IMPLEMENTING L2 AT THE DATA CENTER ACCESS LAYER ON JUNIPER NETWORKS INFRASTRUCTURE

Although Juniper Networks has attempted to provide accurate information in this guide, Juniper Networks does not warrant or guarantee the accuracy of the information provided herein. Third party product descriptions and related technical details provided in this document are for information purposes only and such products are not supported by Juniper Networks. All information provided in this guide is provided "as is", with all faults, and without warranty of any kind, either expressed or implied or statutory. Juniper Networks and its suppliers hereby disclaim all warranties related to this guide and the information contained herein, whether expressed or implied of statutory including, without limitation, those of merchantability, fitness for a particular purpose and noninfringement, or arising from a course of dealing, usage, or trade practice.

Copyright © 2009, Juniper Networks, Inc.

Table of Contents

Table of Figures

Copyright © 2009, Juniper Networks, Inc.

Introduction

This implementation guide explains how to implement a 2-tier LAN design in a single data center using the Juniper Networks® EX4200 Ethernet Switch with Virtual Chassis technology at the access tier, and interconnecting the access tier to the data center core network built on Juniper Networks MX Series Ethernet Services Routers. This 2-tier design delivers a fully fault tolerant network that supports a variety of server technologies applicable to the data center. By following the design principles of Juniper Networks L2 tier data center network, the traffic oversubscription is managed by the number of links that interconnect the access network tier to the core network, unlike traditional designs where the network oversubscription and performance remain dependant on device characteristics. No device oversubscription limitation will be imposed in this design, as Juniper Networks systems use a fully non-blocking architecture.

In conjunction with Juniper Networks overall data center solution, and as described in the *Enterprise Data Center Network Reference Architecture*, we identify the benefits of collapsing the data center tiers from the traditional 3-tier network design (access switching tier connects to aggregation switching that connects to core switching and routing devices) to a 2-tier design that eliminates the aggregation layer. Through the use of Juniper's virtual chassis technology at the access layer combined with high density 10 Gbps platforms like the MX Series, the technical need to implement an aggregation layer is eliminated. Service deployment can best be done at the core and, by reducing the number of access layer uplinks, the need for an aggregation layer to support the shear number of physical connections is eliminated.

Beyond the topic of how many tiers will comprise the data center network, network architects must consider where to implement Open Systems Interconnection (OSI) Model L3 routing, as opposed to where to expand OSI model L2 broadcast domains. Typical benefits for implementing L3 routing in the network infrastructure as close to the server as possible (access switching tier) include fault isolation and more predictable network traffic paths. Typical benefits of expanding L2 broadcast domain sizes relate to making them as large as possible, or pushing the L3 termination to the aggregation or core layers. This technique provides flexibility by locating servers in the network with minimal provisioning overhead.

In this document, we provide implementation guidelines for building a data center LAN constructed of two network tiers only (access and core) in which the IP addressing of network infrastructure is placed at the core network devices. Hence, the default gateway of the servers connected to the access switches reside on the core Ethernet services routers. In this design, L2 broadcast domains can span across multiple access elements (a single physical switch or virtual chassis acting as a single logical switch) to implement multi-node server cluster technologies that require L2 connectivity among the nodes participating in these clusters. Some examples of these technologies can be VMware live migration technology or "vmotion," Microsoft active clusters, and other High Performance Compute (HPC) clusters or grid computing applications.

The access switches are configured to provide full fault tolerance at the uplink level. Hence, there are at least two links connecting up to the core tier from each access switching element. Each link terminates at a different core device. With this design, a core device failure will not impact the availability of a server that connects to the access element. Additionally, we provide link-level redundancy using the connections from the access switching element to each core device over multiple, physical links that act as the link aggregation group (LAG). Needless to say, each access element supports local hosting of multiple VLANs and trunks all VLANs in the uplinks to the core network tier.

Scope

This document presents a best practice approach, demonstrating that the data center LAN can be built reliably. The implementation guidelines address a data center network that includes multiple, separate functional areas that may or may not be separated physically. Hence, we support the use case in which servers that reside in the same rack belong to different data center functional areas (or application tiers). In addition, this document addresses L2 and L3 considerations in implementing a data center LAN, with the access tier operating at L2 and the core tier operating at L3. Connectivity guidelines between the access layer and the core layer are provided.

Note: The following discussions are not covered in this document but will become available as separate implementation guides in the data center series.

- Design and implementation considerations for extending the data center core beyond a single location
- Building the data center access network
- Options for implementing the access connectivity to the core network.

Copyright © 2009, Juniper Networks, Inc.

This paper is structured to first present the data center design assumptions pertaining to physical and logical connectivity of the access and core data center networks. This description will include the connectivity requirements from external elements (e.g., servers) that connect. We will then present the implementation and configuration guidelines that should be followed to achieve a successful design. We will provide failed scenarios in relationship to the devices, specifically detailing the cause of the failure and the effects these failures have on traffic flows between server and client, and vice versa. Lastly, we offer best practices for operating the data center network using the MX Series and EX4200 platforms by describing device configuration and sequence of configuration/upgrade procedures.

Target Audience

This guide is intended for the following audiences:

- Data center network architects evaluating the feasibility of new approaches in network design
- Data center network engineers and operators designing and implementing new data center networks
- Technologists researching and analyzing new approaches for implementing flexible robust networks

Design Considerations

The following bulleted list describes the overall data center network environment in which we will implement L2 connectivity across multiple access elements. The list starts by describing the connectivity requirements at the server level, and walks through the network elements all the way to the network core. Additionally, the list illustrates some of the capabilities of a network implemented according to this guide. The main objective of this section is to provide clarity around the networking environment that was used in developing these implementation guidelines. This implementation may also work under other assumptions, but some corner cases may not be as well addressed.

- Servers connect to each access switch virtual chassis with a single network interface using a single IP address and no VLAN tags. Hence, each server port will be associated with a single L2 broadcast domain (VLAN). Because of this scheme, network interface card (NIC) teaming at the server level will not be covered.

Note: Each server may connect to two separate virtual chassis for high availability (HA) access.

- Multiple VLANs can be configured on each access element. These VLANs will be used by separate servers, and all are trunked from that access element to the core network tier.
- Each virtual chassis connects up to each single MX Series router using a LAG. Although the LAG consists of two link members, it is not limited to two link members and can have a maximum of eight physical link members per LAG.
- Each virtual chassis connects to two MX Series routers. These connections (LAG as previously stated) are identical trunk links that connect to interfaces on the MX Series, and are associated with the same bridge domains on the routers using the same VLAN IDs.
- The core network consists of two MX Series for redundancy purposes. These redundant routers are associated with the same networks and bridge domains.
- The core network tier includes a bridge domain that corresponds to each VLAN at the access tier.
- The core network connects to the network services tier (firewalls, load balancers, caching systems, and application acceleration devices) and manages all connectivity between the different server networks using routing or policies.
- Both core network MX Series routers connect to each other over all bridge domains using either pure L2 Ethernet 802.1q trunk or virtual private LAN service label-switched path (VPLS LSP) labels, keeping the different networks segmented and keeping the access tier VLANs interconnected. The implementation details of interconnecting the MX Series routers are not covered.
- Both of the core network tier MX Series routers will be configured to present the default gateway IP address to the servers with full HA by running Virtual Routing Redundancy Protocol (VRRP) across integrated routing and bridging (IRB) interfaces associated with bridge domains.

Copyright © 2009, Juniper Networks, Inc.

Note: The following diagram will be used as a reference throughout this document for the different configuration code examples, and also as a reference for network interface numbers and device names.

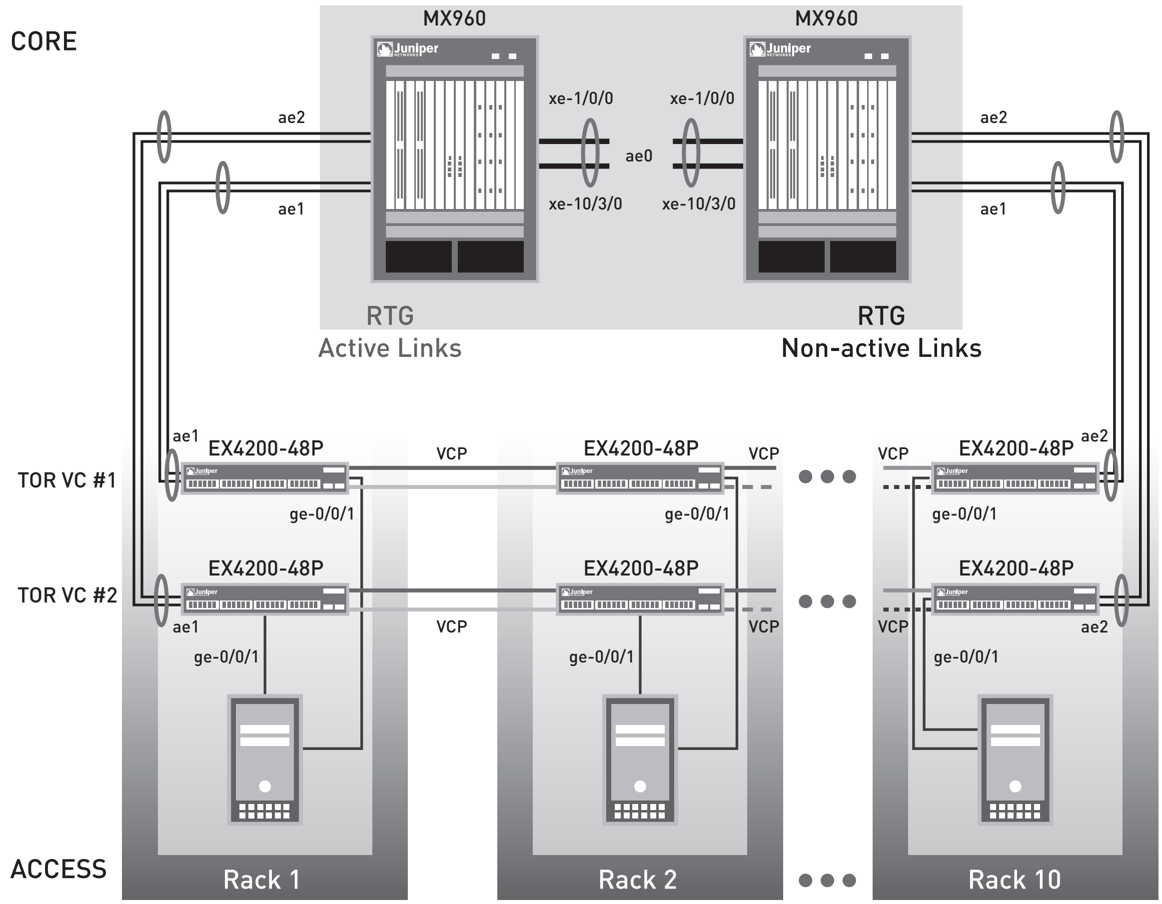

Figure 1: Data center showing EX4200 in Access Layer and MX Series in Core Layer

Implementation and Configuration Guidelines

This section presents solution implementation guidelines by describing device configurations for all associated network components or infrastructures. It also shows readers how to verify operation of the solution.

The primary implementation and configuration guidelines discussed in this section include the following:

- Connecting the access layer to the core layer over L2
- Configuring the core layer over L2
- Configuring L3 on the core MX Series Ethernet Services Routers

Connecting the Access Layer to the Core Layer over L2

The following list outlines the major steps for connecting the access layer to the core layer over L2:

- Connect access layer to core layer via 10 Gbps uplink
- Enable LAG bundling multiple uplinks (from different virtual chassis members)
- Enable dot1q trunking on the LAG
- Enable Redundant Trunk Group (RTG) on the LAG uplinks

Copyright © 2009, Juniper Networks, Inc.

Step 1—Connect the Access Layer to Core Layer via 10 Gbps Uplink

In this section, we describe the best practices for interconnecting the data center access layer EX4200 switches to the core layer MX Series using the 10 Gbps uplinks. On each rack, as described previously in the *Design Considerations* section, we suggest two EX4200 virtual chassis as top-of-rack switches where each virtual chassis connects to the data center core tier that runs MX Series routers.

We suggest that you make the redundant 10 Gbps uplink cable connections on each EX4200 virtual chassis on the **first** and **end** member switches, as illustrated in Figure 1. Juniper Networks takes this approach because these first and end member switches are *not* configured as the virtual chassis primary and backup switch members. This method avoids a simultaneous failover between the uplink and virtual chassis Routing Engine. The EX4200 virtual chassis design, in this case, is also referred to as a "braided-ring" connection (connects every other switch). This provides the lowest intra-virtual chassis and uplink latency using a maximum of 3 meters between member switches.

Note: Virtual chassis connection options in building the data center access network are not discussed, as mentioned in the Scope section of this paper.

In this scenario, the server typically connects to both EX4200 virtual chassis on the access layer through two NICs. Each interface card connects to a different top-of-rack virtual chassis as illustrated in the Figure 1. For further information concerning server configuration, code snippets, and related screenshots, refer to *Implementing VMware Server Virtualization* on Juniper Networks Infrastructure.

The key reason for this deployment is a redundant layout, in which each virtual chassis backs up the other. The access layer EX4200 virtual chassis' default mode of operation uses one of two uplink connections at any time, thereby distributing the traffic across 10 Gbps links based on the LAG and RTG configuration.

Juniper Networks JUNOS® Software implementation for the access port on EX4200 requires that VLAN and interfaces be associated on a one-to-one basis, while the trunk port on EX4200 only needs to define a range of VLAN members under the interfaces (see Step 3). A sample configuration for the access port connecting to the server is listed below:

```
{master}[edit]
root@EX4200# set vlans vlan71 vlan-id 71 interface ge-0/0/1.0  ## define vlan71 and assign
interface ##
root@EX4200# set interfaces ge-0/0/1 unit 0 family ethernet-switching port-mode access  ## set
interface mode ##
root@EX4200# set interfaces ge-0/0/1 unit 0 family ethernet-switching vlan members vlan71
## allow vlan on the access interface ##
root@EX4200# run show vlans vlan71  ## verify vlan71 is allowed on the interface ge-0/0/1.0 ##
```

Step 2—Enable LAG Using Multiple Uplinks from Different Virtual Chassis Members

The LAG increases bandwidth and provides link redundancy. You can create a maximum of 64 aggregated interfaces on EX4200 switches and you can group a maximum of eight physical links into each LAG interface, called the Aggregated Ethernet (AE) interface, in this case. You can create a maximum of 128 aggregated interfaces on the MX Series platforms.

Link aggregation also provides load-balanced link utilization, as hashing is done across the member links in a virtual bundle based on the L2/L3 header. The AE interfaces can be created statically or can use the Link Aggregation Control Protocol (LACP), as defined in the IEEE standard 802.3ad.

The physical ports in an AE virtual bundle interface are not required to be contiguous and can reside on different member switches within a virtual chassis. To properly form a virtual bundle, however, the LAG member ports in an AE interface are required to be of the same physical type, as well as the same speed and duplex.

AE interfaces must be configured correspondingly on both the access layer EX4200 switches and on the core layer MX Series routers. The JUNOS Software implementation of LAG does provide basic error-checking to prevent misconfiguration.

Copyright © 2009, Juniper Networks, Inc.

A sample configuration for enabling LAG with LACP on the EX4200 access switch is listed below:

```
{master}[edit]
root@EX4200# set chassis aggregated-devices ethernet device-count  64 ## set the total number
of AE ports ##
root@EX4200# set interfaces ae1 aggregated-ether-options minimum-links 1 ## define min. link to
keep AE port up ##
root@EX4200# set interfaces ae1 aggregated-ether-options link-speed 10g ## define bandwidth of
LAG member port ##
root@EX4200# set interfaces ae1 aggregated-ether-options lacp active ## activate LACP protocol
on LAG ports ##
root@EX4200# set interfaces xe-0/1/0 ether-options 802.3ad ae1 ## join the LAG on physical
interface ##
root@EX4200# set interfaces xe-1/1/0 ether-options 802.3ad ae1 ## join the LAG on physical
interface ##
root@EX4200# run show interface terse | match ae ## verify that AE ports are defined ##
root@EX4200# run show lacp interfaces ## verify LAG contains correct physical ports when AE
interface is up ##
```

Step 3—Enable dot1q Trunking on the LAG

On the EX4200 switch, ports that are assigned to a VLAN can be configured as either access or trunk ports. To identify which VLAN the traffic belongs to when traversing across trunk ports, all frames on an Ethernet VLAN are identified by a tag, as defined in the IEEE 802.1Q standard. These frames are tagged and encapsulated with 802.1Q tags, sometimes referred to simply as dot1q tags. A dot1q trunk interface is typically used to interconnect switches to one another. Trunk ports use the tag to multiplex traffic among a number of VLANs onto a single link (logical or physical). EX4200 switches support a maximum of 4096 VLANs. Vlan-id number **0** and **4095** are reserved by JUNOS.

Unlike the access port configuration where VLANs and interfaces must be associated on a one-to-one basis as described in Step 1, the dot1q trunk port on the EX4200 switch only needs to list the VLAN members in the interface configuration. This approach requires a minimum number of command-line interface (CLI) command lines when configuring the dot1q trunk ports on EX4200 switches, as a dot1q trunk port can literally allow more than 4,000 VLANs in an interface.

A sample configuration for enabling a dot1q trunk on a LAG interface is listed below:

```
{master}[edit]
root@EX4200# set interfaces ae1 unit 0 family ethernet-switching port-mode trunk ## set
interface mode to trunk ##
root@EX4200# set interfaces ae1 unit 0 family ethernet-switching vlan members [ 71-73 573 ] ##
define allowed vlans ##
root@EX4200# run show vlans ## verify all the vlans are allowed in trunk interface ae1.0 ##
```

Step 4—Enable RTG on the LAG Uplinks

A Redundant Trunk Group is an L2 link failover mechanism supported on the EX4200 switches. It is ideally implemented on an access switch or virtual chassis with a dual home connection, where one link is active and forwards traffic while the other blocks traffic and acts as a backup to the active link. This feature eliminates configuring Spanning Tree Protocol (STP) on the switch if RTG previously has been enabled.

In a typical enterprise network comprised of core and access layers, a redundant trunk link provides a simple solution for network recovery when a trunk port goes down. With a redundant trunk link, traffic is routed to another trunk port, keeping network convergence time to a minimum. You can configure a maximum of 16 redundant trunk groups on a standalone switch or on a virtual chassis.

With RTG, STP can be replaced in the access layer. RTG and STP are mutually exclusive on a given port. Because Rapid Spanning Tree Protocol (RSTP) is enabled by default on EX4200 switches to create a loop-free topology, trunk links that are placed in a redundant group cannot be part of an STP topology. Otherwise, the following error would display, and as a result, the configuration changes would not take effect.

Copyright © 2009, Juniper Networks, Inc.

```
error: XSTP : msti 0 STP and RTG cannot be enabled on the same interface ae1
error: configuration check-out failed
```

Although the JUNOS Software for EX4200 switches does not allow an interface to be in a redundant trunk group and in an STP topology at the same time, STP can continue operating in other parts of the network. At the core layer, you can enable STP to minimize human errors, regardless of whether or not RTG is configured at the access layer. On the access layer, switches where certain uplinks are configured for RTG, the other uplink ports, if any, and all of its downlink ports can still be enabled for STP to prevent any accidental L2 loops in the network.

A sample configuration for disabling STP on the LAG uplink interfaces is listed as below:

```
{master}[edit]
root@EX4200# set protocols rstp interface ae1 disable ## specifically disable STP on interface
defined in RTG ##
root@EX4200# set protocols rstp interface ae2 disable ## specifically disable STP on interface
defined in RTG ##
root@EX4200# run show spanning-tree interface ## verify STP state on AE1 and AE2 interfaces is
DIS (disabled) ##
```

A Redundant Trunk Group configuration also requires both active and non-active links to be trunking the same VLAN members. The following error will prevent the RTG configuration from taking effect if a VLAN mismatch occurs on the active and non-active links in an RTG group:

```
error: RTG : grp DC_RTG primary ae1.0 secondary ae2.0 vlan mismatch vlan573
error: configuration check-out failed
```

To configure a redundant trunk link, first create an RTG. You configure an RTG on the access switch which contains two links: a primary (active) link and a secondary link. If the active link fails, the secondary link automatically starts forwarding data traffic without waiting for normal STP convergence. RTG does not need to be enabled on both the access and core layers.

Data traffic is forwarded only on the active link. Data traffic on the secondary link is dropped and shown as dropped packets when you issue the operational mode command show interfaces interface-name extensive. While data traffic is blocked on the secondary link, L2 control traffic is still permitted. For example, a Link Layer Discovery Protocol (LLDP) session can be run between two EX4200 switches on the secondary link.

A sample configuration for enabling RTG on the LAG uplink interfaces is listed below:

```
{master}[edit]
root@EX4200# set ethernet-switching-options redundant-trunk-group group DC_RTG interface ae1 ##
define RTG ##
root@EX4200# set ethernet-switching-options redundant-trunk-group group DC_RTG interface ae2
root@EX4200# run show redundant-trunk-group ## verify RTG interfaces are correctly defined ##
```

Configuring L2 Bridging on the Core Layer MX Series Ethernet Services Routers

The following list outlines the major steps for configuring the core layer over L2:

- Connect the same VLAN over IRB and multiple physical interfaces
- Configure LAG ports on the MX Series
- Configure virtual router routing instances using IRB interfaces
- Support RTG (RSTP and dot1q trunk on LAG ports between MX Series routers)

Copyright © 2009, Juniper Networks, Inc.

To Perform the Bridge Domain Setup, Connect the Same VLAN over IRB and Multiple Physical Interfaces

Bridge domain (MX Series routers only) is a domain that includes a set of logical interfaces that share the same flooding or broadcast characteristics to perform L2 bridging. (For a physical interface device to function in JUNOS, you must configure at least one logical interface on that device. For each logical interface, you must specify the protocol family that the interface supports.) A bridge domain must include a set of logical interfaces that participate in L2 learning and forwarding. By default, each bridge domain maintains an L2 forwarding database that contains media access control (MAC) addresses learned from packets received on the ports belonging to the bridge domain.

Bridging operates at L2 of the OSI reference model, while routing operates at L3. A set of logical ports configured for bridging can be said to constitute a bridging domain.

Integrated routing and bridging is a routed interface configured for a specific bridge domain that provides simultaneous support for L2 bridging and L3 routing on the same interface. IRB enables you to route packets to another routed interface or to another bridge domain that has an IRB interface configured. IRB performs the following functions:

- Routes a packet if the destination MAC address is the MAC address of the router and the packet ethertype is IPv4, IPv6, or MPLS
- Switches all multicast and broadcast packets within a bridging domain at L2
- Routes a copy of the packet if the destination MAC address is a multicast address and the ethertype is
- IPv4 or IPv6
- Switches all other unicast packets at L2
- Handles supported L2 control packets such as STP and LACP
- Handles supported L3 control packets such as OSPF and RIP

The following diagram illustrates the logical layout of the IRB and bridge domains at the core layer MX Series routers.

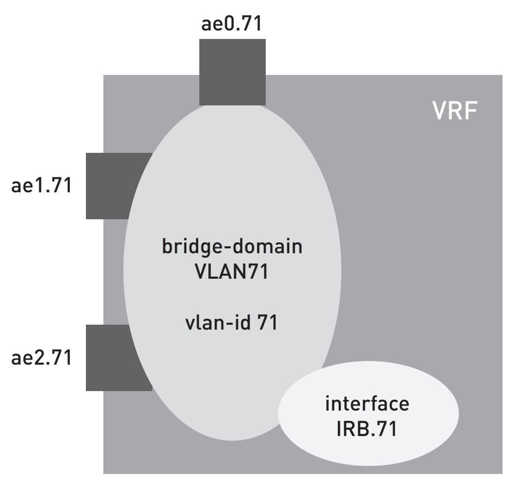

Figure 2: Logical layout of IRB and bridge domains at core layer MX Series router

Copyright © 2009, Juniper Networks, Inc.

A bridging domain can be created by configuring a routing instance and by specifying the instance-type as bridge. A sample configuration to set up bridge domains for the same VLAN over multiple physical interfaces at the core layer MX Series routers is shown below:

```
{master}[edit]
root@MX960# set bridge-domains VLAN71 domain-type bridge ## define VLAN71 as a bridge-domain on
MX Series ##
root@MX960# set bridge-domains VLAN71 vlan-id 71 ## assign vlan-id 71 to bridge-domain VLAN71
##
root@MX960# set bridge-domains VLAN71 interface ae0.71 ## assign interface to bridge-domain
VLAN71 ##
root@MX960# set bridge-domains VLAN71 interface ae1.71 ## assign interface to bridge-domain
VLAN71 ##
root@MX960# set bridge-domains VLAN71 interface ae2.71 ## assign interface to bridge-domain
VLAN71 ##
root@MX960# run show bridge domain interface ae0.71 ## verify VLAN71 is allowed on correct
interfaces ##
```

Configure LAG Ports on the MX Series

The LAG configuration at the core layer MX Series router is exactly the same as the EX4200 access layer. The following configuration shows how to configure a LAG port at the core layer MX Series routers.

```
{master}[edit]
root@MX960# set chassis aggregated-devices ethernet device-count 128 ## define total number of
LAG ports on MX ##
root@MX960# set interfaces ae0 aggregated-ether-options minimum-links 1 ## set min link to keep
AE interface up ##
root@MX960# set interfaces ae0 aggregated-ether-options link-speed 10g ## define bandwidth of
LAG member ports ##
root@MX960# set interfaces ae0 aggregated-ether-options lacp active ## enable LACP protocol on
LAG interface ##
root@MX960# set interfaces xe-1/0/0 gigether-options 802.3ad ae0 ## join the LAG port on
physical interface ##
root@MX960# set interfaces xe-10/3/0 gigether-options 802.3ad ae0 ## join the LAG port on
physical interface ##
root@MX960# run show lacp interfaces ## verify LAG has the correct physical ports when AE
interface is up ##
```

A sample configuration for enabling dot1q trunk on the LAG interface between the core layer MX Series routers is listed below.

For JUNOS code prior to 9.2 release:

```
{master}[edit]
root@MX960# set interfaces ae0 flexible-vlan-tagging ## set flexible vlan tagging on LAG port ##
root@MX960# set interfaces ae0 encapsulation flexible-ethernet-services ## set encapsulation on
LAG port ##
root@MX960# set interfaces ae0 unit 71 encapsulation vlan-bridge ## set encapsulation for sub-
interface ##
root@MX960# set interfaces ae0 unit 71 vlan-id 71 ## allow vlan-id 71 traffic on sub-interface
##
root@MX960# run show bridge domain interface ae0.71 ## verify correct vlans are allowed on ae0
trunk interface ##
```


Copyright © 2009, Juniper Networks, Inc.

For JUNOS code after 9.2 release:

```
{master}[edit]
root@MX960# set interface ae0 unit 0 family bridge interface-mode trunk ## set interface mode
to trunk ##
root@MX960# set interface ae0 unit 0 family bridge vlan-id-list 71-73 ## allow vlan range on
LAG port ##
root@MX960# set interface ae0 unit 0 family bridge vlan-id-list 573 ## allow a vlan on LAG port
##
root@MX960# run show bridge domain interface ae0.0 ## verify vlans are allowed on ae0.0 trunk
interface ##
```

Configure Virtual Router Routing Instances Using IRB Interfaces

A virtual router routing instance, like a virtual routing and forwarding (VRF) routing instance, maintains separate routing and forwarding tables for each instance. However, many of the configuration steps required for VRF routing instances are not required for virtual router routing instances. Specifically, you do not need to configure a route distinguisher, a routing table policy (vrf-export, vrf-import, and route-distinguisher statements), or MPLS between the service provider routers.

A sample configuration for setting up virtual router routing instances using IRB interfaces is listed below.

```
{master}[edit]
root@MX960# set routing-instances VMotionTest instance-type virtual-router ## define a virtual
router on MX ##
root@MX960# set routing-instances VMotionTest interface irb.71 ## assign interface to the
virtual router ##
root@MX960# set routing-instances VMotionTest interface irb.72 ## assign interface to the
virtual router ##
root@MX960# run show route instance VMotionTest detail ## verify virtual router is correctly
defined ##
```

Configuration Required for Supporting RTG (Enable RSTP on LAG Ports Between MX Series Routers)

The traditional approach for preventing L2 loops, while providing device redundancy at L2, requires STP. Both the EX4200 switches and MX Series routers support the three standard versions of STP: 802.1d STP, 802.1w Rapid STP, and 802.1s Multiple-instance STP.

Note: STP, RSTP, and Multiple Spanning Tree Protocol (MSTP) are also interoperable with Cisco's PVST+/RPVST+.

When you enable RTG at the access layer EX4200 switches, RSTP, which is the default version of Spanning Tree on the MX Series, is implemented between the core layer MX Series routers to prevent L2 loops caused by human error.

A sample configuration for enabling RSTP on the LAG interface between the core layer MX Series routers is listed below.

```
{master}[edit]
root@MX960# set protocols rstp bridge-priority 4k ## set spanning-tree bridge priority to MX ##
root@MX960# set protocols rstp interface ae0 mode point-to-point ## enable STP on interface and
set mode to p2p ##
root@MX960# run show spanning-tree bridge ## verify root bridge is the correct MX Ethernet
services router by bridge ID ##
root@MX960# run show spanning-tree interface ## verify port is in correct STP state ##
```

Note: The STP bridge priority can be set only in increments of **4,096**. The range is from 0 through 61,440. The default value is **32,768**.

Copyright © 2009, Juniper Networks, Inc.

Configuring L3 on the Core MX Series

The following list outlines the major steps for configuring L3 on the core MX Series routers:

- Review the IRB and VRF configuration processes
- Configure VRRP between the MX Series

Step 1—Review the IRB's and VRF's Configuration Process

IRB on the core layer MX Series provides simultaneous support for L2 bridging and L3 IP routing on the same interface. IRB enables an operator to route local packets to another routed interface or to another bridging domain that has an L3 protocol configured.

The EX4200 switches also support routed interfaces called Routed VLAN Interfaces (RVIs). Because this implementation guide only discusses L2 at the access layer, RVI is out of the scope of this paper and will not be discussed. As opposed to IRB which routes bridge domains, RVI routes VLAN.

A sample configuration for enabling L3 IRB interface on the core layer MX Series routers is listed below.

```
{master}[edit]
root@MX960# set bridge-domains VLAN71 domain-type bridge
root@MX960# set bridge-domains VLAN71 vlan-id 71
root@MX960# set bridge-domains VLAN71 routing-interface irb.71 ## define L3 IRB routing
interface in VLAN71 ##
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.2/24 ## assign IP address
to the defined IRB interface ##
root@MX960# run show interface irb.71 terse ## verify IRB interface is correctly defined ##
```

Step 2—Configure VRRP between the MX Series

Both the MX Series and EX4200 platforms support VRRP. With VRRP, routers viewed as a redundancy group share the responsibility for forwarding packets as if they owned the IP address corresponding to the default gateway configured on the hosts. At any time, one of the VRRP routers acts as the primary while the other VRRP routers act as secondary (backup) routers. If the primary router fails, a backup router becomes the new primary router. This ensures that router redundancy is always provided, allowing traffic on the LAN to be routed without relying on a single router.

A sample configuration for enabling VRRP on an IRB interface is listed below:

On primary MX Series device:

```
{master}[edit]
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.2/24 vrrp-group 1 virtual-
address 172.16.56.1
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.2/24 vrrp-group 1 priority
190
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.2/24 vrrp-group 1 preempt
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.2/24 vrrp-group 1 accept-
data
root@MX960# run show vrrp interface irb.71 ## verify VRRP is correctly defined ##
```

On secondary MX Series device:

```
{master}[edit]
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.3/24 vrrp-group 1 virtual-
address 172.16.56.1
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.3/24 vrrp-group 1 priority
180
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.3/24 vrrp-group 1 preempt
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.3/24 vrrp-group 1 accept-
data
root@MX960# run show vrrp interface irb.71 ## verify VRRP is correctly defined ##
```

 Copyright © 2009, Juniper Networks, Inc.

Failover Scenario Overview

This section presents failover scenarios that may occur, and in particular describes the cause of the failover and the effects the failovers have upon northbound and southbound traffic flows.

- Access switch/virtual chassis failure
- Active link from access switch/virtual chassis to MX Series failure
- Active MX Series device failure

To increase network uptime, we use redundant links to interconnect access layer switches to the core layer, and to connect dual-homed servers to the access switches in the data center. For path redundancy and fast failover at L2, switches that support 802.3ad link aggregation and the RSTP (802.1w) provide faster recovery from a link failure than the original STP. The EX4200 access switch also offers the RTG feature as a faster, easier-to-implement alternative to the STP. There are three failover scenarios that we would like to discuss in the following sections. We discuss these failover scenarios for both the northbound (server-to-client) and southbound (client-to-server) traffic flows.

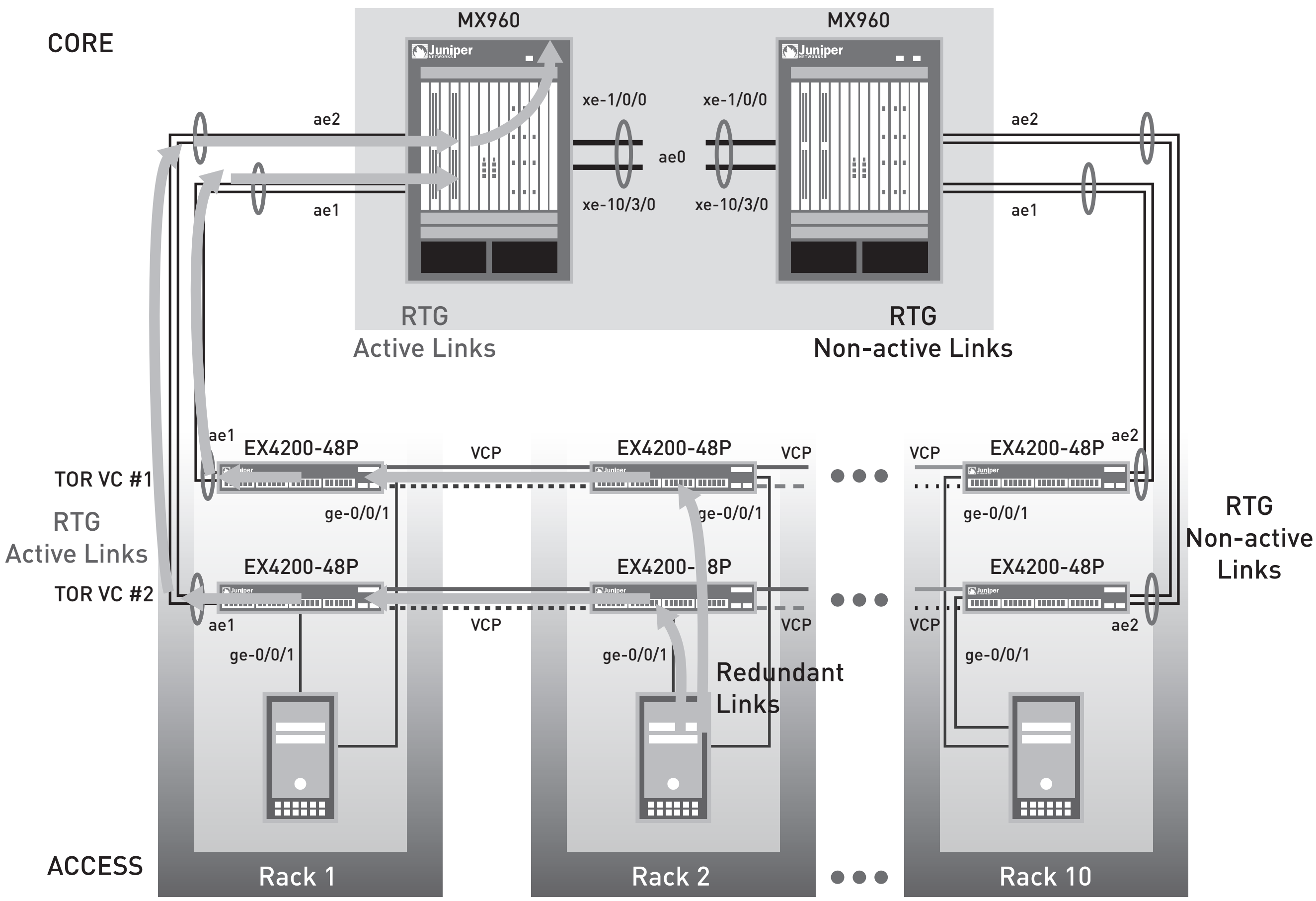

Figure 3: Traffic flow in normal status (northbound from server to client)

Copyright © 2009, Juniper Networks, Inc.

Access Switch/Virtual Chassis Failure

Northbound Traffic Flows from Server to Client

As suggested in Step 1 of *Connecting the Access Layer to the Core Layer over L2*, we connect the server to the access layer switches through two redundant links. When both links are up, the server's operational default mode uses one of the links at any time, thereby distributing the traffic across the active link based on redundant links and port group configuration on the server side.

When one of the directly-connected access switches fails, the northbound traffic from server to client takes the alternative link to the redundant access switch virtual chassis, triggered when the active link connection between the server and the access switch goes down. It then flows through the active uplink on the redundant virtual chassis to the core layer. The traffic is then routed out from the data center network to the client end from the core layer MX Series routers. The traffic flow, in case of an access switch virtual chassis failure, follows the exact same path as the directly-connected access switch failure in this scenario.

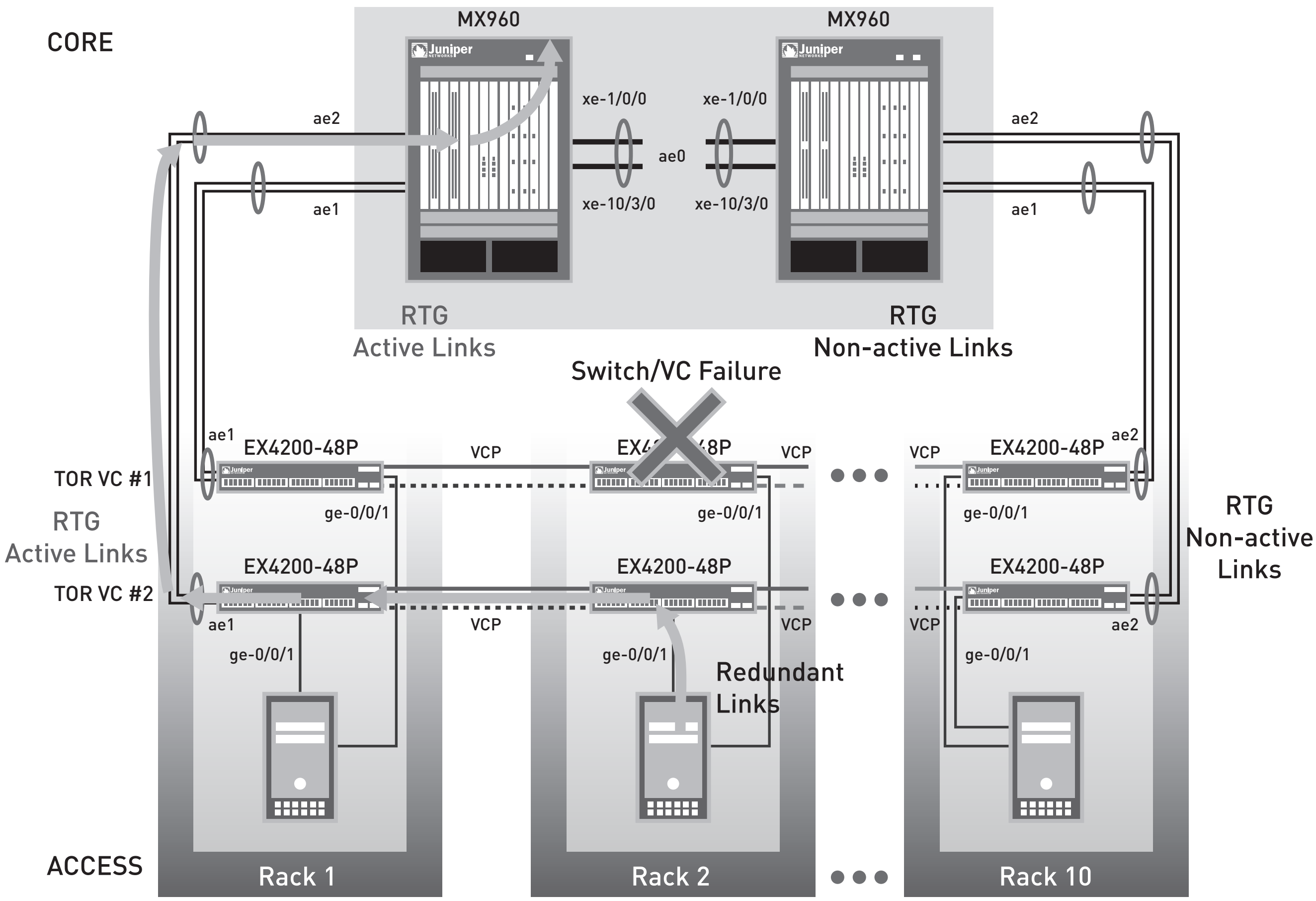

Figure 4: Traffic flow in server's directly-connected Access Switches/Virtual Chassis failure

Southbound Traffic Flow from Client to Server

During normal operation, the southbound traffic from client to server flows over the active RTG LAG connection to the access layer EX4200 virtual chassis. When one of the directly-connected access switches fails, the remaining active switch/virtual chassis becomes active because the connection between the server and the remaining access switch is now the only link in up/up status, The traffic from client to server takes the remaining active LAG connection to the redundant access switch virtual chassis and then flows through the server-access link to the server. The traffic flow in a virtual chassis failure follows the exact same path as the directly-connected access switch failure in this scenario.

Copyright © 2009, Juniper Networks, Inc.

Active Link from Access Switch/Virtual Chassis to MX Series Failure

Northbound Traffic Flow from Server to Client

This failover scenario assumes that the 10 Gbps uplinks interconnecting the access and core layers are configured as RTG on LAG interfaces. In case of a physical member link failure in the active LAG virtual bundle, the northbound traffic flowing on this link is spread to the other member links within this LAG. Up to eight 10 Gbps uplinks can be configured in a LAG interface. This provides nonstop redundancy and minimizes downtime caused by the physical link failure.

In the rare scenario where the active LAG interface is down (caused by either the failure of all physical member links in this LAG connection, or the link failure exceeds the defined LAG threshold), the non-active LAG connection between this access layer EX4200 virtual chassis and the core layer MX Series router will become active, triggered by RTG setup on access switch/virtual chassis, in a relatively faster convergent time than STP. At this point, the traffic will start flowing on this secondary LAG link to the core layer.

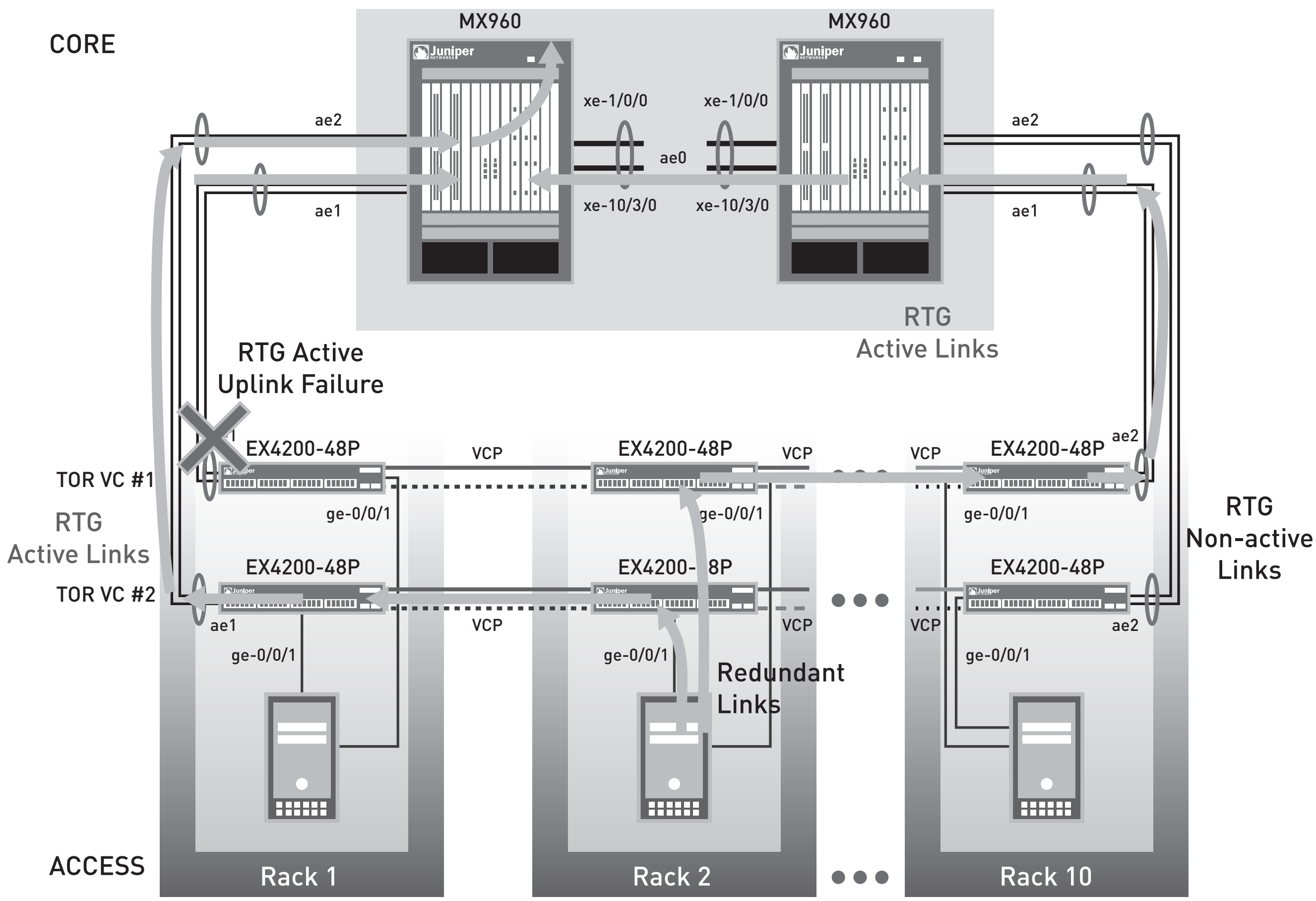

Figure 5: Traffic flow when the RTG Active Link between Access Switches/Virtual Chassis and MX Series fails

Southbound Traffic Flow from Client to Server

During normal operation, the southbound traffic from client to server flows over two active RTG LAG connections to the access layer EX4200 virtual chassis. In a failure scenario where one of the member links in the active LAG interface fails, the traffic from client to server takes the remaining member links in this LAG connection to the access switch virtual chassis, as defined in link aggregation groups, and then flows through the server-access links to the server.

When an active AE interface is down on the access layer EX4200 virtual chassis and the core layer MX Series router, the secondary AE interface becomes active, triggered by RTG setup on access switch/virtual chassis, and starts carrying the routed client traffic from the core layer MX Series router to the server side.

Copyright © 2009, Juniper Networks, Inc.

Active MX Series Device Failure

Northbound Traffic Flow from Server to Client

When one of the core layer MX Series routers fails, the RTG LAG interfaces on this MX Series router becomes unavailable and the RTG LAG interfaces on the other MX Series router becomes active, triggered by RTG setup on access switch/virtual chassis. The northbound traffic from server to client takes the redundant links to the access layer EX4200 virtual chassis and then flows through the active uplinks on both the virtual chassis to the core layer EX4200. The traffic is then routed out from the data center network to the client end from the remaining core layer MX Series router.

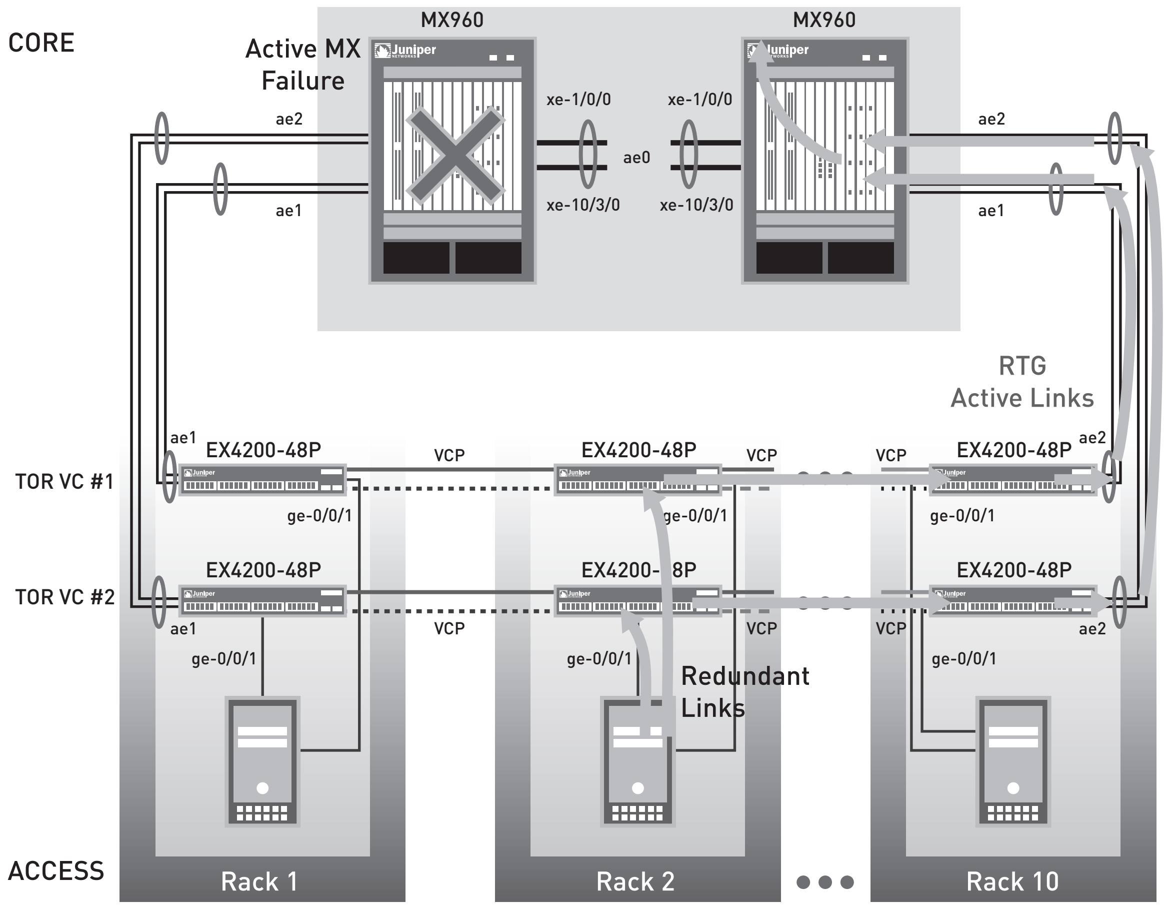

Figure 6: Traffic flow when the active MX Series fails

Southbound Traffic Flow from Client to Server

During normal operation, the southbound traffic from client to server is routed on the active MX Series and then flows over the active RTG LAG connection to the access layer EX4200 virtual chassis. When one of the core layer MX Series routers fails, the traffic from client to server is routed only on the remaining MX Series router as the connections from the remaining MX Series become active, triggered by RTG setup on access switch/virtual chassis. The traffic then flows over the active LAG connection to the access switch virtual chassis, and the client-to-server traffic flows through the server-access links to the server. The traffic flow in a virtual chassis failure follows exactly the same path as the directly-connected access switch failure in this scenario.

Copyright © 2009, Juniper Networks, Inc.

Operational Best Practices

This section provides best practices for operating the data center network using the MX Series and EX4200 platforms, by describing device configuration and sequence of configuration/upgrade procedures.

Carving New VLAN (End-to-End)

Network devices are connected in a data center LAN to provide sharing of common resources such as Domain Name System (DNS) and file servers to enable clients to connect to the servers through external networks. Without bridging and VLANs, all devices on the Ethernet LAN are in a single broadcast domain, and all of the devices detect all of the packets on the LAN. Bridging separates broadcast domains on the LAN by creating VLANs, which are independent logical networks that group together related devices into separate network segments. The grouping of devices on a VLAN is independent of where devices are physically located in the LAN.

Consider the data center network, as illustrated in Figure 1. To connect network devices to a data center LAN which is running EX4200 switches in the access layer and MX Series routers at the core layer, you must, at a minimum, configure bridging and VLANs on both the EX4200 and MX Series devices by following these steps:

1. Create new VLAN, for example vlan71, and associate this VLAN with the access ports connecting to the server on both of the EX4200 top-of-rack virtual chassis.

```
{master}[edit]
root@EX4200# set vlans vlan71 vlan-id 71 interface ge-0/0/1.0
root@EX4200# set interfaces ge-0/0/1 unit 0 family ethernet-switching port-mode access
root@EX4200# set interfaces ge-0/0/1 unit 0 family ethernet-switching vlan members vlan71
```

2. Add this new VLAN into all four LAG uplinks that connect to the core layer MX Series on both EX4200 top-of-rack virtual chassis.

```
{master}[edit]
root@EX4200# set interfaces ae1 unit 0 family ethernet-switching port-mode trunk
root@EX4200# set interfaces ae1 unit 0 family ethernet-switching vlan members vlan71
root@EX4200# set interfaces ae2 unit 0 family ethernet-switching port-mode trunk
root@EX4200# set interfaces ae2 unit 0 family ethernet-switching vlan members vlan71
```

3. Create the bridge domain for this new VLAN on the core layer MX Series routers.

```
{master}[edit]
root@MX960# set routing-instances ServicesSwitch bridge-domains VLAN71 domain-type bridge
root@MX960# set routing-instances ServicesSwitch bridge-domains VLAN71 vlan-id 71
root@MX960# set routing-instances ServicesSwitch bridge-domains VLAN71 interface ae0.71
root@MX960# set routing-instances ServicesSwitch bridge-domains VLAN71 interface ae1.71
root@MX960# set routing-instances ServicesSwitch bridge-domains VLAN71 interface ae2.71
```

4. Add this new VLAN into all LAG trunks on the core layer MX Series, connecting to both the access layer EX4200 top-of-rack virtual chassis and the redundant core layer MX Series routers.

For JUNOS code prior to 9.2 release:

```
{master}[edit]
root@MX960# set interfaces ae0 unit 71 encapsulation vlan-bridge
root@MX960# set interfaces ae0 unit 71 vlan-id 71
root@MX960# set interfaces ae1 unit 71 encapsulation vlan-bridge
root@MX960# set interfaces ae1 unit 71 vlan-id 71
root@MX960# set interfaces ae2 unit 71 encapsulation vlan-bridge
root@MX960# set interfaces ae2 unit 71 vlan-id 71
```

Copyright © 2009, Juniper Networks, Inc.

For JUNOS code after 9.2 release:

```
{master}[edit]
root@MX960# set interface ae0 unit 0 family bridge vlan-id-list 71
```

5. If necessary, create IRB interfaces with VRRP support for this new VLAN on both of the core layer MX Series routers.

On the primary MX Series router:

```
{master}[edit]
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.2/24 vrrp-group 1 virtual-
address 172.16.56.1
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.2/24 vrrp-group 1 priority
190
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.2/24 vrrp-group 1 preempt
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.2/24 vrrp-group 1 accept-
data
```

On the secondary MX Series router:

```
{master}[edit]
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.3/24 vrrp-group 1 virtual-
address 172.16.56.1
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.3/24 vrrp-group 1 priority
180
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.3/24 vrrp-group 1 preempt
root@MX960# set interfaces irb unit 71 family inet address 172.16.56.3/24 vrrp-group 1 accept-
data
```

Upgrading the Two Tiers

Before you upgrade JUNOS for the data center MX Series and EX4200 platforms, it is important to log information about the existing system so that after the upgrade, you can compare the same information to verify that all components are installed and working as expected. For detailed steps on how to log the information about your system before upgrading JUNOS, refer to **www.juniper.net/techpubs/software/nog/nog-baseline/html/upgrade2.html.**

You can download the JUNOS for MX Series and EX4200 platforms from the Download Software menu at **www.juniper.net/customers/support**. To download the software, you must have a Juniper Networks user account. For information on obtaining an account, see **www.juniper.net/entitlement/setupAccountInfo.do**.

To provide a smooth software upgrade for the two tiers in the data center network, you need to install software upgrades on the EX4200 top-of-rack virtual chassis **number 2**, as illustrated in Figure 1.

To upgrade the switch software, perform the following steps.

1. Download the JUNOS package for the related EX4200 switch as described above.
2. Copy the software package to the switch. We recommend that you use FTP to copy the file to the /var/tmp directory.
3. To install the new package on the switch, enter the following command in operational mode:

```
root@EX4200> request system software add source [member member_id] reboot
```

Include the member option to install the software package on only one member of a virtual chassis. Other members of the virtual chassis are not affected.

To install the software on all members of the virtual chassis, do not include the member option.

Copyright © 2009, Juniper Networks, Inc.

Replace source with one of the following paths:

For a software package that is installed from a local directory on the switch: /pathname/package-name

For software packages that are downloaded and installed from a remote location:

```
ftp://hostname/pathname/package-name
http://hostname/pathname/package-name
```

For example, where package-name is jinstall-ex-9.2R1.10-domestic-signed.tgz:

1. Install the software upgrades on the EX4200 top-of-rack virtual chassis **number 1** by following the above steps.

2. Install the JUNOS Software package on the backup MX Series first in the core layer, as illustrated in Figure 1.

To upgrade the router software, follow these steps:

1. Install the new software package using the request system software add command:

```
user@host> request system software add validate /var/tmp/jinstall-9.2R1.8–domestic-signed.tgz
```

2. Reboot the router to start the new software using the request system reboot command:

```
user@host> request system reboot
```

3. The software is loaded when you reboot the system. Installation can take between 5 and 10 minutes. The router then reboots from the boot device on which the software was just installed. When the reboot is complete, the router displays the login prompt. Log in and issue the show version command to verify the version of the software installed.

If the MX Series router in your network has two Routing Engines, perform a JUNOS installation on each Routing Engine separately to avoid disrupting network operations.

Install the new JUNOS Software release on the backup Routing Engine, while keeping the currently running software version on the *primary* Routing Engine.

After making sure that the new software version is running correctly on the backup Routing Engine, switch over to the newly installed Routing Engine to activate the new software.

Finally, install the new software on the new backup Routing Engine. For additional information about how to perform a software upgrade for devices with redundant Routing Engines, refer to **www.juniper.net/techpubs/software/junos/junos92/swconfig-install/installing-the-software-package-on-a-router-with-redundantrouting-engines.html#jN11F4E.**

4. Perform the JUNOS package upgrade on the primary MX Series router in the core layer by following the above steps.

Copyright © 2009, Juniper Networks, Inc.

Summary

Juniper Networks offers customers high-performance routing and switching products to fit the end-to-end solution requirements in today's enterprise network. These products are optimal for running data centers at great scale, and for enabling emerging trends that impact the data center such as energy conservation, location consolidation, and server and application virtualization. Enterprise data center architecture can be simplified into two layers (as compared with the traditional 3-layer model), from devices connected to wire-speed stackable switches at the access layer to the high-density robust wire-rate routers at the data center core. CAPEX and OPEX savings can be realized by collapsing devices and/or layers within the data center. The hierarchical topology segments the data center network into separate functional areas, simplifying operations and increasing availability. Each layer within the hierarchical infrastructure has a specific role.

This document explains how to implement a 2-tier LAN design in a single data center using Juniper Networks EX4200 switches with virtual chassis technology at the access tier, and by interconnecting the access tier to the data center core network built on MX Series Ethernet Services Routers. This document also provides the reader with a middle of the road approach by which the data center LAN can be built reliably. The implementation guidelines addressed in this paper define a data center network that consists of multiple, functionally distinct areas that may or may not be physically separated. Hence, we also provide use cases that show servers residing in the same rack but belonging to different data center functional areas (or application tiers).

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.

8010014-001-EN June 2009

Printed on recycled paper.