing	26.8%	9.2%
Virtual Private Data Center	17.6%	14.1%
Disaster Recovery	16.2%	21.8%
High Performance Computing	10.6%	9.9%

Because storage and computing were the initial set of IaaS services that were brought to market, it was not at all surprising to see that over a quarter of the Survey Respondents indicated that they currently used those services. In addition, given that high performance computing (HPC) is somewhat of a niche application, it was not surprising that there was relatively little interest in acquiring HPC from an IaaS supplier. However it was somewhat of a surprise to see that:

There is strong interest on the part of IT organizations in acquiring both virtual private data center and disaster recovery services from IaaS providers.

Drivers and Inhibitors

This section will discuss the factors that are driving and the factors that are inhibiting the deployment of IaaS solutions.

Drivers

The Survey Respondents were given a set of eleven factors and were asked to indicate the two factors that were the primary drivers of their organization's interest in using Cloud-based IaaS solutions. The responses of the Survey Respondents are shown in **Table 10**. In **Table 10**, the column on the right is labeled *Percentage of Respondents*. That column contains the percentage of the Survey Respondents that indicated that the factor in the left hand column of **Table 10** was one of the two primary drivers of their organization's interest in using Cloud-based IaaS solutions.

Table 10: Factors Driving the Adoption of Cloud-Based IaaS Solutions *N = 171*	
Factor	**Percentage of Respondents**
Lower cost	30.4%
The ability to dynamically add capacity	30.4%
Reduce time to deploy new functionality	26.3%
Obtain functionality we are not able to provide ourselves	22.2%
Deploy more highly available soluti`ons	19.3%
Free up resources	17.0%
Easier to justify OPEX than CAPEX	15.8%
Prefer to only pay for services that we use	14.0%
Satisfy temporary requirements	11.7%
Other	4.7%
Our strategy is to use IaaS providers wherever possible	4.1%
Leverage the security expertise of the provider	4.1%

The conventional wisdom in the IT industry is that lower cost is the primary factor driving the adoption of Cloud-based IaaS solutions and that factors such as the ability to dynamically add new capacity, while important, are nowhere near as important. As the data in **Table 10** highlights, the reality is that the ability to dynamically add new capacity is as important a driver of the adoption of Cloud-based IaaS solutions as is lowering cost. In addition, another very important driver of the adoption of Cloud-based IaaS solutions is the ability to reduce the time it takes to deploy new functionality. It is reasonable to look at the ability to dynamically add capacity and the ability to reduce the time it takes to deploy new functionality as two components of a single factor – agility. Looked at this way,

By a wide margin, agility is the most important factor driving the adoption of Cloud-based IaaS solutions.

Inhibitors

The Survey Respondents were asked to indicate the two primary factors that limit their company's interest in using a Cloud-based IaaS solution. Those factors and the percentage of times that they were indicated by the Survey Respondents are shown in **Table 11**.

Table 11: Inhibitors to the adoption of Cloud-based IaaS Solutions N = 171	
Factor	**Percentage of Respondents**
We are concerned about the security and confidentiality of our data	57.9%
We don't see significant enough cost savings	24.0%
The lack of time and resources to sufficiently analyze the offerings and the providers	19.9%
Uncertainty about the provider living up to their promises	19.9%
We have concerns about the availability of the solutions	16.4%
Our lack of confidence in a shared infrastructure	15.2%
The lack of a meaningful SLA	14.6%
We don't believe that the gains in the agility of these solutions justifies the cost and/or the risk	11.7%
Our policy is to either limit or totally avoid using IaaS providers	8.8%
The provider is not capable of adding capacity in a dynamic enough fashion	4.7%

One conclusion that can be drawn from the data in **Table 11** is:

Concern about the security and confidentiality of data is by a wide margin the number one factor inhibiting the adoption of Cloud-based IaaS solutions

In order to understand the organizational dynamic that underlies the decision to use an IaaS solution from a CSP, the Survey Respondents were asked about the roles of the organizations that are involved in making that decision, Their responses, shown in **Table 12**, indicate how the decision is made.

Table 12: The Decision Making Process N=160	
Role	**Percentage of Respondents**
Largely by the IT organization with some input from the business or functional unit	40.0%
The IT unit and the business or functional unit participate equally	26.3%
Largely by the business or functionaly unit with some input from the IT organization	15.6%
Entirely by the IT organization	11.3%
Entirely by the business or functional unit	6.9%

One obvious conclusion that can be drawn from the data in **Table 12** is:

> ***Roughly 20% of the times that a company is evaluting public IaaS solutions, the company's IT organization is either not involved at all or plays a minor role.***

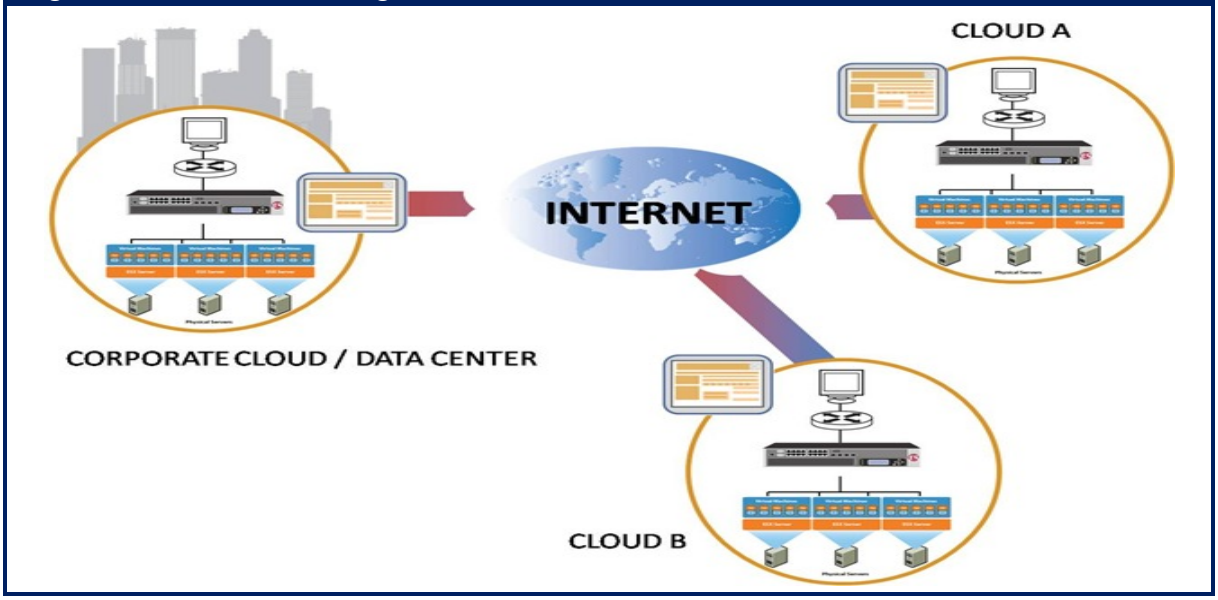
Hybrid Cloud Computing

According to Wikipedia[10], "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 8**, 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 8: 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.

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

However, a cloud balancing solution extends the focus of a GSLB solution to a solution with more of a business focus. Given that extended focus, a cloud balancing solution includes in the criteria that it uses to make a routing decision the:

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

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

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

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

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

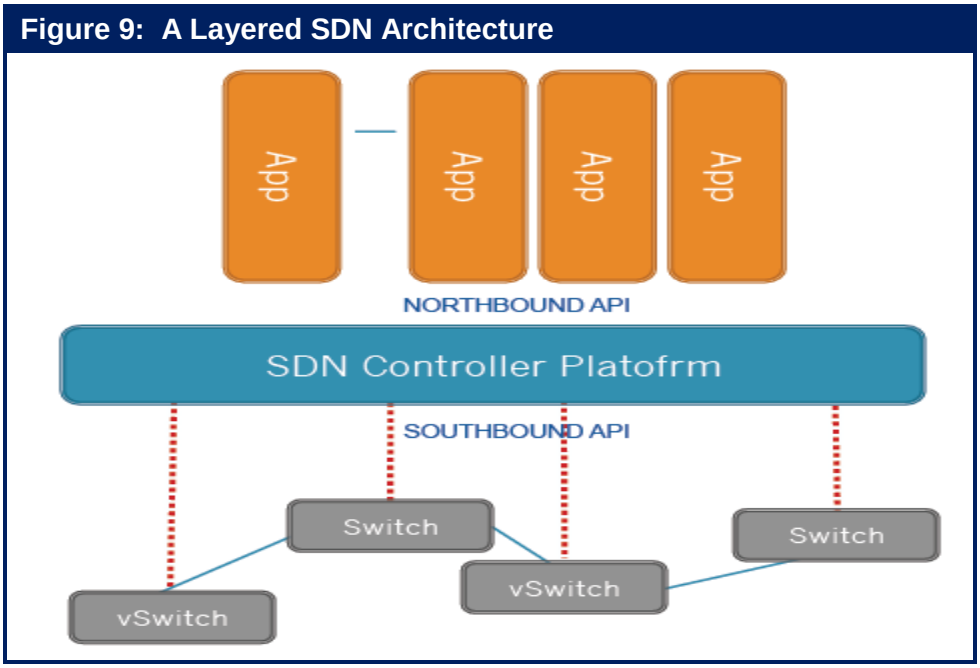
- **Regulatory Compliance**
  For compliance with regulations such as PCI, it may be possible to partition a web services application such that the PCI-related portions remain in the PCI-compliant enterprise data center, while other portions are cloud balanced. In this example, application requests are 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

Software Defined Networking

Unlike the preceding topics in this chapter of The Handbook, Software Defined Networking (SDN) shouldn't make application and service delivery more difficult. To the contrary, some of the advocates of SDN believe that SDN will make application and service delivery easier as it will enable applications to directly signal the network in order to request the types of services that they need. That said, at the current time there isn't a universally agreed to definition as to what is meant by SDN. However, most discussions of SDN include a layered architecture such as the one that is shown in **Figure 9**. In that architecture, the control plane functionality is centralized in the SDN controller's software. Most of the time that SDN is being discussed, the OpenFlow protocol is used to program the forwarding behavior of the switch. There are, however alternative to the use of OpenFlow, including the Extensible Messaging and Presence Protocol (XMPP), the Network Configuration Protocol (Netcong) and OpenStack® from Rackspace and NASA.

Figure 9: A Layered SDN Architecture



In the model shown in **Figure 9**, applications and network functions are written to a set of application programming interfaces (APIs) that are provided by the SDN controller. These northbound APIs are not standardized and so an application that runs on a given SDN controller would have to be modified to run on another SDN controller. Examples of network functions that could run on an SDN controller are given below.

The SDN controller supports a number of drivers that control the behavior of the underlying network elements so that the network will provide the desired network services. The controller provides management plane functionality such as performance and fault management via SNMP and other standard protocols, and it typically handles configuration management of OpenFlow compliant devices in order to provide network topology, forwarding, QoS, and link management.

OpenFlow

The group most associated with the development of a standards based SDN is the Open Networking Foundation (ONF). The ONF was launched in 2011 and has as its vision to make OpenFlow-based SDN the new norm for networks. To help achieve that vision, the ONF has taken on the responsibility to drive the standardization of the OpenFlow protocol. The breadth of the SDN ecosystem is reflected in the fact that the ONF currently has roughly 90 members of varying types including vendors that provide the enabling silicon as well as the switches, network appliances, controllers, test equipment, telecommunications services, hyper-scale data center services and smart phones.

Most modern Ethernet switches and routers contain flow-tables, that are typically supported by TCAMs, that run at line-rate and are used to perform forwarding functions based on Layer 2,3, and 4 packet headers. While each vendor's flow-table is different, there is a common set of functions that is supported by a wide variety of switches and routers. This common set of functions is leveraged by OpenFlow, which is an open protocol that runs between a central OpenFlow controller and an OpenFlow switch and which, as noted, can be used to program the forwarding behavior of the switch. With OpenFlow, a single central controller can program all the physical and virtual switches in a network.

The OpenFlow protocol was developed at Stanford, with v1.0 published at the end of 2009 and v1.1 at the beginning of 2011. In March of 2011, the Open Networking Forum (ONF) was created and the intellectual property rights of OpenFlow were transitioned to it. Part of the ONF charter is to control and commercialize OpenFlow. With that goal in mind, the ONF recently released OpenFlow v1.3 and OpenFlow v1.4 is expected to be released in the June – July 2013 timeframe.

OpenFlow v1.0 defined OpenFlow-only switches and OpenFlow-enabled switches. In an OpenFlow-only switch, all of the control functions of a traditional switch (e.g. the routing protocols that are used to build forwarding information bases (FIBs)) are run in the central OpenFlow controller. An OpenFlow-enabled switch (dubbed a OpenFlow-hybrid switch in V1.1) supports both OpenFlow flow forwarding and traditional Ethernet switch bridging and routing. Hybrid switches allow OpenFlow and traditional bridge/routing to share the same Ethernet infrastructure.

Many existing high functionality Layer2/3 switches can be converted to be OpenFlow-hybrid switches by the relatively simple addition of an OpenFlow agent in firmware supported by the native switch Network Operating System (NOS). Alternatively, once the semiconductor vendors have produced chips that effectively process the OpenFlow protocol, an OpenFlow-only switch would be relatively simple and inexpensive to build because it would have very little resident software and would not require a powerful CPU or large memory to support the extensive control functionality typically packaged in a traditional network operating system (NOS).

There are a number of possible ways that the centralization of control, the programmability, and the flow forwarding characteristics of OpenFlow can be leveraged to provide value to IT organizations. For example, one of the primary benefits of OpenFlow is the centralized nature of the Forwarding Information Base (FIB). Centralization allows optimum routes to be calculated deterministically for each flow leveraging a complete model of the end-to-end topology of the network. Based on an understanding of the service levels required for each type of flow, the centralized OpenFlow controller can apply traffic engineering principles to ensure each flow is properly serviced. One advantage of this capability is that it enables the network to dynamically

respond to application requirements. It also enables notably better utilization of the network without sacrificing service quality.

Another benefit is that OpenFlow switches can filter packets as they enter the network, and hence these switches can act as simple firewalls at the edge of the network. With OpenFlow switches that support the modification of packet headers, an optional feature in OpenFlow v1.0, the OpenFlow controller will also be able to have the switch redirect certain suspicious traffic flows to higher-layer security controls, such as IDS/IPS systems, application firewalls, and Data Loss Prevention (DLP) devices.

OpenFlow switches that support the modification of packet headers will also be able to function as a simple, cost-effective load-balancing device. With modification functionality, a new flow can result in a new flow table entry that is directed to a server that is selected by the OpenFlow controller's load balancing policies. In order to create load-balancing policies based on server load, the OpenFlow controller would have to monitor the pool of servers as they report current load levels.

About the Webtorials® Editorial/Analyst Division

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

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

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

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

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

Silver Peak

Visit the Silver Peak Marketplace Today!

WANop Anywhere. Anytime.

Data acceleration has never been easier!

www.silver-peak.com/marketplace