The 2011
Application & Service Delivery Handbook

Part 4: Network and Application Optimization

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

Platinum Sponsors:

exinda. NETSCOUT. zeus

BlueCoat ipanema Technologies

Gold Sponsors:

A10 Networks Akamai certeon Accelerate Your Business CISCO

CITRIX EXPAND networks VIRTUALLY EVERYWHERE radware riverbed

Produced by: Webtorials

Network and Application Optimization

Executive Summary

The *2011 Application and Service Delivery Handbook* will be published both in its entirety and in a serial fashion. This is the fourth of the serial publications. One of the focus areas of this section of the handbook is the continuation of the discussion of how IT organizations are approaching optimization that was started in a previous section of the handbook. This section will discuss the two primary pieces of functionality that IT organizations can utilize to optimize the performance of applications and services: WAN Optimization Controllers (WOCs) and Application Delivery Controllers (ADCs). Included in that discussion will be the identification of criteria that IT organizations can use to evaluate WOCs and ADCs.

This section will also discuss the growing implementation of virtualized WOCs and ADCs and the development of a new class of optimization device – the data mobility controller (DMC). DMCs are designed to support the growing need to move massive volumes of data between data centers. The section also includes a discussion of the trends in the evolution of WOCs, ADCs and DMCs.

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

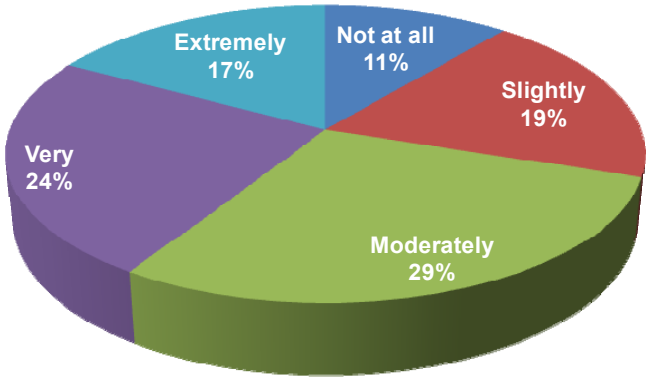
Introduction

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 as part of assuring acceptable application performance 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;
- 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, 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.

As described in a previous section of the handbook, some optimizations tasks, such as optimizing the performance of a key set of business critical applications, has become extremely important to the vast majority of IT organizations. Because of that importance, 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%

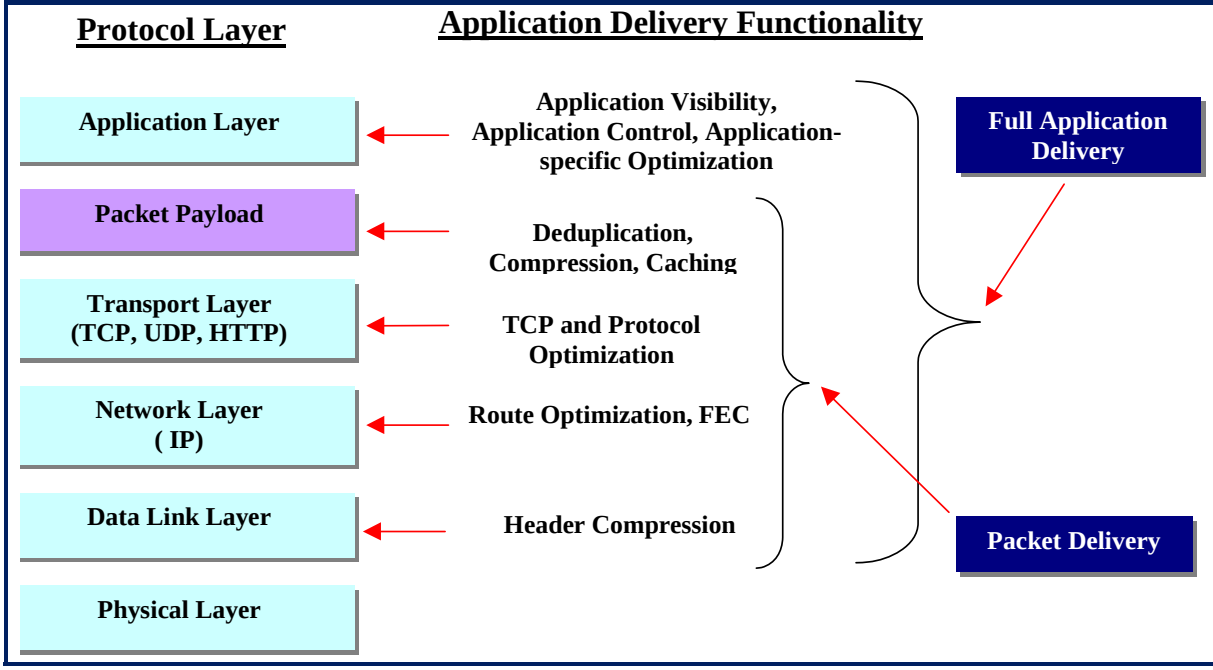
The most common way that IT organizations approach implementing optimization functionality is on a case-by-case basis.

As was previously explained in handbook, some Cloud Computing Service Providers (CCSPs) offer network and application optimization as a service. It is also possible for an IT organization to acquire and implement network and application optimization products such as WOCs and ADCs. In many cases, these two approaches are complimentary.

There are two principal categories of network and application optimization products. One category focuses on mitigating the negative effect that WAN services such as MPLS have on application and service performance. This category of products has historically included WAN optimization controllers (WOCs). However, due to some of the second generation of application and service delivery challenges, this category of products now also contains an emerging class of WAN optimization device - the Data Mobility Controller (DMC). As described in detail later in this section of the handbook, WOCs are focused primarily on accelerating end user traffic between remote branch offices and central data centers. In contrast, DMCs are focused on accelerating 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. WOCs and DMCs are often referred to as *symmetric solutions* because they typically require complementary functionality at both ends of the connection. However, as is explained 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 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 2, WOC solutions that leverage application layer functionality focus on recognizing application layer signatures of end user applications and mitigating application-specific inefficiencies in communicating over the WAN. In contrast to WOCs, DMCs are focused primarily on the packet delivery functionality at the transport layer and with the packet payload. However, like the WOC, the DMC can use QoS functionality at the transport and application layers to classify traffic and identify traffic that requires optimization.

Figure 2: Application Delivery Functionality



In order to choose the most appropriate optimization solution, IT organizations need to understand their environment, including traffic volumes and the characteristics of the traffic they wish to accelerate. For example, the degree of data reduction experienced 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 live environment.

Quantifying Application Response Time

A model is helpful to illustrate the potential performance bottlenecks in the performance of an application. The following model (Figure 3) is a variation of the application response time model created by Sevcik and Wetzel[1]. Like all models, the following is only an approximation and as a result is not intended to provide results that are accurate to the millisecond level. It is, however, intended to provide insight into the key factors impacting application response time. As shown below, the application response time (R) is impacted by a number of factors including the amount of data being transmitted (Payload), the goodput which is the actual throughput on a WAN link, the network round trip time (RTT), the number of application turns (AppTurns), the number of simultaneous TCP sessions (concurrent requests), the server side delay (Cs) and the client side delay (Cc).

Figure 3: Application Response Time Model

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

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

[1] Why SAP Performance Needs Help

WAN Optimization Controllers

The goal of a WOC is to improve the performance of applications delivered from the data center to the branch office or directly to the end user. The Survey Respondents were asked to indicate their company's current deployment of WOCs. They were also asked whether or not they have currently deployed WOCs, to indicate their company's planned deployment of WOCs over the next year. Their responses are shown in Table 2 and Table 3 respectively.

Table 2: Current Deployment of WOCs

Response	Percentage
No deployment	50.6%
Employed in test mode	10.4%
Limited production deployment	18.9%
Broadly deployed	15.9%
Other	4.3%

Roughly half of IT organizations have not made any deployment of WOCs.

Table 3: Planned Deployment of WOCs

Response	Percentage
No plans	45.1%
Will deploy in test mode	8.5%
Will make limited production deployment	20.7%
Will deploy broadly	20.7%
Other	4.9%

Comparing the data in Table 2 and Table 3 yields the conclusion that:

Over the next year, IT organizations plan to make a moderate increase in their deployment of 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 the WAN.

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

Table 4: Techniques to Improve Application Performance	
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[2].

Forward Error Correction (FEC)

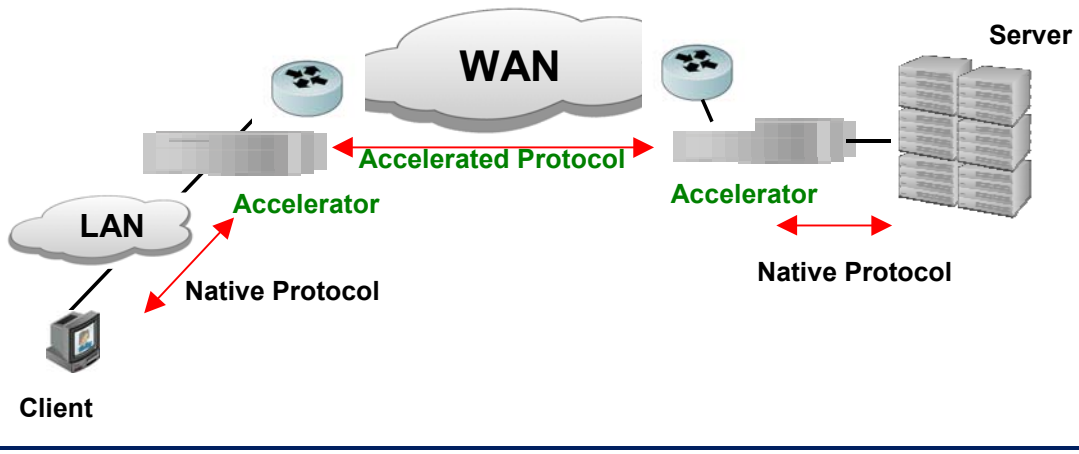
FEC is typically used at the physical layer (Layer 1) of the OSI stack. FEC can also be applied at the network layer (Layer 3) whereby an extra packet is transmitted for every *n* packets sent. This extra packet is used to recover from an error and hence avoid having to retransmit packets. A subsequent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

[2] Transmission_Control_Protocol

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 4. 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 4: 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, 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 and WOC Selection Criteria

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.

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.

Virtual Appliances

The operating system and WOC software can be optimized to run in a virtual machine on a virtualized server. The performance of the resulting virtual appliance is largely determined by the processing power of the underlying physical server and in part by the functionality provided by the hypervisor. Ideally, the performance of a virtual appliance would be identical to the performance of a standalone appliance assuming the same underlying server hardware. One of the primary advantages of a virtual WOC appliance is the ease of deployment and the centralized provisioning via the hypervisor management system. Another advantage is that in many cases a virtual WOC costs considerably less than a hardware-based WOC. Virtual appliances can also be deployed in support of public and private cloud projects, with the virtual WOCs deployed at the IaaS or SaaS cloud service provider sites.

Client software

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

The recommended criteria for evaluating WAN Optimization Controllers are listed in Table 5. 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 5 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 5: 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 5 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. The DMC is the type of product focused on the latter problem.

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 6.

Table 6: Importance of Optimizing Communications Based Traffic

	VoIP	Traditional Video Traffic	Telepresence
Extremely Important	18.7%	8.6%	4.4%
Very Important	42.3%	23.3%	25.4%
Moderately Important	23.6%	30.2%	27.2%
Slightly Important	8.1%	24.1%	23.7%
Not at all Important	7.3%	13.8%	19.3%

One of the conclusions that can be drawn from the data in Table 6 is:

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

The section of the handbook that is entitled "Application and Service Delivery Challenges" discussed the importance of managing communications based traffic. In that discussion the observation was made that it is notably more important to IT organizations to get better at managing VoIP than it is for them to get better at managing either traditional video traffic or telepresence. The data in Table 6 indicates that a similar comment applies to the importance of getting better at optimizing VoIP vs. getting better at optimizing traditional video and telepresence.

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, as previously discussed, 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.

Hybrid WAN Optimization

The traditional approach to providing Internet access to branch office employees is to carry the Internet traffic on the organization's enterprise network (e.g., their MPLS network) to a central site where the traffic is handed off to the Internet. The primary advantage of this approach is that it enables IT organizations to exert more control over the Internet traffic.

A 2010 market research report entitled *Cloud Networking* reported on the results of a survey in which 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 7.

Table 7: Routing of Internet Traffic

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

One of the disadvantages of a centralized approach to Internet access is performance. For example, an IT organization might use WOCs to optimize the performance of the traffic as it flows from the branch office to the central site. However, once the traffic is handed off to the Internet, the traffic is not optimized.

The vast majority of IT organizations have a centralized approach to Internet access.

The preceding section of the handbook entitled *Optimizing and Securing the Use of the Internet* described the use of a Cloud Networking Service to optimize the performance of applications that transit the Internet. One approach that IT organizations can take to optimize the end-to-end performance of Internet traffic is to implement WOCs to optimize the performance of that traffic as it transits the enterprise WAN. This WOC-based solution is then integrated with a CNS 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 optimization solution.

Data Mobility Controllers (DMCs)

Background

Most large IT organizations are experiencing dramatic increases in the volume of business critical traffic traversing the WAN between data centers. The growth in traffic volume stems not only from data proliferation, but also from an increased emphasis on improving IT support for business agility and business continuity.

Inter-data center traffic is generated by a number of server-to-server utilities and applications, including:

Storage Replication
Replicating data on storage arrays and NAS filers is a critical aspect of many disaster recovery strategies because of the need to preserve the critical data that enables business continuity. Disk array and filer volumes are expanding rapidly driven by the combination of business requirements and advances in disk technology. Volume sizes of up to 500 TB –1 PB are becoming increasingly common. Like most inter-data center traffic, storage replication traffic is characterized by a relatively small number of flows or connections, but very high traffic volume per flow.

System Backups
Backups are an important component of a disaster recovery strategy because they focus on the continuity of physical and virtual application servers as well as data base servers. The increased complexity of operating systems and application software is causing server image sizes to grow dramatically. Backup solutions that can minimize the backup window and allow for more frequent backups can help make a backup strategy more effective.

Virtual Machine Migration
Enabled by the wide spread adoption of virtualization and cloud computing, the migration of VMs between data centers is becoming increasingly common. Live migration of production VMs between physical servers provides tremendous value. For example, it allows for the automated optimization of workloads across resource pools. VM migration also makes it possible to transfer VMs away from physical servers that are undergoing maintenance procedures or that are either experiencing faults or performance issues. During VM migration the machine image, which typically runs 10 Gigabytes or more, the active memory, and the execution state of a virtual machine is transmitted over a high-speed network from one physical server to another. For VM migrations that go between data centers, the virtual machine's disk space may either be asynchronously replicated to the new data center or accessed from the original data center over the WAN. A third approach is to use synchronous replication between the data centers. This approach allows the data to reside at both locations and to be actively accessed by VMs at both sites; a.k.a., active-active storage. In the case of VMotion, VMware recommends that the network connecting physical servers involved in a VMotion transfer have at least 622 Mbps of bandwidth and no more than 5 ms of end-to-end latency[3].

[3] VMWare.com

High Performance Computing

A significant portion of supercomputing is performed on very large parallel computing clusters resident at R&D labs, universities, and Cloud Computing Service Providers that offer HPC as a service. Transferring HPC jobs to these data centers for execution often involves the transfer of huge files of data over the WAN. In some cases, HPC applications can be executed in parallel across servers distributed across two or more data centers. For these relatively loosely coupled applications (e.g. those based on Hadoop or MapReduce), inter-processor communications may involve large data transfers of interim results.

The Interest in DMCs

In order to quantify the interest that IT organizations have in some of the challenges that were described in the preceding sub-section of the handbook, The Survey Respondents were asked two questions. One question was how important was it to their organization over the next year to get better at optimizing the transfer of storage data between different data centers. The other question was how important was it to their organization over the next year to get better at optimizing the transfer of virtual machines. Their responses are shown in Figure 5 and Figure 6.

Figure 5: The Interest in Optimizing the Transfer of Storage Data

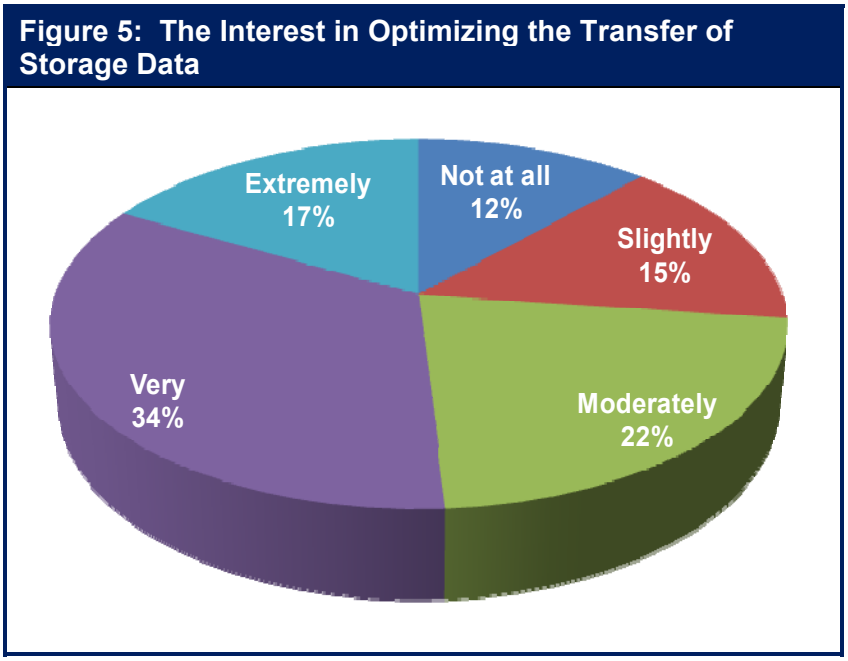
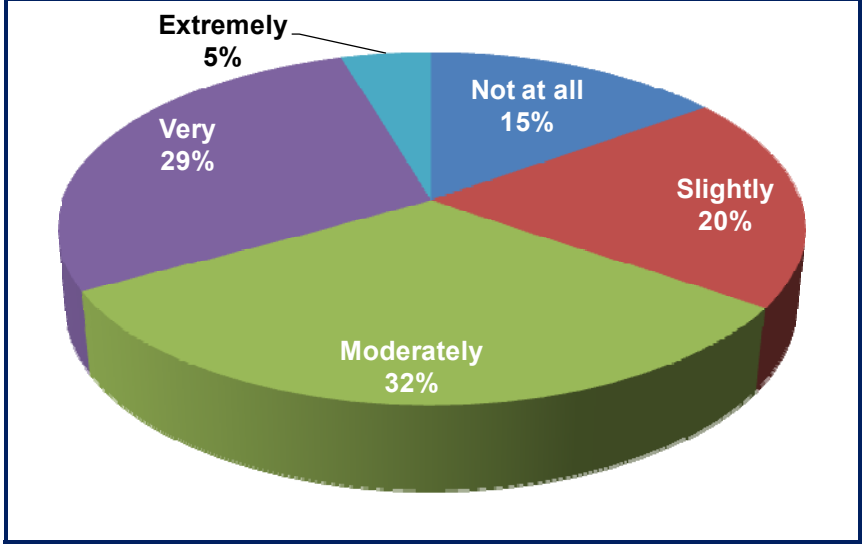


Figure 6: The Interest in Optimizing the Transfer of VMs



For the Majority of IT organizations, getting better at optimizing the transfer of data between data centers is either very or extremely important.

The data in Figure 6 indicates that getting better at optimizing the transfer of VMs between data centers is of at least moderate importance to the majority of IT organizations. That importance should grow as IT organizations make increased use of virtualized servers and as they make increased use of functionality such as VMotion.

DMC Functionality

One way to support the huge inter-data center traffic flows described above is to connect the data centers with WAN links running at speeds of 10 Gbps or higher. One limitation of this approach is that these WAN links are not always available. Another more fundamental limitation is that these WAN links can be inordinately expensive.

A far more practical way to support the huge inter-data center traffic flows is to implement techniques that reduce the amount of data that gets transferred over the WAN and that guarantee performance for critical traffic. Over the last few years, many IT organizations have implemented WAN optimization controllers (WOCs). Two of the primary functions of a WOC are to reduce the amount of data that gets transferred over the WAN and to guarantee performance for business critical traffic. Unfortunately, the traditional WOC may not be able to effectively optimize the huge inter-data center traffic flows. That follows because while the functionality they provide appears to be what is needed, WOCs were designed to support traffic between branch offices and a data center. This traffic is comprised of tens, if not hundreds, of slow-speed connections. As previously mentioned, inter-data center traffic is comprised of a small number of very high-speed connections.

DMCs focus on a subset of the capabilities listed in Table 4**.** This includes QoS and TCP optimization. The TCP optimization provided by DMCs often includes functionality such as packet level FEC and the ability to recovery from out of order packets. It also includes de-duplication, compression and support for specific backup applications and specialized transfer protocols. The primary challenge for this class of device is to be able to perform high levels of data reduction while filling high bandwidth WAN pipes (e.g., 1 Gbps or more) with non-conflicting data flows.

Some DMC implementations are based on narrowing the functionality of an existing high end WOC hardware appliance to focus on the set of capabilities required for bulk data transfers. In addition, software enhancements may be added to provide support for specific replication and backup applications. When these devices are placed in DMC mode, the TCP buffers and other system resources may be tuned to support replication and backup applications. These hardware appliance devices typically rely on multi-core CPUs to achieve elevated performance levels. When the performance limit of a single appliance is reached, it may be necessary to load balance the inter-data center traffic across a cluster of DMC appliances.

Another approach to DMC design is to use network processors and other programmable logic to provide specialized hardware support for compute-intensive functions, such as transport optimization, de-duplication and compression. Hardware support for some of these functions may be the only cost-effective way to fill very high bandwidth (multi-gigabit) WAN links with highly optimized bulk data traffic.

Efficient bulk transfers and data replication are critical requirements to gain many of the potential benefits of both private and public cloud computing.

Techniques for Coping with Packet Loss and Out of Order Packets

The section of the handbook that is entitled "Optimizing and Securing the Use of the Internet" discussed in detail the impact of packet loss on TCP throughput. As discussed in that section, small amounts of packet loss can significantly reduce the maximum throughput of a single TCP session. In addition, while packet loss affects throughput for any TCP stream, it particularly affects throughput for high-speed streams, such as those associated with bulk data transfers.

As a result of the well-known impact of packet loss on throughput, numerous techniques have been developed to mitigate the impact of packet loss and out-of-order packets. This includes Performance Enhancing Proxies (PEPs), which are intended to mitigate link-related degradations and Forward Error Correction (FEC), which is intended to offset the impact of dropped packets.

PEPs are often employed in transport optimization solutions that use a proprietary transport protocol to transfer data between the DMCs. The proprietary transport can support very aggressive behaviors in order to expand the window sizes as new connections are formed. This allows the PEP to efficiently fill a link even though that link has both high bandwidth and high delay. With transparent proxies, TCP is used as the transport protocol between the DMCs and the end systems, allowing the end systems to run unmodified. The DMC's PEPs can intercept and terminate the TCP connections from the end systems. They typically use large buffers to isolate the end systems from having an awareness of packet loss and therefore they eliminate the need for retransmissions and adjustments in the TCP window sizes of the end systems.

FEC[4] has long been used at the physical level to ensure error free transmission with a minimum of re-transmissions. The basic premise of FEC is that an additional error recovery packet is transmitted for every *n* packets sent. The additional packet enables the network equipment at the receiving end to reconstitute one of the *n* lost packets and hence negates the actual packet loss. The ability of the equipment at the receiving end to reconstitute the lost packets depends on how many packets were lost and how many extra packets were transmitted. In the case in which one extra packet is carried for every ten normal packets (1:10 FEC), a 1% packet loss can be reduced to less than 0.09%. If one extra packet is carried for every five normal packets (1:5 FEC), a 1% packet loss can be reduced to less than 0.04%. To put this in the context of application performance, assume that the MSS is 1,420, RTT is 100 ms, and the packet loss is 0.1%. Transmitting a 10 Mbyte file without FEC would take a minimum of 22.3 seconds. Using a 1:10 FEC algorithm would reduce this to 2.1 seconds and a 1:5 FEC algorithm would reduce this to 1.4 seconds.

The preceding example demonstrates the value of FEC in a TCP environment. The technique, however, applies equally well to any application regardless of transport protocol. A negative factor that is associated with FEC is that the use of FEC introduces overhead which itself can reduce throughput. One way to avoid this is to implement a FEC algorithm that dynamically adapts to packet loss. For example, if a WAN link is not experiencing packet loss, no extra packets are transmitted. When loss is detected, the algorithm begins to carry extra packets and it increase the amount of extra packets as the amount of loss increases.

[4] RFC 2354, Options for Repair of Streaming Media

A DMC can also perform re-sequencing of packets at both ends of the WAN link to eliminate the re-transmissions that occur when packets arrive out of order. Packet re-ordering is applicable to all IP data flows regardless of transport protocol.

The criteria for evaluating DMC solutions presented below in Table 8 is to a large degree a subset of the criteria for evaluating WOC solutions summarized in Table 4.

Table 8: Criteria for WAN Optimization Solutions

Criterion	Weight Wi	Score for Solution "A" Ai	Score for Solution "B" Bi
Performance: Throughput and Latency			
Transparency			
Solution Architecture			
Capability to Perform Application Monitoring			
Scalability			
Cost-Effectiveness			
QoS and Traffic Management			
Data Reduction Efficiency			
Security			
Ease of Deployment and Management			
Change Management			
High Availability Features			
Total Score		Σ WiAi	Σ WiBi

Trends in WOC and DMC 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.

DMC Enhancements

As discussed, DMCs facilitate the transfer of high volume data between data centers owned either by enterprise IT organizations or by CCSPs. DMC products are still in an early stage of evolution and a number of developments can be expected in this space in the near term. This includes enhanced hardware support for various functions including encryption and higher speed WAN and LAN interfaces (10 GbE and higher) to support a combination of highly efficient data reduction and high bandwidth WAN services.

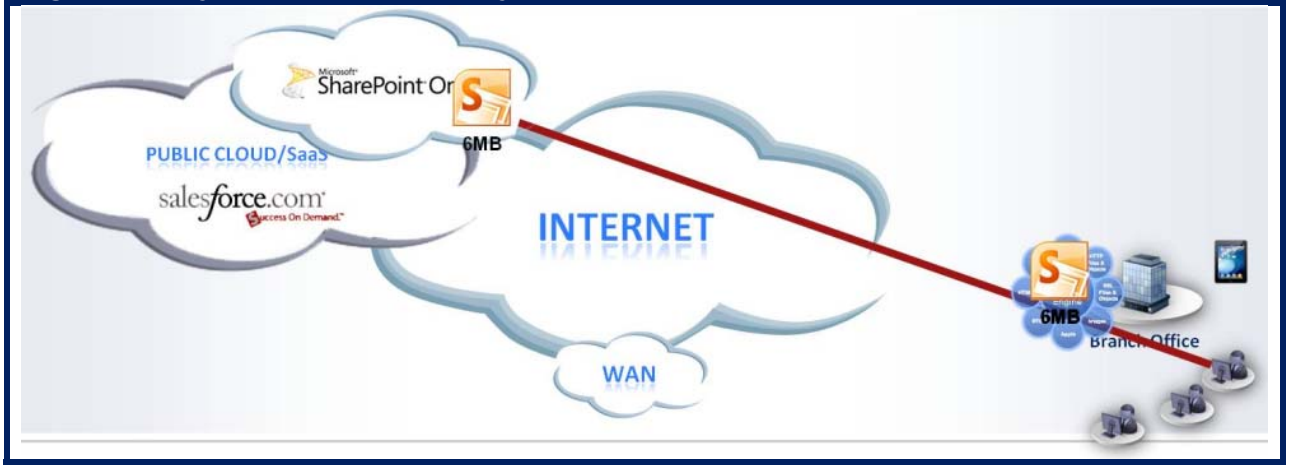
Cloud Networking Services

Not all CCSPs will either provide WOC functionality themselves nor support vWOC instances being hosted at their data centers. A previous section of the handbook described the growing use of Cloud Networking Services (CNSs). Using a CNS that provides optimization can accelerate the performance of services acquired from these CCSPs.

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 7, 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 7: 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.

Application Delivery Controllers (ADCs)

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[5]. 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.

Like the WOC, the ADC is available in a number of form factors including virtual appliances, hardware appliances, and as line card modules for switches and routers. Hardware implementations can be based on multi-core CPUs (possibly with specialized co-processors for compute–intensive operations such as encryption/decryption) or with network processors.

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

[5] 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

The ADC evaluation criteria are listed in Table 8. 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 8: 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.

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.

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.

A10 Networks

The Fastest Growing Application Networking Company

64-bit AX Series

Application Delivery

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

IPv6 Migration

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

Cloud Computing & Virtualization

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

Advanced Core Operating System (ACOS)

AX Series Advantage

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

Application Solutions

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

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

Transforming the Internet into a Business-Ready Application Delivery Platform

Ensuring applications perform to support your business goals

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

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

Key Challenges in Delivering Applications

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

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

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

Akamai's Application Performance Solutions

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

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

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

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

Next-Generation Application Acceleration

Blue Coat

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

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

The Foundation: Optimizing Traditional Applications

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

Next Generation WAN Optimization Requirements

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

Streaming video and rich media

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

Cloud Delivered SaaS applications

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

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

Next-Generation Acceleration

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

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

Figure 1: Performance gains by technology type

Video Optimization

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

Cloud Acceleration

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

Traditional WAN Optimization

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

Get the right acceleration strategy

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

About Blue Coat

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

Software WAN Optimization

certeon®
Accelerate Your Business

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

aCelera™
Secure Automated Optimized

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

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

Microsoft GOLD CERTIFIED Partner

vmware READY

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

APPLICATION DELIVERY PERFORMANCE
From datacenter and cloud to any user anywhere

Maximum Value & Maximum Performance

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

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

aCelera: Built for Global Performance

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

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

aCelera: Built for Global Deployment

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

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

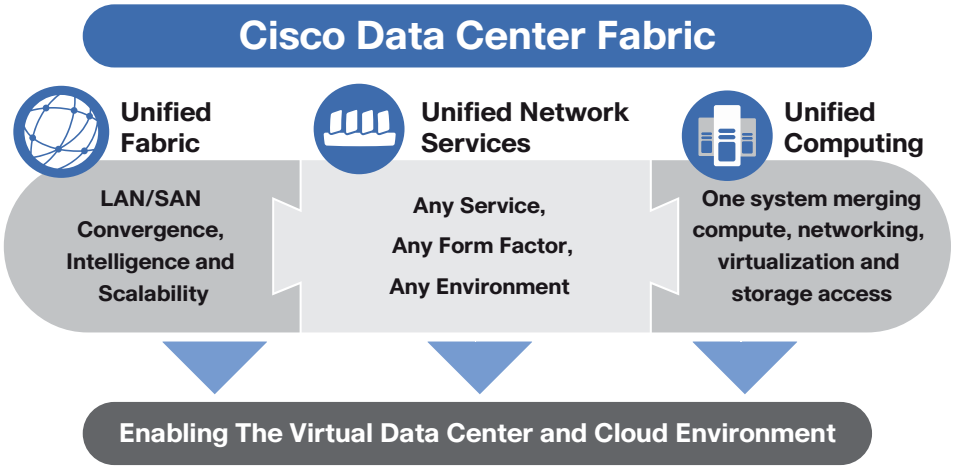
aCelera can be deployed in any private, public, and hybrid cloud computing environment and is poised to meet ANY future performance, scale and connection demands imposed by any enterprise IT environment, private network, private cloud, public cloud or a hybrid of them all.

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

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

Cisco Unified Network Services

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



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

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

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

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

Application Delivery Controllers

Enhanced web application performance, availability, and server scalability

WAN Optimization

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

Network Analysis and Monitoring

Simplifies application performance monitoring

Data Center Security

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

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**.

C96-673827-00 05/11

exinda

UNIFIED PERFORMANCE MANAGEMENT

VISIBILITY | CONTROL | OPTIMIZATION

COMPLETE WAN OPTIMIZATION

Increase the speed and efficiency of your wide area network.

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

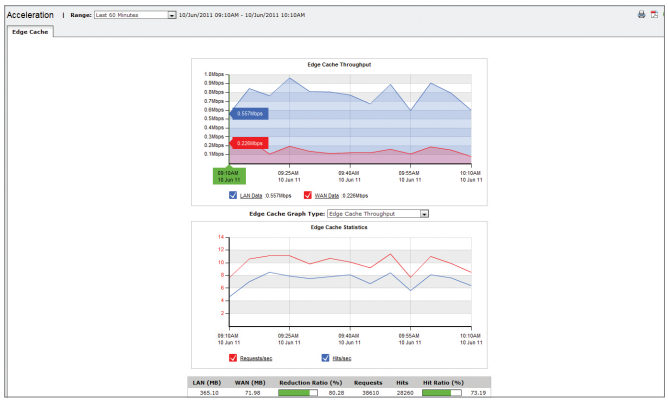
Point solutions lack inter-communication between the functions of visibility, control, and optimization. This creates contention between these independent solutions, as each function is unaware of the effect its actions has on the other.

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

LATEST ADVANCES IN UNIFIED PERFORMANCE MANAGEMENT

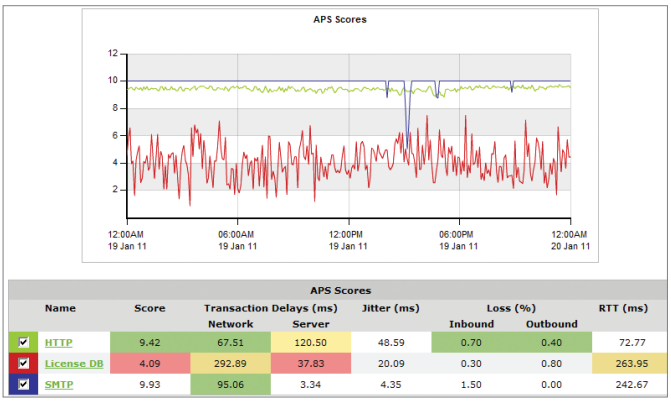
Exinda's development team is continually adding new features and functionality into our unified performance management solution. It is because of our agile development cycle and constant push to add innovation to our product line that Exinda has become the fastest growing WAN optimization vendor in the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

EDGE CACHE



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

APPLICATION PERFORMANCE SCORE



APS Scores

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

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

Edge Cache

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

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

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

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

Application Performance Score

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

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

Unified Performance Management

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

exinda
performance you can see

"Unified Performance Management is driven by improving the quality of a user's experience."

- Ed Ryan, Exinda Vice President of Products

The Best Solution For You.

Identify and Improve Application Performance

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

Offer a Superior User Experience

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

Real-time and Historical Reporting

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

Conserve WAN Resources

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

Leverage Your Investment

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

Features & Benefits

Visibility

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

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

Control

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

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

Optimization

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

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

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

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

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

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

Advertorial

EXPAND networks | VIRTUALLY EVERYWHERE™

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

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

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

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

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

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

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

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

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

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

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

Enabling Strategic Initiatives

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

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

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

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

Advertorial

SOLUTION BRIEF

Software as a Service (SaaS)

A Cloud-Ready Network ensures rollout success

www.ipanematech.com

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

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

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

Using WAN Governance, your organization can:

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

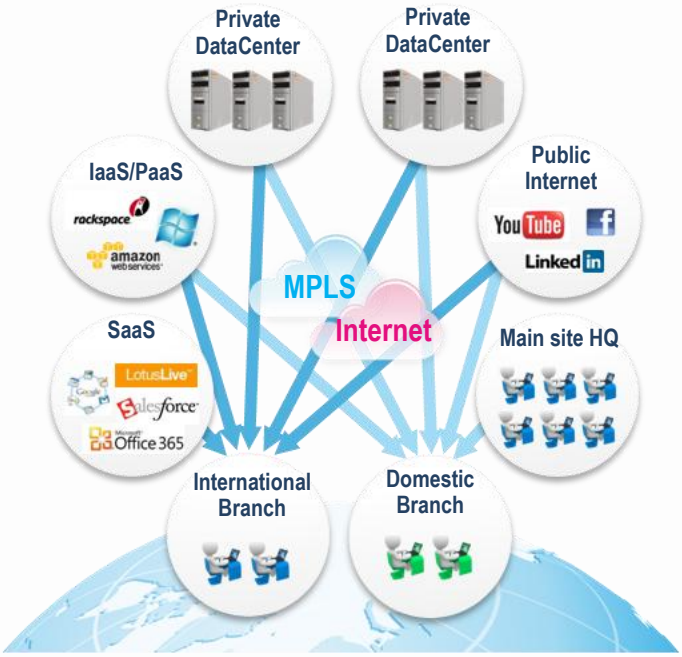
IT infrastructure directors today find themselves in one of two situations: the business side of their organization is planning for SaaS applications that the VPN will need to support, or existing SaaS applications are underperforming or impacting the performance of other business applications.

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

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

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

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



Beyond the Network…

ipanema Technologies

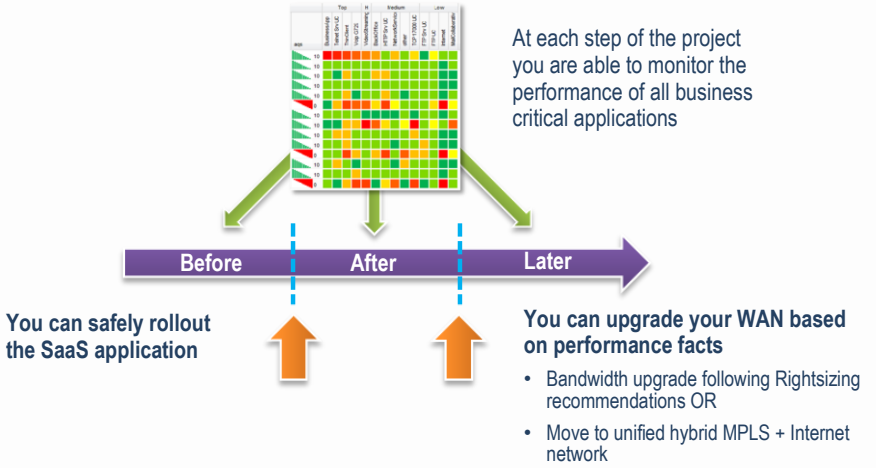
www.ipanematech.com

All-in-One Solution for Guaranteeing Application Performance

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

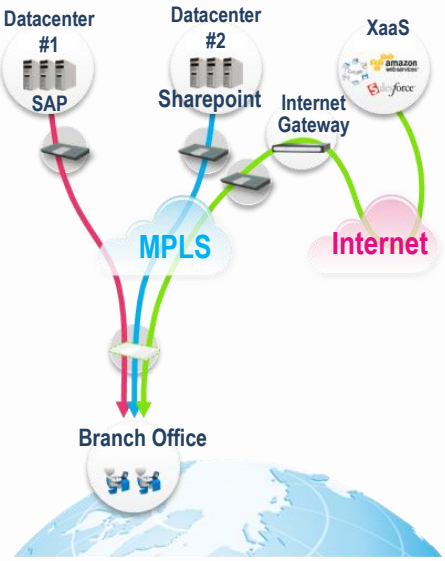
- QoS & Control
- Application Visibility
- WAN Optimization
- Dynamic WAN Selection (hybrid network unification)

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

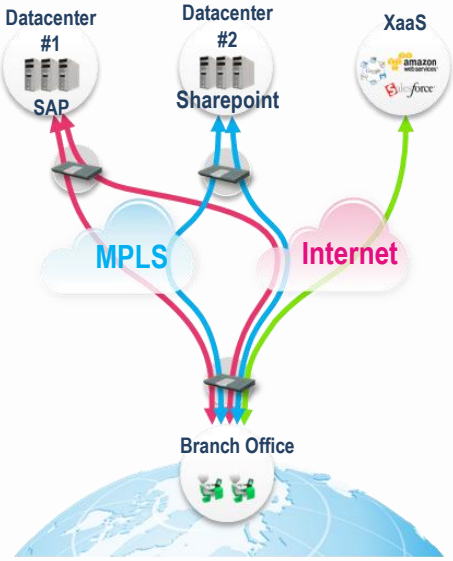


Since every enterprise is different, IT strategy on whether or not to change network architecture for SaaS collaboration varies from one company to another. You do not necessarily need to change your architecture to make your network "cloud-ready".

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



An MPLS CRN



A unified hybrid MPLS + Internet CRN

Valeo Embraces the Cloud and Maximizes Value

Valeo, one of the world's leading suppliers of components, integrated systems and mod-ules for automotive CO_2 emissions reduction, rolled out a hybrid network with MPLS + Internet for its migration from conventional email and collaboration applications to Google Apps.

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

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

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

Return on investment

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

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

Data Center Virtualization and Application Delivery

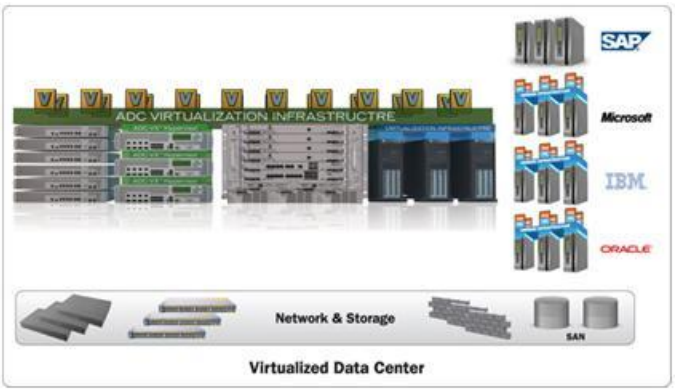
Network infrastructure virtualization/consolidation has had a major impact on the Application Delivery Controller (ADC) role, position and deployment models. For instance, ADCs which were previously tightly coupled with a single application must now be able to service a layer of virtualized applications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

A New Paradigm: Virtual Application Delivery Infrastructure (VADI)

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

Radware's VADI delivers the following business benefits:

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



VADI Key Components

Virtual ADC Instances

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

ADC Computing Resources via Three Form-Factors

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

Virtual Data Center Ecosystem Integration

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

Advanced VADI Services

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

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

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