DELIVERING HPC APPLICATIONS WITH JUNIPER NETWORKS AND CHELSIO COMMUNICATIONS

Ultra Low Latency Data Center Switches and iWARP Network Interface Cards

Table of Contents

Table of Figures

List of Tables

Executive Summary

Ethernet provides a reliable and ubiquitous networking protocol for high-performance computing (HPC) environments, with the option to migrate from Gigabit Ethernet (GbE) through higher performance solutions such as 10, 40, and 100GbE. When used in conjunction with the Internet Wide Area RDMA Protocol (iWARP), Ethernet delivers an ultralow latency, high-performance, and highly scalable interconnect solution for application layer environments ranging from dozens to thousands of compute nodes.

Working together, Juniper Networks and Chelsio Communications offer high-performance, low latency products that deliver a complete end-to-end Ethernet solution for the most demanding HPC environments.

Introduction

For years, InfiniBand was the dominant interconnect technology for HPC applications leveraging Message Passing Interface (MPI) and remote direct memory access (RDMA). Today however, thanks to the rapid adoption of x86 servers in supercomputing and other high-performance parallel computing environments, Ethernet—ubiquitous and widely available—has supplanted InfiniBand as the preferred networking protocol in these environments. A full 48% of the top 500 supercomputers now use Ethernet as their standard networking technology (source: top500.org), while the high-performance, latency sensitive applications required for HPC, financial trading, and modeling environments leverage IP/Ethernet networks to run the same MPI/RDMA applications using iWARP.

Juniper Networks® QFX3500 Switch, when combined with Chelsio's industry-leading iWARP network interface card (NIC), delivers a complete end-to-end solution for the most demanding HPC environments. Juniper's high-performance, low latency 10GbE switch coupled with iWARP provides a scalable, end-to-end network solution that allows HPC clusters to grow from tens to hundreds to thousands of nodes, without being negatively impacted by reduced interconnect bandwidth or higher latency.

What Is iWARP?

iWARP, also called RDMA over Ethernet, is a low latency solution for supporting high-performance computing over TCP/IP. Developed by the Internet Engineering Task Force (IETF) and supported by the industry's leading 10GbE Ethernet adapters, iWARP works with existing Ethernet switches and routers to deliver low latency fabric technologies for high-performance data centers.

In addition to providing all of the total cost of ownership (TCO) benefits of Ethernet, iWARP delivers several distinct advantages for use with Ethernet in HPC environments:

- It is a multivendor solution that works with legacy switches.
- It is an established IETF standard.
- It is built on top of IP, making it routable and scalable from just a few to thousands of collocated or geographically dispersed endpoints.
- It is built on top of TCP, making it highly reliable.
- It allows RDMA and MPI applications to be ported from InfiniBand (IB) interconnect to IP/Ethernet interconnect in a seamless fashion.

Chelsio's iWARP and TCP Offload Engine Solutions

Chelsio's T420-LL-CR 10GbE iWARP adapters improve HPC application performance by leveraging an embedded TCP Offload Engine (TOE), a technology that offloads TCP/IP stack processing from the host to the NIC. A TOE frees up server memory, bandwidth, and valuable CPU cycles to improve the performance of applications that are sensitive to these parameters. When used with higher speed interfaces such as 10GbE, the performance improvement enabled by a TOE is even more dramatic, since 10GbE delivers data rates that are so high, host-based TCP instruction execution can quickly overwhelm even the fastest servers.

With the rapid adoption of 10GbE, and the resulting increase in data flow into and out of multi-core servers, TOEs have become a requirement to deliver the high throughput and low latency needed for HPC applications, while leveraging Ethernet's ubiquity, scalability, and cost-effectiveness.

For example, Chelsio's TOE provides an ideal way to enhance profitability in financial market data environments, where traffic essentially consists of high message rates and small message sizes, and where even the smallest delays can result in significant financial losses. Other HPC environments exhibit similar sensitivities to increased latency or delay.

Chelsio's TOE offers several key benefits:

- Increases Ethernet throughput and reduces latency and jitter, freeing up CPU cycles for application use
- Reduces operational costs by increasing I/O on a per-watt basis where, aggregating thousands of CPU connections, the energy costs can be overwhelming
- Minimizes bottlenecks in the memory subsystem by allowing direct memory access on both send and receive
- Enhances application-level capacity over non-offload adapters while using fewer, lower cost CPUs
- Offloads expensive byte touching operations and protocol functionality at higher layers by implementing a reliable transport layer in hardware, increasing the value of the technology

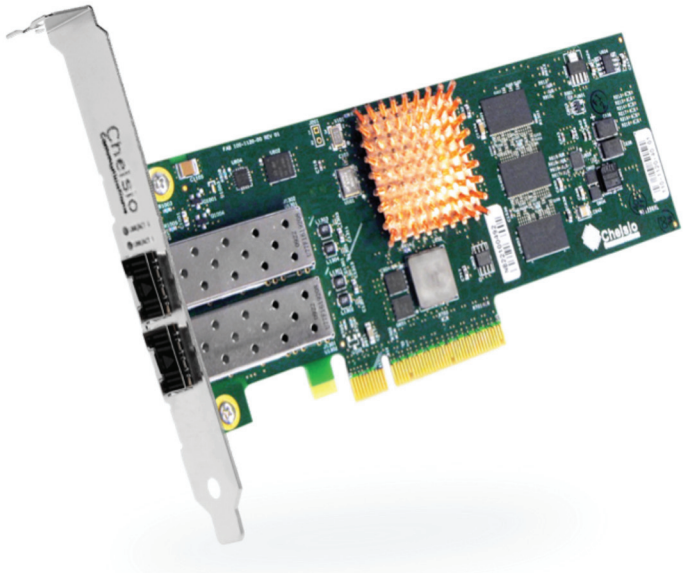

Figure 1. Chelsio's T420-LL-CR 10GbE iWARP network interface adapter

A Portfolio of Storage Offloads

Chelsio's second generation T420 iWARP design builds on the RDMA capabilities of the previous generation with continued MPI support on Linux with OpenFabrics Enterprise Distribution (OFED) and Windows HPC Server 2008. The Chelsio ASIC is already a field proven performer in Purdue University's 1,300-node cluster. The following benchmarks demonstrate the linear scalability of Chelsio's RDMA architecture to deliver comparable or lower latency than InfiniBand double data rate routing (DDR) or quad data rate routing (QDR), and to scale effortlessly in real-world applications as connections are added.

In addition, the Chelsio T420-LL-CR provides protocol acceleration for both file and block-level storage traffic. For file storage, it supports full TOE under Linux and TCP Chimney under Windows with added support for IPv6, which is increasingly prevalent and now a requirement for many government and wide area applications. For block storage, the T420 supports partial or full iSCSI offload of processor intensive tasks such as protocol data unit (PDU) recovery, header and data digest, cyclic redundancy checking (CRC), and direct data placement (DDP), supporting VMware ESX. To broaden Chelsio's already extensive support for block storage, the T420 adds partial and full Fibre Channel over Ethernet (FCoE) offload. With a host bus adaptor (HBA) driver, full offload provides maximum performance as well as compatibility with storage area network (SAN) management software. For software initiators, Chelsio supports the Open-FCoE stack and the T420 offloads for certain processing tasks, much as it does in iSCSI.

The following table and graph compare latency and CPU utilization of the QFX3500 Switch with and without TOE, showing a seven times reduction in CPU utilization and a 25% decrease in average latency when TOE is enabled.

Table 1: Performance of the QFX3500 Switch with Chelsio T420-LL-CR 10GbE iWARP Adapter (measured both with TOE disabled and TOE enabled)

PERFORMANCE MEASUREMENT	TOE DISABLED	TOE ENABLED
Latency (µs)	8.63 µs	6.48 µs
CPU utilization (%)	47%	8%
Bandwidth (Gbps)	9.49 Gbps	9.49 Gbps

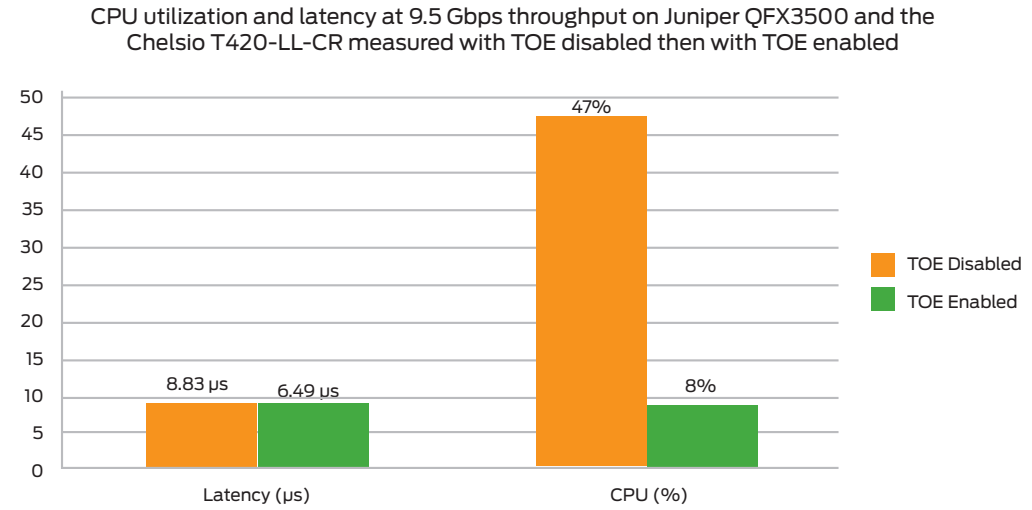

Figure 2. Latency and CPU utilization improvements when using the QFX3500 Switch and Chelsio T420 NIC passing 9.5 Gbps of data

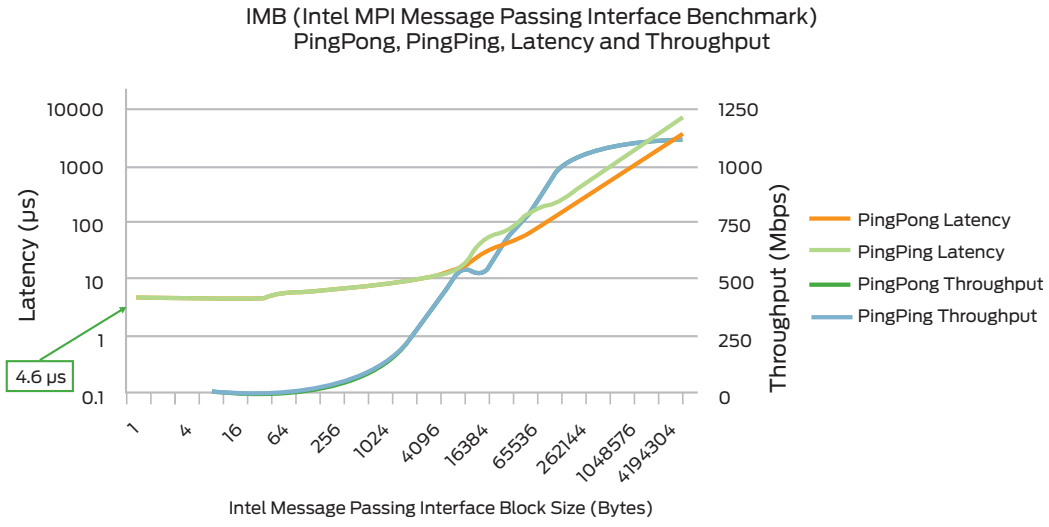

Figure 3. IMB showing latency and throughput for PingPong and PingPing

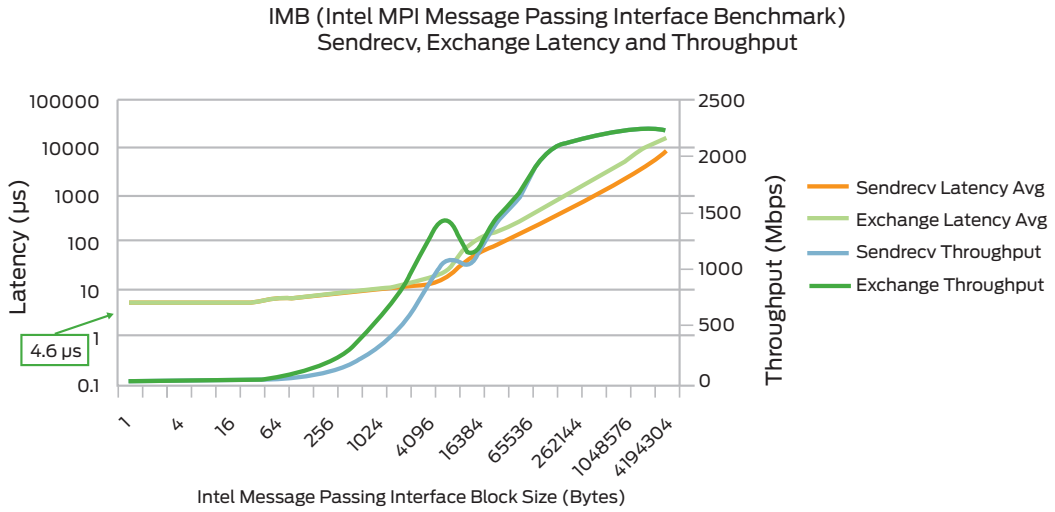

Figure 4. IMB showing latency and throughput for Sendrecv and Exchange

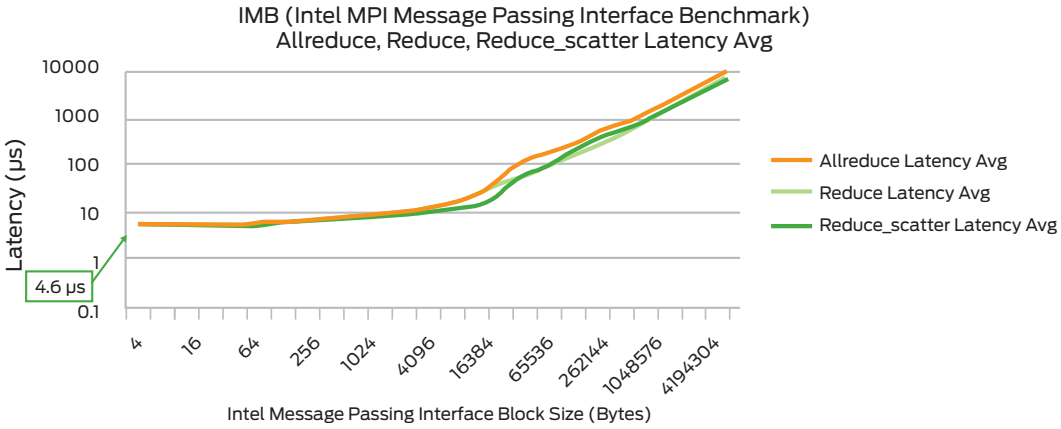

Figure 5. IMB showing allreduce latency average, reduce latency average, and reduce scatter latency average

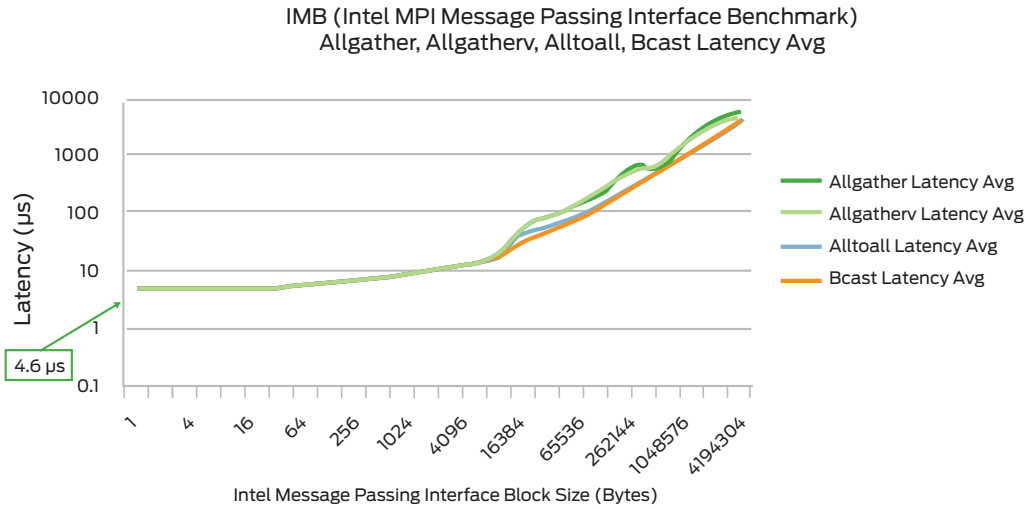

Figure 6. IMB showing allgather latency average, allgatherv latency average, all-to-all latency average, and broadcast latency average

Juniper Networks Data Center Strategy and 10GbE Switching Solutions

Juniper's strategy for the high-performance data center focuses on flattening the network to eliminate complexity and improve overall application performance. Called the "3-2-1" data center network architecture, the strategy asserts that today's data center is far too complex, requiring three layers of switches to provide the required port densities. As a result of this three-tiered "tree" structure, east-west network traffic between servers is first forced to travel north and south up and down the tree, adding latency and negatively impacting application performance.

The Juniper 3-2-1 architecture uses innovative fabric technology that dramatically simplifies and consolidates the network architecture, allowing it to move from its current three-tier design to two tiers and eventually to just one. This "flattening" of the network not only reduces the number of layers and weaves the remaining components into a common fabric that provides reliable, high capacity, any-to-any connectivity, it also enables multiple networking devices such as switches to operate and be managed as a single, logical device. By fundamentally reducing the number of networked devices to manage, fabric technologies dramatically reduce the cost and complexity associated with large data center networks while improving performance and efficiency.

The Juniper Networks QFX3500 Switch delivers a high-performance, ultralow latency, feature-rich Layer 2 and Layer 3 switching solution in a compact form factor designed for the most demanding data center environments. Featuring standards-based Fiber Channel I/O convergence capabilities, the QFX3500 is a versatile, high-density, 10GbE platform that delivers a highly efficient fabric-ready solution for implementing Juniper Networks QFabric™ technology (see Figure 7).

Figure 7. Juniper's ultralow latency 10GbE QFX3500 Switch

The high-performance QFX3500 platform is a perfect solution for a wide range of deployment scenarios. These include traditional data centers, virtualized data centers, high-performance computing, network-attached storage, converged server I/O, and cloud computing. Featuring 48 small form-factor pluggable transceiver (SFP+/SFP) and 4 QSFP+ ports in a 1 U form factor, the QFX3500 Switch delivers feature-rich L2 and L3 connectivity to networked devices such as rack servers, blade servers, storage systems, and other switches used in demanding, high-performance data center environments.

When deployed with other components of the Juniper Networks QFabric product family, which implements a flat, single-tier data center network, the QFX3500 delivers a fabric-ready solution that contributes to a high-performance, low-latency fabric architecture that unleashes the power of the exponential data center. The QFX3500 provides investment protection and architectural migration from the traditional multitier network to a QFabric solution (see Figure 8).

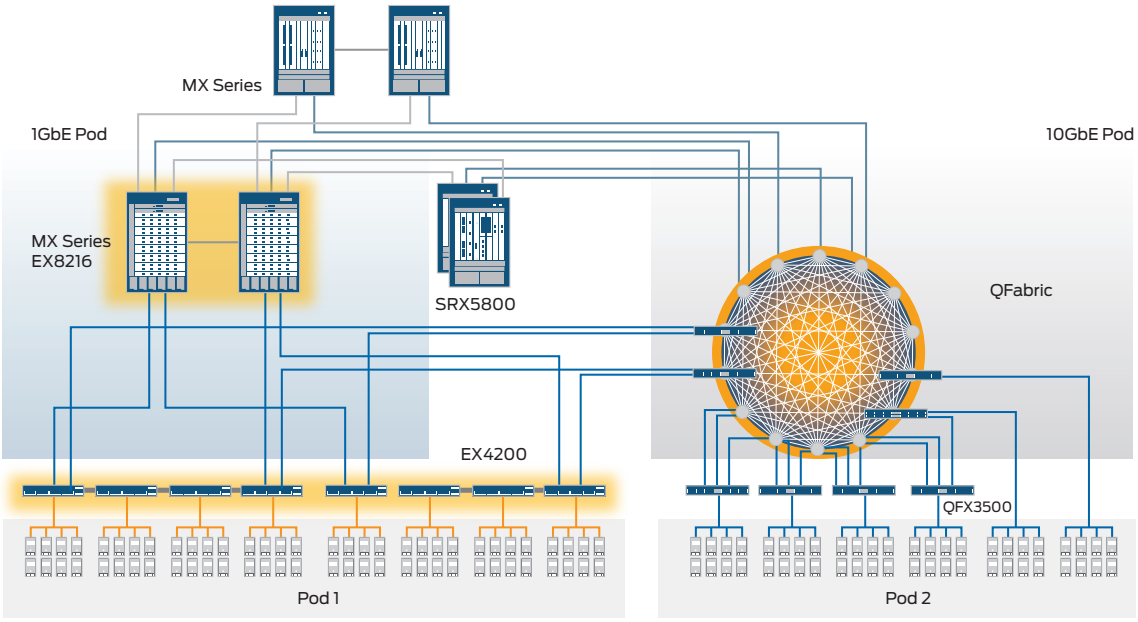

Figure 8. Data center with 1GbE using MX Series 3D Universal Edge Routers, EX4200 and EX8200 Ethernet switches, and 10GbE QFX3500 switches

For small IT data centers with a mixture of 10GbE and 1GbE servers, the QFX3500 can provide access for high-performance 10GbE servers as a two-tier data center architecture, while the Juniper Networks EX8200 line of Ethernet switches with Virtual Chassis technology or the Juniper Networks MX Series 3D Universal Edge Routers deliver a robust, resilient solution for the data center core that eliminates the need to run Spanning Tree Protocol (see Figure 9).

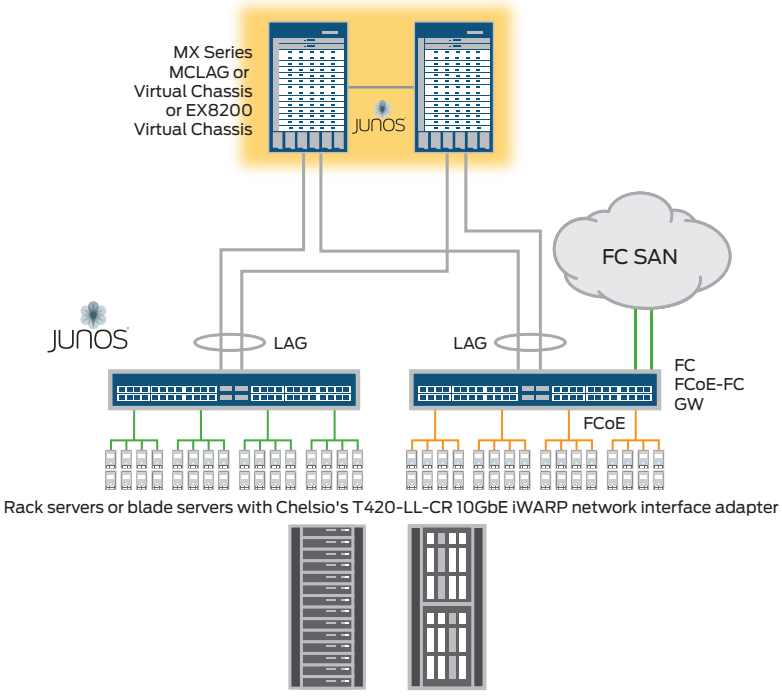

Figure 9. QFX3500 as a 10GbE top-of-rack deployment in the data center

Conclusion

For years, InfiniBand was the dominant interconnect technology for HPC applications, but it has now been eclipsed by Ethernet as the preferred networking protocol where scalability and ultralow latency are required. Juniper Networks QFX3500 Switch is a high-performance, ultralow latency, 10GbE switch specifically designed to address a wide range of demanding deployment scenarios such as traditional data centers, virtualized data centers, high-performance computing, network-attached storage, converged server I/O, and cloud computing.

Working in concert with Chelsio's industry-leading iWARP network interface card (NIC) with TOE technology, the QFX3500 switch and Juniper Networks QFabric technology deliver a complete, scalable, end-to-end solution for today's most demanding environments. When deployed with other components of the QFabric family of products, QFX3500 delivers an industry-leading and cost-effective solution that will unleash the power of the exponential data center.

About Juniper Networks

Juniper Networks is in the business of network innovation. From devices to data centers, from consumers to cloud providers, Juniper Networks delivers the software, silicon and systems that transform the experience and economics of networking. The company serves customers and partners worldwide. Additional information can be found at **www.juniper.net**.

Corporate and Sales Headquarters
Juniper Networks, Inc.
1194 North Mathilda Avenue
Sunnyvale, CA 94089 USA
Phone: 888.JUNIPER (888.586.4737)
or 408.745.2000
Fax: 408.745.2100
www.juniper.net

APAC Headquarters
Juniper Networks (Hong Kong)
26/F, Cityplaza One
1111 King's Road
Taikoo Shing, Hong Kong
Phone: 852.2332.3636
Fax: 852.2574.7803

EMEA Headquarters
Juniper Networks Ireland
Airside Business Park
Swords, County Dublin, Ireland
Phone: 35.31.8903.600
EMEA Sales: 00800.4586.4737
Fax: 35.31.8903.601

To purchase Juniper Networks solutions, please contact your Juniper Networks representative at 1-866-298-6428 or authorized reseller.

2000429-001-EN Aug 2011 Printed on recycled paper