

White Paper

# Accelerating Application Performance Across the WAN

---



Juniper Networks, Inc.  
1194 North Mathilda Avenue  
Sunnyvale, CA 94089 USA  
408 745 2000 or 888 JUNIPER  
[www.juniper.net](http://www.juniper.net)

Part Number: 200137-001

## Accelerating Application Performance

The ultimate goal of all network infrastructures is to deliver applications effectively to users. Continuous performance improvements in PCs and LAN infrastructure have made it easy for IT to enable effective application delivery within a building or campus. But maintaining an acceptable level of application performance across the WAN proves a consistent challenge.

To cope with poor performance across the WAN, enterprises have had to make a number of accommodations. IT organizations have been forced to proliferate data centers and server hardware around the world, install applications locally within branch offices, and endure the high cost of increased WAN bandwidth—typically second only to staffing as IT’s highest expense—just because applications run slowly or simply cannot operate across the typical poor-performing WAN link.

As businesses demand tighter integrated globalization, pulling even the most remote branch-office workers into the core business processes and applications, improving application performance through WAN optimization has become paramount. To meet this requirement, IT needs to address the WAN attributes that have slowed application performance for decades:

- bandwidth limitations
- latency
- application contention

The staggering drop in bandwidth as traffic moves from the LAN to the WAN is obvious and well understood. The effects of latency are sometimes less obvious, but latency often slows application performance even when ample bandwidth is available. According to Gartner’s Mark Fabbi, “Network managers who do not spend time addressing the latency issue will not meet the service levels that global applications and business processes demand.” Finally, application contention becomes far more prevalent on bandwidth-restricted WAN links, sometimes getting worse as a result of addressing the bandwidth limitation.

To improve application delivery, IT must look for tools that increase available WAN bandwidth, accelerate applications despite the presence of latency, and resolve contention. In their search for these tools, IT is best served by looking for options that incorporate support for both general TCP acceleration as well as application-specific acceleration to improve the performance of key business applications and processes.

Ultimately, IT needs a broad-based approach to WAN optimization that will accelerate the widest cross-section of their business applications, improve the flexibility of their delivery options, and provide the necessary monitoring and reporting to track application performance.

“WAN planning and design efforts have traditionally centered around bandwidth utilization. However, latency plays a bigger role in determining application performance, especially in global WANs.”

**Mark Fabbi**  
Gartner

## Bandwidth Limitations—Dealing with the Smaller Pipe Size

The most obvious restriction for application delivery over the WAN is the reduced bandwidth available on WAN links. As a percentage of LAN bandwidth, WAN speeds have actually increased over the past couple decades, but discrepancies of 100- to 200-fold are still routine.

While LAN infrastructure has scaled to 10 Gbps of bandwidth, a 155 Mbps OC-3 link is considered very high bandwidth among WAN connections. Typically, businesses rely on T-1/E-1 links running at 1.5 to 2 Mbps or T-3/E-3 connections running at 45 Mbps or 34 Mbps. The typical 100 Mbps fast Ethernet LAN provides more than 60 times the bandwidth of a T-1 link.

Businesses today run more applications across the WAN, and often the bandwidth requirement per application has increased. Web-enabling applications such as ERP systems can cause a transaction's bandwidth to increase as much as 10 fold compared to the bandwidth required for that same transaction in a client/server application architecture.

To keep up with these increases, enterprises have had to constantly increase the size of their WAN links. But given the high cost of these recurring expenses, enterprises are understandably reluctant to increase them. A compelling alternative is compression technology, which replaces repeated data sequences with short flags for transmission across the WAN link.

Traditional compression techniques provide limited bandwidth increases and often introduce additional latency, clearly at odds with improving application performance. Network infrastructure staff today must look for next-generation compression techniques that can dramatically reduce the transmitted traffic and do not slow application delivery.

IT also needs to consider another element when investigating compression options—the kinds of data patterns that benefit from compression. Various compression algorithms and implementations benefit different data sequence types. Algorithms that run solely in memory and operate on shorter data patterns benefit short, chatty applications such as SQL and HTTP. Other approaches that add hard-disk storage to the solution can store longer data sequences and can store them for longer periods of time. This kind of sequence caching approach eliminates repeated data patterns in larger files, such as a PowerPoint presentation, even if the file itself has changed and even when the last transmission occurred weeks earlier. Sequence caching is an ideal technology for collaboration projects that rely on file sharing and for storage applications.

IT needs a combination approach of compression techniques to achieve the highest overall reduction across a broad range of application types

## Latency and the TCP Bandwidth-Delay Product—When Physics Slows Things Down

The impact of latency has historically been a little less obvious and less well understood than the bandwidth limitation on WANs. Latency, which Gartner calls "the silent killer of application performance," refers to the round trip time (RTT) for a packet to traverse from a sender to a receiver. On

"Network latency is the most important parameter in determining application performance."

**Mark Fabbi**  
Gartner

WAN links that cross the United States, typical latency times are 75 ms to 100 ms. In global networks, that RTT routinely reaches 250 ms or more. Latency on satellite links routinely reaches 320 ms to 430 ms.

Different application types are impacted by differing amounts of latency. While e-mail can still perform reasonable well over links with high latency, terminal services will be significantly impacted on the same link, resulting in dropped sessions in many cases.

Sometimes latency doesn't just degrade application performance – sometimes it limits the overall application throughput. Enterprises that have purchased sizable WAN links often assume they have protected against application performance problems because they have ample bandwidth. But latency can limit throughput regardless of bandwidth.

Reducing latency itself is not possible – latency is simply a result of the physics of the speed of light over longer distances combined with store-and-forward hops across routers. What IT needs to consider, then, is how to reduce the impact that latency has on how enterprise applications behave. Applications based on TCP as the reliable transport protocol are especially susceptible to latency limitations.

“In global networks running a typical Web-based application, the WAN latency is responsible for more than 50 percent (in a 128 Kbps network) to 95 percent (in a T-1/E-1 network) of the total application delay,” says Gartner's Fabbi.

TCP relies on a series of handshakes and acknowledgements to ensure reliable traffic delivery. Applications waiting for these processes to complete cannot fully fill the bandwidth available on the WAN and are likely being slowed down by latency. The easiest way to estimate the impact that latency will have on an application's performance is to calculate the “bandwidth-delay product.” The bandwidth-delay product is an equation that says the capacity of a link is equal to the bandwidth of that link multiplied by latency as measured in round trip time or RTT.

**bandwidth-delay product: capacity = bandwidth \* RTT**

Think of capacity as how much the WAN pipe can hold across its length at any given point in time. If an application is able to fill the available bandwidth, then latency has not restricted it and the enterprise is enjoying full utilization. If an application is not able to fill it, and bandwidth remains available, the enterprise is not benefiting from full utilization and latency is the gating factor.

An example illustrates the equation. Consider a T-1 link running across the United States:

$$1.544 \text{ Mbps} \times 90 \text{ ms RTT} = 138,960 \text{ bits} = 17370 \text{ bytes} = 17.3 \text{ KB}$$

Compare the bandwidth-delay product to the host's TCP window size. When the product is less than the window size, bandwidth is the limiting factor; when the product is greater than the window size, latency is the limiting factor. Since the maximum window size is 64 KB, and for many systems the window size is 16 KB, the impact of latency is often the constraining factor on performance.

Keep in mind, though, that the result of this equation can change dramatically once compression and sequence caching are introduced. For example, that same T-1 link with 90 ms of latency, which is currently bandwidth constrained, would immediately become latency constrained with 4x compression results:

$$6.176 \text{ Mbps} \times 90 \text{ ms RTT} = 555840 \text{ bits} = 69480 \text{ bytes} = 69.5 \text{ KB}$$

Sequence caching, which can increase capacity by up to 50x, can make shorter links latency-constrained. When latency is slowing throughput, WAN optimization technologies that change TCP's behavior provide the only opportunity for accelerating application performance. Techniques that change the window size of older clients, eliminate a RTT from the TCP session startup, or replace TCP with another reliable but more efficient transport can significantly increase application performance.

## Latency and Application-Specific Protocol Tuning

For some applications, the gating factor for performance is not the TCP window size but other behavior within the application's Layer 7 protocol. In those cases, solving the TCP limitation will do nothing to increase performance because the internal workings of the application's protocol will continue to restrict the application's throughput and speed. Only after these shortcomings are addressed with protocol-specific acceleration can the application then benefit from TCP acceleration.

Three common applications affected by this type of Layer 7 protocol constraint are Microsoft Exchange, Microsoft file services, and web-based applications. The protocols underlying these applications all transfer information in small blocks and need acknowledgements after each—so-called “ping-pong” behavior that leaves these applications highly impacted by latency.

### Microsoft Exchange

Exchange relies on the Messaging Application Programming Interface (MAPI) as its underlying protocol. MAPI breaks up Exchange e-mails and attachments into small data blocks—sometimes as small as 8 KB—for transmission. The protocol requires an acknowledgement for each data block before sending the next one. Since it could take hundreds or even thousands of RTTs to complete a transfer, even a WAN link with as little as 20 ms to 30 ms of latency will dramatically slow the sending and receiving of e-mail messages.

Often, the problem worsens when the user's desktop freezes while the Exchange transfer completes. This is typical in versions prior to Exchange/Outlook 2003, where clients always run in “online” mode when a connection to the server is available. In online mode, the message contents and the attachment were not downloaded to the PC until opened by the user. Microsoft introduced a new mode in Exchange 2003 in which, when paired with Outlook 2003 clients, attachments and message contents are downloaded in background to reduce response times from centralized servers across the WAN. However, the new “connected” mode doesn't speed the sending or receiving of large attachments and doesn't reduce the bandwidth load on the WAN when Exchange servers are centralized.

## Microsoft File Services

Microsoft file services rely on the Common Internet File System (CIFS) protocol, which behaves like Exchange in its use of small data blocks for file reads and writes. Again, the “ping-pong” behavior as each transmission requires an acknowledgement delays users in branch offices trying to open, read, or write to files stored on centralized servers.

## Web

Web applications, running over HTTP, are similarly subject to “ping-pong” behavior. In the case of HTTP, the applications can transfer data at the full TCP window size, but the protocol retrieves the individual objects on each page one at a time. Since most web pages have a few dozen objects per page, it can take dozens of RTTs to retrieve all the objects associated with a URL.

Web caching often cannot accelerate this process since, in most cases, the protocol will still check with the web server to confirm the freshness of an object before sending it to the client. As a result, the latency impact remains even when caching reduces the WAN bandwidth.

## Application Contention—Deciding Who Goes First

Once applications hit the restricted bandwidth available on the WAN, they must contend for access to that precious bandwidth. Increasing bandwidth through compression techniques reduces contention, but it’s not possible for IT to fully eliminate the possibility of contention.

IT needs a simple but effective approach to resolving application contention. With straightforward and easy-to-use bandwidth allocation and Quality of Service (QoS) tools, IT can configure which applications should get priority across the WAN. The key is “straightforward” and “easy-to-use” tools—in too many cases, the QoS and other bandwidth-management techniques in network equipment and WAN optimization platforms are too detailed and fine-grained for IT to readily apply. QoS tools that require setting parameters that IT typically does not know, such as flow characteristics, defeat the purpose since those tools are rarely applied.

What IT needs instead are simple, wizard-based GUIs that walk them through the steps needed to set up QoS, using common-sense, application-based information to apply business policies and set prioritization requirements.

## How Juniper Networks Increases Application Performance

To help enable secure and assured delivery of differentiated applications and services within the enterprise intranet, Juniper Networks' application acceleration platforms improve application response times within central sites, to branch offices, and for remote users. The WX™ and WXC™ application acceleration platforms specifically address the issues of running applications across wide-area links, delivering LAN-like performance for branch office users accessing centralized resources.

The WX and WXC platforms are based on the WX Framework™, which defines several interdependent technologies that dynamically respond to each other to optimize their settings. The WX Framework includes techniques for compression and caching, acceleration, application control, and visibility.



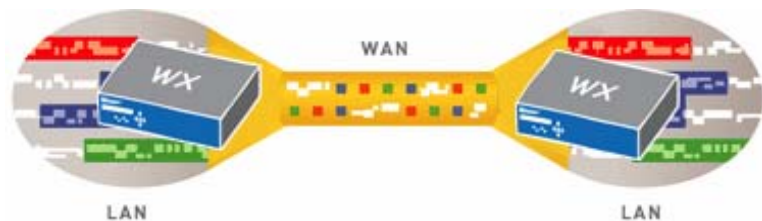
The Integrated WX Framework

## How Juniper Increases Bandwidth

The patented Molecular Sequence Reduction™ (MSR™) technology, which has its roots in DNA pattern matching, is Juniper Networks' flagship compression algorithm that enables enterprises to realize as much as a 10-fold increase in WAN capacity. MSR compression recognizes repeated data patterns and replaces them with labels, dramatically reducing WAN transmissions. MSR operates in memory, and its dictionary can store hundreds of megabytes of patterns.

The MSR compression capabilities benefit a broad cross-section of application types. It effectively reduces both short, chatty applications such as SQL and HTTP, as well as larger data patterns such as Word files.

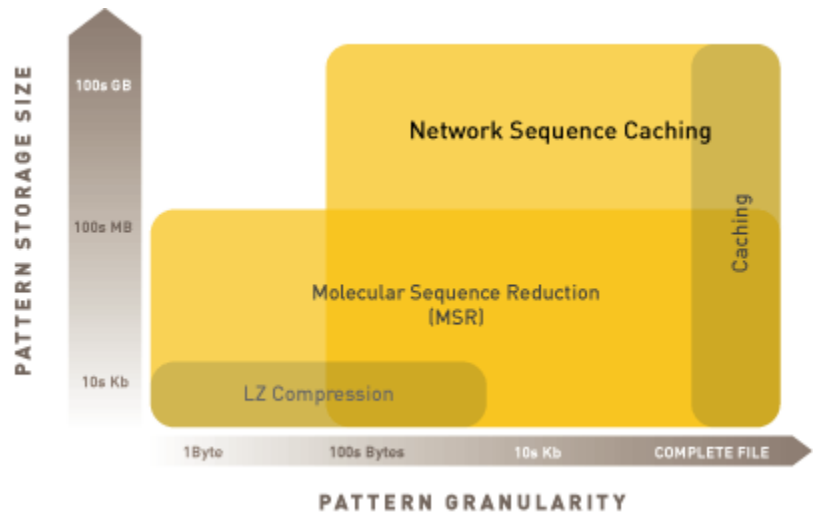
Both the Juniper WX and WXC application acceleration platform families support MSR compression.



The patent-pending Network Sequence Caching technology also looks for repeated data patterns and replaces them with a label to reduce WAN traffic. But unlike MSR compression, which operates in memory, the Sequence Caching technique uses hard drives to store data patterns. As a result, it can store much longer patterns and store them for a longer period of time, enabling detection and elimination of patterns seen days or weeks earlier.

The Sequence Caching feature is highly tuned to accelerate large file transfers, such as engineering drawings and specifications, backup traffic, data replications, virus update files, e-mail attachments, Microsoft Office files, and drawing programs such as CAD/CAM files.

Because businesses run a variety of application types, a combination of reduction technologies provides the greatest overall benefit to application performance and increased WAN capacity. The Juniper WXC application acceleration platforms automatically provide the advantages of both the Sequence Caching and MSR technologies.

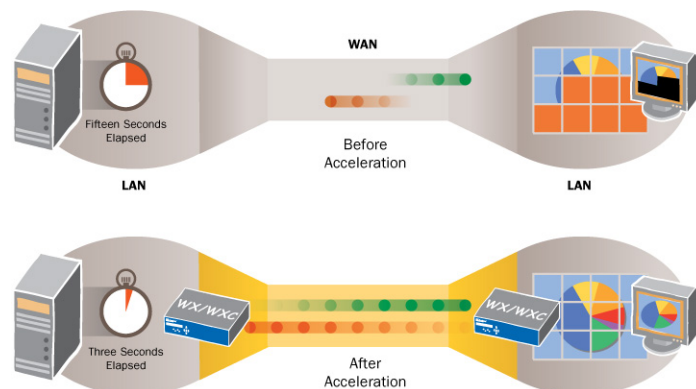


## How Juniper Reduces the Impact of Latency

The Juniper WX and WXC application acceleration platforms also include both TCP acceleration and application-specific acceleration.

The Packet Flow Acceleration™ (PFA™) technology provides general acceleration to any TCP-based application. It features several techniques, including:

- **Fast Connection Setup™**, which improves the performance of short-lived connections by eliminating one round-trip time (RTT) from the TCP connection setup, speeding up applications such as HTTP.
- **Active Flow Pipelining™** (AFP™), which extends the TCP performance improvements by terminating TCP connections locally and using a more efficient transport protocol between WX and/or WXC devices. This feature significantly benefits application performance on high-bandwidth or high-latency connections, and it avoids the inefficient slow-start mechanism within TCP.



- **Forward Error Correction**, which limits the need for retransmissions on lossy networks, such as international IP VPNs or satellite links. It makes use of recovery packets, sent alongside data packets, to allow for reconstruction of lost packets without retransmission delays.

To accelerate applications that are constrained by their own protocol behavior, the WX and WXC platforms include the Application Flow Acceleration™ (AppFlow™) technology. These acceleration techniques are completely transparent to the application and the existing network; they require no changes to clients, they don't interrupt protocol communications between clients and servers; and they fully support any WAN transport and network topology, including high-availability designs.

For Exchange, the AppFlow technique pipelines the hundreds or even thousands of data blocks required to complete a single transmission, sending as many of them in parallel as needed to fill the available WAN capacity and eliminating sequential RTTs. As a result, users receive their messages at LAN speeds, boosting their productivity and improving collaboration.

For Microsoft file services using CIFS, the AppFlow feature pipelines file read- and write-operations, sending them in parallel to fully utilize WAN capacity. By the time a client requests the file's data blocks, they already reside on the nearby WX or WXC platform, which forwards them at LAN speeds. Users are able to access centralized files much faster and collaborate more effectively.

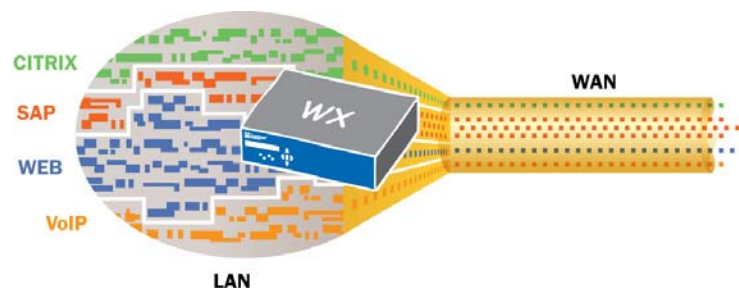
The AppFlow technique for web traffic enables WX and WXC devices to learn and cache objects associated with URLs. The WX and WXC platforms confirm the freshness of each object or pre-fetch them when new or updated versions are needed, in advance of the client's request. With the AppFlow technology, client browsers receive objects at LAN speeds and users' web pages display much faster.

## How Juniper Reduces Application Contention

Bandwidth Management includes both Quality of Service (QoS) capabilities and bandwidth allocation, allowing IT to prioritize business-critical and latency-sensitive applications. Juniper defies the cliché that effective QoS must be difficult to implement—its intuitive wizard/template-based approach enables IT to easily ensure that business policies are met through QoS techniques.

IT can assign priority status and bandwidth allocation metrics to applications. The WX and WXC platforms allow IT to classify traffic by looking not just at IP header or TOS/DiffServ information but also inside the data payload to act on Layer 7 application information. For example, at Layer 3, all Citrix applications look the same, so IT needs Layer 7 information to discern SAP traffic from a print job.

Further, Juniper recognizes that the WX and WXC platforms are not the only QoS-aware devices in the network. Therefore, Juniper's QoS techniques preserve and allow visibility into the QoS markings that other devices apply. For example, the WX and WXC platforms can



preserve an enterprise-based QoS marking, map that QoS policy to an MPLS-based service level for transmission through a service provider's network, and then restore the enterprise-based marking in the remote location.

Because the WX and WXC platforms know both the local WAN link speed as well as all the remote platform's link speeds, IT can perform resource allocation accurately, maximizing a link resource without overrunning either device.

## Customer Success Story—Hines & Associates

### Juniper Enables VoIP Deployment and Accelerates Citrix

Hines & Associates, a healthcare utilization review and case-management firm, relies on high-quality voice communications with its customers and between its employees to help large organizations save on healthcare costs. In an effort to improve its economics, the company deployed VoIP to its two call centers in Illinois and three branch offices in two other states. While the technology saved the company money on its telephony costs, voice quality routinely suffered when call volume increased, and employees experienced printing problems with a Citrix-based application.

The company considered upgrading the WAN links, but the annual cost of \$48,000 was too much to bear. To gain the necessary application acceleration, the company instead turned to Juniper Networks. With Juniper WX application acceleration platforms installed at these five locations, Hines & Associates immediately saw its voice calls improve to toll quality and its Citrix issues disappear.

"Juniper gave four-and-a-half times the capacity without a network upgrade, so it pays for itself in less than a year, and our existing network can now support twice the VoIP traffic with telco-grade quality," says Carl Valiulis, IT director at Hines & Associates.

In contrast to other options the company considered, the Juniper QoS mechanisms were easy to understand and configure. The company has also come to rely on Juniper's visibility tools to monitor application performance and WAN utilization.

#### Business Benefits

- 450% increase in existing WAN capacity
- Toll-quality VoIP over existing WAN
- Doubled the VoIP call volume on existing WAN
- Enabled incremental VoIP migration
- Improved remote printing
- Avoided \$48,000 per year in WAN upgrades
- ROI of less than 12 months

## Customer Success Story—IDEXX Laboratories

### Juniper Speeds Database Access and Large File Transmissions

IDEXX Laboratories provides technology-based products and services for animal health. Located at 30 sites around the world, the company depends on its WAN to enable access to a database located in IDEXX's headquarters in Maine and to transmit electrocardiograms between other sites. The company's 256 Kbps frame relay links were significantly congested, causing application delays. Accessing the database in Maine, for example, took users as long as 20 seconds per transaction.

Rather than increase its WAN links—at an annual cost of \$36,000—IDEXX opted to investigate the Juniper WX application acceleration platforms. Of concern to IDEXX was whether the Juniper devices would preserve data integrity and not add latency. Juniper not only answered those needs, but the WAN application acceleration platforms also increased the available bandwidth nearly three-fold and increased the performance of the company's VoIP deployment. More importantly, response times for the mission-critical database access and file transmissions decreased significantly. Database access, for example, fell from 20 seconds to just nine seconds.

"Prior to purchasing the Juniper products, our warehouse employees had to wait for a barcode reading application that was slow because of traffic-clogged links," says Rob Edwards, senior network engineer at IDEXX Laboratories. "Juniper made a network upgrade unnecessary, and the ROI was just a few months."

#### Business Benefits

- 250% increase in existing WAN capacity
- Decreased response time by more than 50% for mission-critical applications
- Improved VoIP quality
- Avoided \$36,000 per year in additional bandwidth costs
- Three-month ROI

## Customer Success Story—BOC Edwards

### Juniper Deployment Justifies Exchange Server Consolidation Effort

BOC Edwards, a leading supplier of high-capacity vacuum and pressure technology for microelectronics manufacturers, was phasing out end-of-life Exchange servers deployed in remote offices across the U.S.

Rather than replace the aging equipment, the company decided instead to invest in a consolidation effort and centralize all critical servers in their Wilmington, Mass., headquarters, where they have a storage area network (SAN),

#### Business Benefits

- Accelerated remote Exchange transfers by up to 75 percent
- Facilitated strategic consolidation of corporate Exchange servers
- Allowed overall reduction in number of Exchange servers
- Simplified storage, backup, and recovery
- Dramatically reduced maintenance and operational costs

backup and recovery capabilities, and where they could manage and maintain the servers locally.

An engineering facility in Tonawanda, N.Y. was the first to lose its local Exchange server. At first, users complained about poor performance over the WAN. But once the IT team installed Juniper Networks WX application acceleration platforms equipped with Application Flow Acceleration™ (AppFlow™) technology, complaints all but ceased.

The results, according to Martin Cox, Technical Services Manager, Planning and Development, were immediate and dramatic. A typical 1 MB file, which took more than a minute to transfer prior to the WX installation, now takes less than 15 seconds with the AppFlow technology. The initial deployment was such a success, Cox expects to roll out WX devices with AppFlow technology to four more sites in the U.S. and complete the corporate consolidation effort.

“It’s all about centralization today,” says Cox. “The Juniper solutions, which have already had a dramatic impact on e-mail performance over the WAN, will allow us to continue our consolidation effort and realize our corporate objective of centralizing services to reduce costs and simplify administration.”

---

Juniper Networks and the Juniper Networks logo are registered trademarks of Juniper Networks, Inc. in the United States and other countries. All other trademarks, service marks, registered trademarks, or registered service marks in this document are the property of Juniper Networks or their respective owners. All specifications are subject to change without notice. Juniper Networks assumes no responsibility for any inaccuracies in this document or for any obligation to update information in this document. Juniper Networks reserves the right to change, modify, transfer, or otherwise revise this publication without notice.