

APRIL 2004

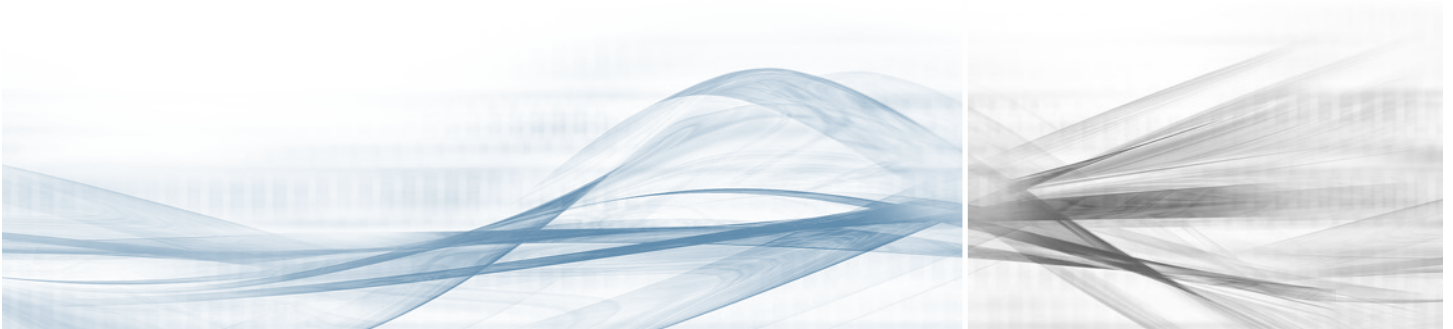


Expand Networks
103 Eisenhower Parkway
Roseland, NJ 07068 USA
TEL +1.888.892.1250
+1.973.618.9000
FAX +1.973.618.9254
www.expand.com

WHITEPAPER

Expand Networks Accelerators

A Technical Overview



INTRODUCTION TO EXPAND ACCELERATORS

Contrary to conventional wisdom, enterprise WAN costs are not going down. In fact, the Gartner Group says that the WAN is the single largest recurring IS cost, other than people, and they predict that WAN costs will increase an average of 7% annually in the future. For this reason network managers are seeking alternatives to expensive WAN upgrades because they don't have the luxury of simply "throwing bandwidth" at performance problems – an approach that is not only costly, but often ineffective.

Expand Networks provides a cost-effective alternative to WAN upgrades. By optimizing the way WAN bandwidth is used, Expand's Accelerators enable enterprises to support increased WAN traffic while guaranteeing the performance of critical applications.

WAN Optimization is the Solution

WAN optimization is a set of techniques and technologies that enable network managers to squeeze additional capacity and better performance out of their existing WAN infrastructures. This allows more users and applications to share enterprise WANs while still protecting the performance of critical applications. Three key elements are required for WAN optimization success.

- **Bandwidth Expansion** increases the effective bandwidth of WANs by eliminating the unnecessary or redundant information that is present in almost all protocols. The freed-up bandwidth can then be used to handle additional applications and user traffic just as if the WAN bandwidth had been upgraded.

- **Application Visibility** provides an understanding of the application traffic flowing on a WAN. It answers key questions such as how is bandwidth being consumed, by which applications, and by which users?
- **QoS management** is the ability to allocate bandwidth based on application requirements and business priorities. QoS enables network managers to take control of their bandwidth rather than relying on IP's best-effort service.

Effective WAN optimization depends on all three of these key elements.

Expanding WAN Capacity

As an alternative to expensive and potentially disruptive WAN upgrades, Accelerators increase the effective bandwidth of existing WAN links. They do this by eliminating the unnecessary or redundant information while maintaining the integrity of all traffic sent across the WAN. Various approaches such as compression and caching have been used over the years to reduce the volume of traffic flowing across WANs but those legacy solutions are of limited value in enterprise networks.

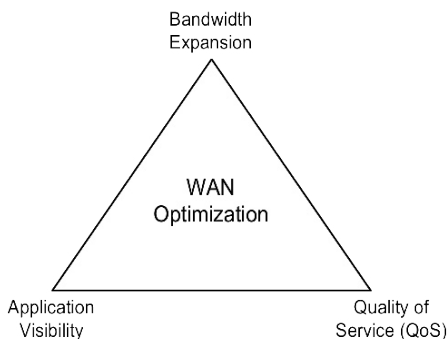
A Brief History of WAN Link Acceleration

The benefits of accelerating WAN links by reducing link traffic volume have been recognized for decades. Early attempts to reduce WAN traffic centered on the use of legacy data compression algorithms. Two of the most widely used compression schemes are:

- **Lempel-Ziv compression** which replaces recurring sequences of bytes with more compact codes that refer to previous occurrences of those byte strings
- **Huffman encoding** which takes advantage of the fact that some characters occur much more frequently than others. Frequently occurring characters such as "a" or "e" are represented by fewer bits while less frequently occurring character such as "x" or "z" are represented by longer bit strings. The result is a reduction in the volume of data sent.

These compression schemes worked quite well on the textual data that was predominant in early data communications networks, but they are ineffective when applied to pre-compressed data, such as GIF/JPEG images or ZIP files.

The emergence of the Web drove vendors to solve the problem of accelerating pre-compressed static objects, particularly GIF and JPEG images. The solution was Web caching which captures pre-compressed static objects and delivers them to users from caching servers distributed throughout the Internet. These Web caching solutions are effective on the Internet because it carries a tremendous amount of Web traffic, but on enterprise networks it's



a very different story. Web applications represent only a small percentage of the traffic on most enterprise networks. This makes legacy compression and caching solutions ineffective. Enterprise-class solutions have to handle a wide range of applications including ERP, CRM, file transfers, Citrix, etc.

Next-Generation WAN Acceleration

In contrast to the legacy solutions, Expand Accelerators employ next-generation compression, caching, and other technologies that handle all application types and protocols, including non-IP protocols. Some enterprise data is compressible and some is not. Some is static and some is dynamic. To accelerate these diverse data types Accelerators employ a combination of techniques. A one-size-fits-all approach just isn't adequate. Accelerators constantly analyze WAN traffic and dynamically select the combinations of techniques that will most effectively reduce traffic volume. The key techniques used by Accelerators are:

- **Selective Caching** is an advanced algorithm that identifies data that is worth caching and stores it for later retrieval. "Data" as defined by Selective Caching may be full objects (such as an entire GIF file), or a partial object such as a colour palette of a GIF file, common JavaScript code that appears in HTML pages, a bitmap that is being repeatedly sent by the Citrix ICA protocol, etc. It is not limited to a specific protocol. Once Selective Caching identifies data that is worth caching it will store a copy of it for later usage. If that data is encountered again, only a reference to the data will be sent rather than the data itself.
- **Vertical Data Analysis (VDA)** dynamically identifies different protocol segments in a packet. Generally, each segment relates to a different part in the protocol stack. The separation into different segments is done in a "semi-dynamic" manner – VDA is programmed with basic rules and with an understanding of some protocols. Using this pre-programmed adaptive understanding, it parses packets into their different segments. Consequently, "header style" information in each packet is reduced significantly. Examples of "header style" information are sequence numbers, checksums, protocol identifiers etc.
- **Adaptive Packet Compression (APC)** is used in conjunction with VDA and Selective Caching to compress data that is not cached or which was not handled by VDA. Different data types are assigned different compression algorithms. For example, HTML, SQL, and JavaScript will be handled in a different manner. By applying a different compression strategy to each of these types of data, better results can be achieved compared to legacy compression.

- **Packet Aggregation** reduces the overhead of sending many short packets across a WAN. Multiple short packets are assembled into a single packet that is sent between a pair of Accelerators. The receiving Accelerator then recreates the original packets.
- **Packet Flow Control** manages the rate at which packets are sent across WAN links to ensure that packets are not dropped due to congestion. This eliminates retransmissions that waste WAN bandwidth and increase end-to-end latency.
- **Packet Loss Recovery** automatically recovers any packets that might be lost on a WAN link between a pair of Accelerators. The number of end-to-end TCP retries that are usually required to recover lost packets is reduced. This is particularly important when operating over high-latency WAN links.

By using a combination of technologies rather than a one-size-fits-all approach, Accelerators are able to routinely achieve 100% to 400% compression rates with peaks of 1,000% across a wide range of traffic types. Accelerators intelligently choose the combination of techniques that will deliver the best results for each application or protocol type.

The benefit of this multi-technology approach is a significant reduction in WAN costs. A 300% capacity increase, for example, is equivalent to upgrading a 128 kbps WAN link to 512 kbps, but with no increase in recurring monthly costs.

Gaining Application Visibility

Application visibility enables network managers to see all traffic flows on a WAN and to identify the applications generating the traffic. Without such traffic monitoring and analysis capability WAN optimization initiatives are just hit-or-miss processes.

Expand's advanced graphic reporting feature monitors all traffic handled by Accelerators, analyzes that traffic and uses a Web interface to display key statistics. The graphs are automatically updated, according to a set frequency. The Accelerator samples the data behind-the-scenes and stores it in a compact way enabling you to view data up to the minute or over the period of a year. The statistics can also be exported to other applications for viewing or analysis.

The reports generated by Accelerators are valuable at the start of WAN optimization initiatives to create the baseline measurements that enable network managers to determine exactly how their bandwidth is being used. The baseline data enables network managers to determine what QoS policies will be needed to align bandwidth utilization with their business objectives. This

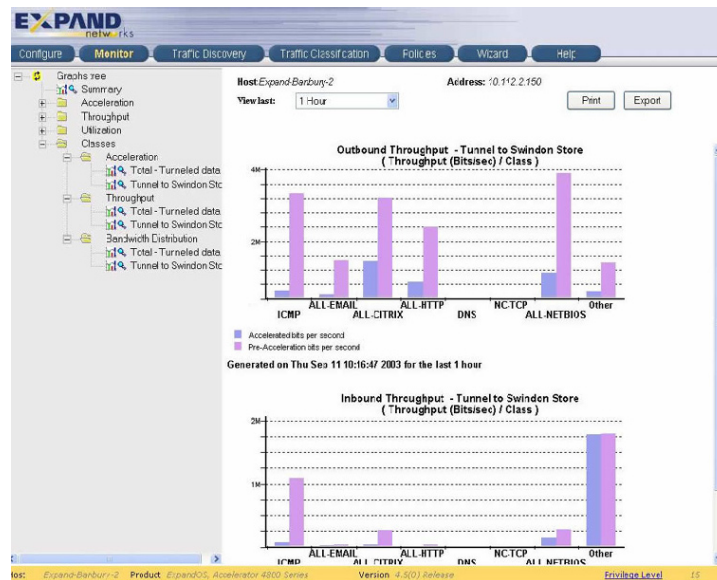
data can also be used later to create a “before and after” view of the effectiveness of both WAN acceleration and QoS policies.

Graphs Generated by Accelerators

Accelerators generate a variety of standard graphs that measure link utilization and the effectiveness of link acceleration and QoS management.

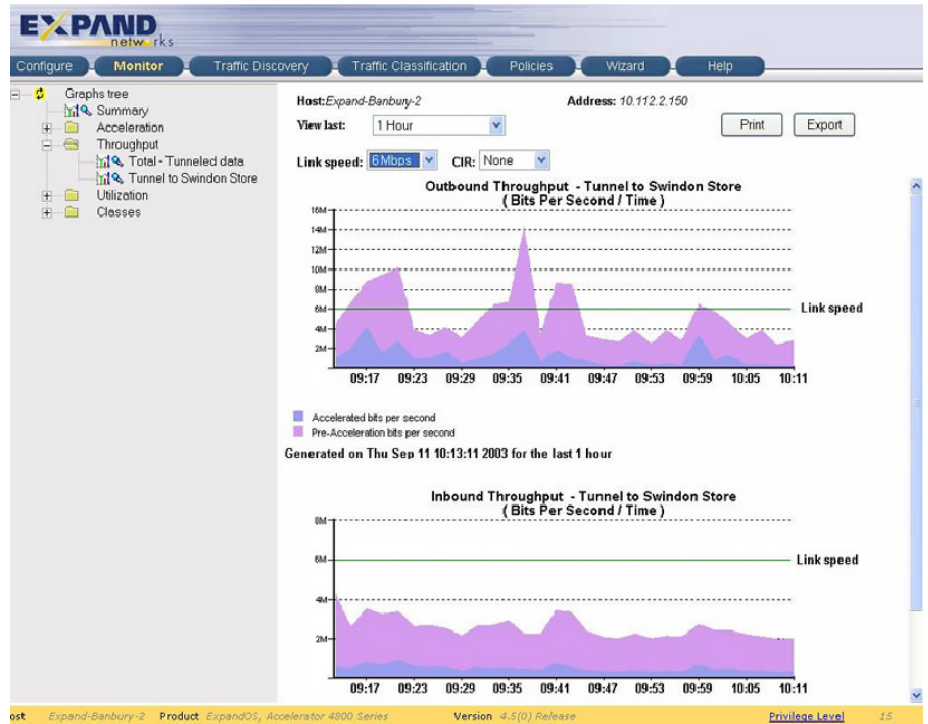
- **Per-class graphs** enable you to view acceleration and throughput statistics for specific applications or traffic classes.

The following graph shows the throughput per class before and after acceleration:



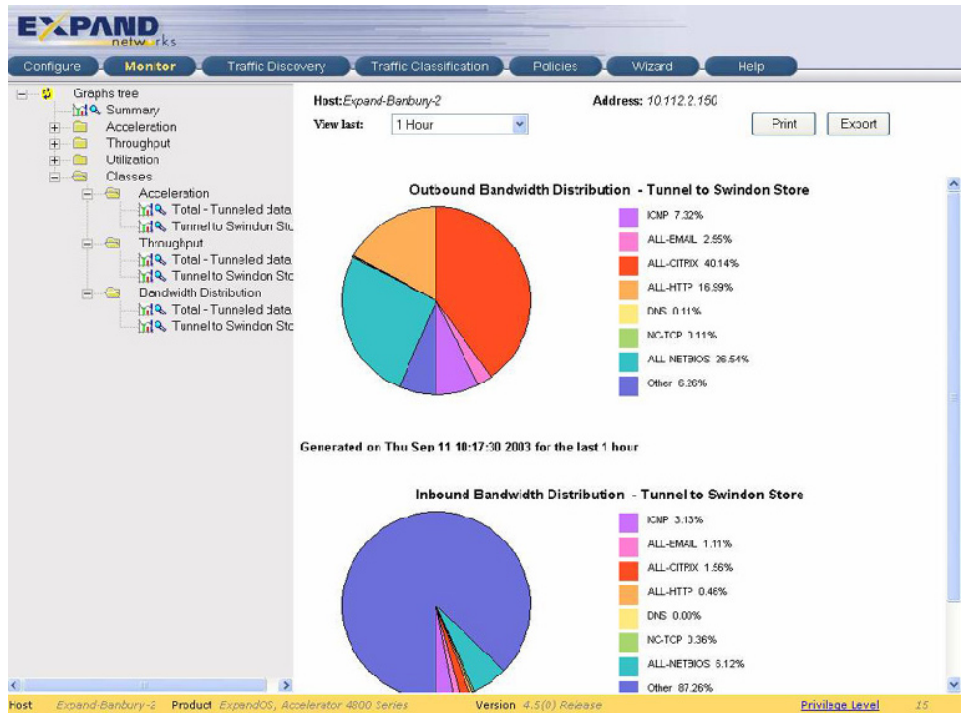
- Graphs of **acceleration percentages** enable you to view the amount of additional effective bandwidth created for both inbound and outbound traffic. These statistics can be per interface/tunnel or for the total for the Accelerator. Users can select the time period for which the acceleration percentages are displayed
- **Throughput graphs** you to monitor how much total throughput passed through the Accelerator. It enables you to compare accelerated throughput that actually goes over your WAN link to the pre-accelerated throughput, which is what your network is able to transmit.

The following graph shows the inbound and outbound throughput for a specific tunnel to an Accelerator at a remote site:



- **Utilization graphs** enable you to monitor how much of the current link connected to the Accelerator, or to a remote site, is being utilized.
- **Bandwidth Distribution Graphs** detail the percentage of bandwidth consumed by each selected traffic class. The distribution of accelerated data is displayed.

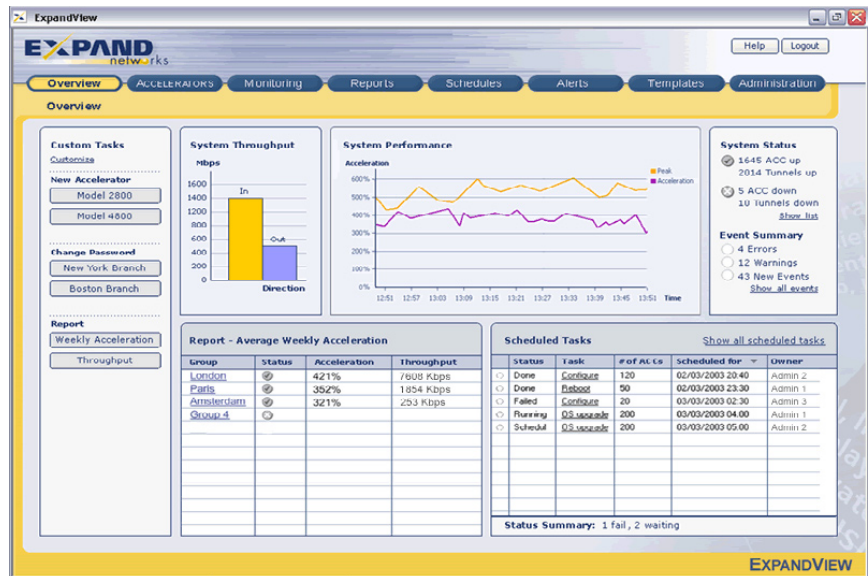
The following graph shows the bandwidth utilization by traffic class:



In addition to their standard reports, Accelerators enable users to customize their reports and export the traffic data to other management and accounting applications.

ExpandView

ExpandView is an optional server-based software package that enhances application visibility by providing centralized management and monitoring capabilities. Network managers can monitor characteristics such as acceleration rates, throughput, and link utilization. ExpandView also simplifies the administration of Accelerators in large networks. The following is an example of an ExpandView screen that provides an integrated view of throughput and acceleration for a set of distributed Accelerators.



RMON/RMON2 Support

Accelerators also fit seamlessly into existing network management systems. They can be managed by widely used SNMP management platforms such as HP OpenView. In addition, each Accelerator has an integrated, industry-standard RMON/RMON2 probe that can provide application-aware traffic statistics to almost any network management platform including ExpandView. RMON/RMON2 differs from basic SNMP management in its ability to provide application-level visibility. For example, RMON2 can identify the “top talkers” and “top listeners”, the devices that are sending and receiving the most data on a link.

Instant QoS

QoS enforcement is the process of proactively controlling the allocation of bandwidth. Without QoS, IP networks are only capable of delivering a best-effort service that provides bandwidth on a first-come first-served basis. But QoS management enables network managers to align bandwidth utilization with business goals and with the technical requirements of critical real-time applications such as VoIP and video conferencing.

Accelerators include a feature set called Instant QoS. As the name implies, Instant QoS is designed to provide the QoS management that enterprises need while minimizing the operational complexity and cost often associated with QoS management.

Discovering Traffic

Accelerators automatically detect and identify traffic for over 100 of the most common applications and protocols found on enterprise WANs. Traffic flows can be classified based on Layer 2, 3, and 4 header information. The layer-4 analysis enables Accelerators to identify traffic associated with specific applications such as Citrix, SAP, Oracle, Siebel, and MS Exchange.

Traffic discover gives you a view of all of the traffic that is contending for the available bandwidth on your WAN.

Classifying Traffic

Traffic categories, or classes, can be based on a variety of factors including application type, protocol, port number, or source and destination IP or MAC address. Each separate category is called a traffic class. For example, SAP traffic to a specific server can be set as a traffic class. The Accelerator's advanced classification capabilities equip you with a large range of predefined classes and also enables you to design your own traffic classes.

Classes are defined via a set of matching rules, which specify traffic types to be included in the class. Incoming packets are checked against these rules and then included within or excluded from the class based on conformity to the rules. Hundreds of classes are pre-defined in the device and user-defined classes can be added simply, by copying attributes and rules from pre-existing classes.

Defining QoS Policies

Once the traffic traveling through the Accelerator has been discovered and classified, decisions can be made as to which queuing policies will provide the desired bandwidth allocation. Accelerators support all of the most widely used queuing policies:

- **First-in-first-out (FIFO)** is the most basic queuing strategy. Packets exit the interface in the order in which they arrived. FIFO enables the creation of basic queues enabling buffering capabilities at the Accelerator's interfaces, but does not prioritize data
- **Priority queuing (PQ)** allocates bandwidth in an absolute manner, assigning unconditional priority to higher priority traffic. This is designed for environments that focus on mission-critical data, excluding or delaying less critical traffic during periods of congestion. Lower priority traffic is only forwarded once the Accelerator has forwarded all high priority traffic, in effect, causing bandwidth starvation to low priority traffic during high-priority traffic transfers.

- **Weighted fair queuing (WFQ)** divides bandwidth fairly among IP traffic. WFQ gives small-packet traffic priority over large-packet traffic and all traffic gets at least a small amount of bandwidth to prevent application time-outs. Small-packet traffic, such as VoIP or any other time-sensitive traffic, enters WFQ's Low Latency Queue (LLQ), meaning that it is allocated a set portion of the total bandwidth. When WFQ is set to be ToS sensitive, the Accelerator prioritizes packets based on their Diffserv ToS bit settings
- **Custom queuing (CQ)** sends packets to one of 17 queues, based on which classes are assigned to which queues. These customizable queues represent 17 sub-pipes that can be used to logically partition the bandwidth of a WAN link. Configuring the ratio for these queues enables bandwidth to be distributed in controlled amounts among these sub-pipes. The logical partitioning capability of CQ can be used to allocate specific amounts of bandwidth to critical applications such as VoIP.

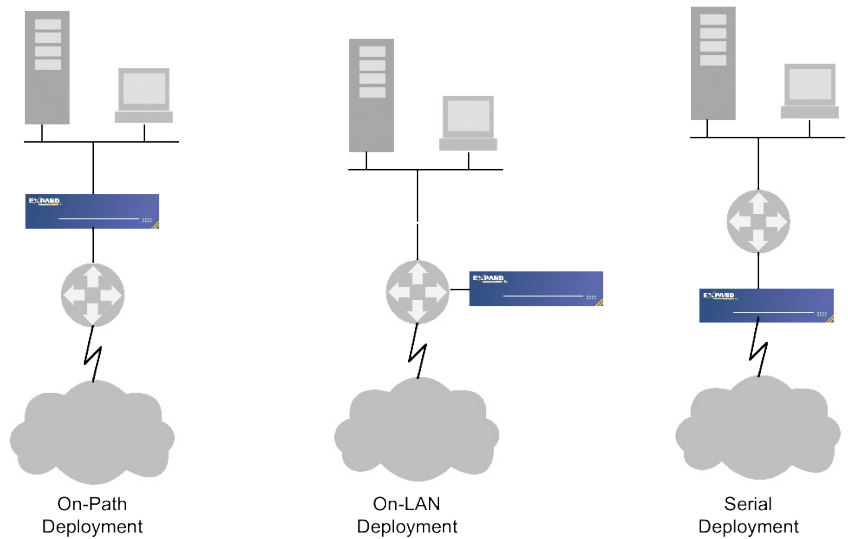
Once traffic classes have been defined and assigned to Instant QoS policies, the Accelerator's scheduling feature enables you to activate these policies at specific times. Policies can be set to be active on a one-time, or a weekly-recurring basis.

Controlling Latency

The final Instant QoS feature is the ability of Accelerators to fragment large packets that can increase the latency of real-time applications such as VoIP and video conferencing. When a large packet is being serialized onto a low-speed WAN link, any successive packets will be held in the output queue until the large packet has been sent. This creates high latency and jitter that can make real-time applications virtually unusable. By using Accelerators to fragment these large packets, network managers can keep latency within the limits of even the most critical real-time applications.

Accelerator Deployment Options

Accelerators can be deployed in any enterprise network without changes to the network infrastructure. They can be used with any WAN technology including private line, frame relay, VPN, IP, ATM, xDSL, ISDN, wireless local loop, or satellite. Accelerators bring the benefits of WAN optimization to all networks. Accelerators can be connected on the LAN side of the router or between the router and any WAN termination device such as a CSU/DSU.



There are two LAN deployment options, On-Path and On-LAN. On-Path configuration places the Accelerator between the LAN and the router on both sides of the IP network. The data from the LAN segment passes through the Accelerator before it reaches the router. The Accelerator changes the destination IP address of the accelerated data, rerouting it via the far-end Accelerator to be reconstructed before it is passed on to its final destination IP address.

On-LAN configuration places the Accelerator directly on the LAN as a host. The Accelerator is considered the next hop for all traffic on the LAN. The accelerated data is redirected to the far-end Accelerator where it is reconstructed before it reaches its final IP address.

In addition to LAN deployment Accelerators have a serial deployment option that places the Accelerator between the WAN router and a WAN termination device, usually a CSU/DSU. Serial deployment accelerates all protocols while the LAN-based options accelerate only IP traffic.

In any of these deployment scenarios, the network operations staff can rely on the Accelerator's Configuration Wizard to simplify the configuration of both local and remote Accelerators, The PC-based wizards be used to create logical tunnels between pairs of Accelerators, set IP addresses and subnet masks, and manage passwords.

The combination of the Configuration Wizard and ExpandView enable network managers to easily deploy, maintain, and monitor Accelerators in large networks with hundreds or even thousands of remote sites.

LAN Resilience

When multiple Accelerators are deployed On-LAN they can take advantage of the Hot Standby Router Protocol (HSRP) and Virtual Router Redundancy Protocol (VRRP) to provide redundancy for IP networks, ensuring that user traffic immediately and transparently recovers from an Accelerator outage.

When using HSRP or VRRP, multiple Accelerators appear to be a single virtual router to the hosts on the LAN.

SUMMARY

Expand Accelerators provide a complete set of WAN optimization capabilities to:

- Increase the effective bandwidth of WAN links (typically 100% to 400% and up to 1,000% depending on the traffic mix).
- Provide application visibility to track bandwidth utilization and the effectiveness of Expand's WAN acceleration techniques
- Use prioritization and bandwidth partitioning to manage QoS based on application requirements and business priorities

All of these capabilities are integrated into each Accelerator and they can be deployed with no changes to existing network topologies. Accelerators are an extremely cost-effective alternative to WAN upgrades because they typically deliver an ROI of only 3 to 9 months.



© Expand Networks 2004. All right reserved. Expand Networks, Accelerator System 9000, Accelerator Server, Accelerator 6800 Series, Accelerator 4800 Series, Accelerator 4000 Series, Accelerator 1800 Series, HTTPS Accelerator Series and ExpandView are trademarks of Expand Networks. All other trademarks are the property of their respective owners.

Expand Networks
103 Eisenhower Parkway
Roseland, NJ 07068 USA

TEL +1.888.892.1250
+1.973.618.9000
FAX +1.973.618.9254
www.expand.com