

J SERIES SERVICES ROUTERS' HIGH AVAILABILITY

Configuring and Deploying the J Series Chassis Cluster Feature

Table of Contents

Introduction	1
Scope.....	1
Design Considerations	1
Hardware Requirements	1
Software Requirements	1
Description and Deployment Scenario	1
Feature Description	1
Redundant Ethernet Interfaces.....	4
Feature Support and Comparison Matrix.....	5
Clustering Configuration	5
Cluster Monitoring	7
Deployment Scenarios	9
Active/Passive Cluster	9
Asymmetric Routing Scenario.....	11
Case I: Failures in the Trust Zone RETH	11
Case II: Failures in the Untrust Zone Interfaces.....	12
Active/Active Full Mesh	13
Special Considerations.....	13
Cluster Upgrade	14
Glossary	14
Summary	15
About Juniper Networks.....	15

Table of Figures

Figure 1: JUNOS Software forwarding model.....	2
Figure 2: Device Clustering.....	3
Figure 3: Active/passive scenario.....	9
Figure 4: Asymmetric routing scenario.....	11
Figure 5: Active/active full-mesh scenario.....	13

Introduction

IT organizations are under continuous pressure to secure their networks in order to meet demanding and ever changing application and service requirements. To keep pace with today's rapidly evolving landscape of internal and external threats, high-performance businesses need highly reliable and resilient security strategies at all layers of the network.

Juniper Networks® strategy is to continually integrate best-in-class services into Juniper Networks JUNOS® Software, enabling businesses to reduce the complexity, cost and risk associated with day-to-day network operations and new feature implementation. Starting with JUNOS version 9.0, Juniper Networks J Series Services Routers can be deployed using a chassis cluster feature that enables stateful device and services failover with an active/backup control plane and active/active data plane, all within a single system view, which provides high availability (HA) between devices and ensures business continuity by providing systems-level resiliency between branch sites and regional or corporate offices.

Scope

The purpose of this application note is to review the chassis cluster 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® 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 chassis cluster feature, enterprises can ensure connectivity in the event of device or link failure.

Hardware Requirements

- Two J Series routers per cluster (Juniper Networks J2320, J2350, J4350 and J6350 Services Routers)

Software Requirements

- JUNOS version 9.0 with enhanced services for J Series

Description and Deployment Scenario

Chassis cluster between devices may be deployed in either active/passive or active/active scenarios. JUNOS allows an HA cluster to also be used in asymmetric routing scenarios.

Throughout the document code examples are provided, and deployment scenarios are discussed towards the end of the paper.

Feature Description

The chassis cluster feature in J Series is modeled after redundancy features first introduced in Juniper Networks M Series Multiservice Edge Router and T Series Core Routers. We will first give a brief overview of how JUNOS redundancy works, so we can better understand how this model is applied when clustering devices.

As JUNOS is designed with separate control and data planes, redundancy must operate in both. The control plane in JUNOS 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 Software-forwarding paradigm is represented in Figure 1.

The JUNOS Redundancy Model

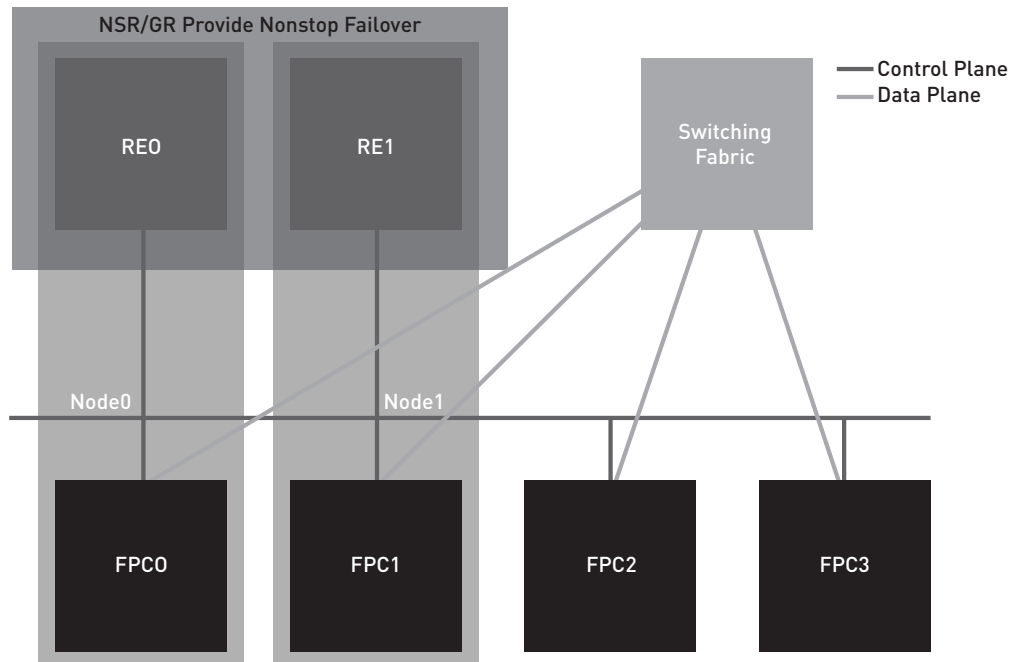


Figure 1: JUNOS Software forwarding model

Control plane failover is provided in JUNOS by using graceful restart or nonstop 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 it 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 the 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 and FPC failovers are handled by diverting traffic through other interfaces, which can be achieved by using conventional routing protocols, VRRP or aggregate interfaces.

When enabling chassis cluster for J Series routers, JUNOS uses a similar model—less the nonstop routing state replication—to provide control plane redundancy as shown in Figure 2.

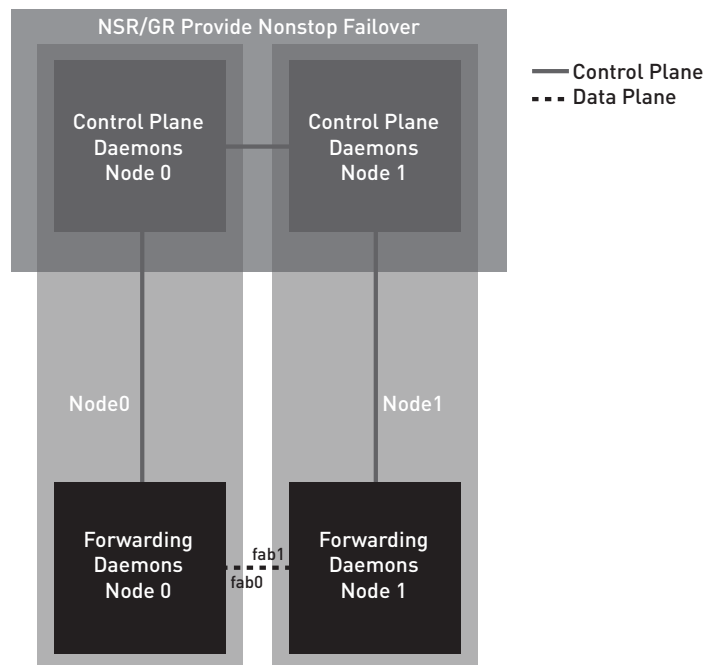


Figure 2: Device clustering

J Series HA 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 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 Gigabit Ethernet 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-5/0/0
J4350	7	ge-0/0/0	ge-7/0/0
J6350	7	ge-0/0/0	ge-7/0/0

After clustering is enabled, interfaces ge-0/0/2 and ge-0/0/3 are renamed to fxp0 and fxp1, respectively. As seen in Figure 2, fxp1 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 Gigabit 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 daemons required for managing clustering, most of the daemons are running only on the master RE. When the primary node fails, the routing daemon and other daemons 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, which was distributed by the active RE. The forwarding table allows each PFE to perform traffic forwarding independent of other system PFEs. On a PFE failure, the rest of the PFEs in the system are unaffected, allowing the control plane to reroute traffic to a non-failed PFE.

In contrast, J Series Services Routers 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's 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 initiate the 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 Software with enhanced services does not support the forwarding of Layer 2 traffic (transparent mode). Chassis cluster 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 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—reaches 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 VSI interfaces on ScreenOS devices. RETH interfaces, just like VSIs, share the same IP and MAC addresses between the different physical interfaces that are members of the VSI/RETH. The redundant interfaces send gratuitous 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 not supported in JUNOS.

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) 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 2: Feature Matrix

Feature	JSRP	NSRP
Session replication	Yes	Yes
ALG replication	Yes	Yes
NAT session replication	Yes	Yes
IPsec session replication (policy-based VPN)	Yes	Yes
IPsec session replication (route-based VPN)	Yes	Yes
Route synchronization	Yes (single system, always synchronized)	Yes
Interface monitoring	Yes	Yes
Zone monitoring	No	Yes
Track IP	No	Yes
Asymmetric routing	Yes	No
Graceful restart	Yes	No
Transparent 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 Services 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 IPs 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 are enabling chassis clustering on a pair of J2320 devices, node left and node right, which are connected back to back using interface ge-0/0/1 for the FAB link and ge-0/0/3 for the control link.

1. Log in to 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 management mode, not in config 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 config 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 node 1.

```
.....
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 config 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 (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 reboot the nodes.

```

set chassis cluster cluster-id 0 node 0

```

Cluster Monitoring

The following commands can be used to verify cluster operation and status. Statistics are not synchronized between the nodes in the cluster. When debugging clusters, it is useful to log in to each member node and analyze the output from each.

```

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

```

This command 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 is responsible for the control plane (it acts as the master RE and runs the control plane daemons such as rpd, kmd, dhcpcd, pppd and others).

```

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

```

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

```

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

```

This command displays the status of the control interface (FXP1) of this particular node.

```

show security flow session [node n]
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

```

This command displays sessions. Synchronized sessions will be seen in both nodes, where they 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 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: 0xffb0006, 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 (that is, 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 numbering 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 using several different topologies. This section provides the equivalent JUNOS configurations 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 controls all forwarding.

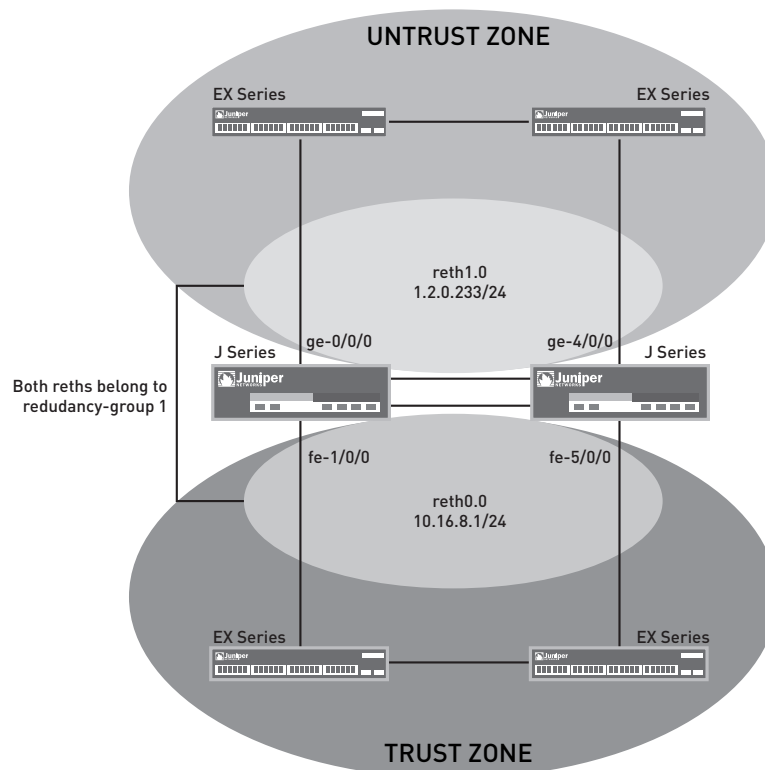


Figure 3: Active/passive scenario

Active/Passive can be achieved using RETH interfaces, similar to 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 over the fabric link, as only one node in the cluster will forward 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 is restored.

```

.....
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 #This command is optional
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 also monitor 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 ge-4/0/0 gigether-options redundant-parent reth1
set interfaces fe-1/0/0 fastether-options redundant-parent reth0
set interfaces fe-5/0/0 fastether-options redundant-parent reth0
set interfaces reth0 redundant-ether-options redundancy-group 1
set interfaces reth0 unit 0 family inet address 10.16.8.1/24
set interfaces reth1 redundant-ether-options redundancy-group 1
set interfaces reth1 unit 0 family inet address 1.2.0.233/24

#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
.....

```

```

#Finally an allow any from Trust to Untrust
.....

```

```

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

```

Asymmetric Routing Scenario

The scenario makes use of the asymmetric routing capability of chassis clustering. Traffic received by a node is matched against that node's session table. The result of this lookup determines if that node processes the session or forwards it to the other node over the FAB link. Sessions are anchored on the node where the first packet that created the session egresses, and the session is synced to the peer. If traffic is received on the node where the session is not anchored, then those packets are forwarded over the fabric link.

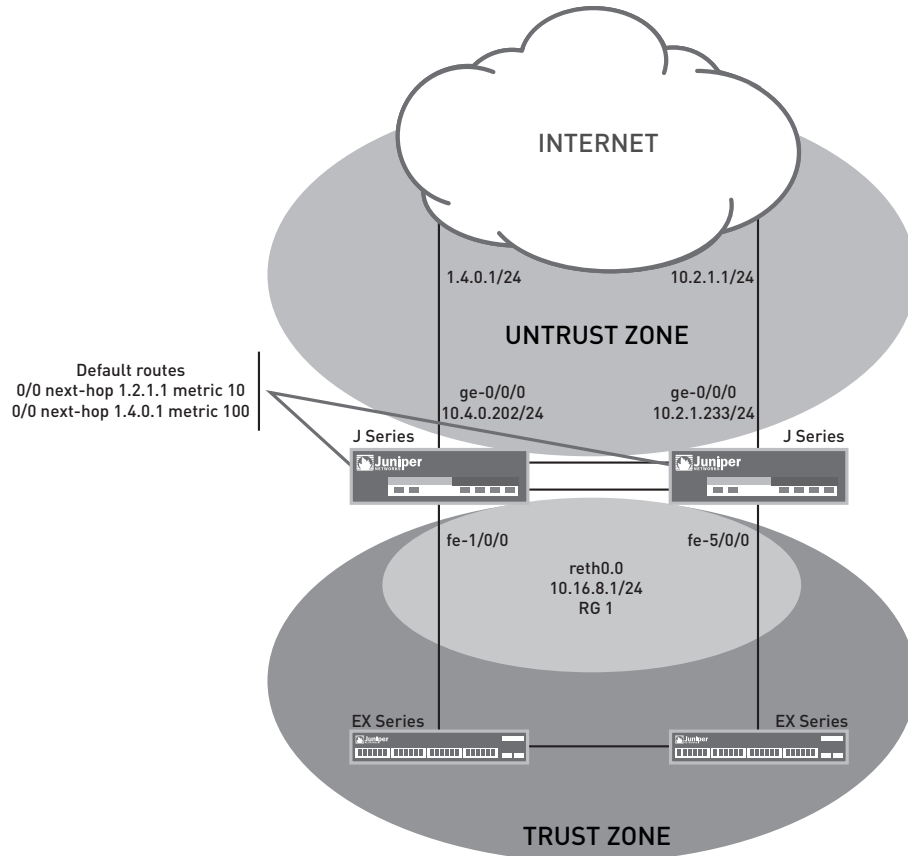


Figure 4: Asymmetric routing scenario

Figure 4 illustrates how asymmetric routing is supported. In this scenario two Internet connections are used with one being preferred. The connection to the Trust Zone is done 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 interface fe-1/0/0 belonging to reth0.0 on 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 these sessions is ge-0/0/0 belonging to node 0).

A failure in the fe-1/0/0 interface will trigger a failover of the redundancy group, causing the interface fe-5/0/0 (fe-1/0/0 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 on this node because the egress interface—ge-0/0/0—is hosted on 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 because the session is anchored to this node, and be sent to node 1 through the FAB interface where node 1 will forward it through the fe-5/0/0 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, under normal operating conditions traffic will be processed only by node 0. A failure of interface ge-0/0/0 connected to the Internet will cause a change in the routing table, which will have a default route after the failure pointing to interface ge-4/0/0 in node 1. After the failure, the sessions in node 0 will become inactive because the egress interface now will reside in node 1, and the passive sessions in node 1 will become active. Traffic arriving from the trust zone will still be received on interface fe-1/0/0, 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 ge-4/0/0 interface.

If this scenario were used with source NAT to accommodate different address spaces assigned by different providers, the aforementioned would not work, as the egress sessions would be NATTED differently after the failover. This limitation is not a consequence of the HA implementation, but due to the fact that if two ISPs are used, and the customer does not own a public address space, a failure in one of the ISPs will result in the loss of connectivity from all the 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 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 nde 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 an allow any from Trust to Untrust
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 services 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 to a particular node. If a failure occurs and return traffic for a session arrives asymmetrically, the FAB link will be used to send the traffic back to the node where sessions are active, which 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, which improves the resiliency of the network while eliminating the need to add extra switches in between the firewalls and routers and reducing the number of network failure points.

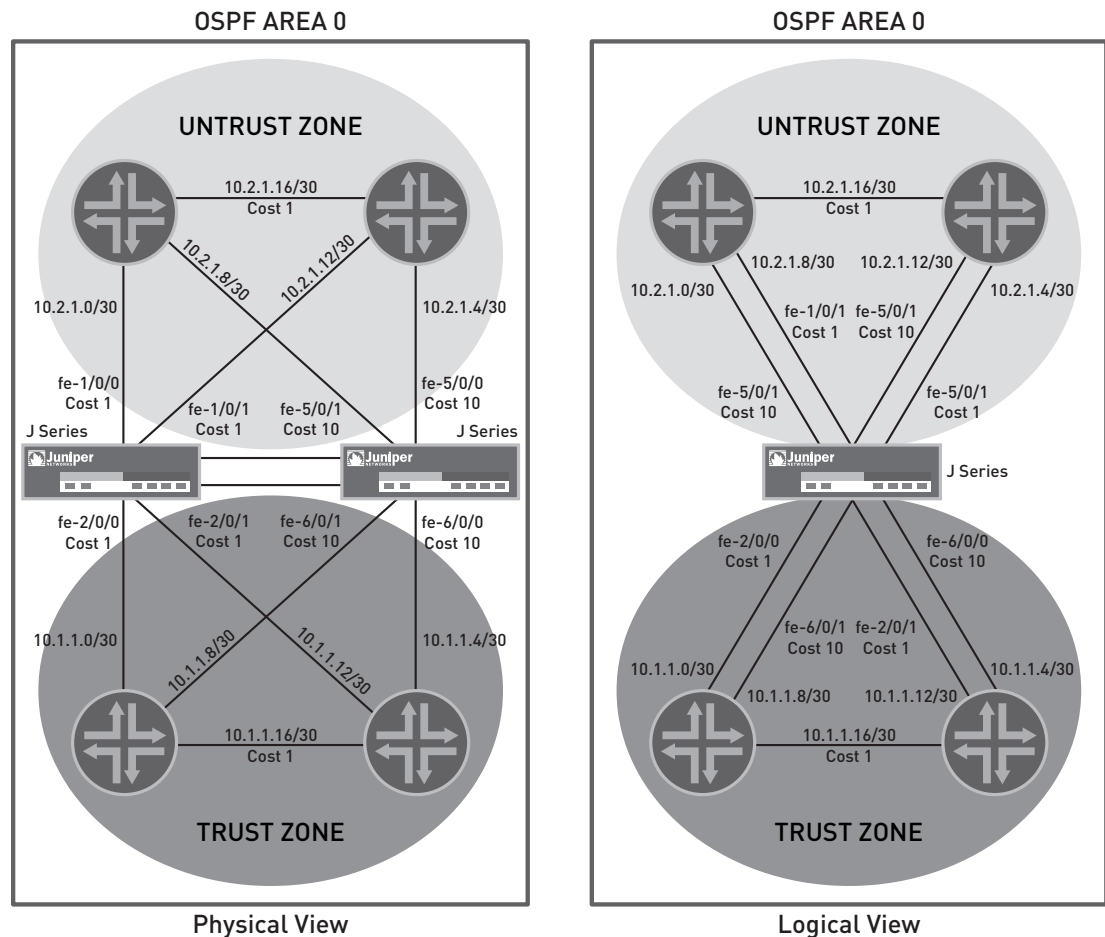


Figure 5: Active/active full-mesh scenario

Special Considerations

The following design considerations should be taken into account when using the chassis cluster feature in JUNOS Software 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 the 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 cluster attempts to avoid a dual mastership scenario by monitoring the fabric link. If hellos are received through 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 as 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 3 seconds or more as the minimum hello time is 1000ms, and the smallest threshold is 3 consecutive lost hellos.
- 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 cluster does not support protocols that require flow bypass mode.
- Juniper Networks ISM200 Integrated Services Modules are not supported with chassis cluster.

Cluster Upgrade

Cluster upgrade is a simple procedure, but please note that a service disruption of about three to five minutes will occur:

1. Load the new image file on 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 on node 1.
4. Perform the image upgrade in node 1, as explained in step 2.
5. Reboot both nodes simultaneously.

Glossary

FAB – Fabric Interface is an interface used to connect two devices in a cluster. The FAB interface is used both to send data traffic between the devices and to perform RTO synchronization.

FPC – Flexible PIC Concentrators in M Series and T Series routers are separate hardware components that house the PICs that connect the router to network media. In J Series routers, FPCs have been virtualized. The onboard Ethernet ports and each PIM slot correspond to a separate FPC.

fxp0 – This management interface is used for out-of-band network access to the router.

fxp1 – This control link is used for communications between the Routing Engines of nodes participating in a chassis cluster.

NSR – Nonstop routing is a mechanism by which redundant routing engines, or nodes in a chassis cluster, provide stateful routing failover that does not require any help from the rest of the network.

NSRP – NetScreen Redundancy Protocol is a Juniper proprietary protocol to implement high availability between Juniper Networks NetScreen Series Security Systems and Juniper Networks SSG Series Secure Services Gateway devices.

PIMs – Physical Interface Modules are used to connect the router to network media.

PFE – The Packet Forwarding Engine Packet processes packets; applies filters, routing policies and other features; and forwards packets to the next hop along the route to their final destination. In M Series and T Series routers, the PFE is implemented in custom ASICs. In J Series routers, the PFE has been virtualized in software.

RE – The Routing Engine provides three main functions:

- Creates the packet-forwarding switch fabric for the Services Router—providing route lookup, filtering and switching on incoming data packets, then directing outbound packets to the appropriate interface for transmission to the network
- Maintains the routing tables used by the router and controls the routing protocols that run on the router
- Provides control and monitoring functions for the router, including controlling power and monitoring system status

RETH – A Redundant Ethernet Interface consists of two interfaces, one from each cluster node, configured to provide redundancy.

VSI – Virtual System Interface is used to provide multiple connections between a firewall cluster and the rest of the network when using NSRP.

Summary

The 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 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

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

Copyright 2009 Juniper Networks, Inc. All rights reserved. Juniper Networks, the Juniper Networks logo, JUNOS, NetScreen, and ScreenOS are registered trademarks of Juniper Networks, Inc. in the United States and other countries. JUNOSe is a trademark of Juniper Networks, Inc. All other trademarks, service marks, registered marks, or registered service marks are the property of their respective owners. Juniper Networks assumes no responsibility for any inaccuracies in this document. Juniper Networks reserves the right to change, modify, transfer, or otherwise revise this publication without notice.

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

