

How Network Transparency Affects Application Acceleration Deployment

By John Bartlett and Peter Sevcik
July 2007

Acceleration deployments should be simple. Vendors have worked hard to make the acceleration appliances network compatible, good network citizens, and easy to deploy and maintain. Thousands of network deployments have occurred without running into the network issues described herein.

Transparency is often raised as a critical consideration by some vendors when evaluating application acceleration devices. This report digs into the meaning of transparency, and explains the relevant issues.

Transparency is not a major concern when evaluating application acceleration appliances. It is a technical issue that occasionally requires setting the right configurations in the appliance or minor network adjustments by the network operations staff. This report explores the cases where problems may occur and details how they are resolved.

Two market leaders in this space, Riverbed and Cisco, have slightly different approaches to the way their accelerators are deployed and the way they address packets flowing between their devices. This report will dive into some detail on how these differences can impact the network, and if problems exist, offer remedies to work around them. The specific issues covered in the document focus on “transparency”, and networks that support QoS, VoIP, traffic monitoring, header and content-based routing, firewall and intrusion detection systems, and cases of asymmetric routing.

In the rest of this report, we use “application accelerators” or similar terms to refer to Riverbed and Cisco products generically. It is important to note that there are additional competitors in the application acceleration market that we do not consider in detail, and that some of our observations and insights will not apply to them.

Application Acceleration Deployment Architecture Alternatives

Application accelerators can be deployed in one of two distinct configurations, either in-path or out-of-path. When deployed in-path, the application appliance is physically in the data stream between the LAN and the WAN router. All traffic going to or from the WAN must pass through the appliance. When deployed out-of-path, the application accelerator is connected to a separate switch or router port near the WAN connection, but is not physically in the data path. In this configuration some additional redirection mechanism, such as Web Cache Communications Protocol (WCCP) or Policy Based Routing (PBR), is required to identify relevant traffic, and redirect it to the application acceleration appliance.

In-Path Deployment: The in-path configuration has the advantage of simplicity. No additional network segments or ports are required. The appliance provides two network interfaces, one for the LAN-side connection, and the other for the WAN-side connection as shown in Figure 1. Auto-discovery of an in-path appliance happens naturally because the traffic must pass through the device on the way to its destination.

Once a traffic stream is recognized as being one that can be optimized by the acceleration appliance, the sending side appliance steps in, performs its optimizations on the sending end and passes modified information on to the receiving end. The receiving end appliance then reverses the process creating the original content, and forwards that on to the

NetForecast Report
5088

©2007
NetForecast, Inc.

receiving client or server. If the stream represents an application that cannot or should not be optimized, the sending-side appliance merely passes that traffic on through without modification. The WAN router continues to serve its initial function of routing and queuing traffic for the WAN link, its functions are not replaced in any way by the acceleration appliance.

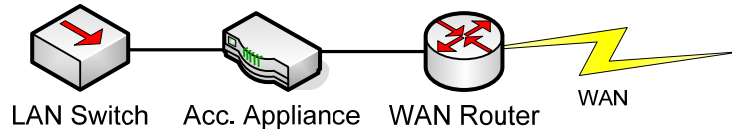


Figure 1 - Acceleration Appliance, In-Path Deployment

Because an in-path solution creates a single point of failure, most acceleration appliances come with a fail-to-wire feature. If the appliance fails for some reason, the network card goes into a pass-thru mode, becoming a wire. All network traffic is passed directly through the device, even if the device has no power, and is sent on to the edge router. This feature allows communications with the remote office to continue, albeit without the optimization features of the appliance, until it can be repaired.

Out-of-path: An out-of-path accelerator is connected to an additional router port near the WAN link, not directly in the path of traffic to the WAN as shown in Figure 2. Traffic which should not or cannot be optimized passes directly to the WAN router as if the appliance were not installed. Traffic which can be optimized is redirected to the appliance by the router, using some additional protocol such as WCCP or PBR.

Using this approach has a number of advantages, but comes with the added complexity of implementing the redirection protocol to allow it to work. The WCCP or PBR mechanism must be explicitly configured to redirect traffic that can be optimized to the acceleration appliance. Once optimized, the traffic will then flow back from the appliance to the router and be forwarded over the WAN link to the remote acceleration device, where the original content is recreated.

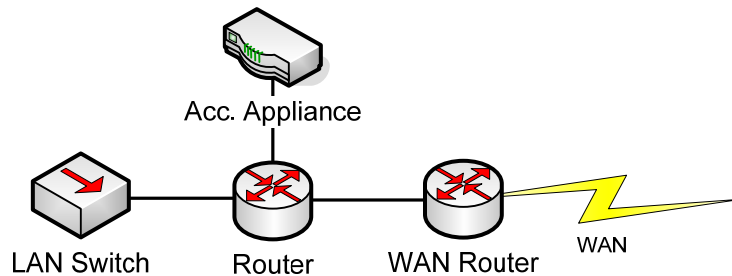


Figure 2 - Acceleration Appliance, Out-of-Path Deployment

As in the in-path case, there is no change in the functionality of the WAN edge router. The router continues to provide routing and queuing functions both for traffic flowing directly to the WAN as well as for traffic that has been optimized by the acceleration appliance.

A failure of the optimization appliance in the out-of-path configuration has no effect on un-optimized traffic. If the network is using WCCP, the protocol will recognize the absence of the optimizer and redirect traffic to another appliance or to the WAN link directly. As in the in-path case, communications will be maintained without the optimization features until the failure is resolved.

Deploying WCCP creates an additional load on routers. The reader is cautioned to check with his/her router vendor to ensure that existing routers are capable of supporting WCCP and have sufficient CPU and memory resources to do so.

Mixed mode operation: There is no requirement that the acceleration appliances all be operating in-path or out-of-path. For the branch office with a single WAN router and WAN link, using an in-path appliance makes deployment simple, minimizing installation time and complexity. For the data center, headquarters location or large branch office being served by multiple or redundant WAN connections, the out-of-path configuration may be more suitable. When WAN traffic switches from one WAN link to another either due to link failure, congestion or usage optimizations, traffic can continue to be directed to an out-of-path appliance and correct operation is maintained. As long as traffic passes through an appliance on both ends of the WAN link, whether they are in-path or out-of-path, optimization will occur.

Deployment Alternatives Summary		
Architecture	Cisco WAAS	Riverbed Steelhead
In path – out of path	Both in-path and out-of-path offered. In-path solution requires additional card, which provides 2 WAN links. ISR has optional chassis WAAS blade, and only operates out-of-path	Both in-path and out-of-path configurations offered. Base product supports a single in-path WAN link; optional card supports a redundant link.
Implement WCCP	Recommended approach. Required with out-of-path mode.	Supported, not required. Both in-path and out-of-path modes can function with or without WCCP.
Redundant core network	WCCP required to support multiple paths from branch locations back to acceleration appliance	Use WCCP to manage redundant links or connection forwarding among multiple in-path appliances
Redundant appliance deployment	WCCP supports redirection of traffic to a group of appliances.	WCCP supports redirection to a group of appliances or vendor offers an Interceptor (virtual in-path) appliance to support large arrays of appliances
Asymmetric Routing	WCCP manages asymmetric routing when implemented. No solution without WCCP.	Built-in support for asymmetric routing (detection and connection forwarding). WCCP manages asymmetric routing when implemented.

Network Transparency

Network transparency has to do with how application acceleration vendors have chosen to address the traffic flowing between the application acceleration appliances. Before we dive into this topic, lets first clear up the definition.

Application Transparency means that the traffic seen by the client and server engaged in a normal application interchange on the network is exactly the same both before and after the installation of the acceleration appliance. No changes to the client software or server software are required to allow the acceleration appliances to work. All solutions

can be configured to maintain the original source and destination address/port values when presented to the end-points as shown in Figure 3.

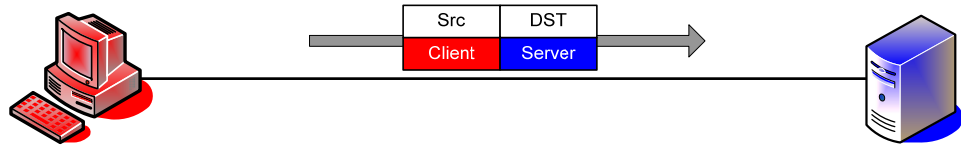


Figure 3 - Source / Destination Addressing, No Acceleration Appliances

Network Transparency has to do with the traffic flowing between the two application acceleration appliances on any given client-to-server path, specifically with what IP addresses and port numbers are used for this traffic. Two camps exist. The Cisco WAAS product primarily uses *transparent addressing*, meaning that the traffic flowing between the two appliances uses the source/destination/port addressing of the client and server involved in the original conversation as shown in Figure 4. Note that the addresses used between the two appliances are spoofing any intermediate network components into thinking that they are the original client and server.

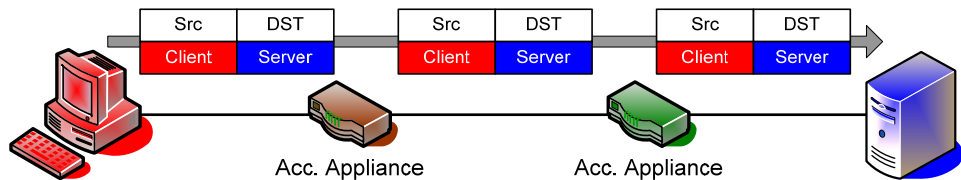


Figure 4 - Source / Destination Addressing, Transparent Addressing

The Riverbed Steelhead product, on the other hand, uses *correct addressing*, meaning that the traffic flowing between the two appliances use the source and destination addresses of the appliances themselves to reflect the optimized nature of the traffic as shown in Figure 5. In this case the intermediate network components see the actual (correct) addresses of the acceleration appliances. The port number may match the original traffic, or may be different depending on the configuration.

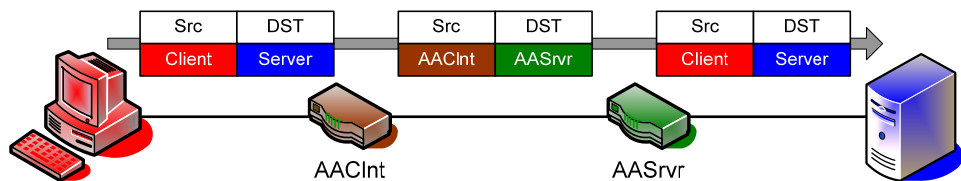


Figure 5 - Source / Destination Addressing, Correct Addressing

We will explore the impact of these two different approaches later in this report. But before we dive into those details, just note that in the vast majority of network configurations, either approach will work and work well. There are a few situations that cause trouble for one approach or the other, and in most cases some simple configuration work can be done to solve the issue. The ‘problems’ we describe should not be seen as show stoppers, and the primary decision on which vendor to choose should be made on the basis of ease of deployment and the functionality the vendor provides. Indeed, the two vendors are not even as far apart as this brief summary might suggest: the optional CIFS acceleration part of the Cisco WAAS product uses a form of *correct addressing*, while the “port mapping” feature of the Riverbed Steelhead product can achieve a weak form of *transparent addressing*.

Tunnels Lastly also note that the addressing schemes we refer to above are not ‘tunnels’. A tunnel takes a packet stream, encapsulates those packets in an additional IP header and

passes them through an intermediate network. In the case of the acceleration appliances, TCP streams are being terminated at the appliance, and a new and separate TCP connection is established between the appliances to pass back and forth the optimized traffic needed to reconstruct the original at the far end. Problems and features associated with tunnels are different, and are not caused by the mechanisms used by acceleration appliances. Neither Cisco nor Riverbed acceleration appliances use tunnels.

Issues Affecting Implementation

There are some areas of concern where the addressing between the appliances may affect the behavior of the network. The report looks at each one of these in the sections below, defines the problem, explains how each addressing scheme affects this issue, and suggests workarounds that will resolve the problem in a simple manner.

Quality of Service (QoS)

QoS enforcement and queuing is primarily the job of the network routers either with or without acceleration appliances installed. For the vast majority of implementations, no change to the QoS policy is required, and no impact occurs when appliances are installed.

There are two parts to the QoS job, classification and queuing. Classification is the job of deciding which streams need to be given priority. Queuing is the job of giving the high priority streams precedence when there is congestion. The queuing job belongs to the router that supports the WAN interface, as this is where the bandwidth steps down from LAN speeds to WAN access link speeds, and congestion is most likely to occur. This doesn't change with the introduction of the acceleration appliance, it is still the right place to do the queuing work, and so the router continues to do this job in any configuration. QoS queuing may also be deployed in the LAN switches and routers.

Classification may occur on the WAN router, on the access switch, or in the endpoint itself. If classification is occurring in the endpoint or on the access switch or router, then packets are classified before reaching the acceleration appliance. If classification does not occur until packets reach the WAN router, classification occurs after they leave acceleration appliance

The following scenarios review the network approach to classification both with and without acceleration to understand the impact of the appliances on each case.

Out-of-path – with no acceleration: If the appliances are implemented out-of-path, and the traffic stream in question is not to be optimized by the appliance (e.g. Voice over IP or VoIP), then the network does not route the traffic to the appliance, but instead directs the traffic to the WAN edge router. Returning traffic proceeds directly from the edge router towards its final destination. This traffic follows the same path and therefore behaves exactly as it does with no acceleration appliance installed, and thus there is no impact. QoS continues to be marked either near the source or at the edge router, and the router continues to support the priority queuing as before.

Out-of-path – with acceleration: If the appliances are implemented out-of-path, and the traffic stream in question should be optimized, the network redirects this traffic to the acceleration appliance. If the traffic has been marked with QoS tags before arriving at the appliance, the appliance only needs to replicate those QoS tags on the optimized traffic being forwarded to the far end appliance. Both Cisco and Riverbed appliances default to this mode of operation. The router will then see these marked packets as priority packets, and provide the prioritized queuing necessary to give them the quality of service needed.

If the QoS marking is not done upstream, but was (before the introduction of acceleration appliances) implemented by the edge router, we have to consider the addressing transparency of the appliances.

- For an appliance that implements *transparent addressing* (such as Cisco WAAS), the traffic flowing between the appliances will have the original IP addresses and port values of the source and destination computers. If the router QoS decision is based on this address and port information, no further configuration is needed. However, if the router looks deeper into the packet to make the classification decision, it may get confused, because the packet contents will have been changed through the optimization process and may not be recognizable to the router. To avoid this problem either move classification of the packet streams to a router on the LAN-side of the acceleration appliance, or configure the acceleration appliance itself to identify and mark the packets.
- If an acceleration appliance that uses *correct addressing* (e.g. the Riverbed Steelhead appliance) is deployed, the addresses on the optimized traffic will not match the original addresses. If the WAN router is making QoS marking decisions either based on addresses and ports, or on deeper packet inspection, it will be affected. The same two solutions exist here, either to mark packets further upstream or to mark them in the acceleration appliance with a simple configuration change.

In-path – with no acceleration: Appliances deployed in-path see all the WAN traffic, whether it is to be accelerated or not. Traffic that is not to be accelerated is identified either by its protocol (UDP versus TCP) or by a specific configuration in the appliance. This traffic is passed through to the WAN router without change. Classification either upstream or downstream of the acceleration appliance will continue to work.

In-path – with acceleration: Classification that occurs before packets reach the acceleration appliance will be optimized, and the resulting traffic will reflect the same QoS markings as the original. Both *transparent addressing* and *correct addressing* systems will work the same way, passing the original QoS markings on to the WAN router. If classification is being handled after the application appliance, the *transparent addressing* system will work without modification. The *correct addressing* system will require a simple configuration change to identify the traffic and mark its packets with the correct QoS level.

We see from this analysis that neither system breaks QoS. Each system type requires some simple configuration work in certain situations to support the existing network QoS configuration.

Queuing in the Appliance: Enterprises running very time sensitive applications such as telepresence may be concerned with an in-path appliance adding latency or jitter to packets passing through the unit. In most environments this effect is very small, since the application appliance has LAN speed connections on either side of the unit.

Traffic Monitoring

Introduction of an application acceleration appliance can affect traffic monitoring tools watching the flow of traffic across the wide area network. NetFlow or sFlow statistics are often being collected from probes in the WAN or the edge routers, showing how much traffic is crossing the WAN, and identifying the source of that traffic by IP addresses and port numbers. Introduction of an application acceleration appliance may hide some of the information previously seen, or may identify much of the traffic as being between the IP addresses of the acceleration appliances, thus losing information about the original source and destination of those streams.

Some changes are required to accommodate the new environment when an application accelerator is introduced. Remember that we expect change in the WAN since one of the primary features of the application acceleration appliance is that it dramatically reduces WAN utilization. These changes are often quick and easy to implement.

Transparent Addressing Architecture: After installing a *transparent addressing* accelerator, the traffic monitoring tools will continue to show network usage by source and destination address and port number. However the statistics shown will be for the optimized traffic, not for the volume of the original traffic. This information will be useful in understanding how specific users or groups impact the WAN, but not too useful in understanding wider traffic patterns since actual usage is masked by the efficiency of the application acceleration appliances.

Correct Addressing Architecture: With no change to the monitoring tools after installing a *correct addressing* appliance, traffic monitoring tools will show all optimized traffic using a single source/destination pair, the addresses of the two acceleration appliances. Again this information is probably not the most useful for managing the network.

In both cases the best solution is either to move the network monitoring probes to the LAN-side of the appliance, or to gather usage information directly from the appliance itself. Most acceleration appliances provide detailed information both on the volume of the traffic before optimization as well as providing the optimized traffic volume (the impact on the WAN link). The advantage of using this information source is that it also shows the compression ratio of the acceleration appliance, e.g. it shows what traffic is being optimized, and how well the appliance is doing at reducing the WAN traffic volume.

Header- and Content-based Routing

Networks that use Access Control Lists (ACLs) or Policy Based Routing (PBR) to enable, block or route traffic through the network can be affected by acceleration appliances. Networks that use Network Based Application Routing (NBAR) look even deeper into the packets, and will also be affected by acceleration appliances. Networks heavily dependent on these techniques may need to configure the acceleration appliances to maintain their current architecture.

Header Based Routing (ACLs and PBR): Networks that block certain types of traffic based on the information in their headers will not be affected by either type of acceleration appliance (*transparent* or *correct addressing*). The reason is that the initial TCP connection setup between appliances is an extension of the original client/server three-way handshake, which uses the client and server addressing information. If the ACLs suppress this traffic, the connection is blocked, independent of the application appliance addressing scheme.

However, networks that *enable* traffic based on the header information, or cause certain traffic to be routed in an alternate direction will be affected by *correct addressing* appliances. Because *transparent addressing* appliances maintain the original addressing, they will continue to work.

If these control functions can be moved to the LAN side of the appliance, this solves the problem. It is also possible to create ACL-like rules in the application appliance itself to replicate the rules implemented in the network. If you have detailed ACLs or PBR operating in the WAN that enables or routes traffic, consult with your acceleration vendor to find the best solution for your environment.

Content-based Routing: Some networks make routing decisions based on deep packet inspection, using features like NBAR to recognize specific applications. Once recognized, these streams can then be routed to different paths in the network, blocked, or selectively enabled.

Since application acceleration appliances change the nature of the information carried in the packets, both *transparent* and *correct addressing* appliances will break this routing approach. The same solutions of moving the functions to the LAN described above apply.

Firewalls and Intrusion Detection Systems (IDS)

Firewalls and intrusion detection systems both look deep into packet streams to determine if traffic should be allowed through or if it is potentially dangerous. In this way they are very similar to the NBAR routing issues described in the last section. Lets look at them one at a time.

Intrusion Detection Systems (IDS) monitor traffic and perform deep packet inspection to determine if the traffic contains a virus, worm or other threat to the network. Because application acceleration appliances change the nature of the traffic crossing the WAN, these systems may be confused by the optimized traffic. Appliances that use *correct addressing* will usually not cause a problem, because they use a TCP port number that is distinct to their traffic type. The IDS will either ignore this port number, or can be easily configured to ignore traffic flowing between the acceleration appliances.

Conversely an acceleration appliance using *transparent addressing* may confuse the IDS because it appears to be on a standard port, but the contents don't match the expected structure of the standard port protocol. This can lead to false positives, meaning alerts from the IDS that there is a problem when no real problem exists.

It is usually easy to change the IDS configuration to avoid these issues. It also may be possible to move the IDS appliance to the LAN-side of the application accelerator, completely avoiding the problem.

Firewalls fall into two categories depending on how they are deployed in the network. Many branch office networks have a firewall between the branch LAN and the WAN link because the single WAN link is providing both WAN access to headquarters, as well as direct access to the Internet. The firewall is required for the Internet access. Usually all traffic flowing to the headquarters router is enabled through the firewall. In this configuration, acceleration appliance traffic may have no impact, since it is just another form of traffic flowing through the firewall to the headquarters office. If the firewall does not allow this connection, it is a simple configuration change to the firewall to identify the two IP addresses of the acceleration appliances, and allow their traffic to pass through.

The second category of firewalls is those that are implemented within an enterprise to maintain security between different trust levels of the network. A typical example is a firewall between the development network and the production network, or between the voice VLAN overlay and the data VLAN overlay.

These internal firewalls cause problems very similar to the IDS described above, and the solutions are the same. Either enable the accelerator to accelerator traffic through these firewalls with a simple configuration change, or move the firewall to the LAN-side of the acceleration appliance.

Asymmetric Routing

Asymmetric routing can occur in networks where more than one path exists between a source and destination. Asymmetric routing means that the packets flowing from source

NetForecast helps enterprises and vendors understand and improve the performance of networked applications.

Additional information is available at:
www.netforecast.com

NetForecast and the curve-on-grid logo are registered trademarks of NetForecast, Inc.

to destination take one network path, and the packets flowing from the destination back to the source take a second path. Multiple paths are common on networks where redundant links are deployed for network resiliency.

As we have seen earlier in this report, acceleration appliances modify the traffic in transit and depend on a partner appliance at the far end to reconstruct the original. The two appliances establish a pairing, and they store information common to the pair which is used to optimize the flow of traffic between them. If traffic destined for acceleration appliance A ends up at acceleration appliance B, or bypasses the appliance and ends up at the destination endpoint, it will not be able to be reconstructed and the connection will be lost. With *transparent addressing*, it may also be possible for optimized traffic to reach destination servers appearing as malformed packets.

Again, these problems can be avoided through some straight forward network configuration work. This problem usually manifests at the headquarters or data center location, since that is where network redundancy exists. If the HQ or data center uses the WCCP protocol, WCCP will handle the asymmetric routing and bring the traffic back to the right appliance automatically.

A second approach is to dig into the characteristics of the network and understand why the asymmetric routing condition occurred. Changes to the costing of links or flattening of the network hierarchy can eliminate the asymmetric routing from occurring.

Correct addressing appliances have an advantage in this environment, since the inter-appliance traffic is addressed to reach the correct other appliance. If asymmetric routing is a permanent feature of your network, using a *correct addressing* appliance may be the simplest resolution to this issue.

Conclusions

Although some specific technical issues arise when deploying application acceleration devices, most are easily overcome either by proper placement in the network or by simple configuration work in the appliances. The differences between a *transparent addressing* system and a *correct addressing* system cause minor differences in the way they are deployed, but not in their usefulness to the enterprise.

We encourage the reader to understand the technical issues involved with their choice of application acceleration vendor, but not to get stuck on the issues raised in this report. Find the right solution for your network and make the changes necessary to get the reliable high performance these appliances can provide.

We suggest that the bigger issues of vendor choice should be based on the usual criteria of how well the solution meets the needs of the enterprise in terms of applications accelerated, cost, ease of deployment, reduction of WAN traffic and of course the performance delivered to application users.

John Bartlett is Vice President of NetForecast, and has 28 years of experience designing at the chip, board, system and network levels, with a focus on performance. John led the team that built the first VLAN implementation, one of the first ATM switches, and he is a leading authority on the behavior of real-time traffic on the Internet. He can be reached at john@netforecast.com.

Peter Sevcik is President of NetForecast and is a leading authority on Internet traffic, performance, and technology. Peter has contributed to the design of more than 100 networks, including the Internet, and holds the patent on application response-time prediction. He can be reached at peter@netforecast.com.