The 2013 Application & Service Delivery Handbook

Part 2: Network and Application Optimization

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

Platinum Sponsors:

cisco

ipanema Technologies

NETSCOUT

Gold Sponsors:

A10 Networks

agility made possible
ca technologies

radware

riverbed

Silver Peak

Produced by:

Webtorials

Network and Application Optimization

Executive Summary

The ***2013 Application and Service Delivery Handbook*** (The Handbook) will be published both in its entirety and in a serial fashion. This is the second of the serial publications. The first publication focused on describing a set of factors that complicate the task of ensuring acceptable application delivery. The goal of this publication is to describe the technologies and services that are available to improve the performance of applications and services.

The third publication in this series will focus on describing the technologies and services that are available to improve 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.

The preceding section of The ***2013 Application and Service Delivery Handbook*** described the surveys that were administered to the subscribers of Webtorials. Throughout this document, the IT professionals that responded to those surveys will be referred to as The Survey Respondents.

Background

The phrase ***network and application optimization*** refers to an extensive set of techniques that organizations have deployed in an attempt to optimize the performance of networked applications and services while also controlling WAN bandwidth expenses. The primary role these techniques play is to:

- Reduce the amount of data sent over the WAN;
- Ensure that the WAN link is never idle if there is data to send;
- Reduce the number of round trips (a.k.a., transport layer or application turns) necessary for a given transaction;
- Overcome the packet delivery issues that are common in shared networks that are typically over-subscribed;
- Mitigate the inefficiencies of protocols and applications;
- Offload computationally intensive tasks from client systems and servers;
- Use multiple paths from origin to destination where appropriate;
- Direct traffic to the most appropriate server based on a variety of metrics.

There are two principal categories of premise-based network and application optimization products: WAN optimization controllers (WOCs) and Application Delivery Controller (ADCs). In addition, there are WAN services that optimize traffic performance and in some instances, reduce cost.

One factor that can have a negative impact on application and service delivery is packet loss. The affect of packet loss on TCP has been widely analyzed[1]. Mathis, et al. provide a simple formula that offers insight into the maximum TCP throughput on a single session when there is packet loss. That formula is:

Figure 1: Factors that Impact Throughput

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

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

The preceding equation shows that throughput decreases as either the RTT or the packet loss rate increases. To illustrate the impact of packet loss, assume that MSS is 1,420 bytes, RTT is 100 ms. and p is 0.01%. Based on the formula, the maximum throughput is 1,420 Kbytes/second. If, however, the loss were to increase to 0.1%, the maximum throughput drops to 449 Kbytes/second. **Figure 2** depicts the impact that packet loss has on the throughput of a single TCP stream with a maximum segment size of 1,420 bytes and varying values of RTT.

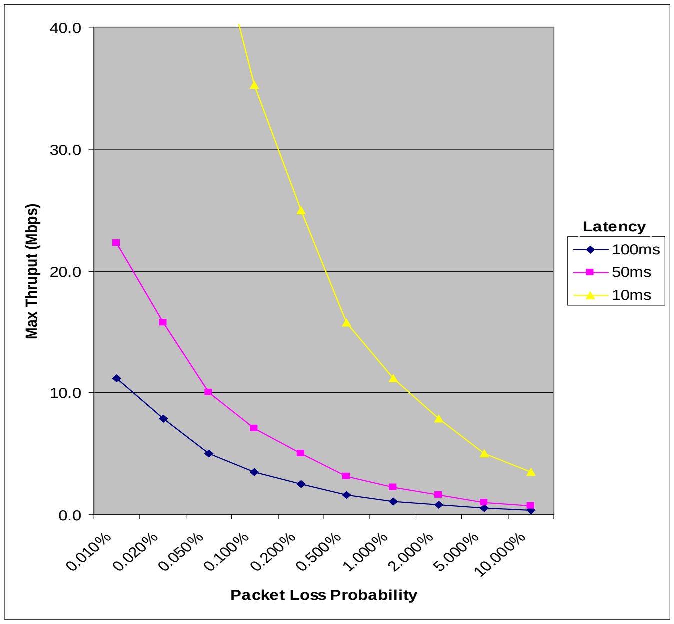
[1] The macroscopic behavior of the TCP congestion avoidance algorithm by Mathis, Semke, Mahdavi & Ott in Computer Communication Review, 27(3), July 1997

One conclusion we can draw from **Figure 2** is:

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

For example, on a WAN link with a 1% packet loss and a round trip time of 50 ms or greater, the maximum throughput is roughly 3 megabits per second no matter how large the WAN link is.

Figure 2: Impact of Packet Loss on Throughput



Because a network and application optimization solution will provide varying degrees of benefit to an enterprise based on the unique characteristics of its environment, third party tests of these solutions are helpful, but not conclusive.

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

Quantifying Application Response Time

A model is helpful to illustrate the potential performance bottlenecks in the performance of an application. The following model (**Figure 3)** is a variation of the application response time model created by Sevcik and Wetzel[2]. Like all models, the following is only an approximation and it is not intended to provide results that are accurate to the millisecond level. It is, however, intended to provide insight into the key factors impacting application response time. As shown below, the application response time (R) is impacted by a number of factors including the amount of data being transmitted (Payload), the goodput which is the actual throughput on a WAN link, the network round trip time (RTT), the number of application turns (AppTurns), the number of simultaneous TCP sessions (concurrent requests), the server side delay (Cs) and the client side delay (Cc).

Figure 3: Application Response Time Model

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

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

Market Research

As was mentioned in the preceding section of The Handbook, in early 2013 two surveys were given to the subscribers of Webtorials. One of the surveys 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

Some of the responses of The Survey Respondents were included in the preceding section of The Handbook. For completeness, **Table 1** shows how The Survey Respondents answered the question about the optimization tasks that are of most interest to their IT organization.

[2] Why SAP Performance Needs Help

Table 1: The Importance of 20 Key Optimization Tasks

	Not at All	Slightly	Moderately	Very	Extremely
Optimizing the performance of a key set of applications that are critical to the success of the business	3.3%	7.2%	17.6%	51.0%	20.9%
Ensuring acceptable performance for VoIP traffic	8.9%	5.1%	20.4%	40.8%	24.8%
Optimizing the performance of TCP	6.9%	13.9%	28.5%	33.3%	17.4%
Improving the performance of applications used by mobile workers	7.8%	13.0%	26.6%	37.7%	14.9%
Ensuring acceptable performance for a business service, such as CRM, that is supported by multiple inter-related applications	9.6%	11.1%	31.1%	34.8%	13.3%
Optimizing the performance of protocols other than TCP; e.g., HTTP and MAPI	7.9%	17.9%	26.4%	35.7%	12.1%
Optimizing the transfer of storage between a data center and a remote user	9.6%	15.1%	30.8%	33.6%	11.0%
Optimizing the transfer of storage between different data centers	11.0%	17.8%	26.7%	30.8%	13.7$
Optimizing the transfer of large files	6.8%	17.6%	33.8%	34.5%	7.4%
Optimizing the performance of specific applications such as SharePoint	10.1%	16.9%	29.1%	33.8%	10.1%
Optimizing the transfer of virtual machines	9.9%	18.4%	29.8%	30.5%	11.3%
Optimizing the performance of servers by offloading SSL and/or TCP processing	15.2%	10.6%	34.1%	28.8%	11.4%
Optimizing the performance of virtual desktops	13.9%	19.4%	21.5%	32.6%	12.5%
Controlling the cost of the WAN by reducing the amount of traffic by techniques such as compression	12.3%	16.4%	34.2%	28.1%	8.9%

Table 1: The Importance of 20 Key Optimization Tasks

	Not at All	Slightly	Moderately	Very	Extremely
Ensuring acceptable performance of traditional video traffic	18.7%	17.3%	24.0%	28.7%	11.3%
Optimizing the performance of applications that you acquire from a SaaS provider such as Salesforce.com	20.9%	14.2%	25.4%	26.9%	12.7%
Ensuring acceptable performance for telepresence traffic	18.6%	20.7%	24.8%	29.0%	6.9%
Optimizing the performance of chatty protocols such as CIFS	15.2%	23.2%	31.2%	25.6%	4.8%
Optimizing the performance of the computing services that you acquire from a third party such as Amazon	29.0%	18.3%	23.7%	21.4%	7.6%
Optimizing the performance of the storage services that you acquire from a third party such as Amazon	32.3%	19.4%	23.4%	20.2%	4.8%

Some of the conclusions that can be drawn from the data in **Table 1** are:

Optimizing the performance of a key set of applications that are critical to the business is the most important optimization task facing IT organizations; followed closely by the need to ensure acceptable performance for VoIP traffic.

Some traditional challenges, such as optimizing the performance of TCP, remain very important while other traditional challenges, such as optimizing the performance of chatty protocols, have become notably less important.

A relatively new challenge, ensuring the performance of applications used by mobile workers, is now one of the most important optimization tasks facing IT organizations.

Another relatively new challenge, optimizing the movement of storage, is becoming important.

Optimizing the performance of services acquired from a public cloud provider such as Salesforce.com or Amazon is relatively unimportant.

WAN Optimization Controllers (WOCs)

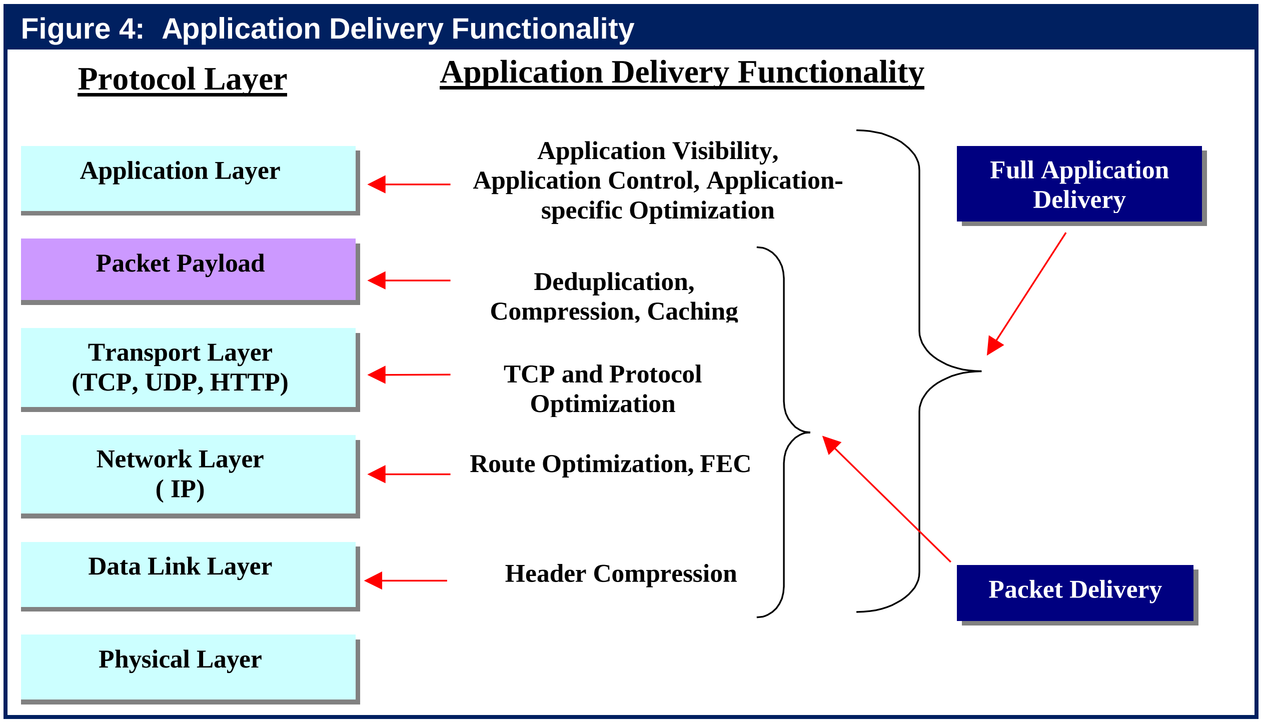
In the vast majority of cases, IT organizations acquire and implement WOCs on a do-it-yourself (DIY) basis. It is also possible for IT organizations to acquire WOC functionality from a managed service provider (MSP). In that scenario, the MSP is responsible for designing, implementing and managing the WOCs. IT organizations have a third option, because some providers offer network and application optimization as part of a WAN service.

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

WOCs are often referred to as *symmetric solutions* because they typically require complementary functionality at both ends of the connection. However, one way that IT organizations can accelerate access to a public cloud computing solution is to deploy WOCs just in branch offices. The WOCs accelerate access by caching the content that a user obtains from the public cloud solution and making that content available to other users in the branch office. Since in this example there is not a WOC at the Cloud Computing Service Provider's (CCSP's) site, this is an example of a case in which a WOC is an asymmetric solution.

When WOCs were first deployed they often focused on improving the performance of a protocol such as TCP or CIFS. However, as shown in **Figure 4**, many WOCs that are available in the marketplace can recognize the application layer signatures of applications and can leverage optimization techniques to mitigate the application-specific inefficiencies that sometimes occur when these applications communicate over a WAN.

Figure 4: Application Delivery Functionality


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

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

WOC Functionality

Table 2 lists some of WAN characteristics that impact application delivery and identifies WAN optimization techniques that a WOC can implement to mitigate the impact of those characteristics.

Table 2: Techniques to Improve Application Performance

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

Below is a description of some of the key techniques used by WOCs:

- ***Caching***
  A copy of information is kept locally, with the goal of either avoiding or minimizing the number of times that information must be accessed from a remote site. Caching can take multiple forms:

  ▪ *Byte Caching*
    With byte caching the sender and the receiver maintain large disk-based caches of byte strings previously sent and received over the WAN link. As data is queued for the WAN, it is scanned for byte strings already in the cache. Any strings resulting in *cache hits* are replaced with a short token that refers to its cache location, allowing the receiver to reconstruct the file from its copy of the cache. With byte caching, the data dictionary can span numerous TCP applications and information flows rather than being constrained to a single file or single application type.

    - *Object Caching*
      Object caching stores copies of remote application objects in a local cache server, which is generally on the same LAN as the requesting system. With object caching, the cache server acts as a proxy for a remote application server. For example, in Web object caching, the client browsers are configured to connect to the proxy server rather than directly to the remote server. When the request for a remote object is made, the local cache is queried first. If the cache contains a current version of the object, the request can be satisfied locally at LAN speed and with minimal latency. Most of the latency involved in a cache hit results from the cache querying the remote source server to ensure that the cached object is up to date.

      If the local proxy does not contain a current version of the remote object, it must be fetched, cached, and then forwarded to the requester. Either data compression or byte caching can potentially facilitate loading the remote object into the cache.

- ***Compression***
  The role of compression is to reduce the size of a file prior to transmitting it over a WAN. Compression also takes various forms.

    - *Static Data Compression*
      Static data compression algorithms find redundancy in a data stream and use encoding techniques to remove the redundancy and to create a smaller file. A number of familiar lossless compression tools for binary data are based on Lempel-Ziv (LZ) compression. This includes zip, PKZIP and gzip algorithms.

      LZ develops a codebook or dictionary as it processes the data stream and builds short codes corresponding to sequences of data. Repeated occurrences of the sequences of data are then replaced with the codes. The LZ codebook is optimized for each specific data stream and the decoding program extracts the codebook directly from the compressed data stream. LZ compression can often reduce text files by as much as 60-70%. However, for data with many possible data values LZ generally proves to be quite ineffective because repeated sequences are fairly uncommon.

    - *Differential Compression; a.k.a., Differencing or De-duplication*
      Differencing algorithms are used to update files by sending only the changes that need to be made to convert an older version of the file to the current version. Differencing algorithms partition a file into two classes of variable length byte strings: those strings that appear in both the new and old versions and those that are unique to the new version being encoded. The latter strings comprise a delta file, which is the minimum set of changes that the receiver needs in order to build the updated version of the file.

      While differential compression is restricted to those cases where the receiver has stored an earlier version of the file, the degree of compression is very high. As a result, differential compression can greatly reduce bandwidth requirements for functions such as software distribution, replication of distributed file systems, and file system backup and restore.

    - *Real Time Dictionary Compression and De-Duplication*
      The same basic LZ data compression algorithms discussed above and proprietary de-duplication algorithms can also be applied to individual blocks of data rather than entire

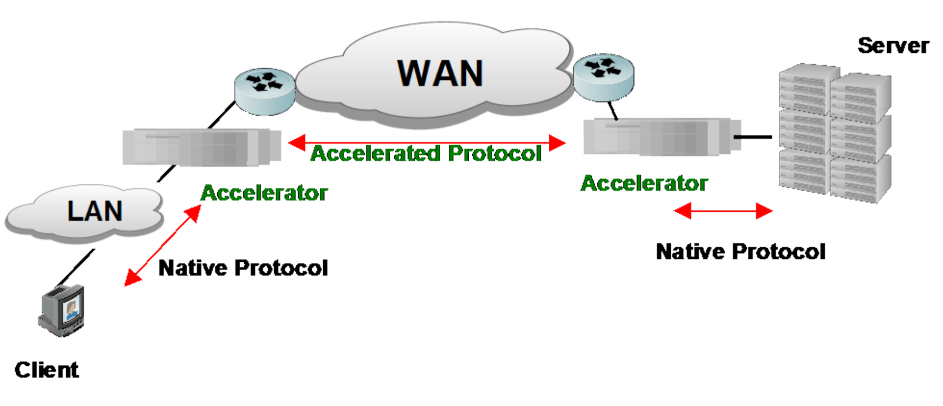
files. This approach results in smaller dynamic dictionaries that can reside in memory rather than on disk. As a result, the processing required for compression and de-compression introduces only a relatively small amount of delay, allowing the technique to be applied to real-time, streaming data. Real time de-duplication applied to small chunks of data at high bandwidths requires a significant amount of memory and processing power.

- ***Congestion Control***
  The goal of congestion control is to ensure that the sending device does not transmit more data than the network can accommodate. To achieve this goal, the TCP congestion control mechanisms are based on a parameter referred to as the *congestion window*. TCP has multiple mechanisms to determine the congestion window[3].

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

- ***Protocol Acceleration***
  Protocol acceleration refers to a class of techniques that improves application performance by circumventing the shortcomings of various communication protocols. Protocol acceleration is typically based on per-session packet processing by appliances at each end of the WAN link, as shown in **Figure 5**. The appliances at each end of the link act as a local proxy for the remote system by providing local termination of the session. Therefore, the end systems communicate with the appliances using the native protocol, and the sessions are relayed between the appliances across the WAN using the accelerated version of the protocol or using a special protocol designed to address the WAN performance issues of the native protocol. As described below, there are many forms of protocol acceleration.

Figure 5: Protocol Acceleration



[3] Transmission_Control_Protocol

- *TCP Acceleration*
  TCP can be accelerated between appliances with a variety of techniques that increase a session's ability to more fully utilize link bandwidth. Some of these techniques include dynamic scaling of the window size, packet aggregation, selective acknowledgement, and TCP Fast Start. Increasing the window size for large transfers allows more packets to be sent simultaneously, thereby boosting bandwidth utilization. With packet aggregation, a number of smaller packets are aggregated into a single larger packet, reducing the overhead associated with numerous small packets. TCP selective acknowledgment (SACK) improves performance in the event that multiple packets are lost from one TCP window of data. With SACK, the receiver tells the sender which packets in the window were received, allowing the sender to retransmit only the missing data segments instead of all segments sent since the first lost packet. TCP slow start and congestion avoidance lower the data throughput drastically when loss is detected. TCP Fast Start remedies this by accelerating the growth of the TCP window size to quickly take advantage of link bandwidth.

- *CIFS and NFS Acceleration*
  CIFS and NFS use numerous Remote Procedure Calls (RPCs) for each file sharing operation. NFS and CIFS suffer from poor performance over the WAN because each small data block must be acknowledged before the next one is sent. This results in an inefficient ping-pong effect that amplifies the effect of WAN latency. CIFS and NFS file access can be greatly accelerated by using a WAFS transport protocol between the acceleration appliances. With the WAFS protocol, when a remote file is accessed, the entire file can be moved or pre-fetched from the remote server to the local appliance's cache. This technique eliminates numerous round trips over the WAN. As a result, it can appear to the user that the file server is local rather than remote. If a file is being updated, CIFS and NFS acceleration can use differential compression and block level compression to further increase WAN efficiency.

- *HTTP Acceleration*
  Web pages are often composed of many separate objects, each of which must be requested and retrieved sequentially. Typically a browser will wait for a requested object to be returned before requesting the next one. This results in the familiar ping-pong behavior that amplifies the effects of latency. HTTP can be accelerated by appliances that use pipelining to overlap fetches of Web objects rather than fetching them sequentially. In addition, the appliance can use object caching to maintain local storage of frequently accessed web objects. Web accesses can be further accelerated if the appliance continually updates objects in the cache instead of waiting for the object to be requested by a local browser before checking for updates.

- *Microsoft Exchange Acceleration*
  Most of the storage and bandwidth requirements of email programs, such as Microsoft Exchange, are due to the attachment of large files to mail messages. Downloading email attachments from remote Microsoft Exchange Servers is slow and wasteful of WAN bandwidth because the same attachment may be downloaded by a large number of email clients on the same remote site LAN. Microsoft Exchange acceleration can be accomplished with a local appliance that caches email attachments as they are downloaded. This means that all subsequent downloads of the same attachment can be satisfied from the local application server. If an attachment is edited locally and then returned to via the remote mail server, the appliances can use differential file compression to conserve WAN bandwidth.

- ***<u>Request Prediction</u>***
  By understanding the semantics of specific protocols or applications, it is often possible to anticipate a request a user will make in the near future. Making this request in advance of it being needed eliminates virtually all of the delay when the user actually makes the request.

  Many applications or application protocols have a wide range of request types that reflect different user actions or use cases. It is important to understand what a vendor means when it says it has a certain application level optimization. For example, in the CIFS (Windows file sharing) protocol, the simplest interactions that can be optimized involve *drag and drop*. But many other interactions are more complex. Not all vendors support the entire range of CIFS optimizations.

- ***<u>Request Spoofing</u>***
  This refers to situations in which a client makes a request of a distant server, but the request is responded to locally.

WOC Form Factors

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

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

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

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

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

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

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

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

To understand the interest that IT organizations have in virtual appliances in general, and virtual WOCs in particular, The Survey Respondents were asked, "Has your organization already implemented, or do you expect that you will implement within the next year, any virtual functionality (e.g., WOC, firewall) in one or more of your branch offices." Just under half responded *yes*. The Survey Respondents that responded *yes* were also given a set of possible IT functionality and asked to indicate the virtual functionality that they have already implemented or that they expected to implement within the next year. Their responses are shown in **Table 3**.

Table 3: Implementation of Virtual Functionality

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

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

One advantage of a vWOC is that some vendors of vWOCs provide a version of their product that is completely free and is obtained on a self-service basis. The relative ease of transferring a vWOC also has a number of advantages. For example, one of the challenges associated with migrating a VM between physical servers is replicating the VM's networking environment in its new location. However, unlike a hardware-based WOC, a vWOC can be easily migrated along with the VM. This makes it easier for the IT organization to replicate the VMs' networking environment in its new location.

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

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

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

Transferring Storage Data

The Challenges of Moving Workflows Among Cloud Data Centers

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

- ***Virtual Machine Migration***

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

  - The SAN or other shared storage system must be extended to the new site;
  - The virtual machine disk space must be migrated to the new data center;
  - The vDisk must be replicated between the two sites.

  In the case of VMotion, VMware recommends that the network connecting the physical servers involved in a VMotion live transfer to have at least 622 Mbps of bandwidth and no more than 5 ms of end-to-end latency[4] [5]. Another requirement is that the source and destination physical servers need to be on the same Layer 2 virtual LAN (VLAN). For inter-data center VM migration, this means that the Layer 2 network must be extended over the WAN.

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

  VXLAN is a scheme to create a Layer 2 overlay on a Layer 3 network via encapsulation. The VXLAN segment is a Layer 3 construct that replaces the VLAN as the mechanism that

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

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

NVGRE is a competing virtual networking proposal before the IETF. It uses GRE as a method to tunnel Layer 2 packets across an IP fabric, and uses 24 bits of the GRE key as a logical network identifier or discriminator, analogous to a VXLAN segment. Another proposal to enable virtual networking is Stateless Transport Tunneling. A detailed comparison of VXLAN, NVGRE and STT can be found in the 2012 Cloud Networking Report[6].

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

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

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

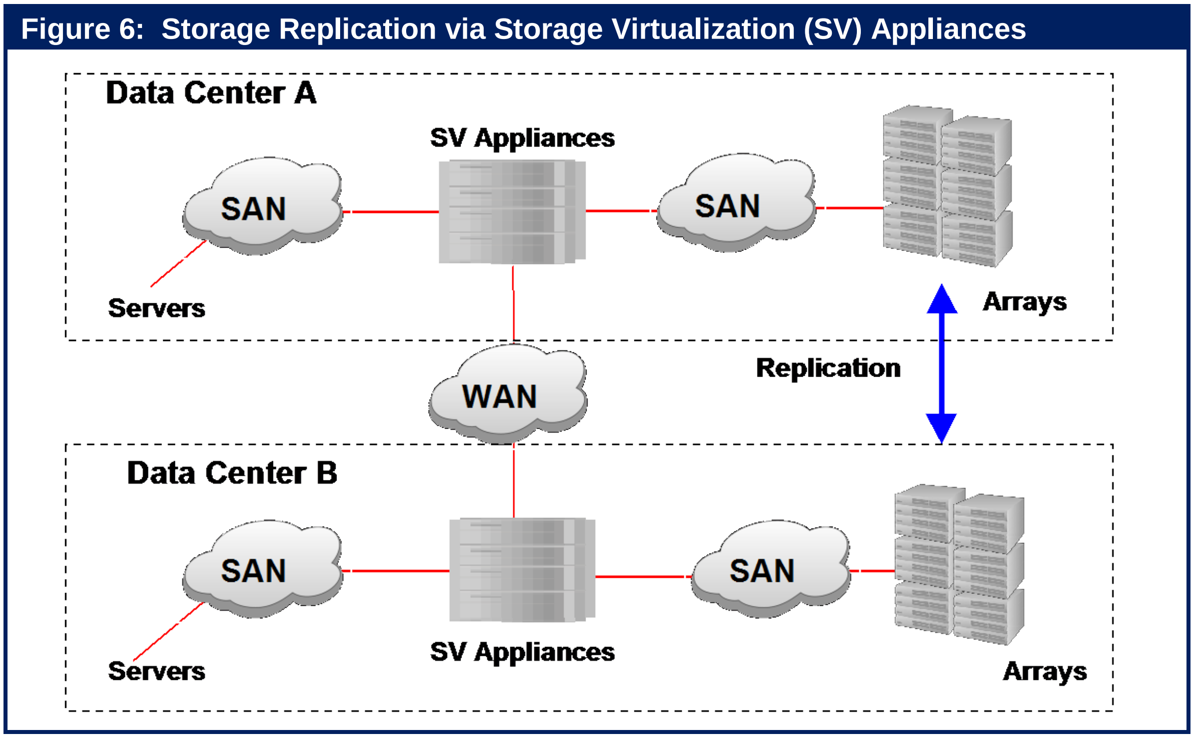
- ***Live Storage Migration***
  Storage migration (e.g., VMware Storage VMotion) can be performed by the server's hypervisor, which relocates the virtual machine disk files from one shared storage location to another shared storage location. The transfer can be completed with zero downtime, with

[6] http://www.webtorials.com/content/2012/12/2012-cloud-networking-report-1.html

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

- ***Storage Replication***
  One way to migrate VMs without the delays associated with storage migration's bulk copy operation is to identify the VMs that are likely to need migration and to replicate the vDisks of those VMs at the remote site in anticipation of an eventual VM migration. **Figure 6** shows in-line server virtualization (SV) appliances performing storage replication over the WAN. Note that storage replication can also be performed by utilities included with some storage devices. In addition to supporting VM migration, storage replication facilitates recovery from data center failures or catastrophic events.

Figure 6: Storage Replication via Storage Virtualization (SV) Appliances



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

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

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

Application Delivery Controllers (ADCs)

Background

The second category of premise-based optimization products is often referred to as an Application Delivery Controller (ADC). This solution is typically referred to as being an *asymmetric solution* because an appliance is only required in the data center and not on the remote end. The genesis of this category of solution dates back to the IBM mainframe-computing model of the late 1960s and early 1970s. Part of that computing model was to have a Front End Processor (FEP) reside in front of the IBM mainframe. The primary role of the FEP was to free up processing power on the general purpose mainframe computer by performing communications processing tasks, such as terminating the 9600 baud multi-point private lines, in a device that was designed specifically for these tasks. The role of the ADC is somewhat similar to that of the FEP in that it performs computationally intensive tasks, such as the processing of Secure Sockets Layer (SSL) traffic, hence freeing up server resources.

While an ADC still functions as a SLB, the ADC has assumed, and will most likely continue to assume, a wider range of more sophisticated roles that enhance server efficiency and provide asymmetrical functionality to accelerate the delivery of applications from the data center to individual remote users. In particular, the ADC can allow a number of compute-intensive functions, such as SSL processing and TCP session processing, to be offloaded from the server. Server offload can increase the transaction capacity of each server and hence can reduce the number of servers that are required for a given level of business activity.

An ADC provides more sophisticated functionality than a SLB does.

ADC Functionality

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

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

- ***SSL Offload***
  One of the primary new roles played by an ADC is to offload CPU-intensive tasks from data center servers. A prime example of this is SSL offload, where the ADC terminates the SSL session by assuming the role of an SSL Proxy for the servers. SSL offload can provide a significant increase in the performance of secure intranet or Internet Web sites. SSL offload frees up server resources which allows existing servers to process more requests for content and handle more transactions.

- ***XML Offload***
  XML is a verbose protocol that is CPU-intensive. Hence, another function that can be provided by the ADC is to offload XML processing from the servers by serving as an XML gateway.

- ***Application Firewalls***
  ADCs may also provide an additional layer of security for Web applications by incorporating application firewall functionality. Application firewalls are focused on blocking the increasingly prevalent application-level attacks. Application firewalls are typically based on Deep Packet Inspection (DPI), coupled with session awareness and behavioral models of normal application interchange. For example, an application firewall would be able to detect and block Web sessions that violate rules defining the normal behavior of HTTP applications and HTML programming.

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

- ***Asymmetrical Application Acceleration***
  ADCs can accelerate the performance of applications delivered over the WAN by implementing optimization techniques such as reverse caching, asymmetrical TCP optimization, and compression. With reverse caching, new user requests for static or dynamic Web objects can often be delivered from a cache in the ADC rather than having to be regenerated by the servers. Reverse caching therefore improves user response time and minimizes the loading on Web servers, application servers, and database servers.

  Asymmetrical TCP optimization is based on the ADC serving as a proxy for TCP processing, minimizing the server overhead for fine-grained TCP session management. TCP proxy functionality is designed to deal with the complexity associated with the fact that each object on a Web page requires its own short-lived TCP connection. Processing all of these connections can consume an inordinate about of the server's CPU resources, Acting as a proxy, the ADC offloads the server TCP session processing by terminating the client-side TCP sessions and multiplexing numerous short-lived network sessions initiated as client-side object requests into a single longer-lived session between the ADC and the Web servers. Within a virtualized server environment the importance of TCP offload is amplified significantly because of the higher levels of physical server utilization that virtualization enables. Physical servers with high levels of utilization will typically support significantly more TCP sessions and therefore more TCP processing overhead.

  The ADC can also offload Web servers by performing compute-intensive HTTP compression operations. HTTP compression is a capability built into both Web servers and Web browsers. Moving HTTP compression from the Web server to the ADC is transparent to the client and so requires no client modifications. HTTP compression is asymmetrical in the sense that there is no requirement for additional client-side appliances or technology.

- ***Response Time Monitoring***
  The application and session intelligence of the ADC also presents an opportunity to provide real-time and historical monitoring and reporting of the response time experienced by end users accessing Web applications. The ADC can provide the granularity to track performance for individual Web pages and to decompose overall response time into client-side delay, network delay, ADC delay, and server-side delay.

- ***Support for Server Virtualization***
  Once a server has been virtualized, there are two primary tasks associated with the dynamic creation of a new VM. The first task is the spawning of the new VM and the second task is

ensuring that the network switches, firewalls and ADCs are properly configured to direct and control traffic destined for that VM. For the ADC (and other devices) the required configuration changes are typically communicated from an external agent via one of the control APIs that the device supports. These APIs are usually based on SOAP, a CLI script, or direct reconfiguration. The external agent could be a start-up script inside of the VM or it could be the provisioning or management agent that initiated the provisioning of the VM. The provisioning or management agent could be part of an external workflow orchestration system or it could be part of the orchestration function within the hypervisor management system. It is preferable if the process of configuring the network elements, including the ADCs, to support new VMs and the movement of VMs within a data center can readily be automated and integrated within the enterprise's overall architecture for managing the virtualized server environment.

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

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

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

IPv6 and ADCs

Background

While it won't happen for several years, IPv6 will replace IPv4 and the entire Internet will be IPv6 only. Gartner, Inc. estimates that 17% of the global Internet users and 28% of new Internet connections will use IPv6 by 2015.[7] This is creating an imperative for enterprises to develop an IPv6 strategy and migration plan. A key component of that strategy and migration plan is ensuring that devices such as firewalls and ADCs that you are implementing today, fully support IPv6.

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

IPv6 was created as an improvement over IPv4 for addressing, efficiency, security, simplicity and Quality of Service (QoS). IPv6's addressing scheme is the centerpiece of its achievement and the main driver behind IPv6 implementation. IPv4 uses 32 bits for IP addresses which allows for a maximum of 4 billion addresses. While this is a large number, rapid increases in Internet usage and growth in Internet devices per person have depleted almost all of the available IPv4 addresses. Network Address Translation (NAT) and use of private IP addresses (IETF RFC 1918) have raised the efficiency of IPv4 addressing, but have also limited Internet functionality. IPv6 addresses quadruples the number of bits used in the network addressesing to 128 bits which provides 4.8 x 10^{28} addresses (5 followed by 28 zeros) for each person on the Earth today. IPv6 eliminates the need to use NAT for IP addresses preservation. NAT will likely continue to be used for privacy or security, but it is not needed for address conservation in IPv6.

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

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

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

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

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

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

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

Enabling Standards and Technologies

IPv6/IPv4 Tunneling

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

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

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

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

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

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

Network Address Translation (NAT)

Network Address Translation (NAT) has been used for several decades with IPv4 networks to effectively extend the amount of available IPv4 addresses. Each IP address can have up to 65,535 connections or ports, but it is rare for this limit to be reached – especially for devices used by end users. In reality, the number of active connections is usually under 100 for end user devices, however behind a home CPE device it may be from 200-500 with multiple devices connected. In addition, connections are typically initiated by the end user device, rather than from the application or server to the end user device. Taking advantage of end user initiated connections with a low connection count, it is quite common to multiplex multiple end user devices' IP addresses together into a few IP addresses and increase the number of connections per IP address. This is accomplished by translating the end user IP address and port number to one of a few IP addresses in each outgoing and returning packet. This is usually accomplished using a network firewall or ADC and this hides the original end user's IP address from the

Internet. Since the end user's original IP address is hidden from the public Internet, end user IP addresses can be duplicated across different networks with no adverse impact. Multiple networks behind firewalls can use the same IP subnets or "private IP subnets", as defined in IETF RFC 1918. NAT has been used extensively in IPv4 to preserve the IPv4 address space and since it translates both IPv4 address and the TCP/UDP port numbers is more correctly called Network Address and Port Translation (NAPT). When NAT is used to translate an IPv4 address to an IPv4 address, it is referred to as NAT44 or NAT444 if these translations are done twice.

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

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

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

Carrier Grade NAT (CGN)

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

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

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

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

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

Dual Stack

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

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

Creating an IPv6 Presence and Supporting Mobile IPv6 Employees

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

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

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

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

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

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

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

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

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

- 6to4, 6rd, 6in4, DS-Lite and Teredo tunneling
- NAT64 and DNS64
- Dual Stack Managed Internet Border Routers
- Dual Stack Managed Firewall Services
- IPv6 addressing, including provider independent IPv6 addressing
- IPv6 BGP
- Network monitoring and reporting for IPv6, including separate IPv6 and IPv4 usage

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

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

- Dual Stack IPv6/IPv4
- How IPv6 to IPv4, IPv6 to IPv6 and IPv4 to IPv6 firewall rules are created and maintained
- Network monitoring and reporting on the firewall for IPv6, including separate IPv6 and IPv4 usage statistics
- Ability to NAT IPv6 to IPv6 for privacy (NAT66)
- Support for VRRP IPv6 (e.g. VRRPv3 RFC 5798) and/or HSPR IPv6 for redundancy

- If the same firewalls are used to screen applications for internal users, then IPv6 compatibility with IF-MAP (TCG's Interface for Metadata Access Points) should be checked if applicable.
- Support for IPv6 remote access VPN (IPsec or SSL or IPsec/SSL Hybrid) termination on firewall

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

Once the data center perimeter firewall supports IPv6, attention can now turn to Application Delivery Controllers (ADCs) that provide load balancing, SSL offloading, WAN optimization, etc. When establishing an IPv6 presence for customers, business partners and suppliers, there are architectures with two or more data centers that benefit from IPv6 ADCs with WAN optimization. ADCs can have the following IPv6 capabilities[8]:

- Ability to provide IPv6/IPv4 Dual Stack for Virtual IPs (VIP)
- Server Load Balancing with port translation (SLB-PT/SLB-64) to IPv4 servers (and the ability to transparently load balance a mix of IPv4 and IPv6 servers)
- 6rd
- NAT64 and DNS64 (to provide IPv6 name resolution services for IPv4-only servers)
- Dual-stack Lite (DS-lite)
- SNMP IPv4 and IPv6 support for monitoring, reporting and configuration
- Ability to provide utilization and usage statistics separated by IPv4 and IPv6

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

Supporting Areas

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

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

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

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

Virtual ADCs

Background

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

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

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

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

The Evolution of Network Appliances

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

As previously mentioned, as network functionality grew, the network evolved from a *packet delivery* service to an *application and service delivery* service. Network appliances evolved from general purpose servers to become the standard building block of the Application and

Service Delivery Network. Network appliances improved upon server technology in two important ways. First, the O/S was changed from a general purpose O/S to one optimized for network operations and processing. Second, the server hardware was updated to include specialized co-processors (e.g. SSL operations and encryption) and network adapters for high performance network operations. This simplified IT operations, as typically only one IT group (e.g. Networks Operations) was involved in changes as opposed to two IT groups (e.g., Network Operations and Server Operations). In general, software updates and security patches are less frequent on network appliances than for general purpose O/Ss and this further reduces the IT operations effort.

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

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

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

The Types of ADC Virtualization

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

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

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

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

Microsoft's System Center. Support for loosely couple systems (e.g. VMWare's VMotioin and Citrix's XenMotion) is common.

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

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

High Availability and Hardware Options

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

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

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

Trends in ADC Evolution

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

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

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

- ***Web Content Optimization (WCO)***
  Two of the challenges that are associated with delivering Web pages are the continually growing number of objects per page, which result in a continually increasing number of round trips per page and the continually growing size of Web pages. Another challenge is the wide range of browsers and mobile devices that access Web pages. Having a range of

browsers and mobile devices makes it very time consuming to manually optimize the Web page for delivery to all the users. WCO refers to efficiently optimizing and streamlining Web page delivery. WCO is available in a number of form factors, including being part of an ADC.

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

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

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

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

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

Developing your ADC Strategy

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

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

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

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

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

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

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

- **IPv4 to IPv6 migration plans** – ADCs are a key point where IPv6 to IPv4 transitions occur as part of an overall IPv6 migration. As such, an organization's IPv6 migration strategy and plans affect the ADC strategy.

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

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

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

After your ADC strategy is developed, an ADC product set needs to be chosen. Some requirements to consider adding to your ADC product selection criteria include the:

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

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

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

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

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

- Completeness and flexibility of IPv6 support.

- Ability to support hybrid and cloud bursting deployments

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

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

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

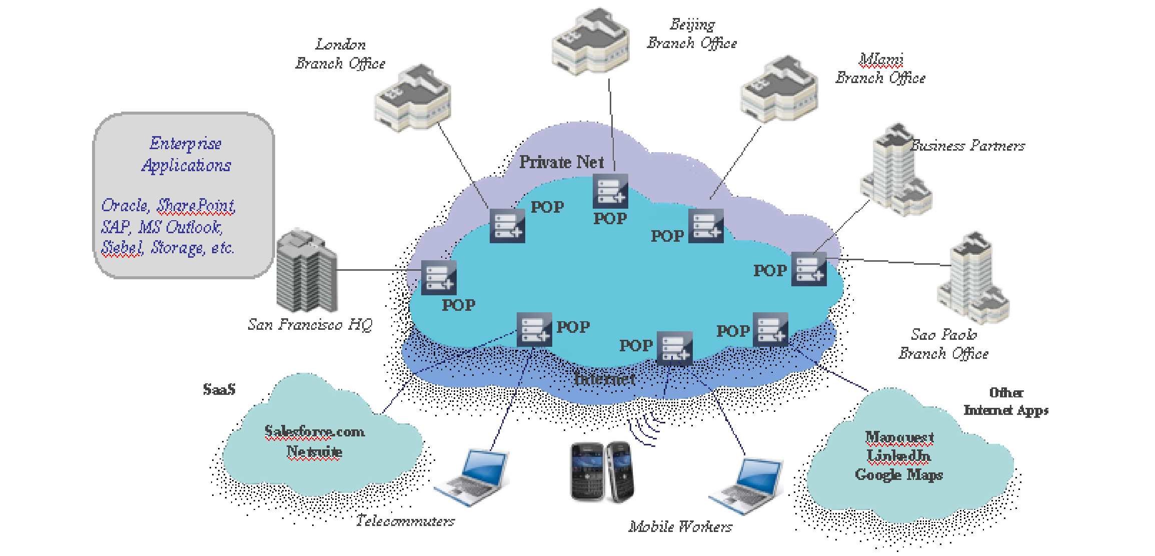
WAN Optimization

This section of The Handbook will discuss WAN services that either provide some form of optimization themselves or which can be modified to provide optimization.

Cloud-Based Optimization Solutions

One form of optimized WAN service is shown In **Figure 7**. In this form of optimized WAN a variety of types of users (e.g., mobile users, branch office users) access WAN optimization functionality at the service provider's points of presence (POPs). Ideally these POPs are inter-connected by a dedicated, secure and highly available network. To be effective, the solution must have enough POPs so that there is a POP in close proximity to the users so as to not introduce unacceptable levels of delay. In addition, the solution should support a wide variety of WAN access services.

Figure 7: WAN and Application Optimization Solution



There are at least two distinct use cases for the type of solution shown in **Figure 7**. One such use case is that this type of solution can be leveraged to solve the type of optimization challenges that an IT organization would normally solve by deploying WOCs; e.g., optimizing communications between branch office users and applications in a corporate data center or optimizing data center to data center communications. In this case, the factors that would cause an IT organization to use such a solution are the same factors that drive the use of any public cloud based services; e.g., cost savings, reduce the time it takes to deploy new functionality and provide functionality that the IT organization could not provide itself

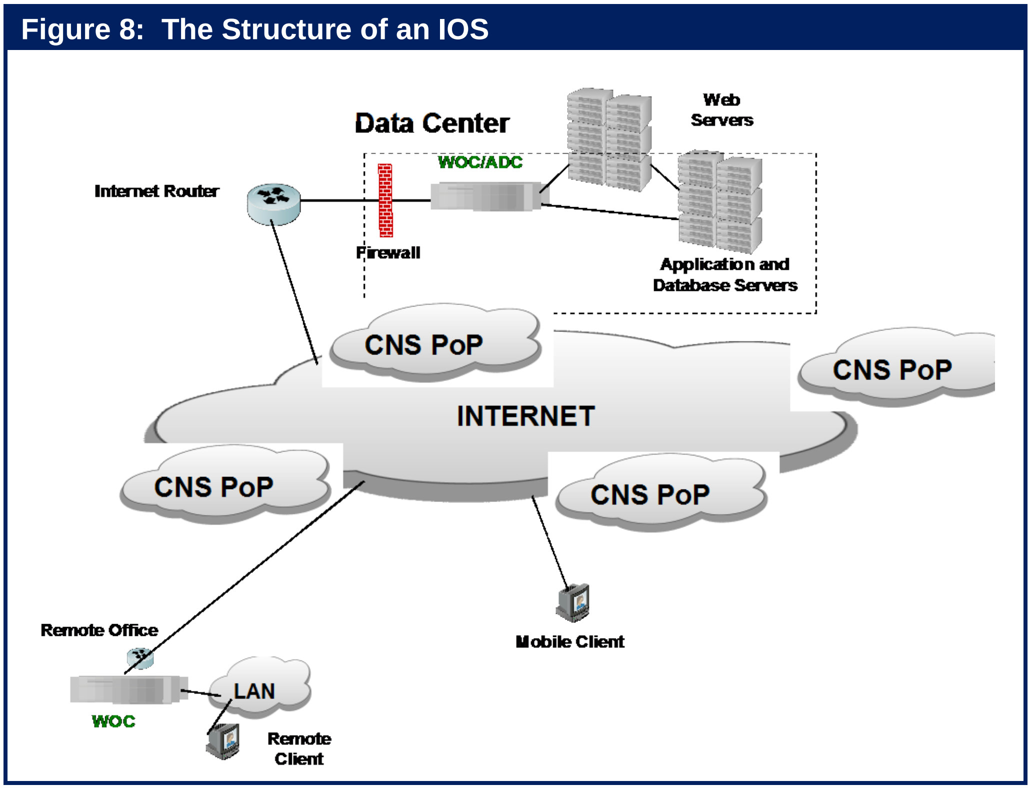
The second use case is the ongoing requirement that IT organizations have to support mobile workers. Some IT organizations will resolve the performance challenges associated with supporting mobile users by loading optimization software onto all of the relevant mobile devices. There are two primary limitations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

The Optimization of Internet Traffic

As previously described, WOCs were designed to address application performance issues at both the client and server endpoints. These solutions make the assumption that performance characteristics within the WAN are not capable of being optimized because they are determined by the relatively static service parameters controlled by the WAN service provider. This assumption is reasonable in the case of private WAN services such as MPLS. However, this assumption does not apply to enterprise application traffic that transits the Internet because there are significant opportunities to optimize performance within the Internet itself. Throughout this section of the handbook, a service that optimizes Internet traffic will be referred to as an Internet Optimization Service (IOS).

An IOS leverages service provider resources that are distributed throughout the Internet. The way this works is that as shown in **Figure 8**, all client requests to the application's origin server in the data center are redirected via DNS to a server in a nearby point of presence (PoP) that is part of the IOS. This edge server then optimizes the traffic flow to the IOS server closest to the data center's origin server.

Figure 8: The Structure of an IOS



The servers at the IOS provider's PoPs perform a variety of optimization functions that are described below. Intelligence within the IOS servers can also be leveraged to provide extensive network monitoring, configuration control and SLA monitoring of a subscriber's application and can also be leveraged to provide security functionality. The management and security functionality that can be provided by an IOS will be discussed in more detail in the next section of the handbook.

Some of the optimization functionality provided by an IOS was described in the preceding discussion of WOCs. This includes optimizing the performance of protocols such as TCP and HTTP. Some of the unique optimization functionality that can be provided by an IOS includes:

- ***Route Optimization***
  Route optimization is a technique for circumventing the limitations of BGP by dynamically optimizing the round trip time between each end user and the application server. A route optimization solution leverages the intelligence of the IOS servers that are deployed in the service provider's PoPs to measure the performance of multiple paths through the Internet and to choose the optimum path from origin to destination. The selected route factors in the degree of congestion, traffic load, and availability on each potential path to provide the lowest possible latency and packet loss for each user session.

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

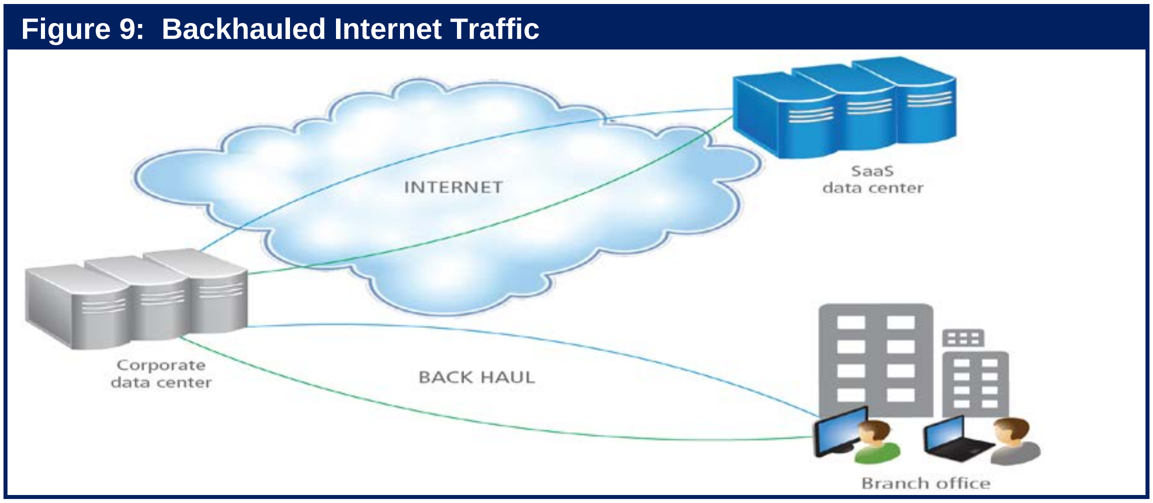
- ***Availability***
  Dynamic route optimization technology can improve the effective availability of the Internet itself by ensuring that viable routes are found to circumvent outages, peering issues or congestion.

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

An Integrated Private-Public WAN

The traditional approach to providing Internet access to branch office employees has been to backhaul that Internet traffic on the organization's enterprise network (e.g., their MPLS network) to a central site where the traffic was handed off to the Internet (**Figure 9**). The advantage of this approach is that it enables IT organizations to exert more control over their Internet traffic and it simplifies management in part because it centralizes the complexity of implementing and managing security policy. One disadvantage of this approach is that it results in extra traffic transiting the enterprise's WAN, which adds to the cost of the WAN. Another disadvantage of this approach is that it usually adds additional delay to the Internet traffic.

Figure 9: Backhauled Internet Traffic



In order to quantify how IT organizations are approaching Internet backhaul, the survey respondents were asked to indicate how they currently route their Internet traffic and how that is likely to change over the next year. Their responses are contained in **Table 4**.

Table 4: Routing of Internet Traffic

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

One of the conclusions that can be drawn from the data in **Table 4** is that:

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

Because backhauling Internet traffic adds delay, one of the disadvantages of this approach to providing Internet access is degraded performance. For example, in the scenario depicted in **Figure 9** (Backhauled Internet Traffic), the delay between users in a branch office and the SaaS application is the sum of the delay in the enterprise WAN plus the delay in the Internet. In order to improve performance, an IT organization might use WOCs to optimize the performance of the traffic as it flows from the branch office to the central site over their enterprise WAN. However, once the traffic is handed off to the Internet, the traffic is not optimized and the organization gets little value out of optimizing the traffic as it flows over just the enterprise WAN.

One way to minimize the degradation in application performance is to not backhaul the traffic but hand it off locally to the Internet. For this approach to be successful, IT organizations must be able to find another way to implement the security and control that it has when it backhauls Internet traffic. One way that this can be done is to use an IOS to carry traffic directly from the branch office to the SaaS provider. With this approach, in addition to providing optimization functionality, the IOS can provide the security functionality that was previously provided in the corporate data center.

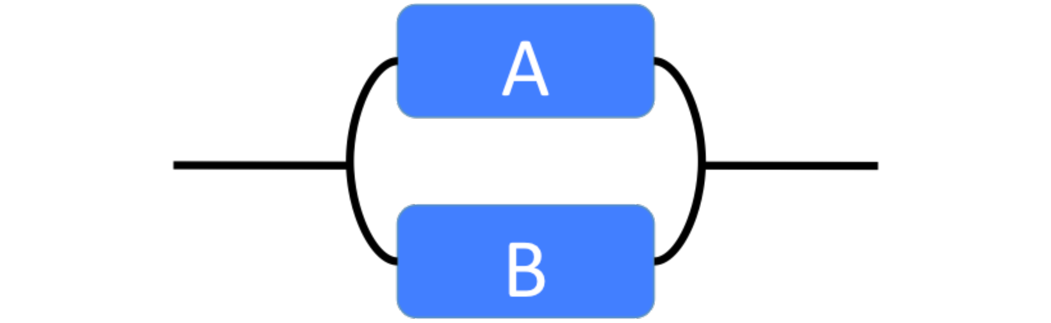
Another approach to optimizing Internet traffic is to implement a form of WAN optimization that enables IT organizations to keep its current approach to backhauling traffic, but which eliminates the performance issues surrounding the fact that once the traffic is handed off to the Internet, the traffic is typically no longer optimized. For this approach to work, the optimization that is in place for enterprise WANs must be integrated with the optimization that is provided by the IOS. As part of this integration, key functionality that is part of the IOS must be integrated into the WOC that sits in the enterprise data center. In addition, WOCs have to be distributed to the PoPs that support the IOS. This integration ensures a seamless handoff of functionality such as TCP optimization between the WOC in the data center and the IOS.

Hybrid WANs with Policy Based Routing (PBR)

The two primary concerns that IT organizations have with the use of the Internet 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[9].

For example, **Figure 10** depicts a system that is composed of two components that are connected in parallel.

Figure 10: Two Components Connected in Parallel



The system depicted in **Figure 10** 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[10]. 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

[9] It is possible to deploy a hybrid WAN with PBR that leverages both MPLS and Internet services.

[10] 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%.

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.

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 discussed in the 2012 Cloud Networking Report, the majority of IT organizations don't place much value in the SLAs that they receive from their network service providers.

Aggregated Virtual WANs

As noted, many IT organizations have concerns about migrating traffic away from MPLS and onto the Internet. As was also discussed, 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. One advantage of a hybrid WAN that is based on both MPLS and the Internet is that the CoS capability of MPLS can be leveraged for delay sensitive, business critical traffic while the Internet VPN can be used both for other traffic and as a backup for the MPLS network.

Independent of whether the PBR-based hybrid WAN is comprised of MPLS and Internet service or just Internet services, 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.

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 the selection 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.

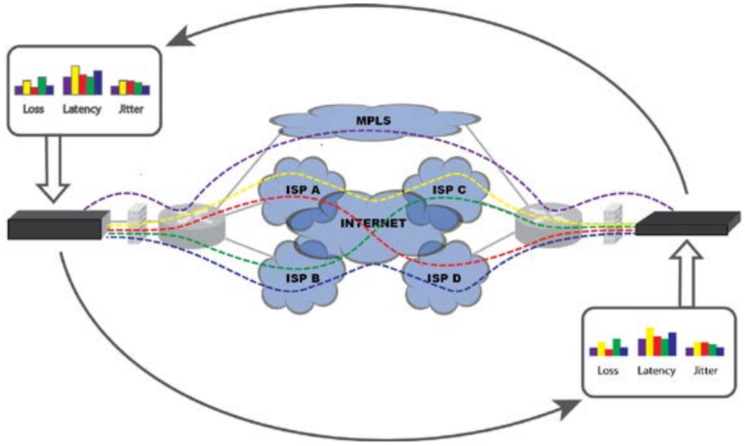
- The instantaneous load for each end-to-end path: The load is weighted based on the business criticality of the application flows. This enables the solution to maximize the business value of the information that is transmitted.

- 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.

As previously noted, 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 11**) 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 11: Aggregated Virtual WANs



As shown in **Figure 11** because the two avWAN appliances work together to continuously measure loss, latency, jitter and bandwidth utilization across all of the various paths between any two 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 10**, 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.

About the Webtorials® Editorial/Analyst Division

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

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

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

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

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

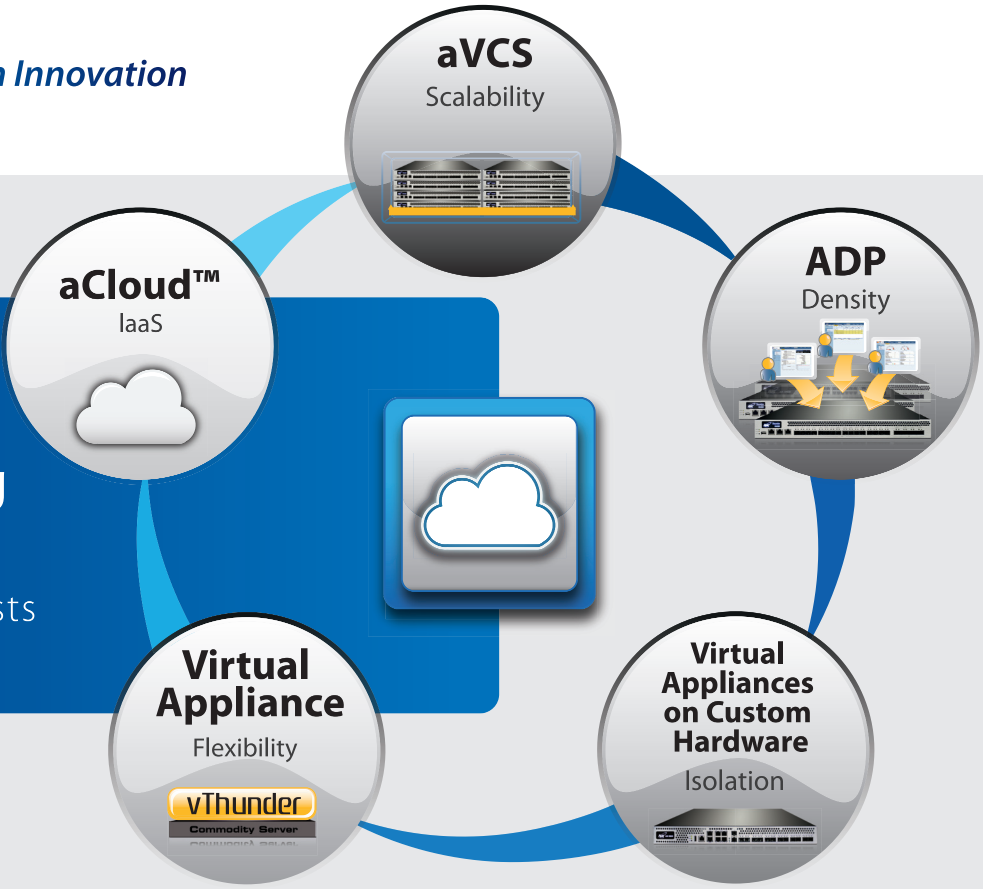
Professional Opinions Disclaimer
All information presented and opinions expressed in this publication represent the current opinions of the author(s) based on professional judgment and best available information at the time of the presentation. Consequently, the information is subject to change, and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

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

Customer Driven Innovation

Cost Effective Cloud Computing

Reduce Your
Application Delivery Costs



A10 Thunder™ and AX Series Products & Solutions

Based on A10's award-winning Application Delivery Controllers (ADCs) and Advanced Core Operating System (ACOS™) architecture, enterprises and service providers will have the flexibility to choose the following scale-as-you-grow virtualization options.

Virtual Appliances

- Virtual machine (VM) on commodity hardware
- Rapidly scale with commodity hardware
- Reduce hardware costs and upload to compatible cloud providers

Virtual Chassis System

- Cluster multiple ADCs to operate as a unified single device
- Scale while maintaining single IP management
- Reduce costs and simplify management while adding devices as you grow

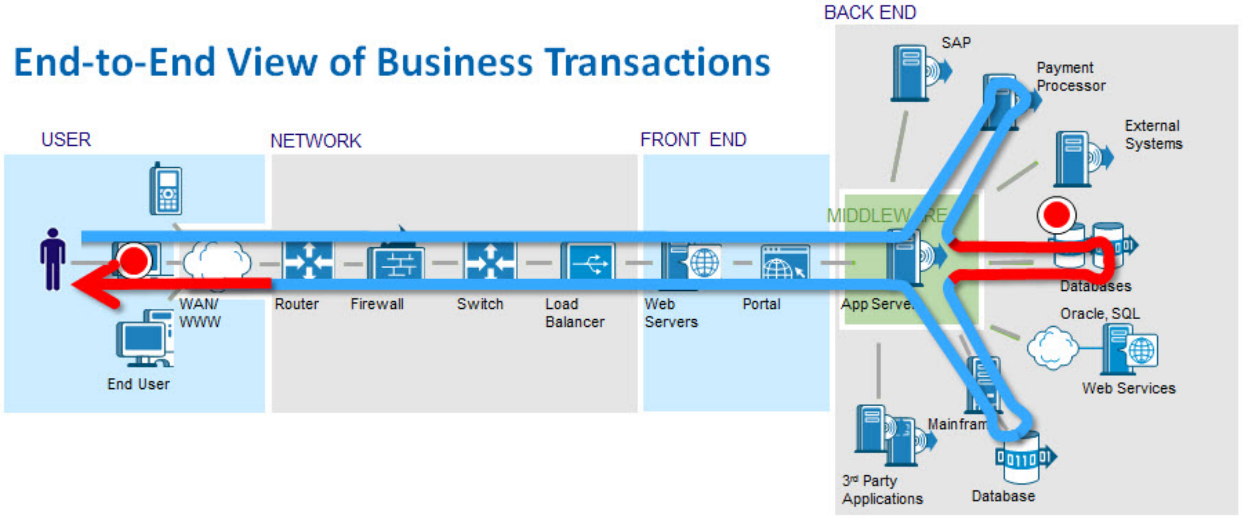
Virtual Appliances on Custom Hardware

- Hardware appliances with embedded hypervisor running virtual appliances
- Flexibility with hardware performance and reliability

Application Delivery Partitions

- Partition the ADC platform resources for individual applications
- Enable quality multi-tenancy with granular resource allocation
- Reduce the number of appliances to host multiple applications

Mission-critical app? You need CA Application Performance Management



What does an hour of an application outage or downtime cost your business? For internal systems, it's lost productivity. For external systems it's lost customers and lost revenue. Let's face it: neither is good for the bottom line. Applications are the lifeblood of virtually all organizations today, meaning they require a robust application performance management solution that helps ensure end users get the experience they expect and business services are reliably delivered. Today's complex, business-critical applications require CA Application Performance Management (CA APM) to ensure business success.

CA APM delivers 360-degree visibility into and analysis of all user transactions across the hybrid-cloud infrastructure – physical, virtual, cloud and mainframe – to understand the health, availability, business impact and end-user experience of critical enterprise, mobile and cloud applications. Advanced Application Behavior Analytics add deeper visibility into the wealth of metric performance collected by CA APM, giving IT operators another set of eyes to look for potential trouble spots. With CA APM, organizations can proactively identify, diagnose and resolve problems throughout the application lifecycle to put organizations firmly in control of the end-user experience and optimize the performance of critical, revenue-generating services.

CA APM is uniquely designed for today's large, complex and heterogeneous production and pre-production environments. CA APM deploys rapidly with little overhead, scales to manage billions of transactions and is easy for both business and IT executives to use, accelerating time to value while increasing return on investment. CA APM is trusted by more than 2,500 large enterprises and service providers to manage the performance and availability of their critical and revenue-generating services. *Learn more at ca.com/apm*.

agility
made possible™

ca technologies

Simplify and Accelerate Virtual Application Deployments with Cisco Cloud Network Services

Cisco and a Multi-vendor Ecosystem Provide Cloud-Ready Application Services

ROLE OF THE NETWORK FOR THE CLOUD

Access to Critical Data, Services, Resources and People

- Core fabric connects resources within the data center and across data centers to each other.
- Pervasive connectivity links users and devices to resources and each other.
- The network provides identity- and context-based access to data, services, resources, and people.

Granular Control of Risk, Performance, and Cost

- Manage and enforce policies to help ensure security, control, reliability, and compliance.
- Manage and enforce service-level agreements (SLAs) and consistent quality of Service (QoS) within and between clouds, enabling hybrid models and workload portability.
- Meter resources and use to provide transparency for cost and performance.

Robustness and Resilience

- Supports self-healing, automatic redirection of workload and transparent rollover.
- Provide scalability, enabling on-demand, elastic computing power through dynamic configuration.

Innovation in Cloud-Specific Services

- Context-aware services understand the identity, location, proximity, presence, and device.
- Resource-aware services discover, allocate, and pre-position services and resources.
- Comprehensive insight accesses and reports on all data that flows in the cloud.

Overview

Cisco has announced the evolution of its network services strategy for virtual and cloud networks, Cisco® Cloud Network Services, a complete portfolio of application networking and security services built on top of the Nexus® 1000V virtual networking portfolio and part of the Cisco Unified Data Center architecture. Cloud Network Services simplifies and accelerates cloud network deployments without compromising the critical security and application delivery services that critical data center applications require.

Introducing Cisco Cloud Network Services

Advances in cloud computing, data center consolidation, mobility, and big data are imposing new demands on the network, along with demands for greater network simplification and automation.

As virtual networking and programmable overlay networks evolve to meet these challenges, a similar evolution needs to take place in Layer 4 through 7 application networking services to support widespread virtualization, application mobility, cloud architectures and network automation.

Cisco's solution to this challenge is Cisco Cloud Network Services, a portfolio of integrated, application-aware network services offerings designed for virtual and cloud environments. The Cloud Network Services framework eliminates the obstacles of physical service appliances to accommodate the requirements of virtual applications and cloud deployments, such as:

- Limited scalability of physical services in fixed locations

CISCO VIRTUAL NETWORK PORTFOLIO

Routing and Switching

- Cisco Nexus 1000V virtual switch
- Cisco Cloud Services Router (CSR) 1000V

Security and VPN

- Cisco Virtual Security Gateway for Nexus 1000V (included in Nexus 1000V Advanced Edition)
- Cisco Adaptive Security Appliance (ASA) 1000V Cloud Firewall
- Imperva SecureSphere Web Application Firewall

WAN Optimization

- Cisco Virtual Wide Area Application Services (vWAAS)

Network Analysis and Monitoring

- Cisco Prime Virtual Network Analysis Module (NAM)

Application Delivery Controllers

- Citrix NetScaler VPX virtual application delivery controller

Virtual Services Deployment Platform

- Cisco Nexus 1100 Series Cloud Services Platform

Cloud Orchestration and Management

- Cisco Intelligent Automation for Cloud (IAC)
- Cisco Prime Network Controller
- OpenStack

To learn more about Cisco's complete virtual networking portfolio, see http://cisco.com/go/1000v

- Inconsistent application performance based on workload location relative to services
- Difficulty in inserting security and network services into virtual networks
- Lack of control over services and policies for applications deployed at cloud service providers

The Cisco Cloud Network Services portfolio includes the Cisco Adaptive Security Appliance (ASA) 1000V Cloud Firewall, Cisco Virtual Security Gateway (VSG) virtual firewall, Cisco Virtual Wide Area Application Services (vWAAS) WAN optimization solution, and Cisco Prime™ Virtual Network Analysis Module (vNAM). This new architecture provides a complete services portfolio while delivering scale-out architecture, elastic instantiation, and multi-tenant operation, all with a common approach to service provisioning and management.

Cloud Network Services also includes best-in-class third-party virtual service offerings that integrate transparently into the framework. It now includes the Citrix NetScaler VPX virtual application delivery controller (ADC), and the Imperva SecureSphere Web Application Firewall (WAF).

Cisco Cloud Network Services form the virtual network services strategy of the larger Cisco Unified Data Center framework, which brings together a seamless architecture of virtualization and cloud-ready compute servers (Unified Compute Servers), network fabric (Unified Fabric) and automation platform (Unified Management).

Cloud Network Services, based on the Nexus 1000V virtual switch, are designed to run across major hypervisors, including VMware vSphere, Microsoft Hyper-V, and Linux Kernel-based Virtual Machine (KVM). It is also designed to support multiple cloud orchestration and virtualization management platforms, such as VMware vCenter and Microsoft Systems Center Virtual Machine Manager (SCVMM), giving customers greater flexibility.

Cisco Cloud Services Platform

With the evolution to Cisco Cloud Network Services as the Layer 4 through 7 framework for virtual and cloud networks, organizations are increasingly looking for a flexible platform on which to deploy virtual service nodes rather than use existing application servers. The Nexus Cloud Services Platform was created to address this need.

The Cisco Nexus 1100 Cloud Services Platform is a group of Cisco Unified Computing System™ (Cisco UCS) appliances dedicated to running Cloud Network Service nodes. In addition to the virtual services listed earlier, the Nexus 1100 series runs the management platforms for the virtual network, the virtual security module (VSM), and the Cisco Data Center Network Manager (DCNM) application. The Cloud Services Platform can be dynamically configured to allocate its virtual CPUs to each service as needed based on current application and performance requirements. Current models of the Nexus 1100 Cloud Services Platform include the Nexus 1110-S and 1110-X.

vPath: Enabling Services in Virtual and Cloud Networks

vPath is a component of the Cisco Nexus 1000V virtual switch that directs traffic to appropriate virtual service nodes, such as firewalls and ADCs, in the correct order for each application, independent of the topology of the network or the location of the network services. This feature allows greater application mobility and more reliable service delivery (Figure 1).

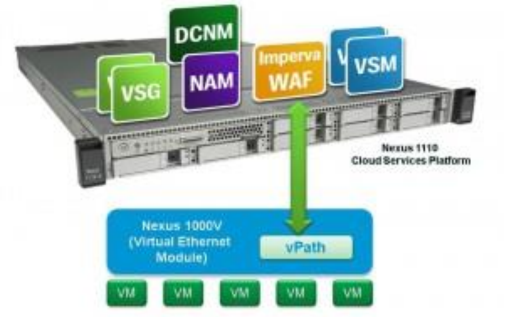


Figure 1 – vPath Connects Virtual Applications to Services Running on the Cisco Cloud Services Platform

Nexus 1000V InterCloud: Enable Hybrid Cloud Connectivity and Cloud-Based Virtual Services

As virtual networks extend from the data center to cloud service providers, organizations are concerned about the consistency of security and application delivery policies and about how these policies are enforced in the cloud. Applications that migrate from the data center to cloud providers can expect different behavior, and organizations may struggle to address compliance issues.

Cisco Nexus 1000V InterCloud complements Cisco Cloud Network Services, allowing seamless hybrid cloud connectivity between data centers and cloud providers and creating one extended network for application and Cloud Network Service deployments.

By deploying Cloud Network Services in all cloud locations, public and private, organizations help ensure consistent policy enforcement, quality of service (QoS) and compliance independent of the location of the virtual applications. Because Cloud Network Services are virtual machines themselves, they are easily deployed within public cloud providers regardless of the infrastructure they provide, and they provide the service consistency required for mission-critical applications.

For More Information

Learn more about Cisco virtual networking portfolio: http://cisco.com/go/1000v

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

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

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

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

About Ipanema Technologies

Application Performance Guarantee

Go Beyond WAN Optimization

About Ipanema

- Selected by worldwide enterprises across all industry sectors.
- One of the largest customer bases (over 150,000 managed sites).
- Visionary in Gartner's WOC Magic Quadrant 2013.
- Leader for Application-Aware Network services (BT, Colt, Vodafone, KDDI, KPN, OBS, Swisscom, Telecom Italia, Telefónica, Easynet…).

79%*

of organizations suffers application performance problems while increasing their IT budget.

*Ipanema Killer Apps survey 2012

Losing 5 minutes

per day for poor application performance means 1% of productivity drop which can turn down profitability by 10%.

"Thanks to Ipanema, our network is totally aligned with our business requirements. With the flexible application-based managed service delivered by e-Qual, we can guarantee the performance of our business critical applications including our ERP and MS Lync, anytime anywhere while reducing our IT costs".

Philippe Faure, Chief Information Officer, Gemalto

Go Beyond WAN Optimization to guarantee your applications performance

Ipanema provides enterprises with a direct connection between application performance and their business requirements.

With Ipanema Technologies, enterprises automatically understand which applications use the network and deliver guaranteed performance to each user. Enterprises can support their strategic IT transformations (like cloud computing and Unified Communications) and control Internet growth while reducing their IT expenses.

Using Ipanema, **enterprises**:

- Guarantee their business application performance;
- Protect unified communications;
- Enable hybrid cloud applications;
- Deploy hybrid networks;
- Control Internet, social media and video traffic;
- Save on IT costs.

ERP, CRM, Financial Reporting
UNIFIED COMMUNICATIONS
CLOUD APPLICATIONS, BIG DATA
INTERNET & HYBRID NETWORKS
INTERNET
MPLS
ETHERNET
VIDEO TRAFFIC & SOCIAL MEDIA
Bring Your Own Device
TRANSPARENT OPERATIONS & IT GOVERNANCE
IT Spending & Efficiency

Ipanema: the only solution that integrates all the features to guarantee application performance

Ipanema's self-learning and self-optimizing Autonomic Networking System™ (ANS) tightly integrates all the features to guarantee the best application performance:

- Application Visibility,
- Application Control,
- WAN Optimization,
- Dynamic WAN Selection
- And Network Rightsizing.

radware

Do You Have Best-in-Class Application Delivery?

Today's data center challenges extend beyond the traditional needs for application availability, performance and security – challenges well-served previously with classic load balancers / application delivery controllers (ADCs). Nowadays, the adoption of data center virtualization, synchronization with dynamic data center changes, true "awareness" of deployed business applications, and the need for end-to-end visibility– all require a new class of advanced (yet cost-effective) ADCs.

Radware **Alteon® 5224** is an advanced ADC specifically targeted to address all of these challenges. Offering the very latest in next generation application delivery technology with ease of operations, it's simply the best-in-class application delivery choice. Here are four reasons why, we know you'll appreciate:

Reason 1: ADC Virtualization and Consolidation

ADC-VX™, part of Radware's Virtual Application Delivery Infrastructure (VADI)™ strategy, is the industry's first ADC virtualization hypervisor, allowing for the most cost-effective ADC consolidation capabilities. ADC-VX is built on a unique architecture that virtualizes the resources of Radware's ADC including CPU, memory, network and acceleration resources. This specialized hypervisor runs virtual ADC instances (vADC) where each delivers full ADC functionality. Each virtual ADC instance contains a complete and separated environment of resources, OS, configurations and management.

In turn, this allows allocation of a separate, fully-isolated vADC instance for each application. Companies can then maximize application availability and meet application SLA requirement with a resource reservation mechanism. Moreover, this deployment model simplifies operations, reduces the ADC infrastructure footprint, and increases business agility with faster roll out of new vADCs and applications. With vADC per application, application lifecycle management is streamlined and its associated cost is significantly reduced compared to traditional ADC deployment models.

Reason 2: Result-Driven Application Acceleration

Radware's FastView™ result-driven acceleration technology adds Web Performance Optimization (WPO) capabilities on top of standard ADC application acceleration features (e.g., caching, compression, SSL acceleration, etc.), to deliver the fastest Web application response time and ensure best application SLA while offloading server processing. This results in increased revenues, higher conversion rates, higher customer loyalty as well as improved employee productivity when using enterprise web applications. It applies to all browsers, all end-user device types and all users, located anywhere. Radware's leading WPO capabilities include:

- Reducing the # of server requests per page
- Accelerate entire web transaction
- Custom optimization templates for each browser
- Static and dynamic, browser-side caching
- Dedicated, mobile caching based on HTML 5 local storage
- Content minification

Reason 3: End-to-End Application QoE & Performance Visibility

Ensuring applications deliver the best quality of experience requires IT administrators to gain maximum visibility on all application delivery chain components, throughout the life cycle of the application. Radware's multilayer approach for monitoring the application delivery infrastructure, coupled with its integrated application

performance monitoring (APM) module, provide a powerful tool to guarantee continuous high application SLA throughout the entire application life cycle, by displaying actual user transactions and errors. The only APM-integrated ADC on the market, the solution enables easy detection and resolution of SLA degradations, while eliminating the need to manually script synthetic transactions. Cross-ADC infrastructure historical reports on resource utilization provide a holistic view enabling better capacity planning when rolling out new applications. In addition, drilldown-able real-time dashboards, that span multiple ADCs, enable instant visibility for spotting problems and a powerful tool for fast and accurate troubleshooting.

Reason 4: Application Awareness with AppShape™ & AppShape™++

Radware's AppShape technology transforms the ADC into a 'smart' device to accelerate, ease and optimize application deployment on the ADC. With Radware's AppShape, each ADC service is tailored to and *aware* of a specific business application (such as SAP, Microsoft, Oracle, IBM and more). In this way, the ADC can be managed from an application-oriented perspective via application specific configuration templates and wizards – resulting in fast application roll-out and simplified application management. Plus, AppShape offers logs and reports for: compliance, per application trends analysis and resources utilization.

Radware's also provides ADC policy scripting capabilities with its AppShape++ technology to further enable the customization of the ADC service per specific application flows and scenarios. By leveraging scripts, examples in Radware's library and dev-community, customers can easily use AppShape++ to refine various layer 4-7 policies including HTTP, HTTPS, TCP, UDP, SSL and more – with no application modifications to further reduce cost and risk.

Simply the Best-in-Class ADC Choice

The combination of these advantages – along with an industry unique 5-year longevity guarantee, "pay-as-you-grow" approach in throughput, # of vADCs and services, plus performance leadership in all layer 4-7 metrics – makes Alteon 5224 simply your best application delivery choice. Want to see for yourself? We invite you to download our Radware ADC Solution white paper here or contact us at: info@radware.com.

DOWNLOAD
FREE TRIAL
TODAY!