APPLICATION NOTE

SRX SERIES SERVICES GATEWAYS CLUSTER DEPLOYMENT ACROSS LAYER 2 NETWORKS

Deployment Requirements for High-End SRX Series Layer 2 Cluster Connectivity

Table of Contents

Table of Figures

Introduction

Stateful firewall clustering has traditionally been deployed in a single location, and often firewalls are in the same rack. This means that distance between devices can be within a few meters, and this short distance allows for a simple and reliable connection between devices for cluster synchronization. Today as networks grow beyond a single location, the requirements of a stateful firewall are changing. It is common that one logical network can be split and placed into multiple locations. In these instances, clustered firewalls need to be able to communicate with each other over greater distances. Juniper Networks® has provided this document to walk the reader through the requirements for deploying firewall clusters that span different locations.

Scope

This document explains how to connect two Juniper Networks SRX Series Services Gateways over a Layer 2 network. The goal is to allow customers to deploy an SRX Series cluster over distances that are longer then three meters. Requirements and deployment scenarios are provided to ensure a successful rollout of the SRX Series across an L2 network.

Design Considerations

Cluster connectivity for high-end SRX Series firewalls has been designed for directly connected links. This was done to reduce possible fault domains for clustering connectivity. However, since directly connected links do not meet the needs of all customers, Juniper Networks has expanded support to include connectivity across L2 domains. Because the connectivity between the two SRX Series devices is critical, specific guidelines for deployment have been created.

In this section, the important design considerations are specified. Supportability for remote cluster connectivity can only be achieved if these guidelines are met. If these guidelines are not specifically followed, anomalous behavior can occur and service through the firewall can be disrupted.

Layer 2 Specifications

In a cluster deployment, communication between the two SRX Series Services Gateways is critical to the cluster's operation, and there are some specific requirements that need to be followed for timely delivery of the data. The foremost consideration is latency. Latency should not exceed more then 100 ms between the two devices. Exceeding this latency can cause the cluster to go into an unstable state. The unstable state can vary between dual mastership or the inability to pass traffic. Most transcontinental Ethernet links should be able to meet this requirement. Dual mastership is when both devices assume that they are the master of the cluster, another name for this is called split brain.

The second consideration is the amount of bandwidth that it takes to communicate between the two devices. Each connection is assumed to be a one gigabit, full duplex connection at a minimum. Each of the cluster members will be operating under the assumption that this bandwidth will be available. Specific messages and communication types for the two link types are detailed in the appropriate sections below.

The control and fabric networks should be free of any traffic except from the two SRX Series Services Gateways. Also the network should be free of any additional hosts, as this can cause instability in the networks. The communications between the two devices use private media access control (MAC) addresses and private IP addresses. These may conflict with other vendors' equipment and/or affect other hosts in an unknown manner. Any foreign broadcasts, packets, or MAC addresses on the control and fabric networks can cause instability to the cluster.

Traffic on the high-end platforms is not tagged with a VLAN. This allows the underlying switching infrastructure to provide the traffic with its own unique broadcast domain. As of Juniper Networks JUNOS® Software 9.5 release, control and fabric communications are not tagged. However, it is possible that this could change in the future, so it is best to be prepared to preserve the VLAN tags for a future release of JUNOS. It is possible to tag the cluster control and fabric traffic with VLANs. This will not have any negative effect on the traffic and will allow the cluster connectivity to pass on a shared switching infrastructure.

While it is possible to use VLAN tagging and have the control and data traffic share the same physical switching infrastructure, it is not recommended. The split-brain avoidance logic assumes that there are two physically separate networks available. If both control and data communication paths are broken simultaneously, the cluster will go into a split-brain condition. This is further detailed in the control and data link sections below.

To support a Layer 2 high availability (HA) environment, no additional configuration is needed on the SRX Series Services Gateways. The same configuration is used on the SRX Series even when connecting to switches. This simplifies the deployment by not requiring specific changes on the SRX Series devices.

High availability L2 requirements are summarized in the table below.

Table 1: Summary of L2 HA Requirements

REQUIREMENT	DETAILS
Latency	Needs to be less than 100 ms between devices.
Bandwidth	Requires a minimum of 1 Gbps per link.
Isolated networks	Each HA network must be isolated from any other hosts.
VLAN preservation	VLAN tags from HA traffic should be preserved.
Redundant networks	Each HA network should be on a physically separate infrastructure.

Control Link Traffic Details

The control link HA connection is the most critical of the two. The control link, as its name implies, is required to control and communicate with all of the components on both of the chassis. A disruption on the control link could cause one or more of the chassis components to exhibit unexpected behavior. Each Services Processing Unit (SPU), Packet Forwarding Engine (PFE) Mezzanine Board, and Network Processing Card (NPC) needs to maintain communication with the primary Routing Engine (RE). The primary RE uses the control link in various ways: kernel state synchronization, configuration synchronization, forwarding command-line interface (CLI) commands, and exchanging JUNOS redundancy protocol daemon (JSRPD) heartbeats. Each communication message is extremely critical and cannot be lost, interpreted, or reordered.

On each Routing Engine, JUNOS redundancy protocol daemons (JSRPDs) communicate with each other. Over the control link, each RE sends a heartbeat packet once every second. The JSRPD service expects also to receive a hello once every second. This ensures that communication is flowing. Currently, the heartbeat cannot be set to lower then once per second. It is possible to miss up to two sequential heartbeat packets without causing a failover.

If communication is missed for both the control and data link simultaneously, a split-brain condition will occur as each device assumes that the other is down. If a control link fails, the secondary Routing Engine will assume mastership of the cluster. Once the secondary RE assumes mastership, it will contact the FPCs in both chassis and establish control over them. The FPCs will now use the new primary RE for configuration updates and management.

In a typical deployment, the two SRX Series Services Gateways will be directly connected. If the control link were to fail, the secondary node begins the process of going into a disabled state. The only way to recover from going into a disabled state is by rebooting the node. If the nodes end up going into dual mastership, one of the nodes will need to be rebooted manually to rejoin the cluster. Alternatively, SRX Series devices can be configured to automatically reboot once they begin to communicate. If the two nodes need to merge together as masters, this will cause all of the line cards to be briefly reset and will cause an outage for the data services. To prevent this outage, reestablish connectivity and then reboot one of the nodes.

If the control path is disrupted, the nodes will still attempt to communicate over the data path. This will prevent the two nodes from going into dual mastership. Another scenario that can occur is that the control link is up but the heartbeat messages are not received over one of the two types of connections. If this occurs, the secondary node will also go into a disabled state. After 180 seconds, the node will first become ineligible, and then, after being ineligible for 180 seconds, the node will become disabled. Even when the two devices are placed in physically separate locations, it is still important to note this behavior. If the secondary node's connection to its local switch fails, it will still go into a disabled state. Once the nodes are connected to the L2 infrastructure, it is important not to disrupt the physical control link.

The control link between the two SRX Series devices is always over a one-gigabit connection. On the Juniper Networks SRX3000 line, the connection can either be copper or fiber. For the Juniper Networks SRX5000 line, the control connection must be a fiber link. This limitation stems from the limited space in the small form-factor pluggable transceiver (SFP) controller on the SRX5000 line's Services Processing Card (SPC). The end-to-end control link path should allow for maximum bandwidth utilization. In practice, the control link will have less bandwidth needs compared to the data plane.

Data Link Traffic Details

The data or fabric link serves two purposes for the SRX Series. Its primary purpose is to synchronize real-time objects (RTOs). RTOs consist of messages used to synchronize information between the two chassis. This type of information includes but is not limited to session state, firewall authentication, application layer gateway state, and IPsec state. There are a few dozen message types that are shared between the two devices. The most common messages will be the *session create* and *session close* RTOs. The second purpose for the data link is to perform what is known as Z-mode forwarding. Z-mode traffic is defined as traffic that enters one node and exits a second chassis. Both applications can be very bandwidth intensive.

RTOs are synchronized directly between the *flowd* daemon running on the SPUs on each node. If an SPU is located in FPC number 0, each SPU will synchronize the RTO to the other node in the same FPC and PIC location. On the SRX3000 line, there is only one SPU per FPC so that places the SPU in PIC 0. In the case of the SRX5000 line, each SPC contains two SPUs and these are numbered zero for the SPU toward the top of the chassis and one for the SPU on the bottom of the chassis. Please see Figure 1 below for a graphical representation of this.

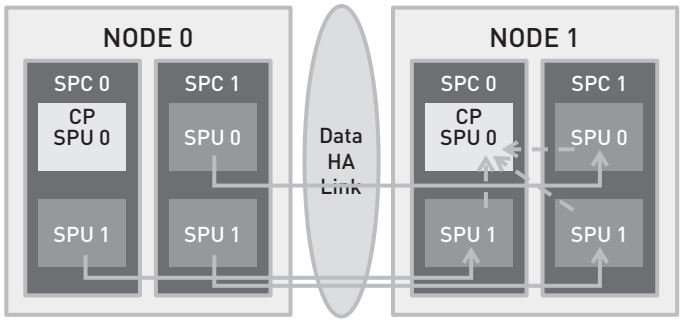

Figure 1: SPU synchronization

RTOs are sent out as needed over the fabric link. The average RTO packet size seen was 320 bytes. The packet consists of 14 bytes for the L2 header, 20 bytes for the IP header, and the remainder of the packet is for the RTO message itself. A *session create* message is a total of 286 bytes. It is possible to have multiple messages in a single packet and message batching will only occur during the cold sync process. Conceptually, this is like a train. Periodically the train will show up to pick up passengers. The train of RTO packets will always pick up a minimum of one message or passenger. If no other passengers are available, then the train will leave. However, if there are additional passengers the train will take them as well. When the remote SPU receives an RTO, it does not acknowledge it, reducing the overhead needed for synchronizing messages by removing the response packet.

It is important to understand the size of the RTO messages to better understand the bandwidth needs of the data link. It can be assumed that the full link bandwidth is always available, but knowing what the link is being used for can help clarify your understanding of link use. Based upon the stated packet sizes, it is possible to saturate a gigabit link[1] with 350,000 new connections per second. Assume that a minimum of one gigabit of bandwidth is available for the data link.

The fabric link is assumed to support jumbo frames. The maximum size of a packet will be no larger then 8800 bytes, but it is best to ensure that the path can support a maximum transmission unit (MTU) of at least 9014. If a lower MTU is set along the path, this could cause packet fragmentation. RTOs must not be fragmented, as doing so would invalidate the packet on the remote end. Because the RTO message cannot be reassembled from fragments, it is best to ensure that the MTU is correctly set end to end between nodes. No changes need to be made to the SRX Series to support the increased MTU for the fabric link.

In the event that traffic needs to be forwarded between the two cluster members, this too is done across the fabric link. This is called Z-mode forwarding. With Z-mode forwarding, a packet is received on a node and the node then forwards this packet to an SPU for processing. The SPU recognizes that it is the backup for the session and forwards the packet over to the second chassis' matching active SPU for that session. The active SPU accepts the packet and processes it before forwarding it out of the correct interface on the node. As this packet is forwarded across the data link, it will consume the necessary bandwidth to forward the packet. There isn't a bandwidth restriction on forwarding traffic between chassis.

[1]This assumes the following formula ((RTO size + inter packet gap)*bytes to bits)*RTO rate) or ((320 + 20) * 8) * 350,000) = 952,000,000/1,000,000 = 952Mbps

To prevent this from occurring, it is best to use the alternate design called line mode. The line mode design ensures that traffic enters and exits the same chassis, and ensures that there are no active egress paths on the secondary node. This design is discussed in the Description and Deployment Scenarios section below. Packets that are forwarded are completely preserved and encapsulated within an HA packet. Because of this, the SRX Series has no control over the packet size. The same MTU restrictions also apply to the data traffic. It is assumed that jumbo frames can be forwarded over the fabric link.

To validate that the fabric link is operating correctly, a fabric probe message is periodically sent. This message originates from the JSRPD daemon and is sent to its peer daemon. The message is routed over the following path: JSRPD -> local CP -> fabric link -> remote CP -> remote JSRPD. Sending the message over this path ensures that each side's data plane is operational. Each node sends a fabric probe once every second. The fabric path is more forgiving for missing messages. If the fabric link physically goes to a down state, the device will recognize this and fail over all of the data plane traffic. It will take up to 66 seconds to detect if the fabric path is disrupted since there was no link down event. The node needs to detect that it has not received a single message over the course of 66 seconds. This is based on the following formula: (60+2)*1(interval)*3(threshold). It is possible to receive only one packet per 66 seconds and for the fabric link to be considered healthy. Both the interval and threshold are configurable.

If the timer has been exceeded, then the secondary node will move into a disabled state. This prevents split brain from occurring, as the secondary node automatically disables itself. Once connectivity between the two nodes is verified, the now disabled node needs to be rebooted. Assuming that all communication is successful, it is now possible for the node to rejoin the cluster upon reboot. Data link connections can be either 1 or 10 GB media types. The same media type is required on both sides of the link, so that nodes do not attempt to send data at a higher rate then what is expected.

Description and Deployment Scenarios

In this section, the deployment scenarios for cluster connectivity across Layer 2 networks are detailed. The initial topic of discussion is high availability deployment modes. These deployment modes move away from the traditional terminology of active/passive and active/active, terms which are focused more around failover behavior than detailing how the devices pass traffic. The two new terms are *line mode* and *Z-mode*. Line mode represents symmetrical traffic passing through one or two nodes, where traffic enters and exits the same chassis. In a Z-mode deployment, traffic enters one chassis and then exits the second chassis. The considerations for both deployment scenarios are discussed in this section.

The second component to this section discusses the physical infrastructure used to connect two SRX Series Services Gateways together. In the beginning of this document, the requirements for L2 infrastructure were shown. As long as those requirements are met, Juniper is not concerned with what is used between each cluster member. However, Juniper Networks has two specific deployment scenarios that are recommended. The first deployment is a dual switch design. The attributes for this design require two physically separate Layer 2 networks connecting the SRX3000 line and SRX5000 line systems. The second design is popular in more advanced customer scenarios. It utilizes the virtual private VLAN service or VPLS. VPLS allows the extension of a VLAN across a WAN and is an acceptable connection type as long as it meets the requirements described in this document.

Line mode deployment

The design goal of a line mode deployment is to ensure that traffic enters and exits the same physical chassis. Doing so reduces the amount of traffic that is sent across the data link. This is desirable in the case where the data link is across a wide area network or WAN. It is possible to use line mode in both an active/passive or active/active deployment. An active/passive deployment only allows for one chassis to actively pass traffic at a time, so by definition this is a line mode deployment. Figure 2 shows an example of an active/passive line deployment.

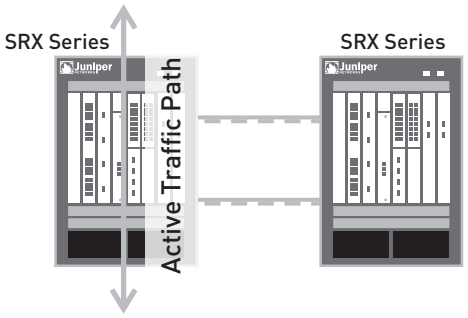

Figure 2: Active/passive line deployment

Active/active HA deployments on the SRX Series are extremely flexible and allow traffic to enter and exit to/from any direction on the cluster nodes. While this is very powerful in a local deployment, it can use excessive critical bandwidth in a WAN deployment. It is possible in an active/active deployment to bind traffic to a specific node. This can be accomplished in one of two ways. The first way is to use local interfaces on each node, which are then bound to a virtual router routing instance. The same is then done on the second chassis. This forces any traffic that enters a chassis to be processed by that chassis. Because the routing instance does not contain interfaces that will egress on the second node, traffic always stays local to a specific node. Figure 3 shows an example of this type of deployment.

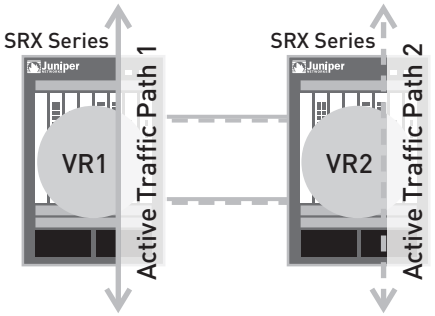

Figure 3: Local interface node binding

The alternate design to using local interfaces is to use redundancy groups. A redundancy group is a logical collection of objects, in this case interfaces, that can fail over between two nodes in a cluster. Redundancy groups can only contain a special interface type called a redundant Ethernet interface. There also is a special redundancy group called redundancy group 0 or RG0. Redundancy group 0 is a special redundancy group that contains the control plane or Routing Engine from each node.

It is possible to create two redundancy groups and then bind each to a specific node. The redundancy group would contain of one or more redundant Ethernet interfaces. Each redundant Ethernet interface would then be bound to its own routing instance. As seen in the local interface design, traffic would always be bound to a specific node. In the event of a node failure, the redundancy group would then fail over to the remaining chassis. This would allow all of the traffic that is being received on the failed node to be processed on the remaining device. The biggest difference between this and the local interface design is that the redundant Ethernet interfaces can fail over between the two nodes. If using local interfaces, another mechanism such as dynamic routing needs to be used to force traffic to the remaining cluster node. An example of this deployment can be found below in Figure 4.

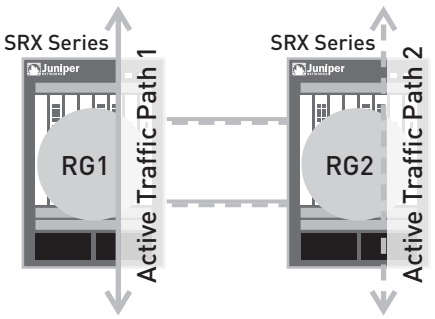

Figure 4: Redundancy group line mode design

Z-Mode Deployment

This alternate design includes what is known as a Z-mode deployment. This type of deployment is typically characterized as an active/active deployment. Specifically, Z-mode is when traffic enters one node and then exits a second node when the best egress path for the traffic is out of the second node. This is an excellent design to use for the SRX Series when extra throughput or redundancy is needed. However, the design loses its appeal in some cases of a multi-site WAN cluster. The most prominent concern is that traffic forwarded between the two nodes in the cluster will be forwarded over the WAN data link connection, requiring additional bandwidth on the data link connection. As long as the WAN can support the additional requirements, then the Z-mode design is an acceptable solution. The Z-mode design is depicted below in Figure 5.

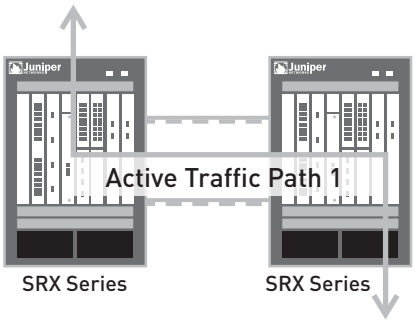

Figure 5: Z-mode chassis cluster deployment

Dual Switch Deployments

When connecting two SRX Series Services Gateways over a Layer 2 network, it is best to keep high availability in mind. The number one item to consider is reliability of the underlying network. Juniper suggests running two separate physical networks for connectivity between the nodes. Doing so separates the control and data link to avoid the failure of a single switch between the two nodes. An example of this deployment can be found in Figure 6 below.

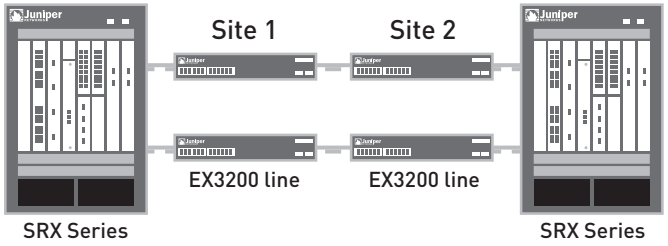

Figure 6: Dual WAN link cluster

There may be times when redundant WAN connections are not available. In these cases, it is acceptable to maintain physical separation between the control and data links, except over the WAN connection. In the event the WAN connection goes down, cluster members will lose connectivity and this will force the cluster to go into a *split-brain* condition.

Figure 7 shows a single WAN connected cluster. Each location has its own local firewall that is still active. This is most likely a desirable behavior as there will be connectivity through the firewall for the local hosts. The risk is then merging the cluster back together, as merging two active units together causes a brief outage. When the master Routing Engine is elected, it must reset communication between itself and the remote FPCs and this will cause traffic to stop for several seconds. This condition can be avoided by rebooting one of the firewalls and then reconnecting the control and data links so that when the firewall comes back up, it can naturally reestablish a cluster with the existing node.

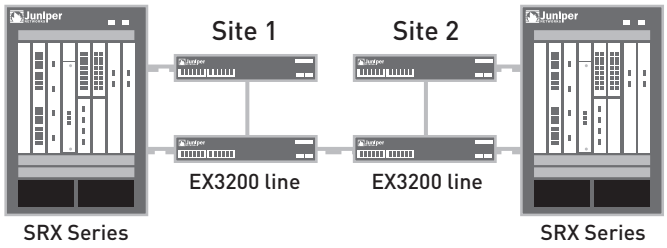

Figure 7: Single WAN link cluster

Non-Ethernet WAN Cluster Deployments

There may be times when an Ethernet connection is not available between two locations. In this case, alternate designs are required to connect the two nodes together. A perfect technology for this deployment is VPLS, as VPLS allows for L2 connectivity across a WAN. In a VPLS deployment, Layer 2 packets can be forwarded across the WAN, and the connectivity looks as if the chassis cluster members are connected directly over a switch. Besides VPLS, there are several other L2 technologies that can be used to accomplish the goal of connectivity between the cluster members. This includes L2VPNs and Circuit Cross-connect (CCC). In the future, Juniper Networks will look to expanding its support of Layer 2 environments to include technologies such as VPLS.

Low Bandwidth Deployments

There may be times when the bandwidth available between sites does not meet the needs specified in this document. In this section, the best practices for overcoming this constraint are outlined. The most critical component is to ensure that control messages are sent and received between the two nodes. Three seconds of missed communication will cause the nodes to break the cluster. It is critical that bandwidth is reserved to ensure that the control messages are able to be sent and received between the two devices. The actual throughput for the control link is less then 10 Mbps at any specific point in time, so bandwidth is never a critical issue for the control link.

Using a low bandwidth deployment is the most detrimental to the fabric link. The fabric link is primarily used to synchronize session data between the two devices. Without synchronization of traffic going across the link, only heartbeat messages are sent. The heartbeat packets are quite small and have low bandwidth requirements. Because of this, it is possible to deploy the SRX Series cluster with less then one gigabit of bandwidth between the two members. This can be done by reducing the number of connections per second (CPS) going through either node. At a lower CPS rate, less bandwidth is actually required for the link. To support 30,000 connections per second, a bandwidth of about 80 Mbps is required.[2] This is a large variance from the 1 Gbps suggested above, and allows less bandwidth to be provisioned between the two nodes for state synchronization.

Because preserving bandwidth is critical, Juniper suggests deploying the firewalls in line mode. As discussed above, a line mode deployment will not forward traffic between the two nodes. Instead, forwarded traffic is pushed over the fabric link. By avoiding the need for pushing the traffic across the fabric link, the requirements for bandwidth to support this are removed. This again reduces the need for dedicated bandwidth on that network.

As seen in this section, it is possible to deploy an SRX Series cluster using a reduced amount of bandwidth. There are some important notes to specify about this deployment. If the amount of RTO traffic is kept to a minimum and no traffic is forwarded between the two nodes, there will not be a need for additional bandwidth. That said, the SRX Series will still be operating under the assumption that the bandwidth is available. Because of this, there aren't any restrictions preventing the SRX Series from attempting to use the additional bandwidth. The only mechanisms available to reduce bandwidth requirements are ensuring that the configuration does not forward traffic between the nodes, and reducing RTOs to a minimum. While it is not possible to ensure that the amount of CPS will operate at a certain level, one can track the traffic rate. If it is tracked and planned correctly, it is possible to support SRX Series clustering over a low bandwidth deployment.

[2]Assuming the following formula ((RTO size + inter packet gap)*bytes to bits)*RTO rate) or ((320 + 20) * 8) * 30,000) = 81,600,000/1,000,000 = 81.5Mbps

Summary

Expanding the physical deployment across multiple locations can be quite simple. The configuration on the SRX Series Services Gateways is the same whether or not they are connected directly or indirectly. This simplifies the deployment strategy for implementing SRX Series Layer 2 connectivity. The most important factor in the deployment of the SRX Series over an L2 network is to understand the requirements for the deployment. This application note lays out clear guidelines specifically designed to ensure a successful deployment in the correct environment.

About Juniper Networks

Juniper Networks, Inc. is the leader in high-performance networking. Juniper offers a high-performance network infrastructure that creates a responsive and trusted environment for accelerating the deployment of services and applications over a single network. This fuels high-performance businesses. 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

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
Fax: 35.31.8903.601

3500165-001-EN July 2009