The 2012
Application & Service Delivery Handbook

Part 3: Network and Application Optimization

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

Platinum Sponsors:

Gold Sponsors:

Produced by:

Network and Application Optimization

Executive Summary

The ***2012 Application and Service Delivery Handbook*** will be published both in its entirety and in a serial fashion. This is the third of the serial publications. The first publication focused on describing a set of factors, such as chatty protocols, that have traditionally complicated the task of ensuring acceptable application delivery. The second publication described a set of emerging challenges, such as the movement to bring your own device to work, that are beginning to impact the ability of IT organizations to ensure acceptable application and service 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 fourth publication will focus on describing the technologies and services that are available to improve the management and security of applications and services. The fifth and final publication will include an executive summary as well as a copy of the complete document.

A preceding section of The ***2012 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.

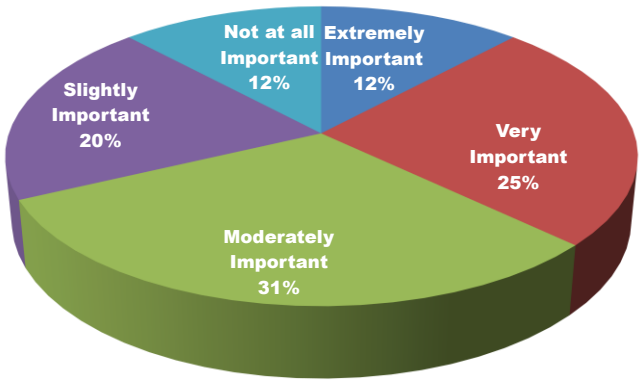
Background

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;
- Direct traffic to the most appropriate server based on a variety of metrics.

The functionality described in the preceding bullets is intended primarily to improve the performance of applications and services. However, as mentioned, another factor driving the use of optimization techniques is the desire to reduce cost. To quantify the impact of that factor, The Survey Respondents were asked to indicate how important it was to their organization over the next year to get better at controlling the cost of the WAN by reducing the amount of WAN traffic by techniques such as compression. Their responses are shown in **Figure 1.**

Figure 1: Importance of Using Optimization to Reduce Cost



The data in **Figure 1** indicates that improving performance is not the only reason why IT organizations implement optimization functionality.

The value proposition of network and application optimization is partly to improve the performance of applications and services and partly to save money.

The Survey Respondents were asked to indicate their company's approach to optimizing network and application optimization. Their responses are shown in **Table 1**.

Table 1: How IT Organizations Approach Network and Application Optimization	
Response	**Percentage**
We implement very little if any functionality specifically to optimize network and application performance	27.4%
We implement optimization functionality on a case-by-case basis in response to high visibility problems	45.7%
We have implemented optimization functionality throughout our environment	21.3%
Other	5.5%

The most common way that IT organizations currently approach implementing optimization functionality is on a case-by-case basis.

The Survey Respondents were given a set of ten viable factors and were asked to indicate the two factors that would likely have the most impact on the evolution of their company's WAN over the next two years. The five factors that were mentioned the most frequently are shown in **Table 2**.

Table 2: Factors Driving WAN Evolution	
Factor	**Percentage of Respondents**
Reduce Cost	34.3%
Improve Application Performance for Business Critical Applications	32.6%
Support video and/or telepresence	20.4%
Support mobile users	18.3%
Provide access to public cloud computing services	17.0%

The data in **Table 2** reflects the responses of all of the 230 IT professionals who responded to the survey. In general, there are only minor differences in the responses of the IT professionals who work for large companies; i.e., 10,000 or more employees. A notable exception to that statement is that whereas the most common factor driving WAN evolution for all companies is reducing cost, which is not the case for large companies. For them it is improving application performance for business critical applications[1].

While historically IT organizations have primarily implemented WAN optimization on a case-by-case basis, that situation is likely to change. One of the key drivers of that change is that as previously explained, the number of business critical applications that the typical business has to support has increased dramatically in the last couple of years. The importance of that driver is enhanced by the fact that, as previously discussed, the most likely impact of poor performance of a business critical application is that the company loses revenue.

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

The growing importance of improving the performance of a growing number of business critical applications is underscored by the data in **Table 2**. That importance will make it increasingly burdensome to implement optimization functionality on a case-by-case basis. In addition, developments that are discussed later in this document, such as the virtualization of WAN Optimization Controllers and the growing deployment of integrated WOCs, will make it easier for IT organization to implement WOC functionality more broadly.

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

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

The role of a WOC is to mitigate the negative effect that the characteristics of WAN services, such as packet loss, have on application and service performance. The affect of packet loss on TCP has been widely analyzed[2]. 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 2: Factors that Impact Throughput

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

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 3** 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.

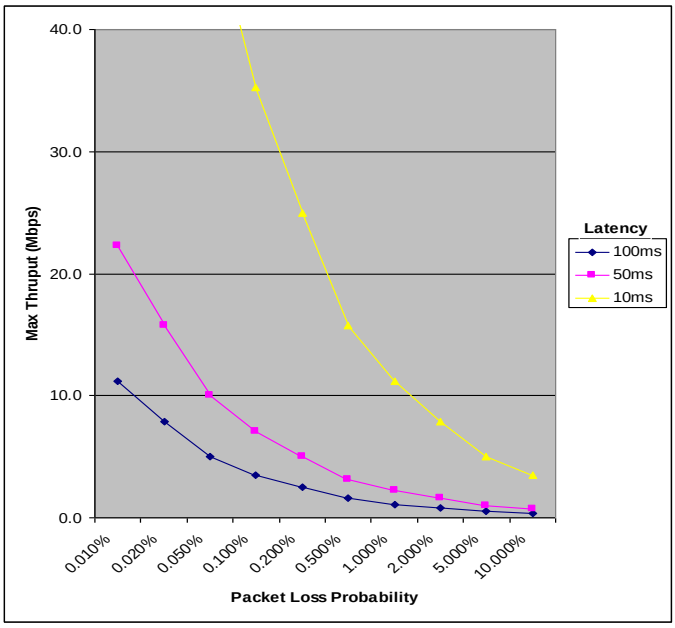
[2] The macroscopic behavior of the TCP congestion avoidance algorithm by Mathis, Semke, Mahdavi & Ott in Computer Communication Review, 27(3), July 1997

One conclusion we can draw from **Figure 3** 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 3: Impact of Packet Loss on Throughput



As described in the next subsection of the handbook, WOCs traditionally focused on accelerating end user traffic between remote branch offices and central data centers. Recently a trend has developed whereby IT organizations use WOCs to accelerate the movement of bulk data between data centers. This includes virtual machine (VM) migrations, storage replication, access to remote storage or cloud storage, and large file transfers.

The Survey Respondents were asked to indicate how important it was to their organization over the next year to get better at optimizing the transfer of storage data between different data centers. Their responses are shown in **Table 3.**

Table 3: The Importance of Optimizing Storage Data	
Extremely Important	13%
Very Important	32%
Moderately Important	27%
Slightly Important	15%
Not at all Important	13%

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

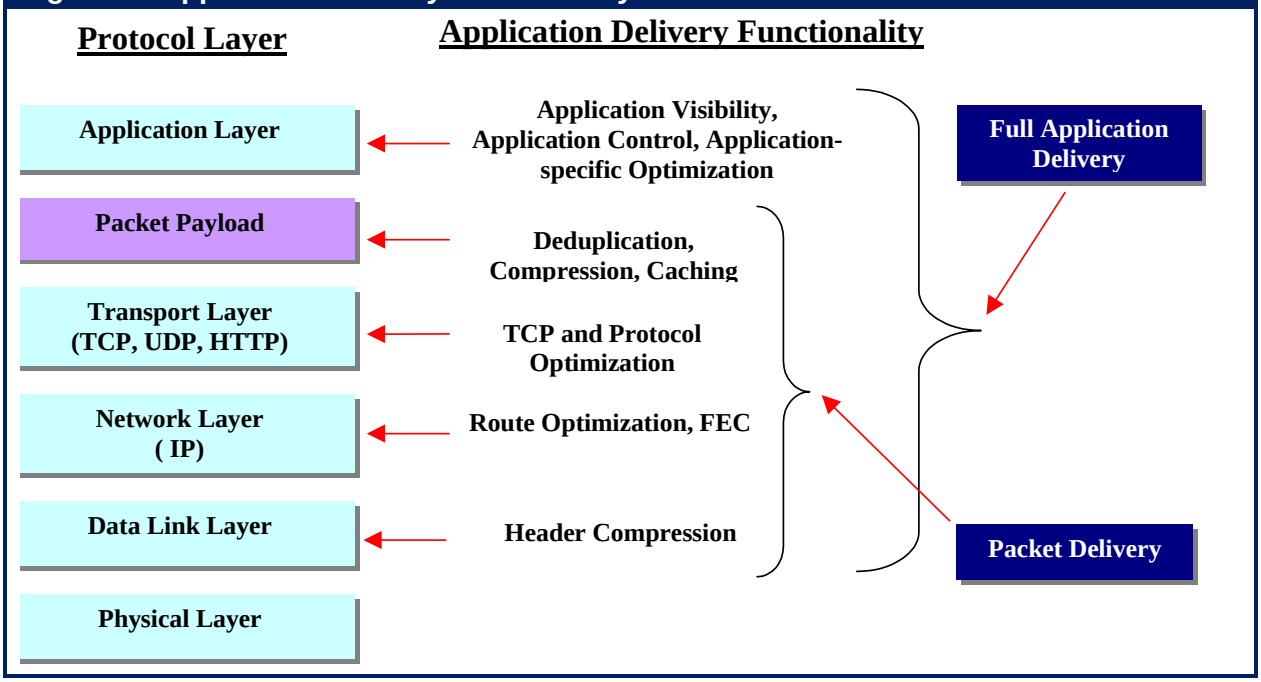
In the vast majority of cases, IT organizations acquire and implement WOCs on a do-it-yourself (DIY) basis. It is also possible for IT organizations to acquire WOC functionality from a managed service provider (MSP). In that scenario, the MSP is responsible for designing, implementing and managing the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

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

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

When WOCs were first deployed they often focused on improving the performance of a protocol such as TCP or CIFS. As discussed in a preceding section of the handbook, optimizing those protocols is still important to the majority of IT organizations. However, as WOCs continue to evolve, much more attention is being paid to the application layer. As shown in **Figure 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



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

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

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

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

Quantifying Application Response Time

A model is helpful to illustrate the potential performance bottlenecks in the performance of an application. The following model (**Figure 5)** is a variation of the application response time model created by Sevcik and Wetzel[3]. 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 5: Application Response Time Model

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

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

[3] Why SAP Performance Needs Help

WAN Optimization Controllers (WOCs)

WOC Functionality

Table 4 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 4: 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

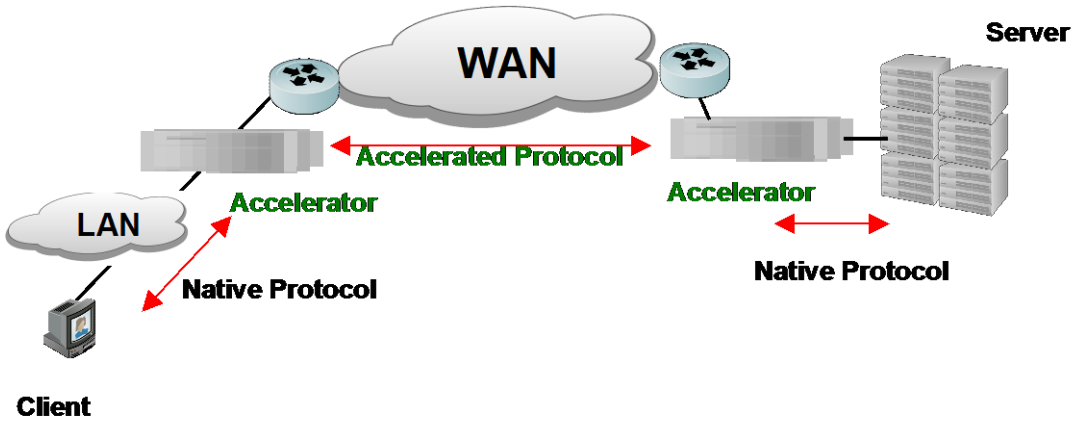
chunks of data at high bandwidths requires a significant amount of memory and processing power.

- ***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[4].

- ***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 6**. 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 6: Protocol Acceleration



  - *TCP Acceleration*
    TCP can be accelerated between appliances with a variety of techniques that increase a session's ability to more fully utilize link bandwidth. Some of these techniques include dynamic scaling of the window size, packet aggregation, selective acknowledgement,

[4] Transmission Control Protocol

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

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

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

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

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

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

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

WOC Form Factors

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

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

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

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

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

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

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

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

To understand the interest that IT organizations have in virtual appliances in general, and virtual WOCs in particular, The Survey Respondents were asked, "Has your organization already implemented, or do you expect that you will implement within the next year, any virtual functionality (e.g., WOC, firewall) in one or more of your branch offices." Just under half responded *yes*. The Survey Respondents that responded *yes* were also given a set of possible IT functionality and asked to indicate the virtual functionality that they have already implemented or that they expected to implement within the next year. Their responses are shown in **Table 5**.

Table 5: Implementation of Virtual Functionality

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

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

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

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

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

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

WOC Selection Criteria

The recommended criteria for evaluating WAN Optimization Controllers are listed in **Table 6**. This list is intended as a fairly complete compilation of all possible criteria, so a given organization may want to apply only a subset of these criteria for a given purchase decision. In addition, individual organizations are expected to ascribe different weights to each of the criteria because of differences in WAN architecture, branch office network design and application mix. Assigning weights to the criteria and relative scores for each solution provides a simple method for comparing competing solutions.

There are many techniques IT organizations can use to complete **Table 6** and then use its contents to compare solutions. For example, the weights can range from 10 points to 50 points, with 10 points meaning not important, 30 points meaning average importance, and 50 points meaning critically important. The score for each criteria can range from 1 to 5, with a 1 meaning fails to meet minimum needs, 3 meaning acceptable, and 5 meaning significantly exceeds requirements.

As an example, consider hypothetical solution A. For this solution, the weighted score for each criterion (WiAi) is found by multiplying the weight (Wi) of each criteria, by the score of each criteria (Ai). The weighted score for each criterion are then summed (Σ WiAi) to get the total score for the solution. This process can then be repeated for additional solutions and the total scores of the solutions can be compared.

Table 6: Criteria for WAN Optimization Solutions

Criterion	Weight Wi	Score for Solution "A" Ai	Score for Solution "B" Bi
Performance			
Transparency			
Solution Architecture			
OSI Layer			
Capability to Perform Application Monitoring			
Scalability			
Cost-Effectiveness			
Module vs. Application Optimization			
Disk vs. RAM-based Compression			
Protocol Support			
Security			
Ease of Deployment and Management			
Change Management			
Bulk Data Transfers			
Support for Meshed Traffic			
Support for Real Time Traffic			
Individual and/or Mobile Clients			
Branch Office Consolidation			
Total Score		**Σ WiAi**	**Σ WiBi**

Each of the criteria contained in **Table 6** is explained below.

- ***Performance***
  Third party tests of an optimization solution can be helpful. It is critical, however, to quantify the kind of performance gains that the solution will provide in the particular environment where it will be installed. For example, if the IT organization is in the process of consolidating servers out of branch offices and into centralized data centers, or has already done so, then it needs to test how well the WAN optimization solution supports CIFS. As part of this quantification, it is important to identify whether the performance degrades as additional functionality within the solution is activated, or as the solution is deployed more broadly across the organization.

A preceding section of the handbook highlighted the fact that the most important optimization task currently facing IT organizations is optimizing a small set of business critical applications. Because of that, IT organizations must test the degree to which a WOC optimizes the performance of those solutions.

- ***Transparency***
  The first rule of networking is not to implement anything that causes the network to break. Therefore, an important criterion when choosing a WOC is that it should be possible to deploy the solution without breaking things such as routing, security, or QoS. The solution should also be transparent relative to both the existing server configurations and the existing Authentication, Authorization and Accounting (AAA) systems, and should not make troubleshooting any more difficult.

- ***Solution Architecture***
  If the organization intends for the solution to support additional optimization functionality over time, it is important to determine whether the hardware and software architecture can support new functionality without an unacceptable loss of performance.

- ***OSI Layer***
  An IT organization can apply many of the optimization techniques discussed in this handbook at various layers of the OSI model. They can apply compression, for example, at the packet layer. The advantage of applying compression at this layer is that it supports all transport protocols and all applications. The disadvantage is that it cannot directly address any issues that occur higher in the stack.

  Alternatively, having an understanding of the semantics of the application means that compression can also be applied to the application; e.g., SAP or Oracle. Applying compression, or other techniques such as request prediction, in this manner has the potential to be highly effective because it can leverage detailed information about how the application performs. However, this approach is by definition application specific and so it might be negatively impacted by changes made to the application.

- ***Capability to Perform or Support Application Monitoring***
  Some WOCs provide significant application monitoring functionality. That functionality might satisfy the monitoring needs of an IT organization. If it does not, it is important that the WOC doesn't interfere with other tools that an IT organization uses for monitoring. For example, many network performance tools rely on network-based traffic statistics gathered from network infrastructure elements at specific points in the network to perform their reporting. By design, all WAN optimization devices apply various optimization techniques on the application packets and hence affect these network-based traffic statistics to varying degrees. One of the important factors that determine the degree of these effects is based on the amount of the original TCP/IP header information retained in the optimized packets.

- ***Scalability***
  One aspect of scalability is the size of the WAN link that can be terminated on the appliance. A more important metric is how much throughput the box can actually support with the desired optimization functionality activated. Other aspects of scalability include how many simultaneous TCP connections the appliance can support, as well as how many branches or users a vendor's complete solution can support. Downward scalability is also important.

Downward scalability refers to the ability of the vendor to offer cost-effective products for small branches or individual laptops and/or wireless devices.

- ***Cost Effectiveness***
  This criterion is related to scalability. In particular, it is important to understand what the initial solution costs, and also to understand how the cost of the solution changes as the scope and scale of the deployment increases.

- ***Module vs. Application Optimization***
  Some WOCs treat each module of an application in the same fashion. Other solutions treat modules based both on the criticality and characteristics of that module. For example, some solutions apply the same optimization techniques to all of SAP, while other solutions would apply different techniques to the individual SAP modules based on factors such as their business importance and latency sensitivity.

- ***Support for Virtualization***
  This criterion includes an evaluation of the support that virtual appliances have for different hypervisors, hypervisor management systems, and VM migration.

- ***Disk vs. RAM***
  Advanced compression solutions can be either disk or RAM-based, or have the ability to provide both options. Disk-based systems can typically store as much as 1,000 times the volume of patterns in their dictionaries as compared with RAM-based systems, and those dictionaries can persist across power failures. The data, however, is slower to access than it would be with the typical RAM-based implementations, although the performance gains of a disk-based system are likely to more than compensate for this extra delay. While disks are more cost effective than a RAM-based solution on a per byte basis, given the size of these systems they do add to the overall cost and introduce additional points of failure to a solution. Standard techniques such as RAID can mitigate the risk associated with these points of failure.

- ***Protocol support***
  Some solutions are specifically designed to support a given protocol (e.g., UDP, TCP, HTTP, Microsoft Print Services, CIFS, MAPI) while other solutions support that protocol generically. In either case, the critical issue is how much of an improvement the solution can offer in the performance of that protocol, in the type of environment in which the solution will be deployed. Also, as previously discussed, the adoption of VDI means that protocols such as ICA, RDP and PCoIP need to be supported. As a result, if VDI is being deployed, WOC performance for remote display protocols should be a significant evaluation criterion.

  In addition to evaluation how a WOC improves the performance of a protocol, it is also important to determine if the WOC makes any modifications to the protocol that could cause unwanted side effects.

- ***Security***
  The solution must be compatible with the current security environment. It must not, for example, break firewall Access Control Lists (ACLs) by hiding TCP header information. In addition, the solution itself must not create any additional security vulnerabilities.

- ***Ease of Deployment and Management***
  As part of deploying a WAN optimization solution, an appliance will be deployed in branch offices that will most likely not have any IT staff. As such, it is important that unskilled personnel can install the solution. In addition, the greater the number of appliances deployed, the more important it is that they are easy to configure and manage.

  It's also important to consider what other systems will have to be modified in order to implement the WAN optimization solution. Some solutions, especially cache-based or WAFS solutions, require that every file server be accessed during implementation.

- ***Change Management***
  As most networks experience periodic changes such as the addition of new sites or new applications, it is important that the WAN optimization solution can adapt to these changes easily – preferably automatically.

- ***Bulk Data Transfers***
  Support for bulk data transfers between branch offices and central data center is a WOC requirement, but in most cases the volume of bulk traffic per branch is quite low compared to the volume of bulk data traffic over WAN links connecting large data centers.

  There are exceptions to the statement that the volume of bulk transfer per branch is small. For example, in those cases in which there are virtualized servers at the branch office that run applications locally, a key benefit of having virtualized the branch office servers is the efficiency it lends to disaster recovery and backup operations. Virtual images of mission critical applications can be maintained at backup data centers or the data centers of providers of public cloud-based backup/recovery services. These images have to transit the WAN in and out of the branch office and can constitute very large file transfers. Client-side application virtualization also involves high volume data transfers from the data center to the remote site.

- ***Support of Meshed Traffic***
  A number of factors are causing a shift in the flow of WAN traffic away from a simple hub-and-spoke pattern to more of a meshed flow. One such factor is the ongoing deployment of VoIP. If a company is making this transition, it is important that the WAN optimization solution it deploys can support meshed traffic flows and can support a range of features such as asymmetric routing.

- ***Support for Real Time Traffic***
  Many companies have deployed real-time applications. For these companies it is important that the WAN optimization solution can support real time traffic. Most real-time applications use UDP, not TCP, as a transport protocol. As a result, they are not significantly addressed by TCP-only acceleration solutions. In addition, the payloads of VoIP and live video packets can't be compressed by the WOC because of the delay sensitive nature of the traffic and the fact that these streams are typically already highly compressed. WOC support for UDP real-time traffic is therefore generally provided in the form of header compression, QoS, and forward error correction. As the WOC performs these functions, it must be able to do so without adding a significant amount of latency.

- ***Individual and/or Mobile Clients***
  As the enterprise workforce continues to become more mobile and more de-centralized, accessing enterprise applications from mobile devices or home offices is becoming a more

common requirement. Accelerating application delivery to these remote users involves a soft WOC or WOC client that is compatible with a range of remote devices, including laptops, PDAs, and smart phones. The WOC client must also be compatible with at least a subset of the functionality offered by the data center WOC. Another issue with WOC clients is whether the software can be integrated with other client software that the enterprise requires to be installed on the remote device. Installation and maintenance of numerous separate pieces of client software on remote devices can become a significant burden for the IT support staff.

- ***Branch Office Platform***
  As previously noted, many enterprises are consolidating servers into a small number of central sites in order to cut costs and to improve the manageability of the branch office IT resources. Another aspect of branch office consolidation is minimizing the number of standalone network devices and hardware appliances in the branch office network. One approach to branch office consolidation is to install a virtualized server at the branch office that provides local services and also supports virtual appliances for various network functions. A variation on this consolidation strategy involves using the WOC as an integrated (or virtualized) platform that supports a local branch office server and possibly other networking functions, such as DNS and/or DHCP. Another variation is to have WOC functionality integrated into the router in the branch office.

Traffic Management and QoS

Traffic Management refers to the ability of the network to provide preferential treatment to certain classes of traffic. It is required in those situations in which bandwidth is scarce, and where there are one or more delay-sensitive, business-critical applications such as VoIP, video or telepresence. Traffic management can be provided by a WOC or alternatively by a router.

To gain insight into the interest that IT organizations have in traffic management and QoS, The Survey Respondents were asked how important it was over the next year for their organization to get better at ensuring acceptable performance for VoIP, traditional video and telepresence. Their responses are shown in **Table 7.**

Table 7: Importance of Optimizing Communications Based Traffic

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

One of the conclusions that can be drawn from the data in **Table 7** is:

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

To ensure that an application receives the required amount of bandwidth, or alternatively does not receive too much bandwidth, the traffic management solution must have application awareness. This often means that the solution needs to have detailed Layer 7 knowledge of the application. This follows because many applications share the same port or hop between ports.

Another important factor in traffic management is the ability to effectively control inbound and outbound traffic. Queuing mechanisms, which form the basis of traditional Quality of Service (QoS) functionality, control bandwidth leaving the network but do not address traffic coming into the network where the bottleneck usually occurs. Technologies such as TCP Rate Control tell the remote servers how fast they can send content providing true bi-directional management.

Some of the key steps in a traffic management process include:

- ***Discovering the Application***
  Application discovery must occur at Layer 7. Information gathered at Layer 4 or lower allows a network manager to assign priority to their Web traffic lower than that of other WAN traffic. Without information gathered at Layer 7, however, network managers are not able manage the company's application to the degree that allows them to assign a higher priority to some Web traffic over other Web traffic.

- ***Profiling the Application***
  Once the application has been discovered, it is necessary to determine the key characteristics of that application.

- ***Quantifying the Impact of the Application***
  As many applications share the same WAN physical or virtual circuit, these applications will tend to interfere with each other. In this step of the process, the degree to which a given application interferes with other applications is identified.

- ***Assigning Appropriate Bandwidth***
  Once the organization has determined the bandwidth requirements and has identified the degree to which a given application interferes with other applications, it may now assign bandwidth to an application. In some cases, it will do this to ensure that the application performs well. In other cases, it will do this primarily to ensure that the application does not interfere with the performance of other applications. Due to the dynamic nature of the network and application environment, it is highly desirable to have the bandwidth assignment be performed dynamically in real time as opposed to using pre-assigned static metrics. In some solutions, it is possible to assign bandwidth relative to a specific application such as SAP. For example, the IT organization might decide to allocate 256 Kbps for SAP traffic. In some other solutions, it is possible to assign bandwidth to a given session. For example, the IT organization could decide to allocate 50 Kbps to each SAP session. The advantage of the latter approach is that it frees the IT organization from having to know how many simultaneous sessions will take place.

Transferring Storage Data

Background

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

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

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

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

The Challenges of Moving Workflows Among Cloud Data Centers

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

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

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

In the case of VMotion, VMware recommends that the network connecting the physical servers involved in a VMotion live transfer to have at least 622 Mbps of bandwidth and no more than 5 ms of end-to-end latency[5] [6]. Another requirement is that the source and destination physical servers need to be on the same Layer 2 virtual LAN (VLAN). For inter-data center VM migration, this means that the Layer 2 network must be extended over the WAN.

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

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

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

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

- ***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

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

WAN. The technologies that are available to provide that mobility are: SAN Extension, Live Storage Migration by the hypervisor, and Storage Replication.

- ***SAN Extension***

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

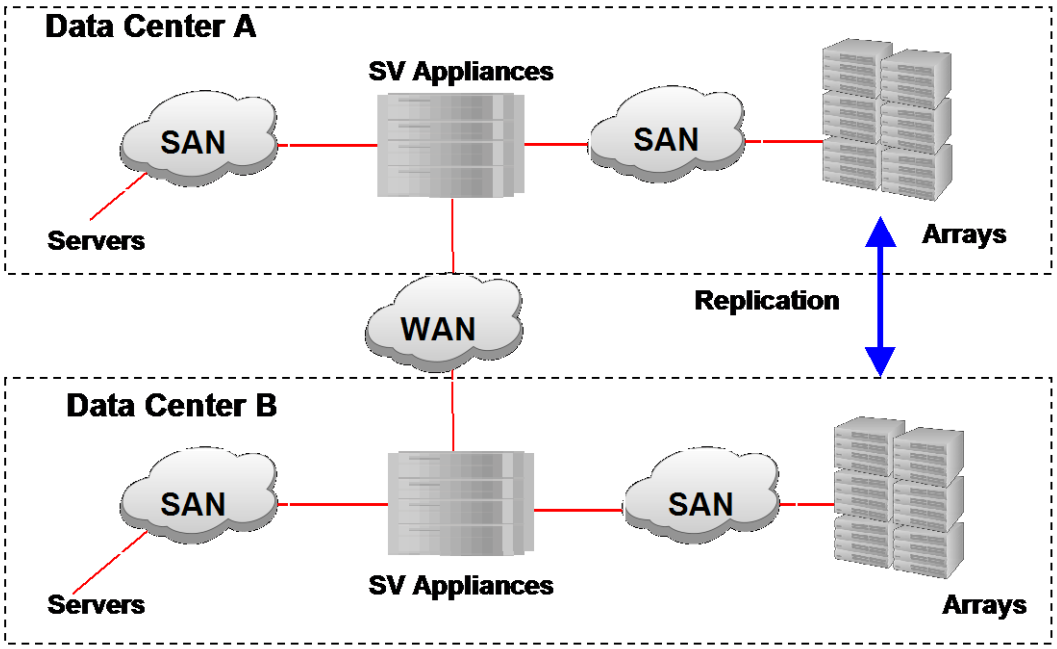
- ***Live Storage Migration***

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

- ***Storage Replication***

  One way to migrate VMs without the delays associated with storage migration's bulk copy operation is to identify the VMs that are likely to need migration and to replicate the vDisks of those VMs at the remote site in anticipation of an eventual VM migration. **Figure 7** 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 7: Storage Replication via Storage Virtualization (SV) Appliances



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

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

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

Resolving the Challenges of Workload Migration

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

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

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

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

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

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

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

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

Trends in WOC Evolution

One of the most significant trends in the WAN optimization market is in the development of functionality that support enterprise IT organizations that are implementing either private cloud strategies or strategies to leverage public and hybrid clouds as extensions of their enterprise data centers. Some recent and anticipated developments include:

- ***Cloud Optimized WOCs***
  This is a purpose-built virtual WOC (vWOC) appliance that was designed with the goal of it being deployed in public and/or hybrid cloud environments. One key feature of this class of device is compatibility with cloud virtualization environments including the relevant hypervisor(s). Other key features include SSL encryption and the acceleration and the automated migration or reconfiguration of vWOCs in conjunction with VM provisioning or migration.

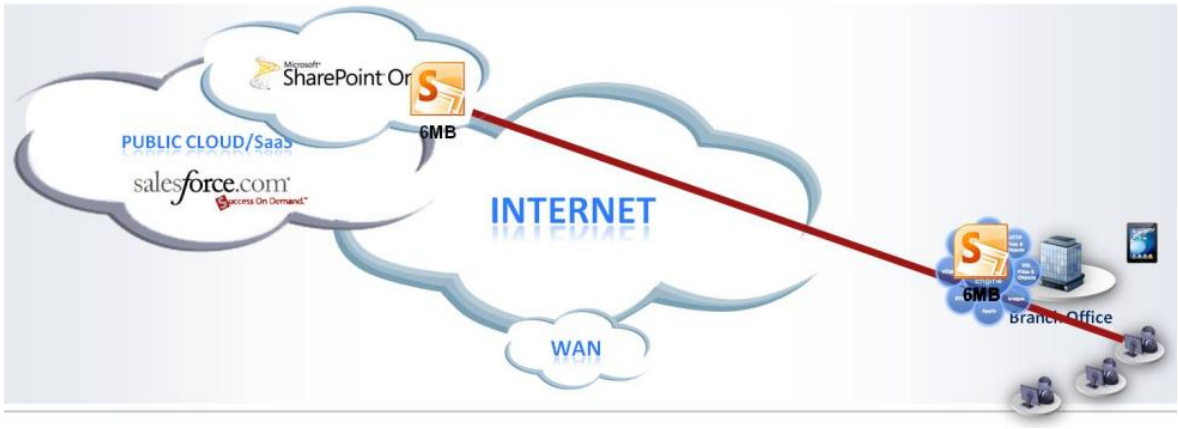
- ***Cloud Storage Optimized WOCs***
  This is a purpose-built virtual or physical WOC appliance that was designed with the goal of it being deployed at a cloud computing site that is used for backup and/or archival storage. Cloud optimized features include support for major backup and archiving tools, sophisticated de-duplication to minimize the data transfer bandwidth and the storage capacity that is required, as well as support for SSL and AES encryption.

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

- ***Asymmetric WOCs***

  Another technique that IT organizations can utilize in those instances in which the CCSP doesn't provide WOC functionality themselves nor do they support vWOC instances being hosted at their data centers is to implement WOCs in an asymmetric fashion. As shown in **Figure 8**, content is downloaded to a WOC in a branch office. Once the content is stored in the WOC's cache for a single user, subsequent users who want to access the same content will experience accelerated application delivery. Caching can be optimized for a range of cloud content, including Web applications, streaming video (e.g., delivered via Flash/RTMP or RTSP) and dynamic Web 2.0 content.

Figure 8: Asymmetric WOC Deployment



- ***IPv6 Application Acceleration***

  Now that the industry has depleted the IPv4 address space, there will be a gradual transition towards IPv6 and mixed IPV4/ IPV6 environments. As applications transition to IPV6 from IPV4, application level optimizations such as those for CIFS, NFS, MAPI, HTTP, and SSL will need to be modified to work in the mixed IPV4/ IPV6 environment. The impact that the adoption of IPv6 has on ADCs will be discussed in detail in the next section of the handbook.

Cloud-Based Optimization Solutions

Background

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

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

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

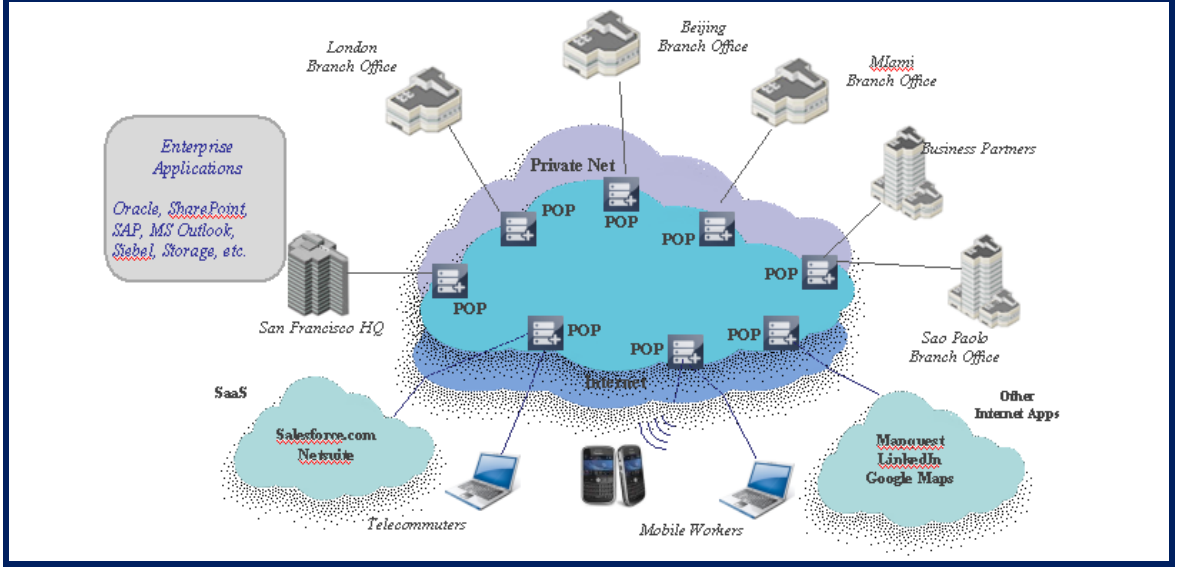
The primary inhibitors are:

- Performance
- Management
- Security

Use Cases

As noted, The Survey Respondents demonstrated significant interest in a network and application optimization solution, such as the one depicted in **Figure 9** that is provided by a CCSP.

Figure 9: WAN and Application Optimization Solution



In **Figure 9**, a variety of types of users (e.g., mobile users, branch office users) access WAN optimization functionality at the service provider's points of presence (POPs). Ideally these POPs are inter-connected by a dedicated, secure and highly available network. To be effective, the solution must have enough POPs so that there is a POP in close proximity to the users. In addition, the solution should support a wide variety of WAN access services. Additional evaluation criteria are described below.

There are at least three distinct use cases for the type of solution shown in **Figure 9**. One such use case is that this type of solution can be leveraged to solve the type of optimization challenges that an IT organization would normally solve by deploying WOCs; e.g., optimizing communications between branch office users and applications in a corporate data center or optimizing data center to data center communications. In this case, the factors that would cause an IT organization to use such a solution are the same factors that drive the use of any public cloud based services; e.g., cost savings, reduce the time it takes to deploy new functionality and provide functionality that the IT organization could not provide itself

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

The third use case for utilizing a solution such as the one shown in **Figure 9** is the expanding requirement that IT organizations have to support access to public cloud services. As previously mentioned, in some instances it is possible for an IT organization to host a soft WOC at an IaaS provider's site. However, that is generally not possible at a SaaS provider's site. In those instances in which it is not possible to host a soft WOC at the CCSP's site, a Cloud-based optimization solution can improve the users access to cloud services by providing to the users the type of functionality typically provided in a WOC.

Evaluating Solutions

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

Evaluating whether or not a given solution provides the required functionality is standard operating procedure for IT organizations. In addition, there is not much difference in terms of how an IT organization would evaluate the functionality provided by a premise based WOC-based solution vs. how it would evaluate the functionality provided by a Cloud-based solution.

What is different about evaluating the later class of solution stems from the fact that they are cloud-based. As such, IT organizations need to closely look at how well the service provider has dealt with the impediments to the use of public cloud computing solutions that were also previously discussed; e.g., the performance of the solution. The performance of a Cloud-based optimization solution is similar to the performance of a standard WOC-based solution – it will vary somewhat based on the requirements of each IT organization. Hence, as was the case with WOC-based solutions, the best way to understand the performance gains that result from using a Cloud-based optimization solution is to test that solution with production traffic.

However, just as important as whether or not the Cloud-based optimization solution mitigates the issues that IT organizations have with public cloud based solutions is whether or not the solution actually provides the benefits (e.g., cost savings) that drive IT organizations to use public cloud computing solutions. While it can be a little tricky to compare the usage sensitive pricing of a Cloud-based optimization solution with the fully loaded cost of a premise based WOC solution[7], the cost information that the IT organization receives from the solution provider should enable the IT organization to do the requisite analysis. The key financial advantages of a Cloud-based solution are that it enables IT organizations to avoid the CAPEX costs that are associated with a typical WOC based solution and also enables IT organizations to migrate away from expensive WAN services such as MPLS.

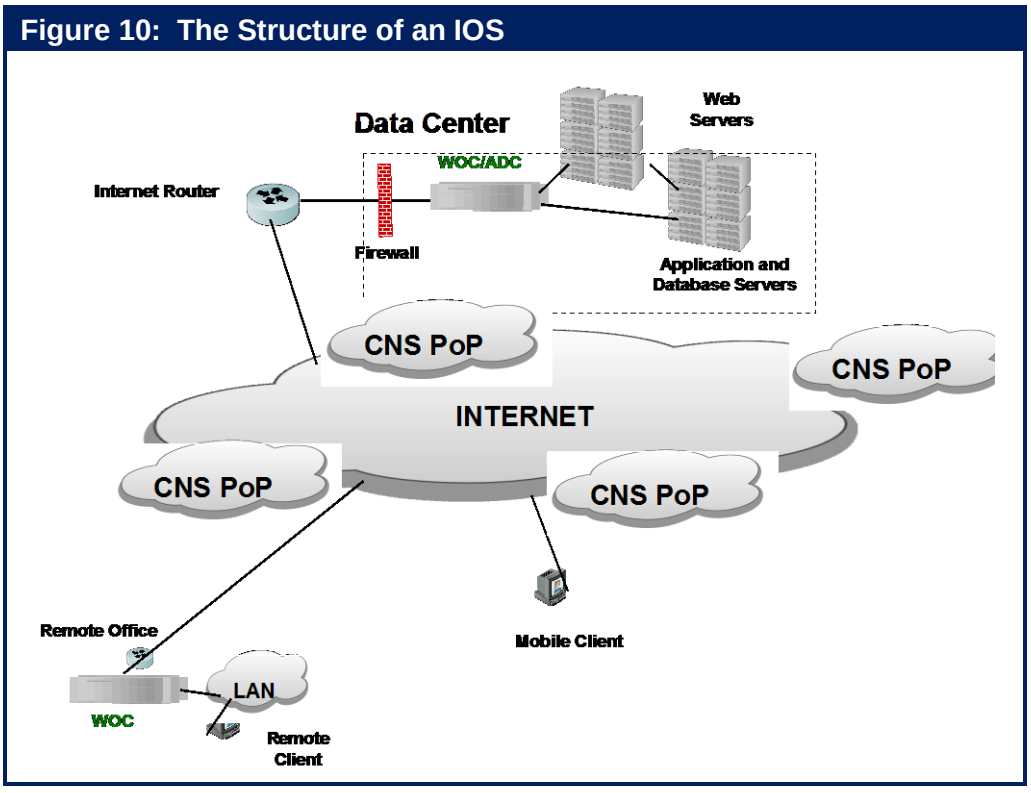
[7] The tricky part is determining the totality of the labor costs associated with the premise based solution.

The Optimization of Internet Traffic

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

An IOS would, out of necessity, leverage service provider resources that are distributed throughout the Internet in order to optimize the performance, security, reliability, and visibility of the enterprise's Internet traffic. As shown in **Figure 10**, 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 10: The Structure of an IOS



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

- ***Route Optimization***
  Route optimization is a technique for circumventing the previously discussed limitations of BGP by dynamically optimizing the round trip time between each end user and the application server. A route optimization solution leverages the intelligence of the IOS servers

that are deployed in the service provider's PoPs to measure the performance of multiple paths through the Internet and to choose the optimum path from origin to destination. The selected route factors in the degree of congestion, traffic load, and availability on each potential path to provide the lowest possible latency and packet loss for each user session.

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

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

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

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

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

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

Visibility and Security

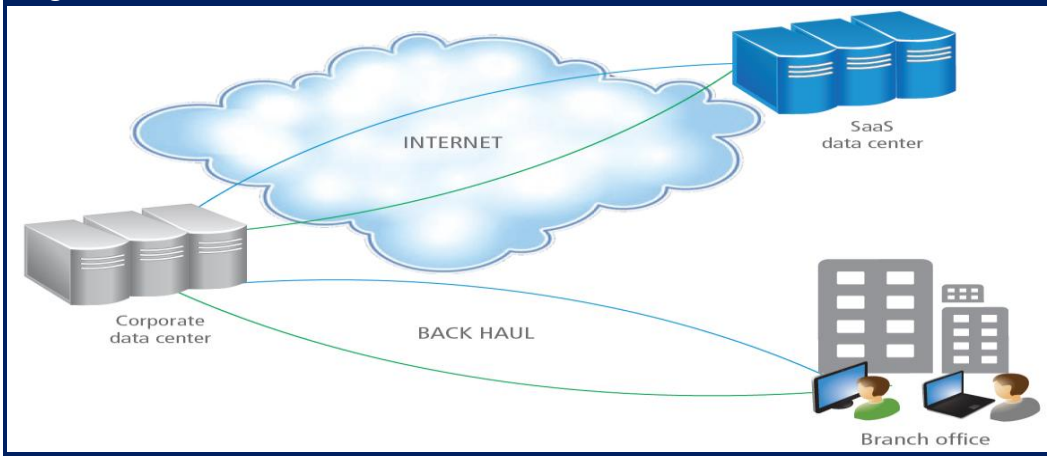
Intelligence within the IOS servers can also be leveraged to provide extensive monitoring, configuration control and SLA monitoring of a subscriber's application with performance metrics, analysis, and alerts made visible to the subscriber via a Web portal.

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

Hybrid WAN Optimization

As shown in **Figure 11**, the traditional approach to providing Internet access to branch office employees has been to backhaul that Internet traffic on the organization's enterprise network (e.g., their MPLS network) to a central site where the traffic was handed off to the Internet. The advantage of this approach is that it enables IT organizations to exert more control over their Internet traffic and it simplifies management in part because it centralizes the complexity of implementing and managing security policy. One disadvantage of this approach is that it results in extra traffic transiting the enterprise's WAN, which adds to the cost of the WAN. Another disadvantage of this approach is that it usually adds additional delay to the Internet traffic.

Figure 11: Backhauled Internet Traffic



The survey respondents were asked to indicate how they currently route their Internet traffic and how that is likely to change over the next year. Their responses are contained in **Table 8.**

Table 8: Routing of Internet Traffic

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

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

Because backhauling Internet traffic adds delay, one of the disadvantages of this approach to providing Internet access is degraded performance. For example, in the scenario depicted in **Figure 11** (Backhauled Internet Traffic), the delay between users in a branch office and the SaaS application is the sum of the delay in the enterprise WAN plus the delay in the Internet. In order to improve performance, an IT organization might use WOCs to optimize the performance of the traffic as it flows from the branch office to the central site over their enterprise WAN. However, once the traffic is handed off to the Internet, the traffic is not optimized and the organization gets little value out of optimizing the traffic as it flows over just the enterprise WAN.

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

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

Application Delivery Controllers (ADCs)

Background

As was mentioned earlier in this section, an historical precedent exists to the current generation of ADCs. That precedent is the Front End Processor (FEP) that was introduced in the late 1960s and was developed and deployed to support mainframe computing. From a more contemporary perspective, the current generation of ADCs evolved from the earlier generations of Server Load Balancers (SLBs) that were deployed to balance the load over a server farm.

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

An ADC provides more sophisticated functionality than a SLB does.

The deployment of an SLB enables an IT organization to get a *linear benefit* out of its servers. That means that if an IT organization that has implemented an SLB doubles the number of servers supported by that SLB that it should be able to roughly double the number of transactions that it supports. The traffic at most Web sites, however, is not growing at a linear rate, but at an exponential rate. To exemplify the type of problem this creates, assume that the traffic at a hypothetical company's (Acme) Web site doubles every year[8]. If Acme's IT organization has deployed a linear solution, such as an SLB, after three years it will have to deploy eight times as many servers as it originally had in order to support the increased traffic. However, if Acme's IT organization were to deploy an effective ADC then after three years it would still have to increase the number of servers it supports, but only by a factor of two or three – not a factor of eight. The phrase ***effective ADC*** refers to the ability of an ADC to have all features turned on and still support the peak traffic load.

ADC Functionality

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

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

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

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

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. The resulting data can be used to support SLAs for guaranteed user response times, guide remedial action and plan for the additional capacity that is required in order to maintain service levels.

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

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

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

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

ADC Selection Criteria

ADC evaluation criteria are listed in **Table 9**. **Figure 4** is intended to describe standard ADC functionality. Subsequent subsections describe in detail how to evaluate an ADCs ability to enable a migration to IPv6 and how to characterize the varying ways to virtualize an ADC. As was the case with WOCs, this list is intended as a fairly complete compilation of possible criteria. As a result, a given organization or enterprise might apply only a subset of these criteria for a given purchase decision.

Table 9: Criteria for Evaluating ADCs

Criterion	Weight Wi	Score for Solution "A" Ai	Score for Solution "B" Bi
Features			
Performance			
Scalability			
Transparency and Integration			
Solution Architecture			
Functional Integration			
Virtualization			
Security			
Application Availability			
Cost-Effectiveness			
Ease of Deployment and Management			
Business Intelligence			
Total Score		Σ WiAi	Σ WiBi

Each of the criteria is described below.

- ***Features***
  ADCs support a wide range of functionality including TCP optimization, HTTP multiplexing, caching, Web compression, image compression as well as bandwidth management and traffic shaping. When choosing an ADC, IT organizations obviously need to understand the features that it supports. However, as this class of product continues to mature, the distinction between the features provided by competing products is lessening. This means that when choosing an ADC, IT organizations should also pay attention to the ability of the ADC to have all features turned on and still support the peak traffic load.

- ***Performance***
  Performance is an important criterion for any piece of networking equipment, but it is critical for a device such as an ADC because data centers are central points of aggregation. As such, the ADC needs to be able to support the extremely high volumes of traffic transmitted to and from servers in data centers.

  A simple definition of performance is how many bits per second the device can support. While this is extremely important, in the case of ADCs other key measures of performance

include how many Layer 4 connections can be supported as well as how many Layer 4 setups and teardowns can be supported.

As is the case with WOCs, third party tests of a solution can be helpful. It is critical, however, to quantify the kind of performance gains that the solution will provide in the particular production application environment where it will be installed. As noted above, an important part of these trails is to identify any performance degradation that may occur as the full suite of desired features and functions are activated or as changes are made to the application mix within the data center.

- ***Transparency and Integration***
  Transparency is an important criterion for any piece of networking equipment. However, unlike proprietary branch office optimization solutions, ADCs are standards based, and thus inclined to be more transparent than other classes of networking equipment. That said, it is still very important to be able to deploy an ADC and not break anything such as routing, security, or QoS. The solution should also be as transparent as possible relative to both the existing server configurations and the existing security domains, and should not make troubleshooting any more difficult.

  The ADC also should be able to easily integrate with other components of the data center, such as the firewalls and other appliances that may be deployed to provide application services. In some data centers, it may be important to integrate the Layer 2 and Layer 3 access switches with the ADC and firewalls so that all that application intelligence, application acceleration, application security and server offloading are applied at a single point in the data center network.

- ***Scalability***
  Scalability of an ADC solution implies the availability of a range of products that span the performance and cost requirements of a variety of data center environments. Performance requirements for accessing data center applications and data resources are usually characterized in terms of both the aggregate throughput of the ADC and the number of simultaneous application sessions that can be supported. As noted, a related consideration is how device performance is affected as additional functionality is enabled.

- ***Solution Architecture***
  Taken together, scalability and solution architecture identify the ability of the solution to support a range of implementations and to be able to be extended to support additional functionality. In particular, if the organization intends the ADC to support additional optimization functionality over time, it is important to determine if the hardware and software architecture can support new functionality without an unacceptable loss of performance and without unacceptable downtime.

- ***Functional Integration***
  Many data center environments have begun programs to reduce overall complexity by consolidating both the servers and the network infrastructure. An ADC solution can contribute significantly to network consolidation by supporting a wide range of application-aware functions that transcend basic server load balancing and content switching. Extensive functional integration reduces the complexity of the network by minimizing the number of separate boxes and user interfaces that must be navigated by data center managers and

administrators. Reduced complexity generally translates to lower TCO and higher availability.

As functional integration continues to evolve, the traditional ADC can begin to assume a broader service delivery role in enterprise data center by incorporating additional functions, such as global server load balancing (GSLB), inter-data center WAN optimization, multi-site identity/access management and enhanced application visibility functions.

- ***Virtualization***
  Virtualization has become a key technology for realizing data center consolidation and its related benefits. The degree of integration of an ADC's configuration management capabilities with the rest of the solution for managing the virtualized environment may be an important selection criterion. For example, it is important to know how the ADC interfaces with the management system of whatever hypervisors that the IT organization currently supports, or expects to support in the near term. With proper integration, vADCs can be managed along with VMs by the hypervisor management console. It is also important to know how the ADC supports the creation and movement of VMs within a dynamic production environment. One option is to pre-provision VMs as members of ADC server pools. For dynamic VM provisioning data center orchestration functionality, based on plug-ins or APIs can automatically add new VMs to resource pools.

  The preceding section of the handbook entitled "Virtualization" described one way of virtualizing an ADC. That was as a virtual appliance in which the ADC software runs in a VM. Partitioning a single physical ADC into a number of logical ADCs or ADC contexts is another way to virtualize an ADC. Each logical ADC can be configured individually to meet the server-load balancing, acceleration and security requirements of a single application or a cluster of applications. A third way that an ADC can be virtualized is that two or more ADCs can be made to appear to be one larger ADC.

- ***Security***
  The ADC must be compatible with the current security environment, while also allowing the configuration of application-specific security features that complement general purpose security measures, such as firewalls and IDS and IPS appliances. In addition, the solution itself must not create any additional security vulnerabilities. Security functionality that IT organizations should look for in an ADC includes protection against denial of service attacks, integrated intrusion protection, protection against SSL attacks and sophisticated reporting.

- ***Application Availability***
  The availability of enterprise applications is typically a very high priority. Since the ADC is in line with the Web servers and other application servers, a traditional approach to defining application availability is to make sure that the ADC is capable of supporting redundant, high availability configurations that feature automated fail-over among the redundant devices. While this is clearly important, there are other dimensions to application availability. For example, an architecture that enables scalability through the use of software license upgrades tends to minimize the application downtime that is associated with hardware-centric capacity upgrades.

- ***Cost Effectiveness***
  This criterion is related to scalability. In particular, it is important not only to understand what the initial solution costs, it is also important to understand how the cost of the solution changes as the scope and scale of the deployment increases.

- ***Ease of Deployment and Management***
  As with any component of the network or the data center, an ADC solution should be relatively easy to deploy and manage. It should also be relatively easy to deploy and manage new applications -- so ease of configuration management is a particularly important consideration in those instances in which a wide diversity of applications is supported by the data center.

- ***Business Intelligence***
  In addition to traditional network functionality, some ADCs also provide data that can be used to provide business level functionality. In particular, data gathered by an ADC can feed security information and event monitoring, fraud management, business intelligence, business process management and Web analytics.

IPv6 and ADCs

Background

June 6th, 2012 was World IPv6 Launch day (http://www.worldipv6launch.org/) and IPv6 is now a permanent part of the Internet. While it won't happen for several years, IPv6 will replace IPv4 and the entire Internet will be IPv6 only. Gartner, Inc. estimates that 17% of the global Internet users and 28% of new Internet connections will use IPv6 by 2015.[9] 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 4.8×10^{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.

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

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

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

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

- **Tunneling** – Transporting IPv6 traffic in IPv4 areas and vice versa.
- **Network Address Translation** – Translating between IPv4 and IPv6 addresses, including DNS support.
- **Dual Stack** – Both IPv4 and IPv6 packets are processed by devices simultaneously.

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

Enabling Standards and Technologies

IPv6/IPv4 Tunneling

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

- **6rd**– Mostly used during initial IPv6 deployment, this protocol allows IPv6 to be transmitted over an IPv4 network without having to configure explicit tunnels. 6rd or "IPv6 **R**apid **D**eployment", is a modification to 6to4 that allows it to be deployed within a single ISP.
- **6in4**–Tunnels are usually manually created and use minimal packet overhead (20 bytes) to minimize packet fragmentation on IPv4 Networks.
- **Teredo**–Encapsulates IPv6 traffic in IPv4 UDP packets for tunneling. Use of UDP allows support of IPv4 Network Address Translation (NAT44 or NAT444) when carrying the IPv6 traffic. This is similar to encapsulating IPSec traffic in UDP to support NAT devices for remote access VPNs.

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

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

Network Address Translation (NAT)

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

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

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

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

Carrier Grade NAT (CGN)

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

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

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

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

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

Dual Stack

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

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

Creating an IPv6 Presence and Supporting Mobile IPv6 Employees

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Once the data center perimeter firewall supports IPv6, attention can now turn to Application Delivery Controllers (ADCs) that provide load balancing, SSL offloading, WAN optimization, etc. When establishing an IPv6 presence for customers, business partners and suppliers, there are architectures with two or more data centers that benefit from IPv6 ADCs with WAN optimization. ADCs can have the following IPv6 capabilities[10]:

- 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

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

data center perimeter firewall for IPv6 does not provide the scalability nor the routing nor security benefits of IPv6.

Supporting Areas

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

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

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

Virtual ADCs

Background

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

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

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

Computers can have more than one CPU that shares memory and I/O ports on a machine and most operating systems can take advantage of multiple CPUs by controlling access to memory blocks with semaphores. Computers with multiple CPUs – sometimes referred to as cores – that share memory and I/O ports are called tightly coupled computing systems. Computers that

do not share memory nor I/O ports but which are interconnected by high-speed communications are called loosely coupled. Several CPUs running appropriate operating systems can cooperate together to form a loosely coupled cluster of CPUs and appear as a single computer. Similarly, hypervisors used for VM technology can take advantage of multiple CPU systems in either tightly coupled or loosely coupled arrangement.

VM technology has many benefits including:

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

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

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

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

The Evolution of Network Appliances

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

As shown in **Figure 4**, as network functionality grew, the network evolved from a *packet delivery* service to an *application and service delivery* service. Network appliances evolved from general purpose servers to become the standard building block of the Application and Service Delivery Network. Network appliances improved upon server technology in two important ways. First, the O/S was changed from a general purpose O/S to one optimized for network operations and processing. Second, the server hardware was updated to include specialized co-processors (e.g. SSL operations and encryption) and network adapters for high performance network operations. This simplified IT operations, as typically only one IT group (e.g. Networks Operations) was involved in changes as opposed to two IT groups (e.g., Network Operations and Server Operations). In general, software updates and security patches are less frequent on network appliances than for general purpose O/Ss and this further reduces the IT operations effort.

Virtualization and Cloud Computing technology challenged network appliances in two important ways and this resulted in a split evolutionary path of the network appliance. The rise of public cloud offerings caused network equipment manufacturers to update their specialized network

appliance operating systems to run under general-purpose hypervisors in CCSP locations. This allowed CCSPs to run specialized network and security functions on their low cost, virtualized server infrastructure filling a much needed functionality gap for their offerings.

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

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

The Types of ADC Virtualization

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

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

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

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

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

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

levels, but provide the least flexibility with limited co-resident applications and virtualization framework support.

High Availability and Hardware Options

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

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

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

Trends in ADC Evolution

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

- ***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 under which ADCs products the CCSP supports and what virtualization management frameworks the CCSP uses.

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

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

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

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

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

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

Perhaps the biggest factor from the above list in developing your ADC strategy is the use of Cloud Computing. Using a CCSP or other outsourcing constrains your ADC options and this helps narrow the field of choices. If your CCSP choice is established and will not change, then

you are constrained to use the ADC products and technologies supported by the CCSP. If you are or will use a hybrid cloud or cloud bursting arrangement, the CCSP's ADC choices can also constrain the ADC choices in the private data center. With a hybrid or cloud bursting approach, you may also be constrained to certain virtualization management frameworks, which in turn will influence your ADC choice.

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

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

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

- Feature Parity between Network Appliance, Virtualized Network Appliance and Virtual products.
- Number of processors and speeds available for network appliance models. Consider any encryption coprocessors and bandwidth (NIC card) partitioning capabilities as well.
- Availability of chassis hardware for scaling and speeds between blades in the chassis as well as external speeds between chassis.
- Ability to virtualize across network appliances, network hardware chassis and virtual instances both locally and across WAN links.
- Aggregate scaling with network appliances, chassis and virtual instances.
- Completeness and flexibility of IPv6 support.
- Ability to support hybrid and cloud bursting deployments
- Flexibility to integrate with virtualization management frameworks including VMware vCenter, Citrix's Xencenter and Microsoft's System Center.
- Overall functionality including load balancing, load detection flexibility, SSL offloading, security processing, proxy support, TCP optimization, WAN Optimization and reporting.

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

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.

Customer Driven Innovation

AX Series Application Delivery and Advanced Server Load Balancing

Flexibility to Solve Critical Business Challenges

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

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

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

Any App

Web Scalability and Availability

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

IPv4 Exhaustion and IPv6 Migration

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

Any Cloud

Enterprises, Web Giants, Service Providers

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

Multi-tenancy and Virtual Clustering

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

Any Size

On-demand Virtual Appliances

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

Scalable and Faster Appliances

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

Contact us

Contact us today to discuss how A10's AX Series Application Networking platforms can solve critical business challenges within your mission-critical IT infrastructure: for any app, any cloud or any size.

Tel: +1 888 A10-6363 | inquire@a10networks.com | www.a10networks.com

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

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

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

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

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

The "Cloud" Defined,
WAN Architecture Redefined

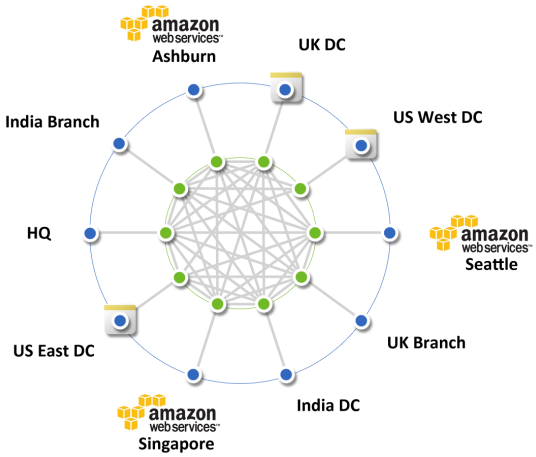
The term "cloud" is intriguing and varied in its description. Vendors within the WAN optimization space and other service providers are trying to find a way to optimize access to the cloud. The only way they can achieve this is by installing another appliance where possible – a virtual appliance – in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

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

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

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

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



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

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

By SONAL PURI

ABOUT ARYAKA

Aryaka is the world's first cloud-based WAN optimization as-a-Service company solving application and network performance issues faced by the distributed enterprise. Aryaka has been named to the Dow Jones VentureWire FASTech 50 innovative startups for 2011, a "Cool Vendor" by a leading analyst firm and a GigaOM Structure 50 company that will shape the future of cloud computing. Aryaka eliminates the need for expensive and complex WAN optimization appliances as well as long-haul connectivity, and enhances collaboration across corporate locations, data centers and cloud services. It offers significant cost benefits, ease-of-use, instant deployment, performance advantages, dramatic productivity gains and real-time insight into WAN applications, locations and performance while providing 24/7 world-class support. To learn more, visit www.aryaka.com. Follow us at Twitter, Facebook, YouTube and on LinkedIn.

Aryaka
691 S. Milpitas Blvd.
Milpitas, CA 95035
Tel: 1-877-727-9252
www.aryaka.com

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

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

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

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

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

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

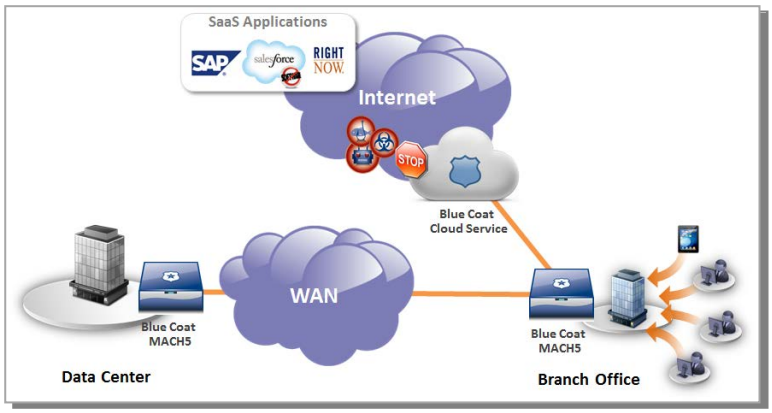
Blue Coat Cloud Service allows you to provide the same enterprise policies and technology to branch and mobile users. By leveraging Blue Coat WebPulse™, a collaborative defense powered by a global community of 75 million users, the Cloud Service is able to deliver real-time protection against the latest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

WebPulse is based on sound analysis-system design principles:

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

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

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



Second: Optimize Performance

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

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

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

If this is you... We need to talk!

- ☐ Require maximum application performance
- ☐ Planning to move applications into a cloud
- ☐ Virtualizing your Applications and Storage
- ☐ Backups or replications don't complete overnight
- ☐ Need affordable acceleration for SOHO & remote users
- ☐ Need WAN Opp for any hardware platform or hypervisor

aCelera™

Get the WAN Optimization solution with the "Strongest Virtualized Architecture" *

Download for yourself: info.certeon.com/certeon-marketplace/

Request a Demo: www.certeon.com/demo

Certeon aCelera software - accelerated access for ANY User, ANY Application, ANY Network, ANY Device.

Deploy in any mix of hardware, virtualization platforms, storage technologies, networking equipment and service providers. Supporting any custom or off the shelf application.

www.certeon.com | 781 425 5200 | 5 Wall Street, Burlington, MA 01803

certeon
Accelerate & Broaden Application Access

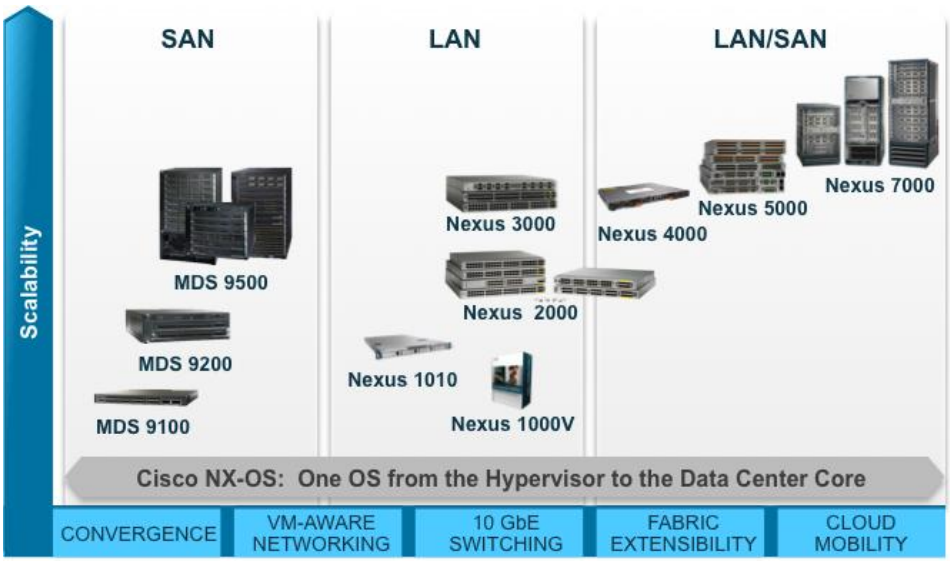
* Enterprise Management Associates

Cisco Unified Fabric

Converged. Scalable. Intelligent.

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

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



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

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

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

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

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

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

Application Performance for Business Efficiency

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

Ipanema Technologies – Fact Sheet 2012

Business Overview

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

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

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

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

Ipanema at a Glance

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

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

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

Solution Overview

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

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

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

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

Business efficiency requires guaranteed application performance

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

With Ipanema, control all your IT transformations

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

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

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

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

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

What our customer say about us

Do more with less

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

Guarantee Unified Communications and increase network capacity

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

Reduce costs in a cloud environment

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

ABOUT IPANEMA TECHNOLOGIES

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

For more information www.ipanematech.com

www.ipanematech.com

Do You Have the Best Choice in Application Delivery?

Overview

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

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

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

Reason 1: Unmatched OnDemand Scalability

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

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

Reason 2: Highest Performance in Class

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

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

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

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

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

Reason 4: Virtualization Ready for Any Enterprise Size

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

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

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

Simply Your Best Application Delivery Choice

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