APPLICATION NOTE

BRANCH SRX SERIES AND J SERIES CHASSIS CLUSTERING

Configuring Chassis Clusters on Branch SRX Series Services Gateways and J Series Services Routers

Table of Contents

Table of Figures

Introduction

Starting with the 9.0 release of Juniper Networks® Junos® operating system, Juniper Networks J Series Services Routers may be deployed using the chassis cluster feature to provide high availability (HA) between devices. This feature is only available on J Series devices with the flow-enabled Junos OS. With the introduction of Juniper Networks SRX Series Services Gateways for the branch in Junos OS release 9.5, HA is supported on all branch SRX Series devices.

Scope

The purpose of this application note is to review the HA chassis clustering feature, together with its limitations and design considerations. We will also discuss some common use cases and how they relate to their Juniper Networks ScreenOS® Software NetScreen Redundancy Protocol (NSRP) counterparts.

Design Considerations

High availability between devices is easily incorporated into enterprise designs and is particularly relevant when architecting branch and remote site links to larger corporate offices. By leveraging the HA feature, enterprises can ensure connectivity in the event of device or link failure.

Hardware Requirements

- Two identical J Series secure routers per cluster (Juniper Networks J2320 Services Router, J2350 Services Router, J4350 Services Router, or J6350 Services Router)
- Two identical SRX Series gateways per cluster (Juniper Networks SRX100 Services Gateway, SRX210 Services Gateway, SRX240 Services Gateway, or SRX650 Services Gateway)

Software Requirements

- Flow-enabled Junos OS 9.0 or later for J Series secure routers
- Junos OS release 9.5 and later for SRX Series Services Gateways

Description and Deployment Scenario

Chassis clustering between devices may be deployed in either active/passive or active/active scenarios. Junos OS allows an HA cluster to additionally be used in asymmetric routing scenarios. Code examples are provided throughout this document, and deployment scenarios are discussed towards the end of the paper.

Feature Description

The HA feature is modeled after redundancy features first introduced in Juniper Networks M Series Multiservice Edge Routers and Juniper Networks T Series Core Routers. We will first give a brief overview of the way Junos OS redundancy works, so that we can better understand how this model is applied when clustering devices. As Junos OS is designed with separate control and data planes, redundancy must operate in both. The control plane in Junos OS is managed by Routing Engines (REs), which perform all the routing and forwarding computations (among many other functions). Once the control plane converges, forwarding entries are pushed to all Packet Forwarding Engines (PFEs), which are virtualized on J Series routers. PFEs then perform route-based lookups to determine the appropriate destination for each packet independent of the REs. This simplistic view of the Junos OS forwarding paradigm is represented in Figure 1.

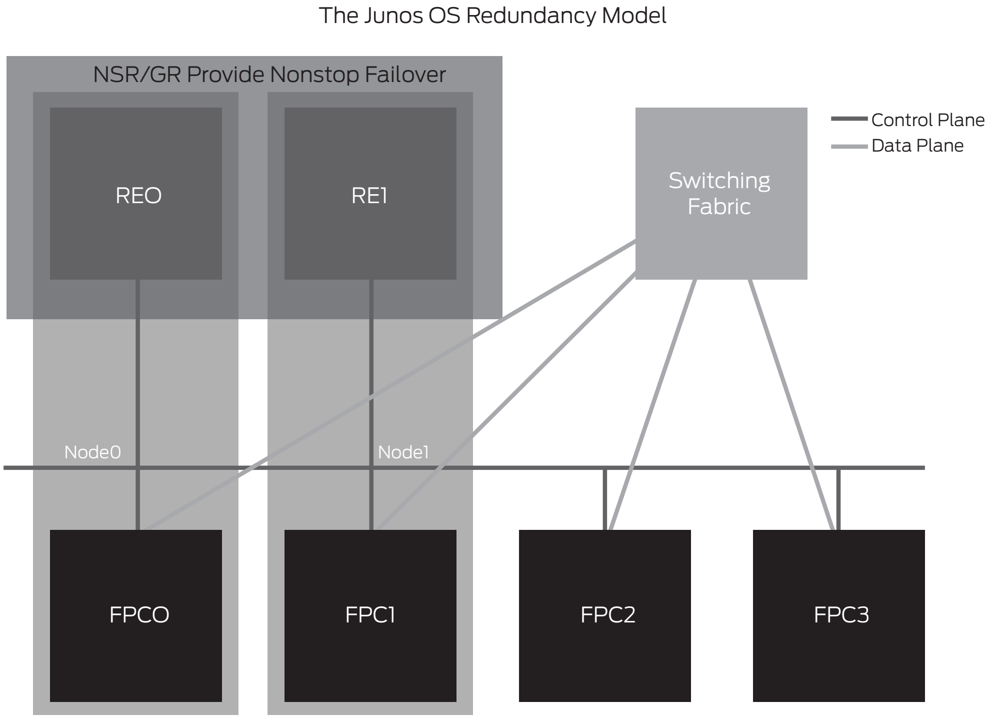


Figure 1: Junos OS redundancy model

Control plane failover is provided in Junos OS by using graceful restart or nonstop active routing (NSR). In the former, the router signals a control plane failure to the rest of the network, while continuing to forward traffic on the data plane (since a control plane failure doesn't affect the forwarding plane). The rest of the network will continue to use the restarting router (for a grace period), while the restarting router forms new adjacencies. The backup RE in this scenario detects the entire configuration, but not the runtime state of the control plane. In a failure, the backup RE has to recalculate all routing/forwarding tables. Nonstop routing leverages state replication between Routing Engines. In this case, a restarting router handles control plane failures transparently, as the backup RE takes control of the router without any assistance from the rest of the network. Routing protocols handle data plane failures, while interface, PFE, or FPC failovers are handled by diverting traffic through other interfaces, which can be achieved by using conventional routing protocols, Virtual Router Redundancy Protocol (VRRP), or aggregate interfaces. When enabling a chassis cluster for J Series routers, Junos OS uses a similar model—less the nonstop routing state replication—to provide control plane redundancy as shown in Figure 2.

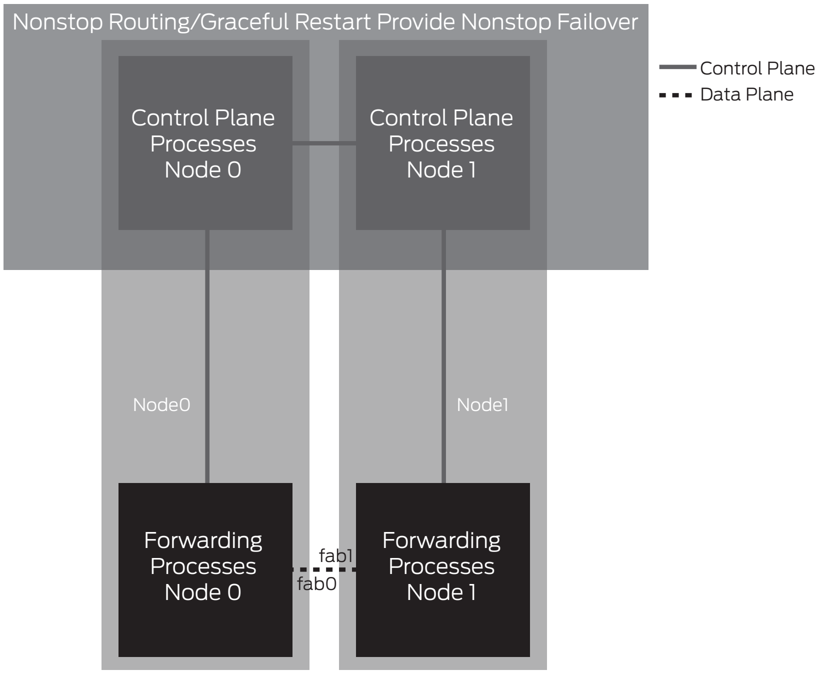

Nonstop Routing/Graceful Restart Provide Nonstop Failover
Control Plane
Data Plane
Control Plane Processes Node 0
Control Plane Processes Node 1
Node0
Node1
Forwarding Processes Node 0
Forwarding Processes Node 1
fab1
fab0

Figure 2: Device clustering

The chassis clustering feature supports clustering of two devices and requires two connections between the devices as previously illustrated. The chassis cluster is seen as a single device by both external devices and administrators of the cluster. When clustering is enabled, node 1 of the cluster will renumber its interfaces to avoid collisions with node 0. Depending on the model used (only two devices of the same model can be clustered), node 1 will renumber its interfaces by adding the total number of system FPCs to the original FPC number of the interface. (On a J Series router, the onboard ports and each Physical Interface Module (PIM) slot correspond to an FPC.) Accordingly, when clustering two J2320 routers, node 1 will renumber its interfaces as ge-4/0/0 to ge-7/0/0, because a J2320 has three PIM slots and four standard GbE ports on the system board acting as FPC0. The following table summarizes the renumbering schema.

Table 1: Interface Renumbering

DEVICE	RENUMBERING CONSTANT	NODE 0 INTERFACE NAME	NODE 1 INTERFACE NAME
J2320	4	ge-0/0/0	ge-4/0/0
J2350	5	ge-0/0/0	ge-6/0/0
J4350	7	ge-0/0/0	ge-7/0/0
J6350	7	ge-0/0/0	ge-7/0/0
SRX100	1	fe-0/0/0	fe-1/0/0
SRX210	2	ge-0/0/0	ge-2/0/0
SRX240	5	ge-0/0/0	ge-5/0/0
SRX650	9	ge-0/0/0	ge-9/0/0

After clustering is enabled, the system creates fxp0, fxp1, and fab interfaces. Depending on the platform, fxp0 and fxp1 are mapped to a physical interface. This is not user configurable. The fab interface is user configurable. The following table summarizes the fxp0 and fxp1 mappings.

Table 2: Mapping of Interfaces fxp0 and fxp1

DEVICE	FXP0 INTERFACE	FXP1 INTERFACE	FAB INTERFACE
J2320	ge-0/0/2	ge-0/0/3	User defined
J2350	ge-0/0/2	ge-0/0/3	User defined
J4350	ge-0/0/2	ge-0/0/3	User defined
J6350	ge-0/0/2	ge-0/0/3	User defined
SRX100	fe-0/0/6	fe-0/0/7	User defined
SRX210	ge-0/0/0	fe-0/0/7	User defined
SRX240	ge-0/0/0	ge-0/0/1	User defined
SRX650	ge-0/0/0	ge-0/0/1	User defined

As seen in figure 2, fxp1 (the HA link) provides control plane communication between the nodes in the cluster, and fxp0 provides management access and is limited to host traffic only. Traffic received through the fxp0 interface will not be forwarded to any other interface in the system. Fab interfaces are used to exchange data plane information and traffic between devices. As opposed to the fxp0 and fxp1 interfaces, the fab interface can be mapped to any Ethernet interface in the system.

The control plane redundancy of the cluster is similar to that used within single M Series and T Series routers. Each device acts as a Routing Engine in a system with redundant REs. Graceful restart is used to provide control plane failover with minimal traffic impact on the network. The control plane redundancy model is active/passive, where a node in the cluster is designated as the active device and performs all cluster routing calculations. Except for a few key processes required for managing clustering, most of the processes are running only on the master RE. When the primary node fails, the routing process and other processes in the backup device will become active and assume control plane operations.

Data plane redundancy is somewhat more involved. M Series and T Series routers perform traffic forwarding on a packet by packet basis. There is no concept of flow, and each PFE maintains a copy of the forwarding table that was distributed by the active RE. The forwarding table allows each PFE to perform traffic forwarding independent of other system PFEs. If a PFE fails, the rest of the PFEs in the system are unaffected, allowing the control plane to reroute the traffic to a working PFE. In contrast, J Series secure routers and the SRX Series gateways inspect all traffic and keep a table of all active sessions. Whenever a new connection is allowed through the system, the device makes note of the 5-tuple that identifies a particular connection (source and destination IP addresses, source and destination ports as applicable, and protocol) and updates the table with session details such as next hop, session timeouts, sequence numbers (if the protocol is TCP), and other session-specific information required to guarantee that no packets are forwarded from unknown or undesired protocols (or users). Session information is updated as traffic traverses the device and is required on both devices in a cluster to guarantee that established sessions are not dropped when a failover occurs.

As shown in Figure 1, the control plane REs function in active/backup mode while the data plane (PFEs) function in active/active mode. With active/active PFEs, it is possible for traffic to ingress the cluster on one node and egress from the other node, which means that both nodes need to be able to create and synchronize sessions. For example, when return traffic arrives asymmetrically at the node that did not record the initial session, the chassis cluster feature gracefully forwards the traffic to the original node for processing, which prevents security features from being compromised. Please be aware that the previous discussion applies only to routed traffic. Junos OS with enhanced services does not support the forwarding of Layer 2 traffic (transparent mode). Chassis clustering supports unicast IPv4 traffic only.

Redundant Ethernet Interfaces

As previously discussed, control plane failures are detected by member nodes, causing the backup node to take control of the cluster. Conversely, data plane failures rely on routing protocols to reroute traffic or redundant Ethernet interfaces to overcome interface failures. The concept of redundant Ethernet is fairly simple; two Ethernet interfaces (one from each node in a cluster) are configured as part of the same redundant Ethernet interface (RETH interface in Junos OS terminology). The RETH interface is then configured as part of a redundancy group. A redundancy group is active only on one of the nodes in the cluster, and the redundant Ethernet interfaces that are members of that group will send (and normally receive) traffic only through the physical interfaces on the active node.

A redundancy group can be configured to monitor one or more physical interfaces. Each monitored interface is given a weight, which is subtracted from the redundancy group threshold if the interface fails. If the threshold—due to interface failover—becomes less than zero, the redundancy group transitions state, causing the other node in the cluster to become active for the group. Consequently, all the redundant Ethernet interfaces that are part of this redundancy group will use the interfaces on the new node to send (and normally receive) traffic, thus routing traffic around the failure. Readers familiar with NSRP will note that RETH interfaces are analogous to virtual security interfaces (VSI) on ScreenOS devices. RETH interfaces, just like VSIs, share the same IP and media access control (MAC) addresses between the different physical interfaces that are members of the VSI/RETH. The redundant interfaces send gratuitous Address Resolution Protocol (ARP) messages when failing over and appear as a single interface to the rest of the network. There are, however, a few significant differences between RETHs and VSIs:

- RETH interfaces always contain the same type of physical Ethernet interfaces—for example, fe-fe or ge-ge.
- VSIs will always force a failover when the physical interface of the active VSI goes down. The state of the redundant Ethernet interface is purely a function of the state of the redundancy group with which the RETH is associated. A RETH interface will go down if its active physical interface is down.
- RETH interfaces will only fail over based on the monitoring of physical interfaces.
- IP tracking and zone monitoring are currently not supported in Junos OS.

To be clear, RETH interfaces are not required to provide HA. Session information will be synchronized regardless of the ingress or egress interface type. Traditional routing protocols can be used to route around failures, but when connecting to simple devices that do not support routing protocols, redundant Ethernet interfaces can be useful to overcome this limitation.

Feature Support and Comparison Matrix

Although both protocols were designed to provide the same services, NSRP and JSRP (the protocol used in Junos OS) do not operate in the same manner and do not provide the same set of features. The following table summarizes the main differences between the protocols.

Table 3: Feature Comparison

FEATURE	JSRP	NSRP
Session replication	YES	YES
Application-level gateway (ALG) replication	YES	YES
Network Address Translation (NAT) session replication	YES	YES
IPsec session replication (policy-based VPN)	YES	YES
IPsec session replication (route-based VPN)	YES	YES
Route synchronization	N/A	YES
Interface monitoring	YES	YES
Zone monitoring	NO	YES
Track IP	NO	YES
Asymmetric routing	YES	NO
Load balancing	YES	NO
Graceful restart	YES	NO
Nonstop routing (future)	YES	NO
Layer 2 mode	NO	YES

Clustering Configuration

This section outlines the steps required to configure J Series chassis clustering. Steps 1 through 3 are the minimum required. After this minimal configuration, two J Series secure routers will appear as a single device controlling all interfaces in both nodes. Steps 4 through 6 detail the configuration statements needed to specify the IP addresses of the management interface (fxp0) and the host name of each cluster node (node 0 and node 1 will have different management IPs and hostnames). Step 7 describes the configuration needed to add redundant Ethernet interfaces and the associated redundancy groups.

In this example, we will assume that we are enabling chassis clustering for a pair of J2320 devices, node left and node right, which are connected back-to-back using interfaces ge-0/0/1 and ge-0/0/3.

1. Log into each device and enable clustering by setting the appropriate cluster ID in the EEPROM. A reboot is required for this setting to take effect. Only node 0 and node 1 can be configured, as the current implementation is limited to two nodes in a cluster. In this example, node 0 (left) and node 1 (right) will be renumbered as illustrated in Table 1.

```
set chassis cluster cluster-id <n> node <m> reboot
On node left:
root@left> set chassis cluster cluster-id 1 node 0 reboot
On node right:
root@right> set chassis cluster cluster-id 1 node 1 reboot
```

Note: Step #1 must be performed in operational mode, not in configuration mode.

After the nodes reboot, they will form a cluster. From this point forward, the configuration of the cluster is going to be synchronized between the node members. The following commands are entered from the configuration mode on either of the devices.

2. Define the interfaces used for the fab connection. These interfaces must be connected back to back, or through a Layer 2 infrastructure, as shown in Figure 2. As expected, fab0 is the fabric interface of node0, while fab1 is the fabric interface of node1.

```
set interface fab0 fabric-options member-interfaces <interface>
set interface fab1 fabric-options member-interfaces <interface>
```

3. Configure the management interface on each device using configuration groups.

```
set groups node0 system host-name <node0 hostname>
set groups node0 interfaces fxp0 unit 0 family inet address <node0 mgmt IP>
set groups node1 system host-name <node1 hostname>
set groups node1 interfaces fxp0 unit 0 family inet address <node1 mgmt IP>
```

4. (Optional) Configure device-specific options.

```
set groups node0 snmp description <node0 snmp sysDesc>
set groups node1 snmp description <node1 snmp sysDesc>
```

5. Apply the group configuration.

```
set apply-groups “${node}”
```

6. (Optional) Define the redundancy groups and RETH interfaces if using redundant Ethernet interfaces.

```
set chassis cluster reth-count <n>
set chassis cluster redundancy-group 1 node 0 priority <n>
set chassis cluster redundancy-group 1 node 1 priority <n>
set interface <interface name> gigether-options redundant-parent reth.<n>
```

The resulting sample configuration is shown below:

```
#The following declares int ge-0/0/1 in node 0 as the fab interface for the node
set interface fab0 fabric-options member-interfaces ge-0/0/1
#The following declares int ge-4/0/1 in node 1 as the fab interface for the node
set interface fab1 fabric-options member-interfaces ge-4/0/1
#Groups configuration. Configuration parameters specific to each node are set here.
set groups node0 system host-name left
set groups node0 interfaces fxp0 unit 0 family inet address 192.168.3.10/24
set groups node1 system host-name right
set groups node1 interfaces fxp0 unit 0 family inet address 192.168.3.11/24
set apply-groups “${node}”
#Define a single RETH interface for the cluster
set chassis cluster reth-count 1
#Define node 0 as the primary node for reth0
set chassis cluster redundancy-group 1 node 0 priority 100
set chassis cluster redundancy-group 1 node 1 priority 1
#Add interfaces ge-0/0/0 (in node 0) and ge-4/0/0 (ge-0/0/0 in node 1) to the
reth
set interface ge-0/0/0 gigether-options redundant-parent reth0
set interface ge-4/0/0 gigether-options redundant-parent reth0
set interfaces reth0 unit 0 family inet address <reth0-ip-address>
set interfaces reth1 redundant-ether-options redundancy-group <rg-id>
#Define node 0 as the primary node for the control path
set chassis cluster redundancy-group 0 node 0 priority 100
set chassis cluster redundancy-group 0 node 1 priority 1
Disabling clustering is a very simple process—first set the cluster id of each
node to 0 and then reboot the nodes.
set chassis cluster cluster-id 0 node 0
```

Cluster Monitoring

The following commands can be used to verify the status of a cluster and present a view of the cluster from a node’s perspective. Statistics are not synchronized between the nodes in the cluster. When debugging clusters, it is useful to log into each member node and analyze the output from each.

Viewing the Chassis Cluster Status

The command below shows the different redundant groups configured in the cluster, together with their specified priorities and the status of each node. This command is useful when trying to determine which RETH interfaces are active on each node. The special redundancy group 0 refers to the status of the control plane. In this example, node 0 is the primary node for this group and, therefore, it is in charge of all control plane calculations (it acts as the master RE and runs the control plane processes like rpd, kmd, dhcpd, pppd, and others).

```
show chassis cluster status
Cluster: 1, Redundancy-Group: 0
    Device name             Priority    Status    Preempt  Manual failover

    node0                   100         Primary   No       No
    node1                   1           Secondary No       No

Cluster: 1, Redundancy-Group: 1
    Device name             Priority    Status    Preempt  Manual failover

    node0                   100         Primary   Yes      No
    node1                   1           Secondary Yes      No
```

Viewing the Cluster Statistics

The command below displays the statistics of the different objects being synchronized, the fabric and control interface hellos, and the status of the monitored interfaces in the cluster.

```
show chassis cluster statistics
Initial hold: 5

Reth Information:
    reth     status      redundancy-group
    reth0    up          1
Services Synchronized:
    Service-name                              Rtos-sent    Rtos-received
    Translation Context                       0            0
    Incoming NAT                              0            0
    Resource Manager                          10           0
    Session-create                            225          10592
    Session-close                             222          10390
    Session-change                            0            0
    Gate-create                               0            0
    Session-Ageout-refresh-request            149          1
    Session-Ageout-refresh-reply              0            0
    VPN                                       0            0
    Firewall User Authentication              0            0
    MGCP Alg                                  0            0
    H323 Alg                                  0            0
    SIP Alg                                   0            0
    SCCP Alg                                  0            0
    PPTP Alg                                  0            0
    RTSP Alg                                  0            0
Interface Monitoring:
    Interface         Weight    Status    Redundancy-group
    ge-4/0/0          255       up        1
    ge-0/0/0          255       up        1
    fe-5/0/0          255       up        1
    fe-1/0/0          255       up        1
```

```
chassis-cluster interfaces:
    Control link: up
    244800 heart beats sent
    244764 heart beats received
    1000 ms interval
    3  threshold
chassis-cluster interfaces:
    Fabric link: up
    244786 heartbeat packets sent on fabric-link interface
    244764 heartbeat packets received on fabric-link interface
```

Viewing the Control Link Status

This command displays the status of the control interface (fxp1) of this particular node

```
show chassis cluster interface
Physical Interface: fxp1.0, Enabled, Control interface , Physical link is Up
```

Viewing the Session

The command shown below displays the sessions in the session table of each node by specifying the node number. Synchronized sessions will be seen in both nodes, where they will appear as active in one node and backup in the other. A detailed view of a session can be obtained by specifying the session id

```
show security flow session node0

Session ID: 2, Policy name: self-traffic-policy/1, State: Active, Timeout: 1800
  In: 172.24.241.53/50045 --> 172.19.101.34/22;tcp, If: ge-0/0/0.0
  Out: 172.19.101.34/22 --> 172.24.241.53/50045;tcp, If: .local..0

1 sessions displayed


show security flow session session-identifier 2
Session ID: 2, Status: Normal, State: Active
Flag: 0x40
Virtual system: root, Policy name: self-traffic-policy/1
Maximum timeout: 1800, Current timeout: 1800
Start time: 1900, Duration: 256
   In: 172.24.241.53/50045 --> 172.19.101.34/22;tcp,
    Interface: ge-0/0/0.0,
    Session token: 0xa, Flag: 0x4097
    Route: 0x20010, Gateway: 172.19.101.1, Tunnel: 0
    Port sequence: 0, FIN sequence: 0,
    FIN state: 0,
   Out: 172.19.101.34/22 --> 172.24.241.53/50045;tcp,
    Interface: .local..0,
    Session token: 0x4, Flag: 0x4112
    Route: 0xfffb0006, Gateway: 172.19.101.34, Tunnel: 0
    Port sequence: 0, FIN sequence: 0,
    FIN state: 0,

1 sessions displayed
```

TCP sequence numbers are not synchronized. However, the active node for a given session will keep track of the sequence numbers. When a session is migrated due to a failure (for example, failures that cause the egress interface of a session/group of sessions to be in a different node than prior to the failure), the sequence number counting will resume on the new node based on the sequence numbers of the packets going through the new active node for the session(s).

Deployment Scenarios

NSRP has been used in multiple networks with several topologies. This section provides the equivalent SRX Series services gateway or J Series router for these typical scenarios.

Active/Passive Cluster

In this case, a single device in the cluster is used to route all traffic, while the other device is used only in the event of a failure. When a failure occurs, the backup device becomes master and takes over all forwarding tasks.

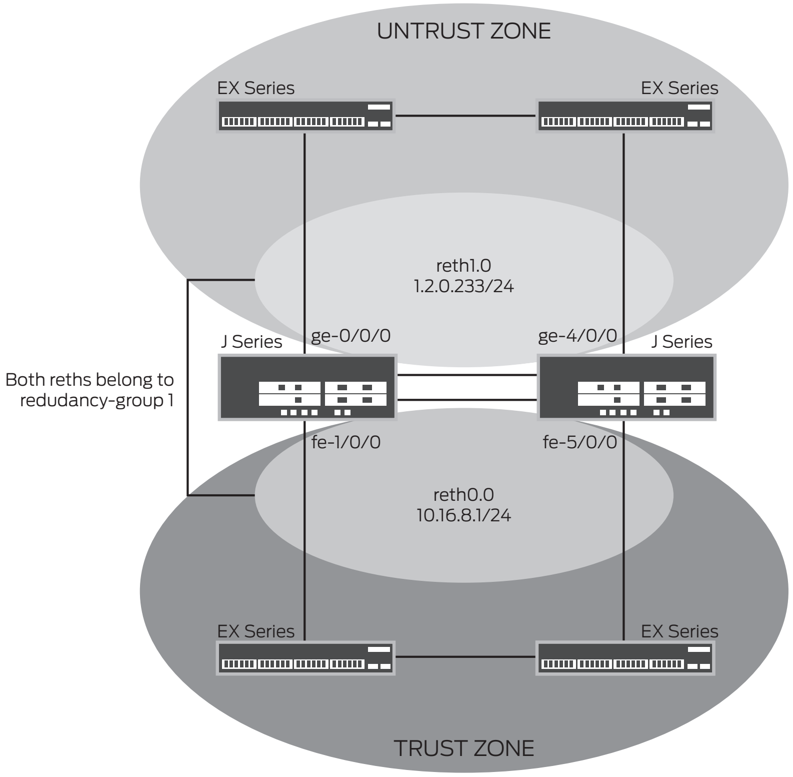


Figure 3: Active/passive cluster

Active/passive can be achieved using RETH interfaces, just as one would do using VSIs. The redundancy group determines the RETH state by monitoring the state of the physical interfaces in reth0 and reth1. If any of these interfaces fails, the group is declared inactive by the system that hosts the failing interface. On a failure, both RETH interfaces will fail over simultaneously, as they belong to the same redundancy group. This configuration minimizes the traffic around the fabric link, as only one node in the cluster will be forwarding traffic at any given time.

```
#Groups Definitions
set groups node0 system host-name J2320-A
set groups node0 interfaces fxp0 unit 0 family inet address 192.168.3.110/24
set groups node1 system host-name J2320-B
set groups node1 interfaces fxp0 unit 0 family inet address 192.168.3.111/24
set apply-groups “${node}”

#Cluster Configuration, redundancy-group 0 determines the status of the RE
mastership, while redundancy-group 1 is used to control the reth interfaces
set chassis cluster reth-count 2
set chassis cluster heartbeat-threshold 3
set chassis cluster node 0
set chassis cluster node 1
set chassis cluster redundancy-group 0 node 0 priority 100
set chassis cluster redundancy-group 0 node 1 priority 1

#The ge-0/0/0 interface on each node is used as the fabric interface between the
nodes
set interfaces fab0 fabric-options member-interfaces ge-0/0/1
set interfaces fab1 fabric-options member-interfaces ge-4/0/1

#Note how the redundancy-group 1 is configured to monitor all the physical
interfaces forwarding traffic. The preempt keyword causes the mastership to be
reverted back to the primary node for the group (node 0, which has a higher
priority) when the failing interface causing the switchover comes back up
set chassis cluster redundancy-group 1 node 0 priority 100
set chassis cluster redundancy-group 1 node 1 priority 1
set chassis cluster redundancy-group 1 preempt
set chassis cluster redundancy-group 1 interface-monitor fe-1/0/0 weight 255
set chassis cluster redundancy-group 1 interface-monitor fe-5/0/0 weight 255
set chassis cluster redundancy-group 1 interface-monitor ge-0/0/0 weight 255
set chassis cluster redundancy-group 1 interface-monitor ge-4/0/0 weight 255

#(Optionally) If both data processing and control plane functions want to be
performed in the same node, then redundancy-group 0 must monitor also the
physical interfaces. If control and data planes are allowed to fail over
independently, the following four commands should not be set.
set chassis cluster redundancy-group 0 interface-monitor fe-1/0/0 weight 255
set chassis cluster redundancy-group 0 interface-monitor fe-5/0/0 weight 255
set chassis cluster redundancy-group 0 interface-monitor ge-0/0/0 weight 255
set chassis cluster redundancy-group 0 interface-monitor ge-4/0/0 weight 255

set interfaces ge-0/0/0 gigether-options redundant-parent reth1
set interfaces fe-1/0/0 fastether-options redundant-parent reth0
set interfaces ge-4/0/0 gigether-options redundant-parent reth1
set interfaces fe-5/0/0 fastether-options redundant-parent reth0
set interfaces reth0 redundant-ether-options redundancy-group 1
set interfaces reth1 redundant-ether-options redundancy-group 1

#Just as regular interfaces, reth interfaces must be part of a security zone
set security zones security-zone Untrust interfaces reth1.0
set security zones security-zone Trust interfaces reth0.0
```

Asymmetric Routing Scenario

This scenario makes use of the asymmetric routing capability of Junos OS with enhanced services. Traffic received by a node is matched against that node's session table. The result of this lookup indicates whether that node processes the session or forwards it to the other node through the fabric link. Sessions can then be anchored to any device in the cluster; and, as long as the session tables are replicated, the traffic will be correctly processed. To minimize fabric traffic, sessions are always anchored to the node hosting the egress interface for that particular connection.

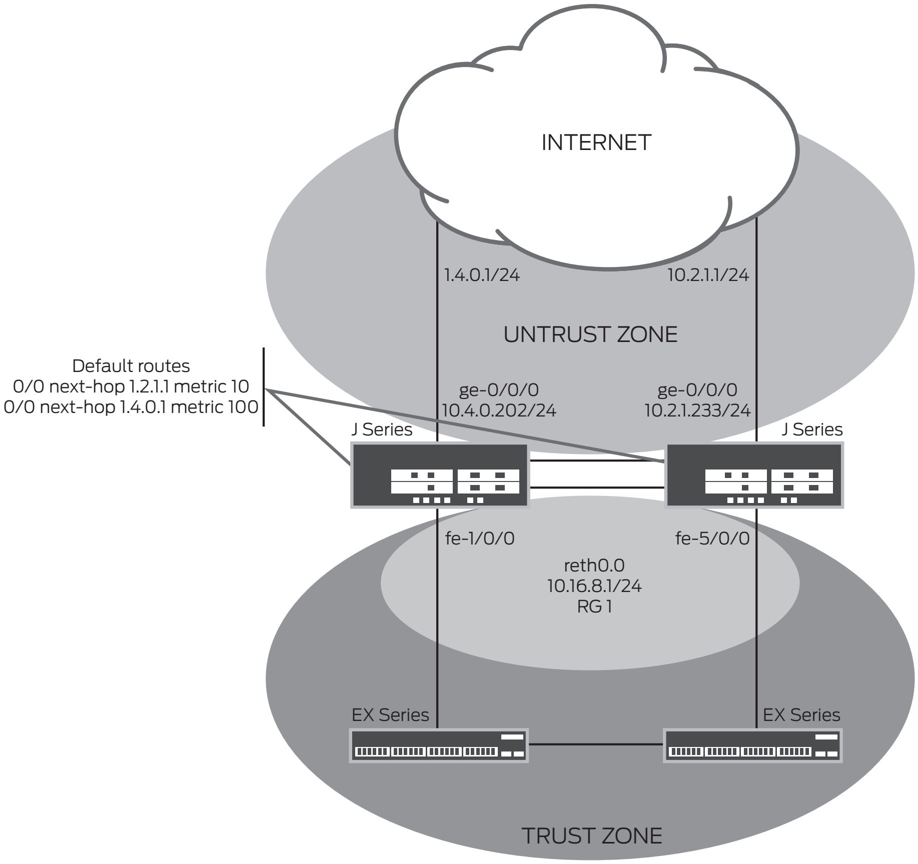


Figure 4: Asymmetric routing scenario

Error! Reference source not found. shows an example of how asymmetric routing is supported. In this scenario two Internet connections are used with one being preferred. The connection to the trust zone is made using a RETH interface to provide LAN redundancy for the devices in the trust zone. For illustrative purposes, we will describe two failover cases in which sessions originate in the trust zone with a destination of the Internet (untrust zone).

Case I: Failures in the Trust Zone RETH

Under normal operating conditions, traffic will flow from the trust zone to the interface ge-0/0/0 (belonging to reth0.0) in node 0. Since the primary Internet connection resides in node 0, the sessions will be created in both node 0 and node 1 but will only be active in node 0 (since the egress interface for all of these sessions is fe-1/0/0 belonging to node 0).

A failure in the ge-0/0/0 interface will trigger a failover of the redundancy group, causing the interface ge-4/0/0 (ge-0/0/0 in node 1) to become active. After the failover, traffic will arrive at node 1. After session lookup, the traffic will be sent to node 0 as the session will be active in this node (since the egress interface, fe-1/0/0 is hosted in this node 0). Node 0 will then process the traffic and forward it to the Internet. The return traffic will follow a similar process. Traffic will arrive at node 0, be processed at node 0 (since the session is anchored to this node), and be sent to node 1 through the fabric interface where node 1 will forward it through the ge-4/0/0 interface.

Case II: Failures in the Untrust Zone Interfaces

This case differs from the previous one in that sessions will be migrated from node to node. As in the previous case, traffic will be processed only by node 0 under normal operating conditions. A failure of interface fe-1/0/0 connected to the Internet will cause a change in the routing table, which will have a default route after the failure pointing to interface fe-5/0/0 in node 1. After the failure, the sessions in node 0 will become inactive (since the egress interface now will reside in node 1), and the backup sessions in node 1 will become active. Traffic arriving from the trust zone will still be received on interface ge-0/0/0, but will be forwarded to node 1 for processing. After traffic is processed in node 1, it will be forwarded to the Internet through the fe-5/0/0 interface.

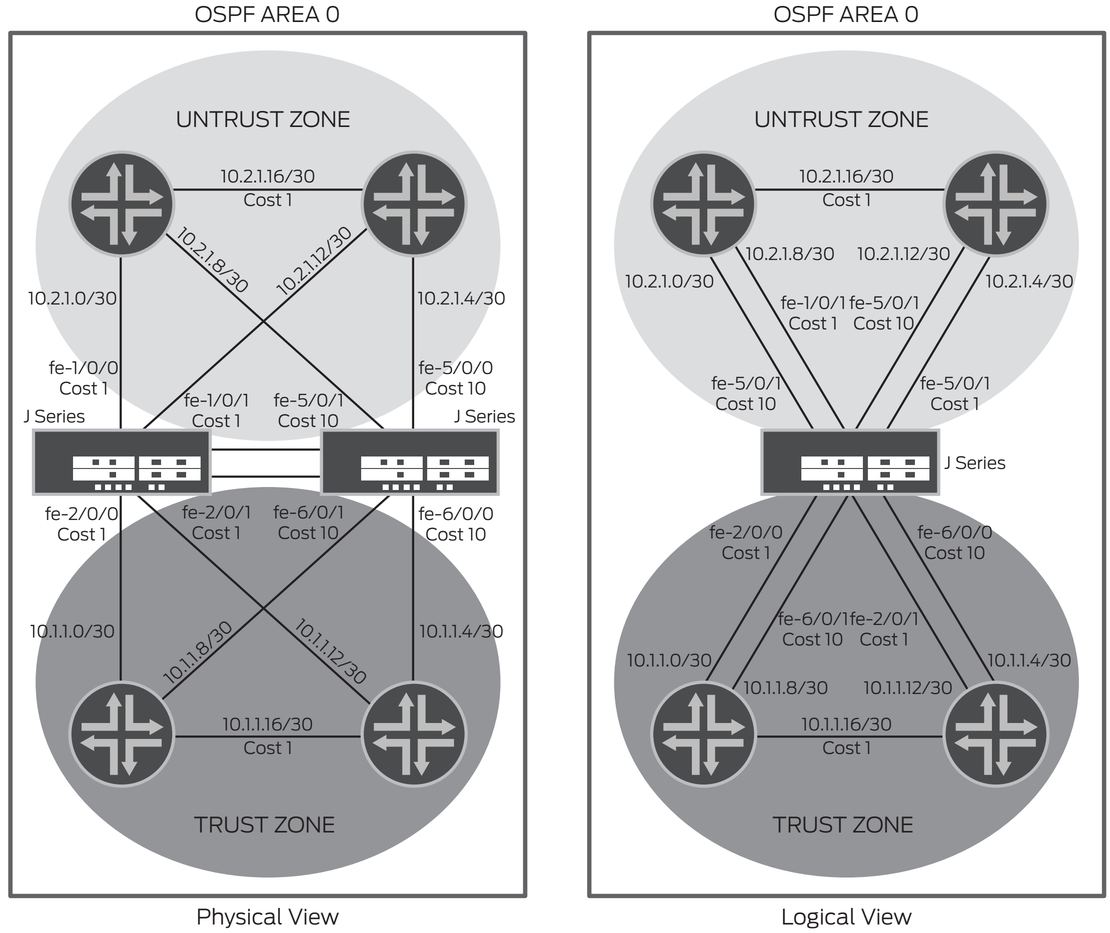
Note that if this scenario were used with source NAT, to accommodate different address spaces assigned by different providers, the above would not work as the egress sessions would be NATed differently after the failover (this is not a limitation of the HA implementation, but a consequence of the fact that if two Internet service providers (ISPs) are used, the customer doesn't own a public address space, and a failure in one of the ISPs will result in the loss of connectivity from all IPs belonging to the failed service provider).

```
#Cluster Configuration, redundancy-group 1 is used to control the RETH interface
connected to the trust zone. Note how the redundancy group (and therefore reth0)
will only failover if either fe-1/0/0 or fe-5/0/0 fail, but not if any of the
interfaces connected to the Internet fails.
set chassis cluster reth-count 1
set chassis cluster node 0
set chassis cluster node 1
set chassis cluster redundancy-group 1 node 0 priority 100
set chassis cluster redundancy-group 1 node 1 priority 1
set chassis cluster redundancy-group 1 preempt
set chassis cluster redundancy-group 1 interface-monitor fe-1/0/0 weight 255
set chassis cluster redundancy-group 1 interface-monitor fe-5/0/0 weight 255

#Interface Definitions
set interfaces ge-0/0/0 unit 0 family inet address 1.4.0.202/24
set interfaces fe-1/0/0 fastether-options redundant-parent reth0
set interfaces fe-1/0/1 disable
set interfaces ge-4/0/0 unit 0 family inet address 1.2.1.233/24
set interfaces fe-5/0/0 fastether-options redundant-parent reth0
set interfaces reth0 unit 0 family inet address 10.16.8.1/24

#ge-0/0/1 one each node will be used for the fab interfaces
set interfaces fab0 fabric-options member-interfaces ge-0/0/1
set interfaces fab1 fabric-options member-interfaces ge-4/0/1

#We have two static routes, one to each ISP, but the preferred one is through ge-
0/0/0
set routing-options static route 0.0.0.0/0 qualified-next-hop 1.4.0.1 metric 10
set routing-options static route 0.0.0.0/0 qualified-next-hop 1.2.1.1 metric 100
#Zones Definitions
set security zones security-zone Untrust interfaces ge-0/0/0.0 host-inbound-
traffic system-services dhcp
set security zones security-zone Untrust interfaces ge-4/0/0.0 host-inbound-
traffic system-services dhcp
set security zones security-zone Trust interfaces reth0.0

#Finally a permit all security policy  from Trust to Untrust zone
set security policies from-zone Trust to-zone Untrust policy ANY match source-
address any
set security policies from-zone Trust to-zone Untrust policy ANY match
destination-address any
set security policies from-zone Trust to-zone Untrust policy ANY match
application any
set security policies from-zone Trust to-zone Untrust policy ANY then permit
```

Active/Active Full Mesh

This scenario is found in medium to large deployments where secure routers are placed between two pairs of routers. OSPF is used to control the traffic flow through the nodes in the cluster, and JSRP is used to synchronize the sessions between the two nodes. Since asymmetric routing is supported, it is not required to force the traffic in both directions to a particular node. If a failure occurs and return traffic for a session arrives at a node different from the session creating node, the fab link will be used to send the traffic back to the node where sessions are active (this will be the node hosting the egress interface for that particular session).

This scenario benefits from the use of full mesh connectivity between the devices (thus improving the resiliency of the network), while eliminating the need to add extra switches in between the firewalls and routers, which reduces the points of failure in the network.



Physical View

Logical View

Figure 5: Active/active full mesh scenario

Special Consideration

The following design consideration should be taken into account when using the chassis cluster feature in Junos OS with enhanced services:

- Errors in either fab or fxp1 links (but not both) will cause the backup node to become disabled (single failure point). If a backup node detects errors in both fab and fxp1 links, it will become master (dual failure point).
- In the event of a control link failure, the system tries to avoid a dual mastership scenario by monitoring the fabric link. If hellos are received though this link, the secondary becomes disabled, while the primary remains active. If neither control link nor fabric link hellos are received, the backup node transitions to active.
- When a fabric link failure is detected, the nodes perform the split-brain avoidance procedure just like in the case of a control link failure. If the fabric link fails but the control link is still operational, the backup node will become disabled, thus avoiding a two master conflict.
- Failover times are in the order of a few seconds. A failure will be detected in three seconds or more (as the minimum hello time is 1000 ms, and the smallest threshold is three consecutive lost hellos).
- As of version 9.0, only a physical interface can be part of a RETH interface. Load balancing based on the use of logical interfaces (one active and one passive in each node) is not supported. Load balancing is supported, however, by using two physical interfaces in each node.
- Unified in-service software upgrade (ISSU) is not supported (please refer to the next section for a description of the upgrade procedure when using the HA feature).
- Chassis clustering does not support packet mode-based protocols (e.g., MPLS, Connectionless Network Service, and IPv6 are not supported).
- Pseudo interfaces are not supported when using the chassis cluster feature. The following services that require pseudo interfaces will not work in a cluster configuration:
  - Link services such as Multilink Point-to-Point Protocol (MLPPP), Multilink Frame Relay (MLFR), and compressed RTP (CRTP)
  - Generic routing encapsulation (GRE) tunnels
  - IP/IP tunnels
  - IPv4 multicast
  - Real-time performance monitoring (RPM)
  - Route-based IPsec tunnels
  - WAN interfaces are supported with the following exceptions:
    - CH-T1, ISDN, and xDSL
    - ISM 200 modules are not supported in HA mode*

Note: ISM modules are only supported on the J Series

Cluster Upgrade

Cluster upgrade is a simple procedure, but please note that a service disruption of about 3 to 5 minutes will occur during this priocess:

1. Load the new image file in node 0
2. Perform the image upgrade, without rebooting the node by typing “request system software add <image name>”
3. Load the new image file in node 1
4. Perform the image upgrade in node 1, as explained in step 2
5. Reboot both nodes simultaneously

Summary

The branch SRX Series services gateway and J Series chassis cluster is a simple to implement feature that ensures reliable enterprise connectivity between branch sites and corporate headquarters or regional offices. It provides stateful traffic failover between two Juniper routers while maintaining the abstraction of a single device, which simplifies network design. The feature has been carefully designed to address many common connectivity challenges such as asymmetric traffic, VPNs, and mixed LAN/WAN environments. Juniper Networks SRX Series for the branch and J Series Services Routers employing chassis cluster provide a foundation for reliable and high-performance network deployments.

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
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.

3500132-002-EN Jan 2010

♻ Printed on recycled paper