

SRX Services Gateway Cluster Deployments Across Layer Two Networks

Deployment requirements for SRX cluster connectivity across layer two networks

Introduction

Stateful firewall clustering has traditionally been deployed in a single location, often with the firewalls located in the same rack a few meters apart. This short distance allows for a simple and reliable connection between devices used for cluster synchronization. Today, as networks grow beyond a single location, the requirements of a stateful firewall have changed. It is common for one logical network to be split and placed into geographically dispersed sites. To accommodate this need, clustered firewalls are required to communicate to each other over greater distances. The aim of this document is to walk the reader through the requirements for such deployments

Scope

This document walks through the different deployment scenarios of clustered SRX service gateways over a layer two network. The goal is to allow users to deploy an SRX cluster over a layer-2 transport network. This application note discusses the requirements and common deployment scenarios needed to ensure a successful roll out of the SRX.

Design Considerations

In order to reduce the number of link errors, the cluster connectivity feature for SRX service gateways was designed to be used with directly connected links. However, since this implementation does not meet all customer needs, Juniper Networks expanded the support to include connectivity across layer two domains creating specific guidelines for such deployments.

In this section, the design considerations are specified. Supportability for remote cluster connectivity can only be achieved if these guidelines are met and, if not followed, anomalous service-impacting behavior can result.

Layer Two Specifications

The communication between the two SRX services gateway is critical to the cluster operation. To ensure timely delivery, there are some specific requirements that need to be followed. The foremost consideration is latency: latency shall not exceed more than one hundred milliseconds between the two devices. Exceeding this latency can cause the cluster to go into an unstable state, with effects varying from dual mastership and or the inability pass traffic. Most transcontinental Ethernet links should be able to meet this requirement.

The second consideration is the amount of bandwidth that it takes to communicate between the two devices. For high-end SRX service gateways, each connection is assumed to be a one gigabit full-duplex connection at a minimum. Each of the cluster members will be assuming that this bandwidth will be available. The specific messages and communication types for the two link types are detailed in the appropriate sections below.

Branch SRX service gateways have bandwidth requirements that vary based on the environment. For the fabric connection, the two most important factors are the maximum sustained number of sessions/s and the amount of data traffic across the link.

The network connecting both nodes in a cluster should be free of any traffic, except for from the traffic generated by the SRX service gateways. Also, the network should be free of any additional devices as a single host could cause

instabilities in the network by flooding traffic. Additionally, the communications between the two devices utilizes private MAC and private IP addresses, which may conflict with other vendor's equipment and/or affect other hosts in an unknown manner. Any foreign packets, or MAC address in the HA networks could potentially cause instabilities to the cluster.

Traffic on the high-end platforms is not tagged with a VLAN. While as of the JUNOS Software 9.5 release the HA communications are not tagged, future releases might change this. Whenever possible, it is recommended to preserve VLAN tags both across control and fabric links. It is possible to mark the cluster's HA traffic with VLAN tags, as long as they are removed before traffic is passed to the nodes. This will not affect the traffic in a negative manor and would allow the inter-node traffic connectivity to be transported using a shared switching infrastructure.

Branch SRX platforms tag control-plane traffic with VLAN ID 4094, this tag should not be modified or removed by the transport network. Options to transport this traffic are discussed in the deployment scenarios section.

While it is possible to use VLAN tagging and have both control and data traffic share the same switching infrastructure, it is not recommended to do so. The split-brain avoidance logic assumes that whenever both control and data communication paths are broken simultaneously the failure is due to some problem in the remote node and not in the transport network; in such cases, cluster instability can occur, as both nodes will become primary.

To support a layer two HA environment no additional configuration needs to be done on the SRX services gateways. The same configuration is used on the SRX services gateways even if they are connecting to switches, therefore simplifying such deployments.

HA summarization of the layer two requirements can be found in table one below.

Table 1. Summary of layer two requirements

Requirement	High-end SRX	Branch SRX
Latency	Less than 100ms	Less than 100ms
Bandwidth	Control Link	???
	Fabric Link	2.8 Mbps per 1000 sessions/s plus any asymmetric traffic resulting from Z-mode deployments
Isolated Networks	Each HA network must be isolated from any other hosts	Each HA network must be isolated from any other hosts
VLAN Preservation	VLAN tags from HA traffic should be preserved	VLAN tags from HA traffic MUST be preserved. Control link traffic is tagged with the VLAN-ID 4094 and it must not be modified by the transport network
Redundant Networks	Each HA network should be on a physically separate infrastructure	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 states, is required to control and communicate the different components on both of chassis. In high-end SRX platforms, each services processing unit (SPU), pic management board (PMB), and network processing unit (NPU) need to maintain communications with the master routing engine. Branch SRX platforms use the control link to interconnect the daemons running on the control plane hardware thread. These communication messages are extremely critical and cannot be lost, altered, or reordered.

Each routing engine (or control plane core) runs an instance of the JUNOS Software redundancy protocol daemon (JSRPD). These daemons use the control link to communicate with each other sending, amongst other things, a heartbeat packet once every second. Link failures over the control link are assumed when three or more consecutive heartbeat packets are lost.

If both data and control links are down or loose consecutive heartbeats a failover will be triggered, causing the secondary routing engine (or control-plane hardware thread in the case of Branch SRX devices) to take mastership of the cluster. Once mastership is assumed, the new primary routing engine will take control of all the components in both chassis.

In a typical deployment the two SRX services gateways will be directly connected. If only one of the links were to fail or lose more keepalives than what are allowed, the secondary node would begin the process of going into disabled state, preventing both nodes from becoming master simultaneously. Initially, the secondary node transitions into ineligible state and relinquish mastership of any redundancy group in the cluster. If after 180 seconds the heartbeat messages are not resumed, the secondary node will transition into disabled state. While the fabric link is allowed to miss all but one heartbeat message every 66 seconds, the control plane is not that forgiving, triggering a failover when more than 3 consecutive probes are lost. The only way to recover a node from disabled state is by rebooting it.

When the two devices are placed in physically separate locations, it is important to note this behavior since a failure in the connection to one of the local switches would result in the transition of the secondary node to disabled state. Therefore, once the nodes are connected through a layer two infrastructure, it is important to not disrupt neither control nor fabric links.

The control link between the two high-end SRX services gateways is always transported over a one-gigabit connection. While in the SRX 3000 series the connection can be either copper or fiber, the SRX 5000 series require a fiber link. This limitation stems from the limited space in the SFP controller on the SRX 5000's services processing card. Branch SRX devices either use a fast-ethernet (SRX100 and SRX210) or a gigabit-ethernet (SRX240 and SRX650) connection, always over copper. The end-to-end control link path should allow for maximum bandwidth utilization. In practice, the control link will have minimal bandwidth needs compared to the data plane.

If the nodes end up going into dual mastership mode, one of them must be rebooted prior to rejoining the cluster and before the connectivity is reestablished. If the two nodes are merged together as masters, it will cause all of the line cards to be briefly reset, resulting in a service outage.

As a final note, please keep in mind that, because it was designed for a direct link, the control-link and data-link traffic is not fully conformed. The switching layer must be configured to allow traffic with invalid checksum and length.

Data Link Traffic Details

The data or fabric link serves two purposes for the SRX services gateway. Its primary purpose is to synchronize real time objects or RTOs, which consist of messages used to synchronize session information between the two chassis. There are a few dozen-message types that are shared between the two devices. The most common messages utilized will be the "session create" and "session close" RTOs.

The second purposes for the data link is to forward what is known as Z-mode traffic, defined as traffic that enters one node and exits the second chassis. Both applications can be very bandwidth intensive.

In high-end platforms, RTOs are synchronized directly between SPUs on each node. An SPU will synchronize RTOs with its corresponding SPU, located in the same FPC and PIC location of the other node. The CP is the only SPU that does not synchronize any sessions. On the SRX 3000 there is only one SPU per FPC, while the SRX 5000 SPC's contains two SPUs numbered from top to bottom. Please see figure one below.

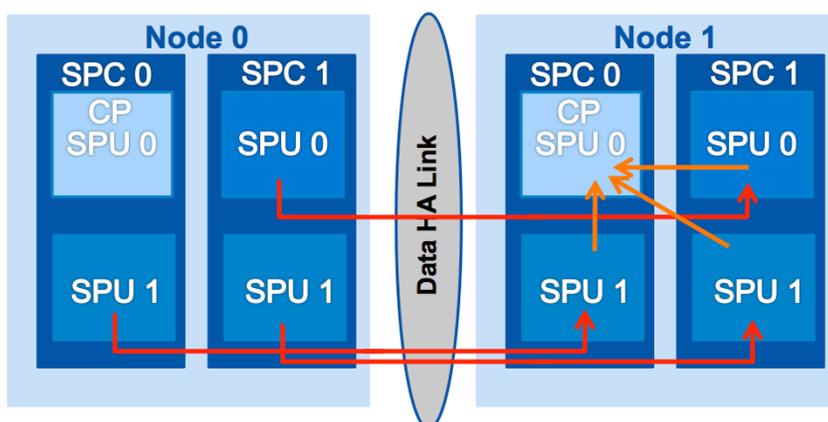


Figure 1. High-end SRX RTO synchronization

Each branch SRX platforms use a single multi-core processor. RTO synchronization messages are exchanged between the flow threads running in this processor, resulting in a situation analogous to the one for high-end systems with only one SPU.

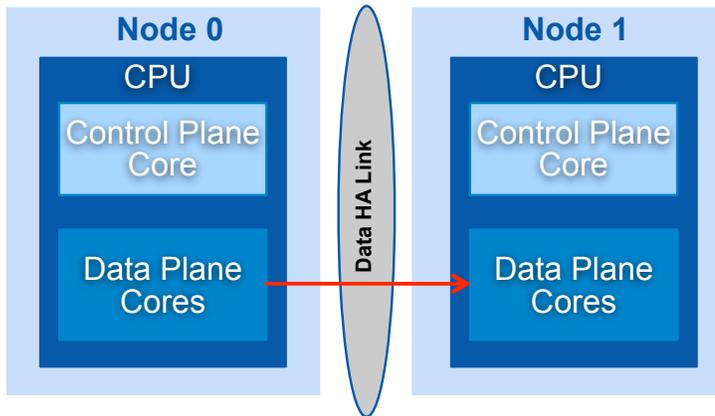


Figure 2. Branch SRX RTO synchronization

RTOs are sent out as needed over the fabric link. Each RTO message has the minimum size of 320 bytes. The packet consists of fourteen bytes for the layer two header, twenty bytes for the IP header, and the remainder used for the RTO message itself. A session create message is 286 bytes long. It is possible to have multiple messages in a single packet. This will happen based upon the rate at which the messages need to be synchronized. A good analogy of this process is the one of a railroad, where the train represents the synchronization packet and the passengers represent the RTOs. Periodically, the train will show up to pick up passengers and wait until a minimum of one passenger (RTO message) arrives. If no other passengers are available at the time of departure, the train leaves with a single passenger. However, if at this time there are additional passengers, the train will take them as well. Because of the CPU processing requirements, RTO messages are not acknowledged, thus reducing both the CPU and bandwidth overhead associated with the synchronization messages.

The size of the RTO packets is important to understand the bandwidth needs of the data link. Based upon the stated packet sizes, and assuming a worse case scenario where RTO messages are not bundled, it is possible to saturate a gigabit link¹ with 350,000 new connections per second. Due to their processing capabilities it is assumed that, when clustering High-end SRX service gateways, a minimum of one gigabit of bandwidth is available for the data link.

In the event that data traffic needs to be forwarded between the two cluster members, this too is done across the fabric link. As previously explained, this is referred to as Z-mode forwarding and occurs when a packet is received on a node in which the session is not active. When a SPU in a backup node for a particular session receives a packet, it forwards the packet over to the other chassis' matching SPU, where the session is active. The SPU of the active node for that session will then receive, process and forward the packet out 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 are no bandwidth restrictions, other than the physical link limitations, for this type of traffic.

An alternate design, called line processing, can prevent Z-mode traffic. Line processing ensures that traffic enters and exits the same node in a chassis cluster, by ensuring that there are no active egress paths on the secondary node. This design is discussed in the deployment scenarios section below.

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

To validate that the fabric link is correctly operating a heartbeat message is periodically sent. The 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 heartbeat message once every second. The fabric path is more forgiving for missing messages. If the fabric link physically goes to the down state the device will recognize this and failover 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. It is possible to receive only one packet per 66 seconds and for the fabric link to be considered healthy.

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 all communication is successful it now would be possible for the node to rejoin the cluster upon reboot. Data link connections can be either 1G or 10G media types. It is advised that the same media type is used on both sides of the link. This way, the nodes would not send data at a higher rate than what either of the links can carry.

MTU Considerations

Inter-cluster messages cannot be fragmented, requiring the transport network to have the ability accommodate them. The minimum MTU required for all platforms is 9014, with the exception of the SRX100s that require an MTU of 1632.

Due to the extra fabric headers added to the packets before they are sent through the fabric link, the MTU of the interfaces used in Z-mode deployments should not exceed 1500 bytes on SRX100 platforms and 8900 bytes on any other platform.

Table 1: Local Interfaces and Fab link MTUs

Requirement	Branch and High-End SRX	SRX100
Maximum MTU of locally connected interfaces	8900	1500
Minimum MTU of the transport network used to carry the Fabric link traffic	9014	1632

Deployment Scenarios

In this section, the deployment scenarios for cluster connectivity across layer two 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 (A/P) and active /active (A/A). These terms are focused more around the failover behavior than detailing how the devices pass traffic. The more appropriate terms are line mode and z-mode. Line mode is used when traffic passing through one or both nodes is symmetric, that is, it ingresses and egresses the same node.. In a Z-mode deployment traffic enters one node of the cluster, while exits the other node. The considerations for both deployment scenarios are discussed in this section.

The second part of this section discusses the physical infrastructure used to interconnect two SRX services gateways. In the beginning of the document the requirements for the layer two infrastructures were denoted. As long as those requirements are met the characteristics of the network used for these connections are or no relevance. However, two specific deployment scenarios are recommended. The first deployment consists of a dual switch design. The attributes for this design requires two physically separate layer two networks connecting the SRX 5800s. 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 traffic enters and exits the same physical chassis, reducing the amount of traffic sent across the data link. This might be desirable in cases where the data link is carried across a wide area network or WAN. Both active/passive or active/active deployments are possible in this mode. An active/passive deployment only allows for one chassis to forward traffic at any given time, making it a line mode deployment. Below in figure two is an example of such deployment.

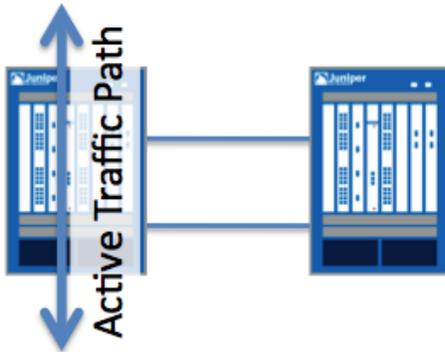


Figure 3. Active/Passive line deployment

Active / Active HA deployments on the SRX are extremely flexible and allow traffic to enter and exit any node in a cluster. While powerful in a local deployment, it can lead to excessively bandwidth usage in a WAN deployment. It is still possible, in an active/active deployment, to bind traffic to a specific node. The first way is to use local interfaces on each node, where the interfaces on a given node are bound to a virtual router (VR) routing instance. This would force any traffic that would enter a chassis to be processed by that chassis only. Because the routing instances contain interfaces belonging to a single node, the traffic would always stay local to that particular node. Routing protocols should be used to divert traffic around failures.

Figure three shows an example of this type of deployment.

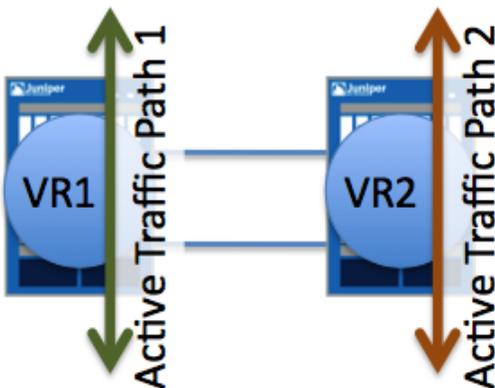


Figure 4. Local interface node binding

The alternate design makes use of redundancy groups. A redundancy group is a logical collection of objects, in this case interfaces, that can failover simultaneously between two nodes in a cluster. It is possible to create two redundancy groups, each group normally active on a different node. The redundancy groups would contain one or more redundant Ethernet interfaces (RETH) with member interfaces on both nodes. In the event of a failure in a node, the redundancy group active on that node would failover, forcing all traffic to be processed by the remaining node. The biggest difference between local and RETH interfaces is that, while RETH interfaces are only active in a single node at a given time, they can span both nodes and do not require the use of any routing protocol to route traffic around failures. Instead, the use of local interfaces

can only be used in conjunction with dynamic routing protocols or some other means to diver traffic around failures. An example of this deployment can be found below in figure 4.

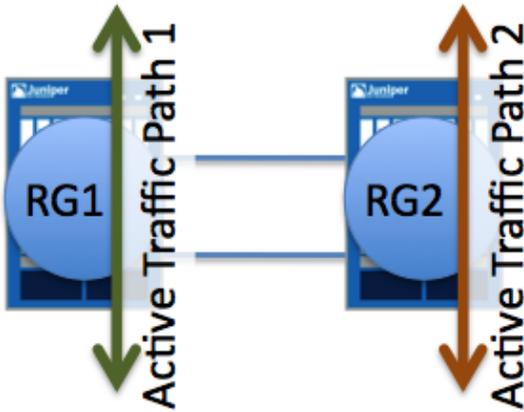


Figure 5. Redundancy group line mode design

Z-mode deployment

The alternate design includes what is known as a z-mode deployment, also characterized as an active / active deployment. On z-mode traffic enters one node but exits the second node. This will occur only when the best egress path for the traffic is through second node. This is an excellent design to utilize for the SRX when extra throughput or redundancy is needed. In the case of a multisite WAN cluster, the design loses its appeal. 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 accommodate the extra traffic, it is acceptable to utilize the Z-mode design. The Z-mode design is depicted below in Figure 6.

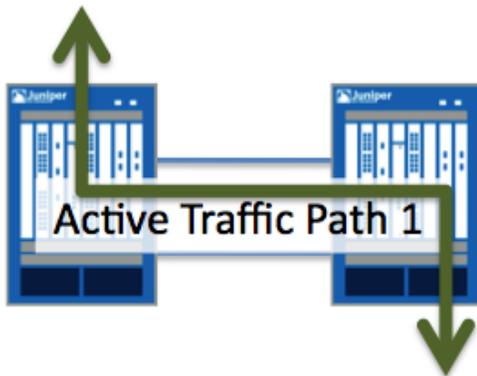


Figure 6. Z-mode chassis cluster deployment

Dual Switch Deployments

When connecting two SRX services gateways together over a layer two network its best to keep high availability in mind. The number one item to consider is reliability of the underlying network. The suggestion of Juniper Networks is to run two separate physical networks for connectivity between the nodes. Doing so would separate the control and data link avoiding a single point of failure between the two nodes. An example of this deployment can be found in Figure 7.

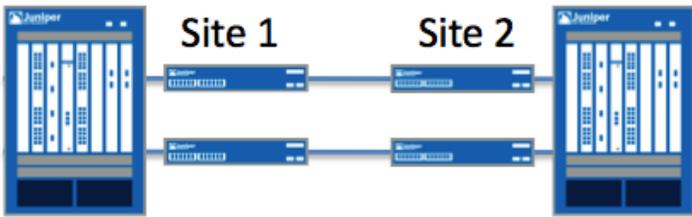


Figure 7. Dual WAN link cluster

There may be times when redundant WAN connections are not available. In these cases its acceptable to maintain physical separation between the control and data links except over the WAN connection. In the event the WAN connection goes down the cluster members will lose connectivity between each other. Below in figure seven a single WAN connected cluster is shown. This will force the cluster to go into dual mastership. Each location will have 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, which will cause a brief outage. When the master route engine is elected it must reestablish communication between itself and the remote FPCs, causing traffic to stop for several seconds. This can be avoided by rebooting one of the firewalls and then reconnect the control and data links.

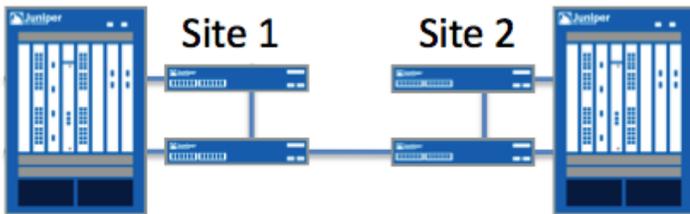


Figure 8. Single WAN link cluster

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 virtual private LAN service or VPLS. VPLS allows for layer-2 connectivity across a WAN. In a VPLS deployment layer two packets can be forwarded across the WAN. The connectivity looks as if the chassis cluster members are connected directly over a switch. In this non-Ethernet deployment the same layer two requirements such as latency need to be followed. As long as the layer two requirements are followed then it is acceptable to use VPLS for connectivity.



Figure 9. VPLS connected chassis cluster

Besides VPLS there are several other layer two technologies that can be utilized to accomplish the goal of connectivity between the cluster members. This includes L2VPNs and circuit cross-connect (CCC). When using alternate technologies keep in mind the requirements stated in requirements for the layer-2 connectivity.

Branch SRX devices Control-Plane Connection

As previously noted, the control-plane link on branch SRX devices tags traffic with VLAN ID 4094. This traffic must be carried unaltered by the transport network. The following sections will describe some configuration options to transport this traffic across an ethernet-based layer-2 network.

Using VLAN Trunks for the Transport of Control Plane Traffic

The simplest way to transport tagged traffic across an ethernet WAN network is to configure the switch port that connects to the Branch SRX control ports as a trunk port, member of the VLAN 4094. Traffic received by the trunk port will be forwarded without modifying the existing VLAN tag. This option can only be used if no other VLAN in the network is using the 4094 VLAN ID.

Using 802.1q Tunneling for the Transport of Control Plane Traffic

The most flexible alternative is the use of 802.1q tunneling, which uses nested VLAN tags to carry the traffic over the transport network, thus avoiding any possible conflicts with the VLAN tags used by the transport network. When this alternative is used, the switch ports insert a user configurable VLAN ID on the traffic received. The new VLAN ID is used by the transport network to forward the traffic, regardless of the inner labels the packets might carry, thus avoiding any possible VLAN conflicts. After traffic reaches its destination the outer VLAN tag is removed, making the operation transparent to the endpoints.

A sample configuration for EX switches that uses the VLAN ID 150 to transport the traffic over the layer-2 WAN is shown below

```
vlans {  
  SRX210-control-port {  
    vlan-id 150;  
    interface {  
      ge-3/0/44.0;#Access interface connected to the branch SRX device control-port  
      ge-3/0/43.0;#Trunk port connected to the ethernet WAN  
    }  
    dot1q-tunneling;  
  }  
}
```

Summary

Expanding the physical deployment across multiple locations can be quite simple. The configuration on the SRX services gateway is the same whether or not it is connected directly or indirectly. This simplifies the deployment strategy for implementing layer two connectivity for the SRX. The most important factor to consider in the deployment of the SRX over a layer two network is to understand the requirements for the deployment. This application note provides clear guidelines that will ensure a successful deployment in the correct environment.