

## Understanding Network-Transparent Application Acceleration and WAN Optimization

IT organizations face constant pressure to provide employees with adequate access to business-critical applications and information while also centralizing costly infrastructure resources, such as servers and storage, to lower costs, facilitate management and data protection. This pressure has resulted in the rise of application acceleration and WAN optimization technologies and products. Although often considered “new”, techniques for accelerating applications over the WAN have been in use for some time. Additionally, integration of these services over the existing WAN is often a key consideration in the deployment. The Cisco Wide-Area Application Services (WAAS) solution seamlessly integrates application acceleration and WAN optimization with the existing WAN infrastructure, ensuring consistency and ease of system deployment and management. This document explores the value of using network-integrated and network-transparent application acceleration and WAN optimization to provide both infrastructure centralization and appropriate application delivery performance.

### Overview

Application acceleration and WAN optimization products are designed to provide IT organizations with the tools necessary to consolidate costly remote-office infrastructure and improve the performance of applications and services that operate over the WAN. These solutions typically include an array of technologies and features, each designed to overcome a specific type of barrier to application delivery performance:

- **Compression**—Provides standards-based compression for data in transit to minimize the amount of bandwidth consumed on a link during transfer
- **Data suppression**—By maintaining a history of previously seen data segments in devices deployed at each end of the link, suppresses transmission of data that has previously been seen
- **Flow optimization**—Uses a WAN-optimized transport protocol to overcome the performance and efficiency limitations of commonly used transport protocols such as TCP
- **Application proxy**—Provides a transparent or nontransparent proxy that understands application messaging so that unnecessary messages can be handled locally, batched for parallel operation, predicted (read ahead and transactional), or forwarded to the originating server; primarily used to overcome application latency
- **Application caching**—Provides a local repository of application-specific information and data to safely and locally serve validated content that has been previously accessed when requested by an authorized user

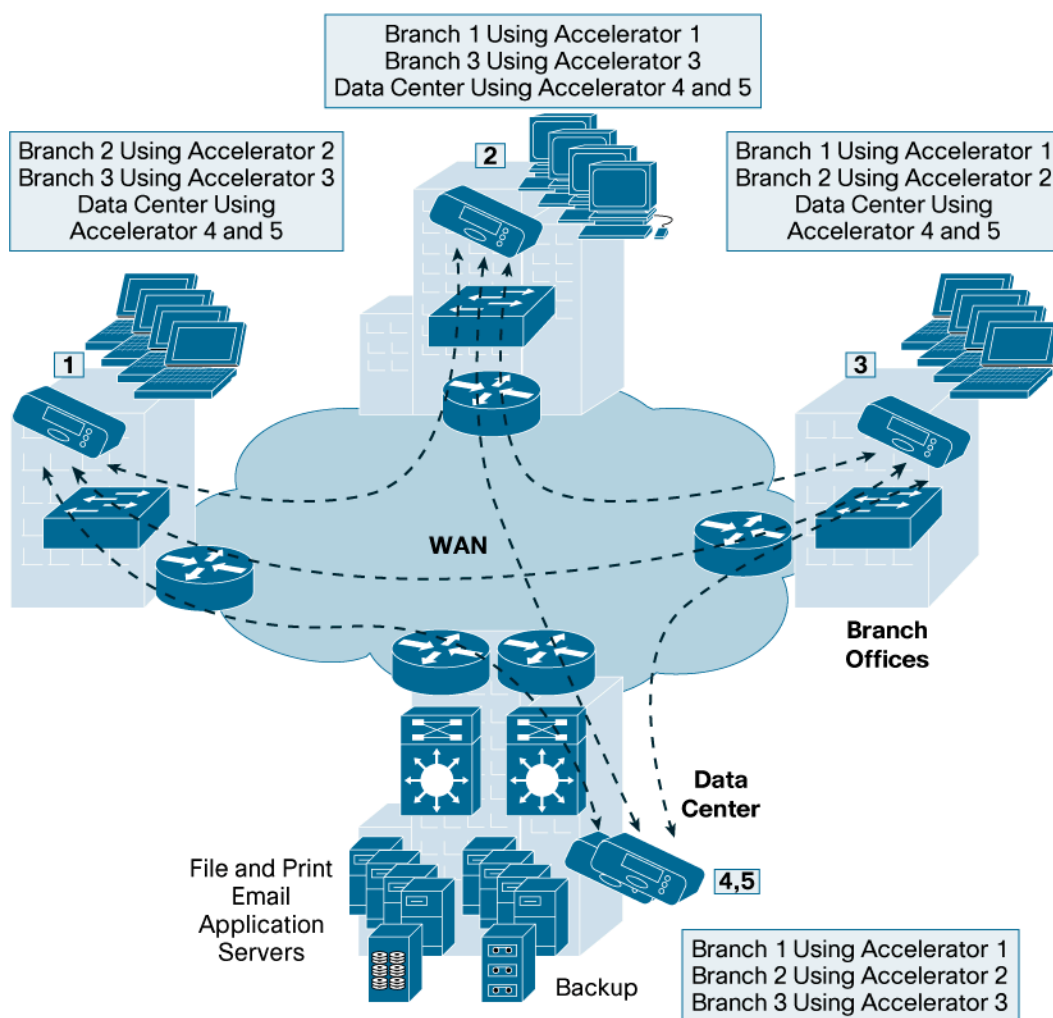
Two categories of solutions that perform the functions above are available today:

- Non-transparent—otherwise known as ‘tunneling’ systems. These systems either use a static tunnel (previously defined by the administrator) or a dynamic tunnel (automatically established upon discovery of a peer) and obfuscate packet header information from the network, thereby disrupting network services. Traffic transmitted across the WAN is done over the tunnel established between accelerator devices.
- Transparent—otherwise known as ‘non-tunneling’ systems. These systems do not use a tunnel, and ensure that packet header information is retained to allow services running in the network to continue to operate. Cisco Wide Area Application Services (WAAS) version 4.0 is an example of a transparent application acceleration and WAN optimization solution that holistically integrates with the underlying network framework and infrastructure devices.

### **Eliminate Complex Overlay Networks with Transparency and Automatic Discovery**

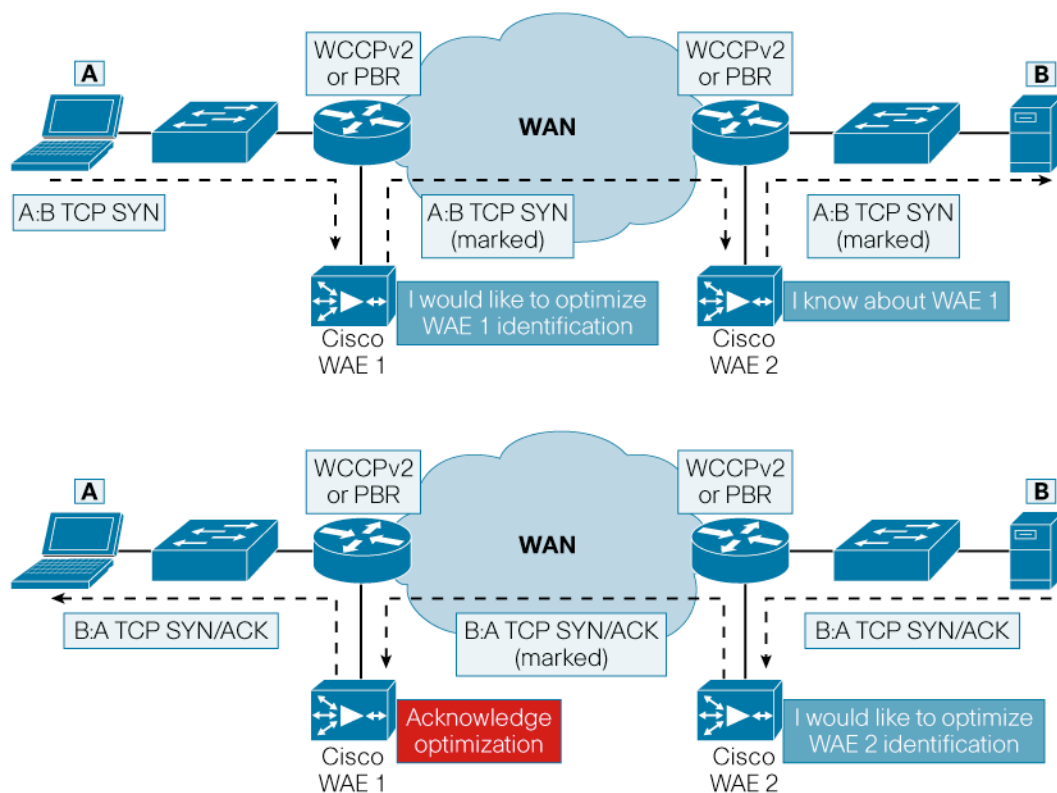
Application acceleration and WAN optimization solutions that are transparent and integrated into the network fabric do not require network administrators to design complex overlay networks to control the direction of packet flow for accelerated traffic (on top of existing routed topologies). In addition, such transparent solutions automatically discover one another and maintain crucial pieces of information so that existing services can run uninterrupted and unchanged. Yet most application acceleration and WAN optimization products nevertheless require administrators to deploy overlay networks, fundamentally undermining the investment of time and money spent designing multilayer, resilient, enterprise-class routed networks. Networks today are designed with multiple routes between end nodes, and the network carefully calculates the best path to take based on existing load conditions, distance, bandwidth, and many other metrics. They also apply features to improve the efficiency, security, and control of data at each hop in the path.

An overlay network, shown in Figure 1, is a network path-control mechanism that sits on top of an existing routed topology. With an overlay network, static paths must be defined that explicitly control the route that a packet coming from an accelerator must take. When an overlay network is used, rather than allowing the underlying network to determine the path that the flow between two communicating nodes should take, the flow must follow the path that has been explicitly defined between two acceleration devices. It is important to point out that some solutions, while they support auto-discovery when the accelerator is deployed in-line of the traffic flow, they often might not when deployed out-of-path. Thus, the acceleration device undermines the routing decisions that would otherwise have been made by the network. Furthermore, deployment of accelerators that use complex overlay networks makes support for networks with asymmetric or any-to-any routing nearly impossible to provide.

**Figure 1.** Example of an Overlay Network

A network-integrated and transparent solution eliminates the need for complex overlay networks, allowing automatic discovery of accelerators. With automatic discovery, shown in Figure 2, accelerators, such as Wide-Area Application Engine (WAE) appliances, first check to see if a peer accelerator exists in the path of packet flow between the source and destination. If an accelerator exists, an optimization policy is transparently negotiated and then applied to the application flow. If a peer accelerator does not exist, the application flow passes through normally, unchanged.

This provides deployment flexibility in that some locations may require optimization and some may not. Understanding the peering relationships across the WAN ensure that no data is optimized when no peer exists. By using automatic discovery, network infrastructure teams can deploy services to improve application performance over the WAN without having to implement complex overlay networks that require as much, or more, administration than the routed network, maximizing their investment in network and routing protocol design. By using network interception mechanisms such as Web Cache Communication Protocol Version 2 (WCCPv2), physical inline, dedicated load-balancing hardware, or policy-based routing (PBR), along with transparent accelerators that automatically discover one another, IT teams can easily implement application acceleration and WAN optimization within the network.

**Figure 2.** Example of Automatic Discovery

Automatic discovery allows the accelerators deployed on the network to identify each accelerator in the network path, even when users are working with remote resources that span multiple locations. With automatic discovery, intermediary accelerators can switch to pass-through mode for the connection and allow the accelerators closest to the communicating nodes to handle optimization of the flow. This function is not available in networks that deploy nontransparent accelerators and complex overlay networks.

### Bridging the Network Application Gap

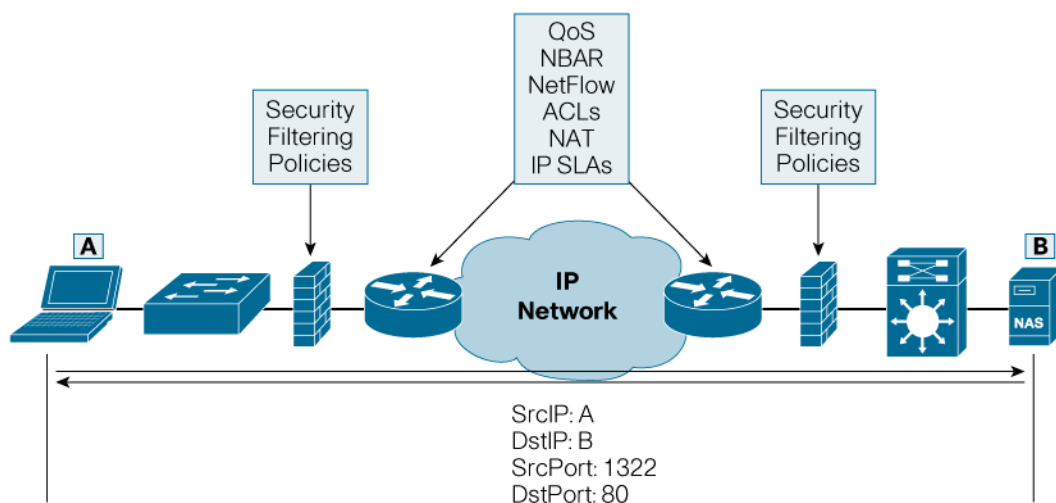
Often times, the network and application teams operate without real knowledge of each others requirements. Application teams see the network as “plumbing” and network teams see the application as “bits” or “just data”. In reality, the network and application are symbiotic and work best when optimized together. It makes no sense for the application to be optimized if, for instance, the network can’t give it priority over other applications on the WAN link.

When deploying WAN optimization and application acceleration solutions, WAN and application teams should assess the effect of the network on the application and vice versa.. Many solutions provide acceleration through the use of optimized connections between accelerators, called tunnels, whereby optimized traffic is routed directly through the network to the distant accelerator. Accelerators that use these tunnels are not transparent to the network, because critical packet header information is manipulated to help route the optimized packets to the distant accelerator. Such solutions mask IP address and TCP port information through a proxy, thereby rendering many networking services running on intermediary devices (routers) useless. Any network feature that relies on visibility to packet header information can be affected by the use of tunnels, including the following features:

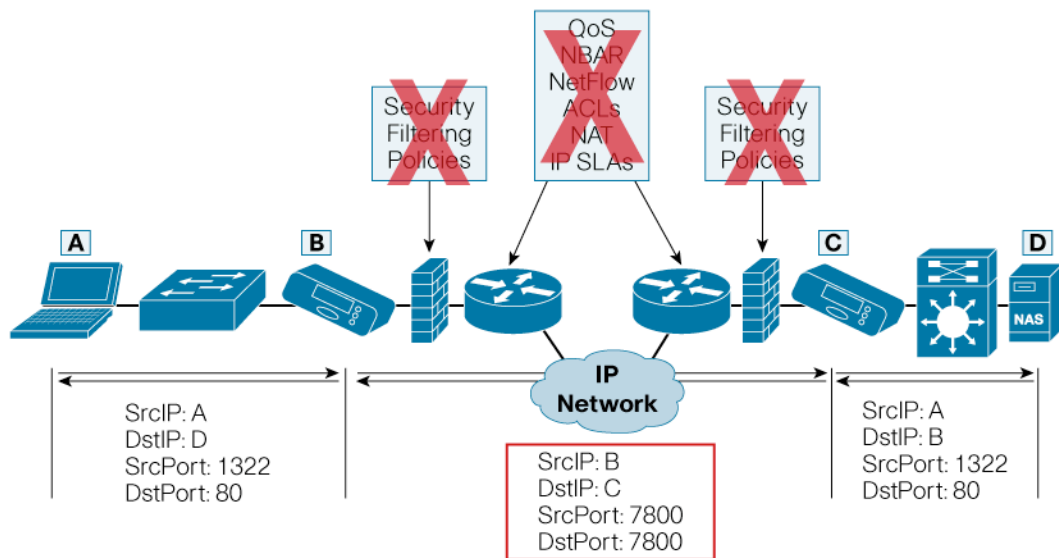
- **NetFlow**—Flows that are cached and sent to the collector no longer show the end nodes that are communicating or the TCP port that is being used. Instead, the collector receives information about flows between accelerators. Administrators examining NetFlow data from the edge routers see only flows between accelerators and can no longer see which users are working with which servers on which application ports.
- **Access control lists (ACLs)**—ACLs deployed on intermediary network devices may not function correctly because the accelerator may overwrite IP and TCP header information. The router where the ACL is configured then no longer sees the actual flows between the users and servers.
- **Firewall policies**—Firewall policies built using IP and TCP header information may not function correctly because the firewall no longer sees the end nodes that are communicating or the TCP ports that are being used. Instead of the actual flows between users and servers, the firewall sees the tunnels between accelerators.
- **Quality of service (QoS)**—QoS and associated features may need to be reconfigured if the packet header information is overwritten. QoS functions that are easily affected by accelerators include traffic shaping, policing, rate limiting, and queuing.
- **Network Based Application Recognition (NBAR)**—NBAR is a protocol discovery and classification technique that relies on visibility to application data. Any accelerator that overwrites the packet header and payload information prevents functions such as NBAR from correctly identifying and classifying data.

Figures 3 and 4 show how a typical network with value-added network features configured can be immediately impacted by accelerators that are non-transparent and manipulate packet header information.

**Figure 3.** Typical Network Without Tunnels

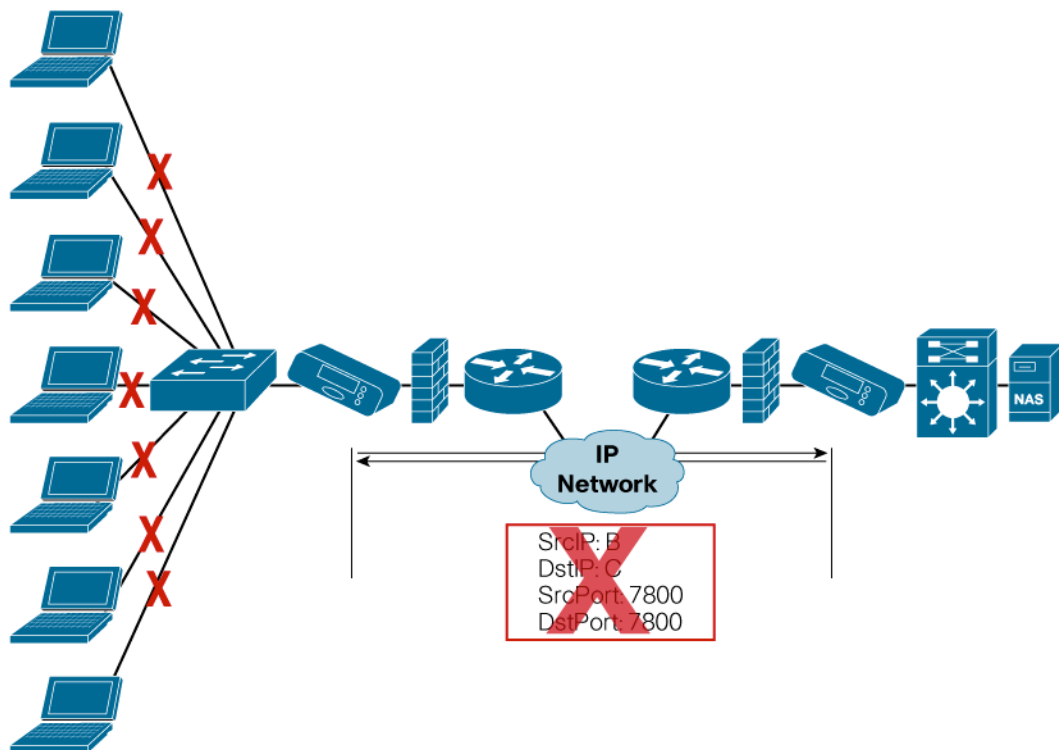


**Figure 4.** Effect on Network Features of Accelerators That Use Tunnels



The use of tunnels also affects connection stability. The accelerator devices aggregate all optimized flows for all communicating nodes over these optimized connections. Thus, if the network encounters a situation that compromises the stability of the tunnel and the tunnel connection fails, every connection traversing that tunnel is reset, as shown in Figure 5.

**Figure 5.** Effect of Tunnel Stability on All User Connections



Network infrastructure organizations do not need to deploy application acceleration or WAN optimization technologies that affect the visibility, monitoring, control, security, and efficiency of the network and decrease the stability of connections and associated applications. Instead, they can deploy an application acceleration and WAN optimization solution that does not rely on tunnels, thereby gaining the benefits of greater stability and compatibility with existing network features.

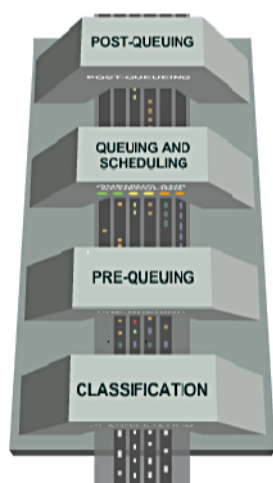
Cisco WAAS Version 4 provides such a solution.

### End to End Quality of Service

Quality of Service is one example of a service framework that needs to be examined from end-to-end and not just in a couple of places in the network. Deploying QoS in LAN-only solutions (such as non-integrated, non-transparent accelerators) implements QoS based on only the flows that traverse the device—which is all over the Ethernet network only. In essence, LAN-only QoS have no visibility to WAN conditions or traffic that is sourced in the WAN or by an intermediary network device. Many of the current accelerator products that attempt to provide this QoS functionality are deployed as LAN-side devices and have no visibility to the WAN because they have no physical or logical connection to it. Given that they have no visibility to the WAN and do not act as the network point managing the bandwidth disparity, they can not provide an adequate solution to meet the needs of the demanding enterprise customer. Nor can they provide end-to-end Quality of Service as they don't act as intermediary nodes at congestion points between communicating nodes.

Network-based QoS delivers much greater flexibility and provides an end-to-end architecture for aligning network resources with application requirements and business priority. The WAN is not merely a “connection”—rather, it is a system in which many applications, some mission-critical, some latency-sensitive, some jitter-sensitive, must all interleave together. Therefore, the network ecosystem from end-to-end must behave as a single entity and common policy and traffic handling characteristics must be ensured.

Figure 6.



Quality of Service (QoS) is designed to ensure that network resources are aligned with business priority relative to each of the applications being serviced, and configured to handle traffic based on the application requirements. For QoS to be effective, it must be deployed using four key stages:

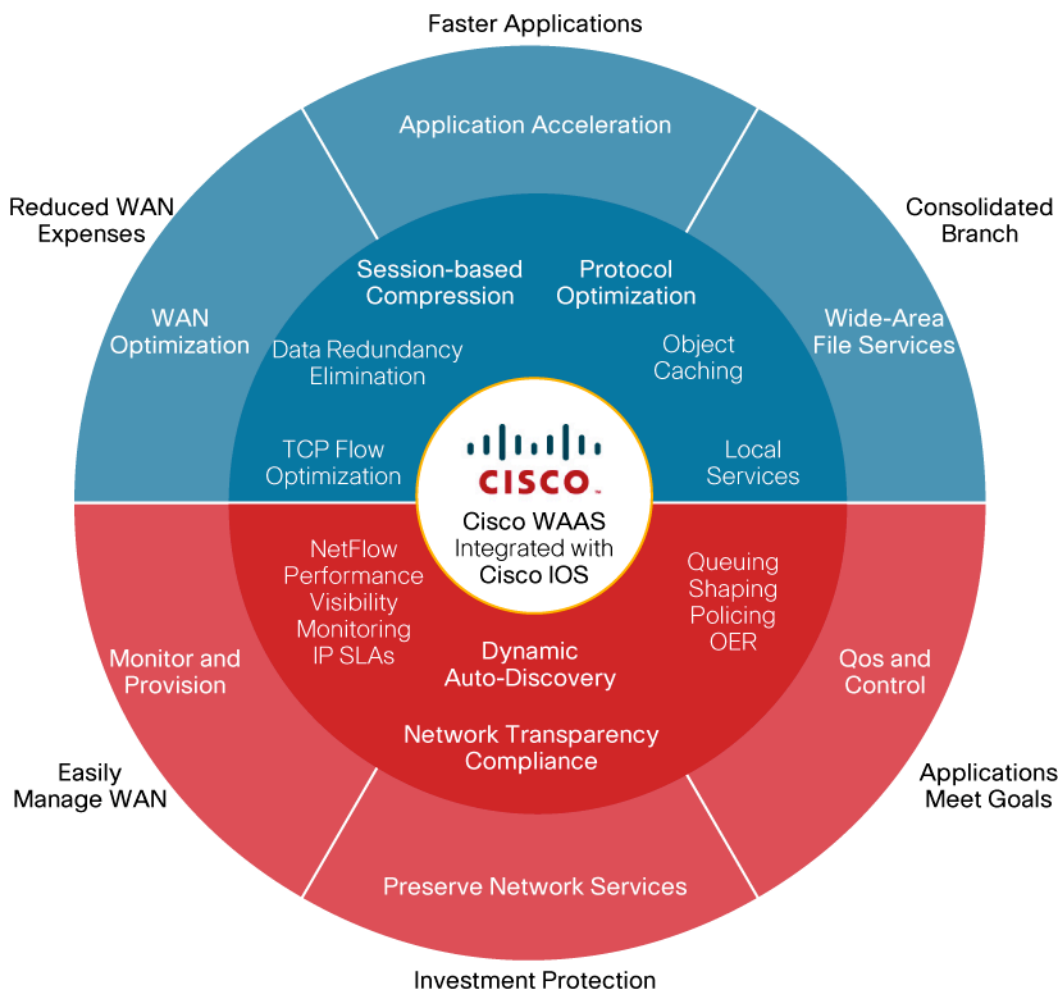
- **Classification**—the network must be able to identify and classify application traffic based on identifiers that span from the data link layer (layer 2) through the application layer (layer 7) using traditional classification and also deep-packet inspection. This includes data beyond the typical IP address and port information, and can include source interface, VLAN, application data, packet length, and more. When classification is deployed in the WAN router (or any network device for that matter), with or without the presence of a transparent accelerator, it is then able to adequately identify traffic based on one or more unique identifiers. This identification is then used for the remaining functions of QoS. With non-transparent accelerators, as shown above in figure 4, intermediary network devices such as WAN routers that may be attempting to classify traffic will have no ability to do so. While such non-transparent accelerators may provide some level of classification, they can not provide the same level of classification that is provided within network devices such as WAN routers. Furthermore, as LAN-only devices, they can only classify traffic that they see, which does not include traffic that is generated at the router such as session border control, DLSW, and many other commonly-used protocols.
- **Pre-Queuing Operators**—pre-queuing operators are functions applied to traffic before they are queued for service at the point of bandwidth disparity—the WAN. Any time traffic is attempting to traverse a connection, especially those that exist between two links of non-proportionate bandwidth, the traffic must wait in a queue for service. Pre-queuing operators allow the network to apply immediate actions to these flows to improve service and conserve network resources. These operators, such as traffic policing (throttling an application flow to a specified throughput level), packet dropping, counting, bandwidth estimation, and DSCP marking all rely on the ability of the network device itself to adequately identify the application based on one or more classifiers. For transparent accelerators, these features work seamlessly. For non-transparent accelerators, all optimized traffic will be funneled through one or more optimized connections or tunnels, which means that pre-queuing operators such as policing are applied against the tunnel rather than on a per-application basis.
- **Queuing and Scheduling**—not all applications or traffic types are created equal. Some applications require different means of being handled to ensure proper operation or performance. This can be measured in terms of perceived latency, packet loss, jitter, throughput, and other metrics. Queuing and scheduling refers to how packets are handled by devices managing bandwidth disparity and how they are serviced. Cisco IOS has a robust queuing and scheduling architecture that allows for traffic type to be handled based on application requirement and business priority—without compromise. LAN-attached accelerator devices commonly only allow a single type of queuing to be enabled, which can cause significant challenges in converged networks that contain a mixture of voice, video, and data, where multi-stage queuing and scheduling are necessary. Furthermore, non-transparent accelerators create the additional challenge of rendering network-based classification useless, which means the powerful queuing and scheduling architecture of the network can not be leveraged because traffic flows can not be differentiated.
- **Post-Queuing Optimization**—post-queuing optimization implemented in the network allows for optimized delivery of certain traffic types that require expedited handling under any condition. For instance, link fragmentation and interleaving allows small packets from latency and delay sensitive applications such as Voice over IP (VoIP) to be interleaved between fragments of larger application packets, thus ensuring predictable performance under varying load conditions.

With robust QoS deployed in the network, there is no need to deploy QoS on LAN-attached acceleration devices. By introducing accelerator QoS into the network, I/T organizations run the risk of having to manage two disparate QoS policies to account for scenarios where the accelerator fails and is unable to function, leaving the network itself to manage QoS. This creates the possibility of having overlapping or underlapping policy configurations, not to mention the additional administrative burden of managing QoS in two locations. Cisco WAAS provides the transparency necessary to ensure compliance with existing network QoS configurations.

**Cisco WAAS: Network Integrated and Service Transparent**

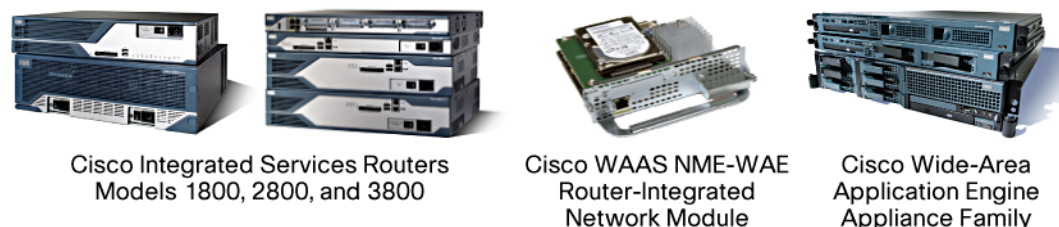
Cisco WAAS provides an outstanding network-integrated, transparent application acceleration and WAN optimization solution. By offering compliance with existing network configurations and features, Cisco WAAS allows IT organizations to safely deploy application acceleration and WAN optimization without compromising the enterprise’s investment in value-added networking services, including NetFlow, QoS, IP SLAs, ACLs, firewall configurations, and almost any network topology. As shown in figure 6, this helps IT organizations to control WAN expenses, consolidate distributed infrastructure, improve performance for centralized services, and more confidently deploy new applications and services.

**Figure 7.** Cisco WAAS and IOS—A Complete Acceleration and Optimization Framework



Cisco WAAS is also designed to physically integrate into the Cisco Systems® access routing platforms, such as the award-winning Cisco Integrated Services Router platforms. By deploying Cisco Integrated Services Routers with the Cisco WAAS router-integrated network module, IT organizations can minimize the remote-office footprint by consolidating server and storage infrastructure by embedding application acceleration and WAN optimization capabilities directly into the network fabric. This results in the need for managing fewer devices in the branch. Figure 7 shows the Cisco WAAS hardware platforms.

**Figure 8.** Cisco Integrated Services Routers and the Cisco WAAS Router-Integrated Network Module



Cisco WAAS provides a comprehensive set of application acceleration and WAN optimization services. By coupling these industry-leading acceleration and optimization services with true network feature compliance, transparency, and integration, Cisco WAAS provides an excellent solution for enterprises that want to consolidate costly infrastructure and improve application delivery.

- **Advanced network compression**—Cisco WAAS provides persistent Lempel-Ziv (LZ) compression for data in transit, as well as Cisco WAAS Data Redundancy Elimination (DRE). DRE compiles a history of application-independent segments that have been previously seen on the network. As redundant segments are seen, DRE intelligently replaces them with instructions that tell the distant Cisco WAE how to rebuild the original message in its entirety, without compromising data integrity or application accuracy. With DRE and persistent LZ compression, Cisco WAAS can achieve up to 100:1 compression.
- **Cisco WAAS Transport Flow Optimization (TFO)**—Cisco WAAS TFO provides a set of optimizations to improve the behavior of TCP in WAN environments. By overcoming TCP limitations such as inefficient handling of congestion or packet loss resulting from the throughput constraints associated with TCP, TFO can help applications use the WAN more efficiently while also shielding users and servers from problematic WAN conditions.
- **Application-specific acceleration**—Cisco WAAS provides application-specific acceleration capabilities that help reduce unnecessary messaging through intelligent message suppression, data and metadata caching and validation, message prediction, and read ahead. Through application-specific acceleration, Cisco WAAS can address WAN performance problems associated with application protocols that cannot be optimized through compression or flow optimization alone.
- **Application traffic policy**—Cisco WAAS Application Traffic Policy provides a comprehensive default policy list that specifies how Cisco WAAS should optimize application flows. Administrators have the flexibility to change existing policies, create new policies, or delete policies to configure Cisco WAAS to provide optimum optimization for their specific applications.

In comparison to other WAN optimization and application acceleration products, Cisco WAAS offers these benefits:

- Cisco WAAS autodiscovery and transparency eliminates the need to configure complex overlay networks.
- Cisco WAAS offers exceptional network-friendly WAN optimization and application acceleration capabilities.
- Cisco WAAS preserves critical packet header information and does not negatively affect value-added network features.
- Cisco WAAS integrates physically into network infrastructure devices to provide low total cost of ownership (TCO) for the branch office.

### Summary

IT organizations considering application acceleration and WAN optimization products to centralize costly distributed infrastructure and improve application delivery for remote users should assess the effects of such products on the existing network. Many products require complex administration through the configuration of overlay networks and mask the original packet header information, thereby making many network value-added features and services unusable. Cisco WAAS provides industry-leading application acceleration and WAN optimization capabilities without compromising on simplicity or compatibility. By using Cisco WAAS, IT organizations can effectively consolidate costly infrastructure, improve application performance over the WAN, and preserve the investment that has already been made in critical network services.



**Americas Headquarters**  
 Cisco Systems, Inc.  
 170 West Tasman Drive  
 San Jose, CA 95134-1706  
 USA  
[www.cisco.com](http://www.cisco.com)  
 Tel: 408 526-4000  
 800 553-NETS (6387)  
 Fax: 408 527-0883

**Asia Pacific Headquarters**  
 Cisco Systems, Inc.  
 168 Robinson Road  
 #28-01 Capital Tower  
 Singapore 068912  
[www.cisco.com](http://www.cisco.com)  
 Tel: +65 6317 7777  
 Fax: +65 6317 7799

**Europe Headquarters**  
 Cisco Systems International BV  
 Haarlerbergpark  
 Haarlerbergweg 13-19  
 1101 CH Amsterdam  
 The Netherlands  
[www-europe.cisco.com](http://www-europe.cisco.com)  
 Tel: +31 0 800 020 0791  
 Fax: +31 0 20 357 1100

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](http://www.cisco.com/go/offices).

©2007 Cisco Systems, Inc. All rights reserved. CCVP, the Cisco logo, and the Cisco Square Bridge logo are trademarks of Cisco Systems, Inc.; Changing the Way We Work, Live, Play, and Learn is a service mark of Cisco Systems, Inc.; and Access Registrar, Aironet, BPX, Catalyst, CCDA, CCDP, CCIE, CCIP, CCNA, CCNP, CCSP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Cisco Unity, Enterprise/Solver, EtherChannel, EtherFast, EtherSwitch, Fast Step, Follow Me Browsing, FormShare, GigaDrive, GigaStack, HomeLink, Internet Quotient, IOS, iPhone, IP/TV, iQ Expertise, the iQ logo, iQ Net Readiness Scorecard, iQuick Study, LightStream, Linksys, MeetingPlace, MGX, Networking Academy, Network Registrar, Packet, PIX, ProConnect, RateMUX, ScriptShare, SlideCast, SMARTnet, StackWise, The Fastest Way to Increase Your Internet Quotient, and TransPath are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the United States and certain other countries.

All other trademarks mentioned in this document or Website are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (0701R)