Nortel Ethernet Routing Switch 8600

Planning and Engineering — Network Design

Release: 5.1
Document Revision: 02.02

www.nortel.com

NN46205-200

Nortel Ethernet Routing Switch 8600
Release: 5.1
Publication: NN46205-200
Document release date: 27 August 2010

Contents

Safety messages

This section describes the different precautionary notices used in this document. This section also contains precautionary notices that you must read for safe operation of the Nortel Ethernet Routing Switch 8600.

Notices

Notice paragraphs alert you about issues that require your attention. The following sections describe the types of notices.

Attention notice

ATTENTION
An attention notice provides important information regarding the installation and operation of Nortel products.

Caution ESD notice

CAUTION
ESD
ESD notices provide information about how to avoid discharge of static electricity and subsequent damage to Nortel products.

CAUTION
ESD (décharge électrostatique)
La mention ESD fournit des informations sur les moyens de prévenir une décharge électrostatique et d'éviter d'endommager les produits Nortel.

CAUTION
ACHTUNG ESD
ESD-Hinweise bieten Information dazu, wie man die Entladung von statischer Elektrizität und Folgeschäden an Nortel-Produkten verhindert.

CAUTION
PRECAUCIÓN ESD (Descarga electrostática)
El aviso de ESD brinda información acerca de cómo evitar una descarga de electricidad estática y el daño posterior a los productos Nortel.

CAUTION
CUIDADO ESD
Os avisos do ESD oferecem informações sobre como evitar descarga de eletricidade estática e os conseqüentes danos aos produtos da Nortel.

CAUTION
ATTENZIONE ESD
Le indicazioni ESD forniscono informazioni per evitare scariche di elettricità statica e i danni correlati per i prodotti Nortel.

Caution notice

CAUTION
Caution notices provide information about how to avoid possible service disruption or damage to Nortel products.

CAUTION
ATTENTION
La mention Attention fournit des informations sur les moyens de prévenir une perturbation possible du service et d'éviter d'endommager les produits Nortel.

CAUTION
ACHTUNG
Achtungshinweise bieten Informationen dazu, wie man mögliche Dienstunterbrechungen oder Schäden an Nortel-Produkten verhindert.

CAUTION
PRECAUCIÓN
Los avisos de Precaución brindan información acerca de cómo evitar posibles interrupciones del servicio o el daño a los productos Nortel.

CAUTION
CUIDADO
Os avisos de cuidado oferecem informações sobre como evitar possíveis interrupções do serviço ou danos aos produtos da Nortel.

CAUTION
ATTENZIONE
Le indicazioni di attenzione forniscono informazioni per evitare possibili interruzioni del servizio o danni ai prodotti Nortel.

Software license

This section contains the Nortel Networks software license.

Nortel Networks Inc. software license agreement

This Software License Agreement ("License Agreement") is between you, the end-user ("Customer") and Nortel Networks Corporation and its subsidiaries and affiliates ("Nortel Networks"). PLEASE READ THE FOLLOWING CAREFULLY. YOU MUST ACCEPT THESE LICENSE TERMS IN ORDER TO DOWNLOAD AND/OR USE THE SOFTWARE. USE OF THE SOFTWARE CONSTITUTES YOUR ACCEPTANCE OF THIS LICENSE AGREEMENT. If you do not accept these terms and conditions, return the Software, unused and in the original shipping container, within 30 days of purchase to obtain a credit for the full purchase price.

"Software" is owned or licensed by Nortel Networks, its parent or one of its subsidiaries or affiliates, and is copyrighted and licensed, not sold. Software consists of machine-readable instructions, its components, data, audio-visual content (such as images, text, recordings or pictures) and related licensed materials including all whole or partial copies. Nortel Networks grants you a license to use the Software only in the country where you acquired the Software. You obtain no rights other than those granted to you under this License Agreement. You are responsible for the selection of the Software and for the installation of, use of, and results obtained from the Software.

1. **Licensed Use of Software.** Nortel Networks grants Customer a nonexclusive license to use a copy of the Software on only one machine at any one time or to the extent of the activation or authorized usage level, whichever is applicable. To the extent Software is furnished for use with designated hardware or Customer furnished equipment ("CFE"), Customer is granted a nonexclusive license to use Software only on such hardware or CFE, as applicable. Software contains trade secrets and Customer agrees to treat Software as confidential information using the same care and discretion Customer uses with its own similar information that it does not wish to disclose, publish or disseminate. Customer will ensure that anyone who uses the Software does so only in compliance with the terms

of this Agreement. Customer shall not a) use, copy, modify, transfer or distribute the Software except as expressly authorized; b) reverse assemble, reverse compile, reverse engineer or otherwise translate the Software; c) create derivative works or modifications unless expressly authorized; or d) sublicense, rent or lease the Software. Licensors of intellectual property to Nortel Networks are beneficiaries of this provision. Upon termination or breach of the license by Customer or in the event designated hardware or CFE is no longer in use, Customer will promptly return the Software to Nortel Networks or certify its destruction. Nortel Networks may audit by remote polling or other reasonable means to determine Customer's Software activation or usage levels. If suppliers of third party software included in Software require Nortel Networks to include additional or different terms, Customer agrees to abide by such terms provided by Nortel Networks with respect to such third party software.

2. Warranty. Except as may be otherwise expressly agreed to in writing between Nortel Networks and Customer, Software is provided "AS IS" without any warranties (conditions) of any kind. NORTEL NETWORKS DISCLAIMS ALL WARRANTIES (CONDITIONS) FOR THE SOFTWARE, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OF NON-INFRINGEMENT. Nortel Networks is not obligated to provide support of any kind for the Software. Some jurisdictions do not allow exclusion of implied warranties, and, in such event, the above exclusions may not apply.

3. Limitation of Remedies. IN NO EVENT SHALL NORTEL NETWORKS OR ITS AGENTS OR SUPPLIERS BE LIABLE FOR ANY OF THE FOLLOWING: a) DAMAGES BASED ON ANY THIRD PARTY CLAIM; b) LOSS OF, OR DAMAGE TO, CUSTOMER'S RECORDS, FILES OR DATA; OR c) DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES (INCLUDING LOST PROFITS OR SAVINGS), WHETHER IN CONTRACT, TORT OR OTHERWISE (INCLUDING NEGLIGENCE) ARISING OUT OF YOUR USE OF THE SOFTWARE, EVEN IF NORTEL NETWORKS, ITS AGENTS OR SUPPLIERS HAVE BEEN ADVISED OF THEIR POSSIBILITY. The forgoing limitations of remedies also apply to any developer and/or supplier of the Software. Such developer and/or supplier is an intended beneficiary of this Section. Some jurisdictions do not allow these limitations or exclusions and, in such event, they may not apply.

4. General

1. If Customer is the United States Government, the following paragraph shall apply: All Nortel Networks Software available under this License Agreement is commercial computer software and commercial computer

software documentation and, in the event Software is licensed for or on behalf of the United States Government, the respective rights to the software and software documentation are governed by Nortel Networks standard commercial license in accordance with U.S. Federal Regulations at 48 C.F.R. Sections 12.212 (for non-DoD entities) and 48 C.F.R. 227.7202 (for DoD entities).

2. Customer may terminate the license at any time. Nortel Networks may terminate the license if Customer fails to comply with the terms and conditions of this license. In either event, upon termination, Customer must either return the Software to Nortel Networks or certify its destruction.
3. Customer is responsible for payment of any taxes, including personal property taxes, resulting from Customer's use of the Software. Customer agrees to comply with all applicable laws including all applicable export and import laws and regulations.
4. Neither party may bring an action, regardless of form, more than two years after the cause of the action arose.
5. The terms and conditions of this License Agreement form the complete and exclusive agreement between Customer and Nortel Networks.
6. This License Agreement is governed by the laws of the country in which Customer acquires the Software. If the Software is acquired in the United States, then this License Agreement is governed by the laws of the state of New York.

New in this release

The following sections detail what's new in *Nortel Ethernet Routing Switch 8600 Planning and Engineering — Network Design* (NN46205-200) for Release 5.1.

Features

See the following sections for information about feature changes:

DWDM XFPs

The Ethernet Routing Switch 8600 R/RS modules now support DWDM XFPs. For more information, see "DWDM XFPs" (page 68).

HA feature support and synchronization information

The HA feature support and synchronization information is updated for release 5.1. See Table 22 "Feature support for HA in specified software release versions" (page 84) and Table 23 "Synchronization capabilities in HA mode" (page 85).

MinLink for LACP

The Ethernet Routing Switch 8600 now supports MinLink for LACP. See "LACP and Minimum Link" (page 93).

Bidirectional Forwarding Detection

The Ethernet Routing Switch 8600 now supports Bidirectional Forwarding Detection. See "Bidirectional Forwarding Detection" (page 95).

Multicast and VRF-lite

The Ethernet Routing Switch 8600 now supports PIM-SM, PIM-SSM, and IGMP in VRF-lite configurations. See "Multicast and VRF-lite" (page 194) and "PIM-SM and PIM-SSM scalability" (page 219).

IGMPv3 backward compatibility

The Ethernet Routing Switch 8600 now supports IGMPv3 backward compatibility with IGMPv1/v2. See “IGMPv3 backward compatibility” (page 206).

MSDP

The Ethernet Routing Switch 8600 now supports Multicast Source Discovery Protocol (MSDP). See “MSDP ” (page 238).

Static mroute

The Ethernet Routing Switch 8600 now supports static mroute. See “Static mroute” (page 240).

TACACS+

The Ethernet Routing Switch 8600 now supports TACACS+. See “TACACS+” (page 325).

Other changes

See the following sections for information about changes that are not feature-related:

R series modules and global FDB filters

This document is updated to state that Release 4.1.5 and later provide the same global forwarding database filter (FDB) operations for R series modules as for Classic modules. For more information, see “R series modules and global FDB filters” (page 35).

Supported scaling capabilities

The supported scaling capabilities table is updated to include DHCP relay scaling information (CR Q01949340). See Table 5 "Supported scaling capabilities" (page 38).

SMLT full-mesh recommendations

SMLT full-mesh recommendations are added (CR Q01971344). See “SMLT full-mesh recommendations with OSPF” (page 114).

VLACP information consolidation

VLACP information and recommendations that were originally included under the Split Multi-Link Trunking section are now moved to “End-to-end fault detection and VLACP” (page 77).

SLPP and Extended CP-Limit information consolidation

SLPP and Extended CP-Limit information and recommendations that were originally included under the Split Multi-Link Trunking section are now moved to “SLPP, Loop Detect, and Extended CP-Limit” (page 148).

MPLS IP VPN and IP VPN Lite

The MPLS IP VPN and IP VPN Lite sections are updated and expanded with new deployment scenarios. See “MPLS IP VPN and IP VPN Lite” (page 257).

Introduction

This document describes a range of design considerations and related information that helps you to optimize the performance and stability of your Ethernet Routing Switch 8600 network.

ATTENTION
This document describes the Nortel recommended best practices for network configuration. If your network diverges from the recommended best practices, Nortel cannot guarantee support for issues that arise.

Navigation

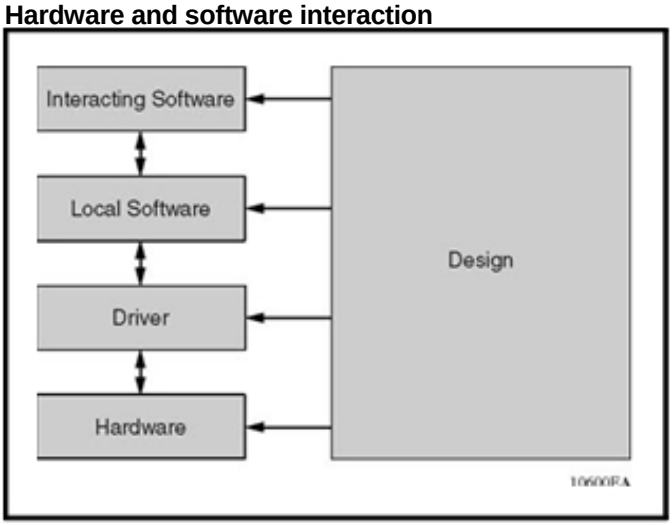
Network design fundamentals

To efficiently and cost-effectively use your Nortel 8000 Series routing switch, you must properly design your network. Use the information in this section to help you properly design your network. When you design networks, you must consider:

- reliability and availability
- platform redundancy
- desired level of redundancy

A robust network depends on the interaction between system hardware and software. System software can be divided into different functions as shown in the following figure.

Figure 1
Hardware and software interaction



These levels are based on the software function. A Driver is the lowest level of software that actually performs a function. Drivers reside on a single module and do not interact with other modules or external devices. Drivers are very stable.

MultiLink Trunking (MLT) is a prime example of Local Software because it interacts with several modules within in the same device. No external interaction is needed, so its function can be easily tested.

Interacting Software is the most complex level of software because it depends on interaction with external devices. The Open Shortest Path First (OSPF) protocol is a good example of this software level. Interaction can occur between devices of the same type or with devices of other vendors than run a completely different implementation.

Based on network problem-tracking statistics, the following is a rough stability estimation model of a system using these components:

- Hardware and drivers represent a small portion of network problems.
- Local Software represents a more significant share.
- Interacting Software represents the vast majority of the reported issues.

Based on this model, network design should attempt to off-load the interacting software level as much as possible to the other levels, especially to the hardware level. Therefore, Nortel recommends that you follow these generic rules when you design networks:

1. Design networks as simply as possible.
2. Provide redundancy, but do not over-engineer your network.
3. Use a toolbox to design your network.
4. Design according to the product capabilities described in the latest Release Notes.
5. Follow the design rules provided in this document and also in the various configuration guides for your switch.

Hardware fundamentals and guidelines

This section provides general hardware guidelines that you should be aware of when designing your network. Use the information in this section to help you during the hardware design and planning phase.

Navigation

Chassis considerations

This section discusses chassis power and cooling considerations. You must properly power and cool your chassis, or nonoptimal switch operation can result.

Chassis power considerations

Each Ethernet Routing Switch 8600 chassis provides redundant power options, depending on the chassis and the number of modules installed.

The 8006 and 8010 chassis support up to three power supplies. You must install at least one power supply per chassis.

To determine the number of power supplies required for your switch configuration, use the *Power Supply Calculator for ERS 8600* (NN48500-519) . This is available on the Nortel Technical Support Web site at www.nortel.com/documentation. Choose Routers & Routing Switches, and then Ethernet Routing Switch 8600.

To support a full configuration of RS modules, an 8004 or 8005 power supply is required. Do not mix 8004 and 8005 power supplies in the same chassis.

Power supply circuit requirements

The Ethernet Routing Switch 8600 AC power supplies require single-phase source AC.

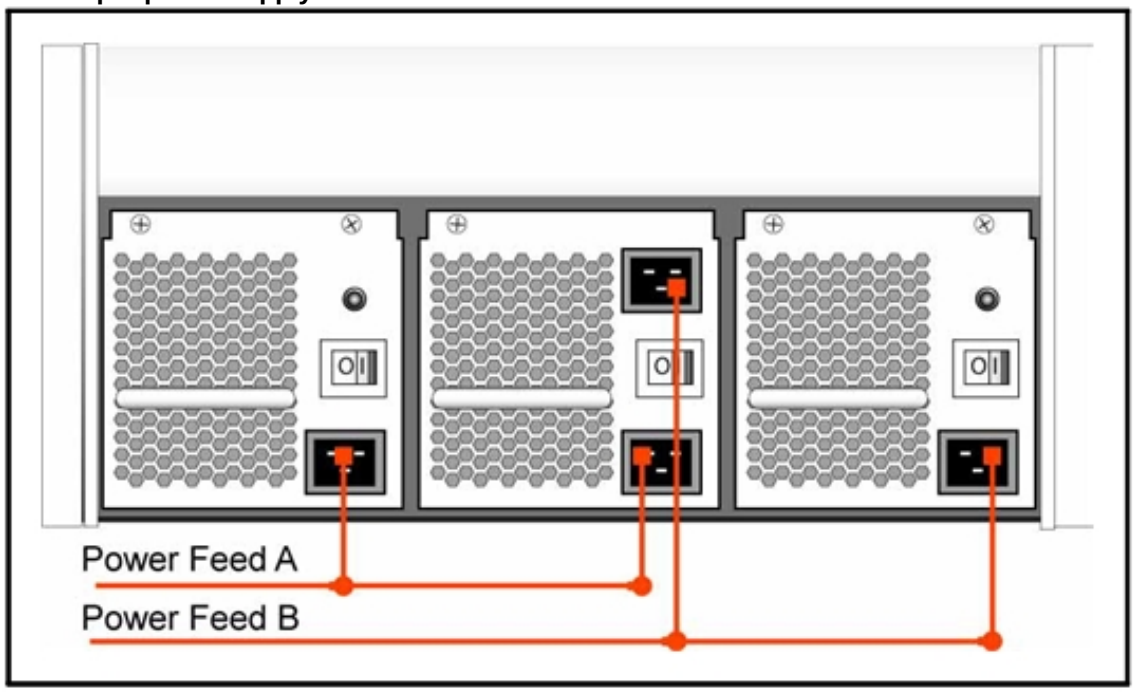
Do not mix AC and DC power supplies in the same chassis.

The source AC can be out of phase between multiple power supplies in the same chassis. Therefore, power supply 1 can operate from phase A, and power supply 2 can operate from phase B.

The source AC can be out of phase between AC inputs on power supplies that are equipped with multiple AC inputs. Therefore, power cord 1 can plug into phase A, and power cord 2 can plug into phase B.

Release 5.0 introduces the dual-input 8005DI AC power supply. You can use this dual-input supply with two other single-phase AC sources of different power feeds. To share the two power feeds with the dual input supply, connect AC source Power Feed 1 to input 1 on the dual-input supply, and connect AC source Power Feed 2 to input 2 on the dual-input supply. See the following figure. Nortel recommends this configuration to provide full power feed redundancy.

Figure 2
Dual-input power supply connections



On the 8005DI AC power supply, the two AC input sources can be out of synchronization with each other, having a different voltage, frequency, phase rotation, and phase angle as long as the power characteristics for each separate input AC source remain within the range of the manufacturer's specifications.

Chassis cooling

Two basic methods can be used to determine the cooling capacity required to cool the switch. You can use the Nortel Power Supply Calculator Tool to determine power draw in watts, or you can use a worse-case power draw.

You can use the Nortel Power Supply Calculator Tool to determine the power draw for a chassis configuration. Use this power draw in the following cooling capacity formula:

Cooling capacity (BTU) = power draw (W) x 3.412

The chassis configuration can affect the switch cooling requirements. If you change the switch configuration, the cooling requirements can change as well.

The alternative method is to determine a worse-case power draw on the power supply and then use this value in the cooling capacity formula.

When using the second method, take into consideration the number of power supplies and redundancy. The worse-case power draw is the maximum power draw plus the number of supplies required to operate the system without redundancy.

For example, if two 8005AC power supplies power a chassis, and a third is added for redundancy, the worse-case value is the maximum power draw of a single 8005AC power supply times two (the total of two power supplies, not three). For the 8005AC power supplies, the actual draw depends on the input voltage. For a nominal input voltage of 110 VAC, the draw is 1140 W. For 220 AC volts (VAC), the draw is 1462 watts (W). For a three-power supply system running at 110 VAC, the maximum worse-case power draw is 1140 W x 2, or 2280 W. Therefore this system requires a cooling capacity of 7164 British thermal units (BTU).

You also need to consider the cooling requirements of the power supplies themselves. For these specifications, see *Nortel Ethernet Routing Switch 8600 Installation — AC Power Supply* (NN46205-306) and *Nortel Ethernet Routing Switch 8600 Installation — DC Power Supply* (NN46205-307) . Add these values to the cooling capacity calculation. For a multi-power supply system, you need to factor into the calculation the maximum nonredundant number of power supplies.

You must also consider the type of module installed on the chassis. If you install an RS module in the chassis, you must install the high speed cooling modules. If you do not install the high speed cooling modules, the software cannot operate on the RS module. For information about installing high speed cooling modules, see *Nortel Ethernet Routing Switch 8600 Installation — Cooling Module* (NN46205-302) .

Design a cooling system with a cooling capacity slightly greater than that calculated to maintain a safe margin for error and to allow for future growth.

Modules

Use modules to interface the switch to the network. This section discusses design guidelines and considerations for Ethernet Routing Switch 8600 modules.

SF/CPU modules

The switch fabric/CPU (SF/CPU) module performs intelligent switching and routing. Every chassis must have at least one SF/CPU; for redundancy, install two SF/CPUs.

The use of dual 8692 SF/CPU modules enables a maximum switch bandwidth of 512 Gbit/s. Dual modules provide redundancy and load sharing between the modules. By using Split Multilink Trunking (SMLT)

in the core in a resilient cluster configuration (redundant switch with two 8692 SF/CPU modules) can provide over 1 terabit per second (Tbit/s) of core switching capacity.

You can install the 8692 SF/CPU module in slots 5 or 6 of the 8006, 8010, or 8010co chassis. The 8692 SF/CPU module is not supported in the 8003 chassis.

Systems that have R or RS modules must use the 8692 SF/CPU. Nortel does not support R or RS modules with an 8690 or 8691 switch fabric.

The Ethernet Routing Switch 8600 software does not support configurations where the 8692 SF/CPU and the 8690 SF/CPU or 8691 SF/CPU module are installed in the same chassis during an upgrade process.

You must use an 8692 SF/CPU with the Enterprise Enhanced CPU Daughter Card (SuperMezz) for certain features, for example, IPv6, VRF Lite, MPLS, IP VPN, and the 100 millisecond failover feature. In a dual CPU switch, if one CPU uses SuperMezz, the other must also use SuperMezz.

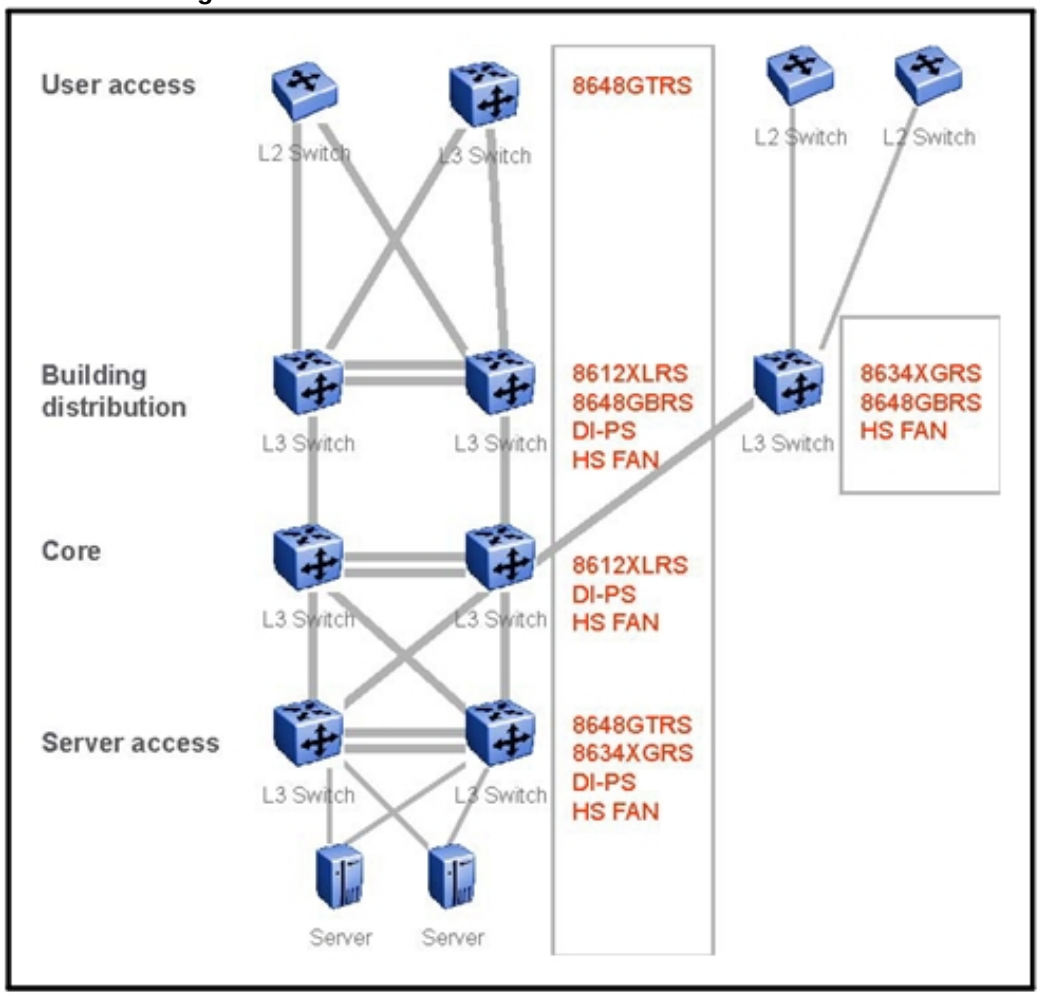
RS module

RS modules include the 8648GTRS, the 8612XLRS, the 8634XGRS, and the 8648GBRS. RS modules provide support for a variety of technologies, interfaces, and feature sets and provide 10 Gbit/s port rates. RS modules require the high-speed cooling module and the 8692 SF/CPU.

In chassis equipped with RS modules, you can use 8005AC, 8005DI AC, 8004AC, or 8004DC power supplies. RS modules are interoperable with R modules.

The following figure shows typical uses for RS modules.

Figure 3
RS module usage



The 8612XLRS, 8648GBRS, and 8634XGRS modules use a three-lane Distributed Processing Module (DPM) based on Route Switch Processor (RSP) 2.6 architecture. The 8648GTRS uses a two-lane DPM. The following table provides details about oversubscription rates for each module. Typical network designs use oversubscribed modules at the building distribution layer and nonoversubscribed links to core. Using oversubscribed modules at the distribution layer are cost-effective as long as the module provides strong built-in packet QoS capabilities—RS modules do so.

Table 1
RS module lane oversubscription

Module	Lane oversubscription
8612XLRS	4:1 (each group of ports [1-4, 5-8, and 9-12] share a 10GE lane)

Table 1
RS module lane oversubscription (cont'd.)

Module	Lane oversubscription
8648GBRS	1.6:1 (each group of ports [1-16, 17-32, and 33-48] share a 10GE lane)
8634XGRS	Lane 1: 1.6:1 Lane 2: 1.6:1 Lane 3: 2:1 (each group of ports [1-16, 17-32, and 33-34] share a 10GE lane)
8648GTRS	2.4:1 (both lanes) (each group of ports [1-24, and 25-48] share a 10GE lane)

The following XFPs are supported on the 8612XLRS module (DS1404097-E6):

- 10GBASE-SR
- 10GBASE-LR/LW
- 10GBASE-LRM
- 10GBASE-ER/EW
- 10GBASE-ZR/ZW
- 10GBASE DWDM

For XFP specifications and information, see *Nortel Ethernet Routing Switch 8600 Installation — SFP, XFP, and GBIC and OADM Hardware Components* (NN46205-320) .

R modules

R modules provide support for a variety of technologies, interfaces, and feature sets and provide 1 and 10 Gbit/s port rates. The following R modules are supported by the Ethernet Routing Switch 8600 and require the use of the 8692 SF/CPU:

- 8630GBR—30 port 1000BASE-X SFP baseboard
- 8648GTR—48 port 10/100/1000BASE-T
- 8683XLR—3 port 10GBASE-x XFP baseboard (LAN phy)
- 8683XZR—3 port 10GBASE-x XFP baseboard (LAN/WAN phy)

R modules are compatible with the existing 8010 and 8006 chassis.

When installed in a standard slot, R modules offer increased port density. When installed in a high-performance slot or chassis, R modules offer increased port density as well as increased performance over existing E and M modules.

R modules inserted in slots 2 to 4 and slots 7 to 9 of the 8010 10-slot chassis, and in slots 2 to 4 of the 8006 six-slot chassis, operate at high-performance. R modules inserted into slots 1 and 10 of the 8010 chassis, and slot 1 of the 8006 chassis, operate at standard performance. For information about relative performance per slot with two fabrics installed in the existing 8010 and 8006 chassis, see the following table.

Table 2
8010 and 8006 chassis data performance

Module type	Standard slot (1 and 10) full-duplex	High-performance slot (2-4, 7-9) full-duplex
E and M	16 Gbit/s	16 Gbit/s
8630GBR	16 Gbit/s	60 Gbit/s
8683XLR	16 Gbit/s	60 Gbit/s
8648GTR	16 Gbit/s	32 Gbit/s
8683XZR	16 Gbit/s	60 Gbit/s
8612XLRS	16 Gbit/s	60 Gbit/s
8648GTRS	16 Gbit/s	40 Gbit/s
8648GBRS	16 Gbit/s	60 Gbit/s
8634XGRS	16 Gbit/s	60 Gbit/s

For maximum switch performance, Nortel recommends that you place R modules in chassis slots 2 to 4 or 7 to 9, as available.

A chassis revision with an upgraded High-performance Backplane (HPB), compatible with existing E and M modules, as well as new R modules supporting high-performance in all slots, is available. You can identify the High-performance Backplane by the chassis revision number. Use the command line interface (CLI) command **`show sys info`** or the NNCLI command **`show sys-info`** to display the revision number. A revision number of 02 or higher in the H/W Config field indicates that the chassis is the high-performance chassis. R series modules and Classic modules can interoperate within the same chassis in a mixed-mode configuration. Chassis Revision A indicates that the chassis is not a high performance chassis and must be upgraded.

R series modules and global FDB filters

Release 4.1.5 and later provides the same global forwarding database filter (FDB) operations for R series modules as for Classic modules (E and M modules). The global FDB filter for R series modules does not use the **`config vlan <vid> fdb-filter`** command. Instead, the command is **`config fdb fdb-filter add <mac-address>`**.

Global FDB filters are not supported on interswitch trunk switches. Instead, manually configure global MAC filters on both aggregation switches.

For more information about the FDB filters, see *Nortel Ethernet Routing Switch 8600 Configuration — VLANs and Spanning Tree* (NN46205-517) .

8648GTR recommendations

Nortel supports the 8648GTR module in a high-performance slot only. Nortel does not support the 8648GTR in a standard slot.

Release 4.1.1 and later allows MLT to run between ports between an 8648GTR and other module types. MLT ports must run at the same speed with the same interface type, even if using different Input/Output (I/O) module types.

8683XLR and 8683XZR information and recommendations

The 8683XLR provides 10 Gigabit LAN connectivity, while the 8683XZR module provides both 10 Gigabit LAN and 10 Gigabit WAN connectivity. A synchronous optical network (SONET) frame encloses WAN Ethernet frames; embedding WAN Ethernet packets inside SONET frames requires support for SONET-like management, configuration, and statistics.

Unlike the WAN 10 GbE module, the LAN version does not use SONET as its transport mechanism. You cannot program WAN and LAN modes of operation. Due to different clock frequencies for LAN and WAN modes of operation, the LAN and WAN versions of the 10 GbE module use different module IDs, and are fixed in one mode of operation.

10 GbE modules support only full-duplex mode. As per the IEEE 802.3ae standard, Auto-Negotiation is not supported on 10 GbE links. The following table provides details about the differences between 1 GbE modules and 10 GbE modules.

Table 3
1 GbE and 10 GbE module comparison

1 GbE	10 GbE
Carrier Sense Multiple Access with Collision Detection (CSMA/CD) and full-duplex	Full-duplex only, no Auto-Negotiation

Table 3
1 GbE and 10 GbE module comparison (cont'd.)

1 GbE	10 GbE
802.3 Ethernet frame format (includes min/max frame size)	802.3 Ethernet frame format (includes min/max frame size)
Carrier extension	Throttle MAC speed (rate adapt)
One physical interface	LAN and WAN physical layer interfaces
Optical or copper media	Optical media only
8B/10B encoding	64B/66B encoding

The 8683 modules have three forwarding engine lanes and three bays for installing 10 Gigabit Small Form Factor Pluggable (XFP) transceivers. Each lane supports 10 Gbit/s bidirectional traffic. All three ports can run concurrently at 10 Gbit/s.

Although the 10GBASE-LR, -ER, and -ZR XFPs support both LAN and WAN modes, the 8683XLR module supports only the LAN mode. The 8683XZR module supports both the LAN and WAN (SONET) modes.

The 8683 modules supports the following XFPs:

- 10GBASE-SR
- 10GBASE-LR/LW
- 10GBASE-LRM
- 10GBASE-ER/EW
- 10GBASE-ZR/ZW
- 10GBASE DWDM

For XFP specifications and information, see *Nortel Ethernet Routing Switch 8600 Installation — SFP, XFP, and GBIC and OADM Hardware Components* (NN46205-320) .

10 GbE clocking

Whether you use internal or line clocking depends on the application and configuration. Typically, the default internal clocking is sufficient. Use line clocking on both ends of a 10 GbE WAN connection (line-line) when using SONET/Synchronous Digital Hierarchy (SDH) Add-Drop Multiplexing (ADM) products, such as the Optical Cross Connect DX. This allows the 10 GbE WAN modules to synchronize to a WAN timing hierarchy, and minimizes any timing slips. Interworking 10 GbE WAN across an Add-Drop Multiplexer requires the use of an OC-192c/VC-4-64c payload cross-connection device.

When connecting 10 GbE modules back-to-back, or through metro (OM5200) or long haul (LH 1600G) dense Wavelength Division Multiplexing (DWDM) equipment, you can use the timing combinations of internal-internal, line-internal, or internal-line on both ends of the 10 GbE WAN connection. In these scenarios, at least one of the modules provides the reference clock. DWDM equipment does not typically provide sources for timing synchronization. For DWDM, Nortel recommends that you avoid using a line-line combination because it causes an undesired timing loop.

The following table describes the recommended clock source settings for 10 GbE WAN interfaces. Use these clock settings to ensure accurate data recovery and to minimize SONET-layer errors.

Table 4
Recommended 10GE WAN interface clock settings

Clock source at both ends of the 10 GbE WAN link	Back-to-back with dark fiber or DWDM	SONET/SDH WAN with ADM
internal-internal	Yes	No
internal-line	Yes	No
line-internal	Yes	No
line-line	No	Yes

For more information about WAN modules, see *Nortel Ethernet Routing Switch 8600 Configuration — Ethernet Modules* (NN46205-503) .

Classic modules

In addition to R or RS modules, the Ethernet Routing Switch 8600 also supports Classic modules (E and M). M modules, or extended memory modules, are designed to support large Layer 2 (bridging and/or multicast) and large Layer 3 (more than 20 000 route) environments.

As shown in Table 7 "Modules and feature availability per module" (page 43), E modules support up to 32K records, whereas M modules support up to 128K records.

M modules are based on the E module architecture and support all E module features and characteristics. The only difference between M and E modules is the amount of memory required to support 128K records. R modules and Classic modules can interoperate within the same chassis in mixed-mode.

For a complete description of scaling limitations based on module type, see Table 6 "Operational modes versus module type" (page 42).

Modules, operational modes, features, and scaling

When you decide which modules you want to use, consider the following information about feature, modes, and scaling supported by each module. The following tables show modules, modes, scaling information, and features available on the Ethernet Routing Switch 8600 modules. For the most recent scaling information, always consult the latest version of the Release Notes.

CAUTION
Risk of traffic loss
SuperMezz is required for configurations that carry over 4000 multicast routes.

Table 5
Supported scaling capabilities

	Maximum number supported 8692SF without SuperMezz (R or RS series modules)	Maximum number supported 8692SF with SuperMezz (R or RS series modules)	Maximum number supported (E and M modules)
Layer 2			
MAC address table entries	64000 32000 when SMLT is used	64000 32000 when SMLT is used	E modules: 25 000 M modules: 119 000
VLANs (port-, protocol-, and IEEE 802.1Q-based)	4000 when max VLAN feature enabled	4000	1972
IP subnet-based VLANs	800	800	E modules: 200 M modules: 800
Ports per Link Aggregation Group (LAG, MLT)	8	8	8
Aggregation groups 802.3ad aggregation groups Multi Link Trunking (MLT) group	NonR mode: 32 R mode: 128	NonR mode: 32 R mode: 128	32
SMLT links	R mode: 128	R mode: 128	32
SLT (single link SMLT)	382	382	382
VLANs on SMLT/IST link	R mode with Max VLAN feature enabled: 2000	R mode with Max VLAN feature enabled: 2000	See "SMLT scalability" (page 111)

Table 5
Supported scaling capabilities (cont'd.)

	Maximum number supported 8692SF without SuperMezz (R or RS series modules)	Maximum number supported 8692SF with SuperMezz (R or RS series modules)	Maximum number supported (E and M modules)
RSMLT per VLAN	32 SMLT links with RSMLT-enabled VLANs	32 SMLT links with RSMLT-enabled VLANs	32 SMLT links with RSMLT-enabled VLANs
RSTP/MSTP (number of ports)	384, with 224 active. Configure the remaining interfaces with Edge mode	384, with 224 active. Configure the remaining interfaces with Edge mode	384, with 224 active. Configure the remaining interfaces with Edge mode
MSTP instances	32	32	32
Advanced Filters			
ACLs for each system	4000	4000	N/A
ACEs for each system	1000	1000	N/A
ACEs for each ACL	1000	1000	N/A
ACEs for each port	2000: 500 inPort 500 inVLAN 500 outPort 500 outVLAN	2000: 500 inPort 500 inVLAN 500 outPort 500 outVLAN	N/A
IP, IP VPN/MPLS, IP VPN Lite, VRF Lite			
IP interfaces (VLAN- and brouter-based)	1972	1972	1972
VRF instances	N/A	255	N/A
ECMP routes	5000	5000	5000
VRRP interfaces	255	255	255
IP forwarding table (Hardware)	120000	250000	E modules: 20 000 forwarding routes M modules: 119 000 forwarding routes
BGP/mBGP peers	10	250	10
iBGP instances	on GRT	on GRT	on GRT
eBGP instances	on GRT	on 256 VRFs (including GRT)	on GRT

Table 5
Supported scaling capabilities (cont'd.)

	Maximum number supported 8692SF without SuperMezz (R or RS series modules)	Maximum number supported 8692SF with SuperMezz (R or RS series modules)	Maximum number supported (E and M modules)
BGP forwarding routes BGP routing information base (RIB) BGP forwarding information base (FIB)	BGP FIB 120 000 BGP RIB 250 000	BGP FIB 250 000 BGP RIB 500 000	E modules: BGP FIB 20000; BGP RIB 250 000; M modules: BGP FIB 119000, BGP RIB 250000
IP VPN routes (total routes for each system)	N/A	180000	N/A
IP VPN VRF instances	N/A	255	N/A
Static ARP entries	2048	2048 per VRF 10000 per system	2048
Dynamic ARP entries	32000 in R mode	32000 in R mode	16000
DHCP Relay instances (total for all VRFs)	512	512	512
Static route entries	2000	2000 per VRF 10000 per system	2000
OSPF instances for each switch	on GRT	on 64 VRFs (including GRT)	on GRT
OSPF areas for each switch	5	5 per VRF 24 per system	5
OSPF adjacencies for each switch	80	80 200 per system	80
OSPF routes	20000	20000 per VRF 50000 per system	E modules 15 000 M modules 20 000
OSPF interfaces	238	238 500 per system	238
OSPF LSA packet maximum size	3000 bytes	3000 bytes	3000 bytes
RIP instances	on GRT	64	on GRT
RIP interfaces	200	200	200
RIP routes	2500	2500 per VRF 10000 per system	2500

Table 5
Supported scaling capabilities (cont'd.)

	Maximum number supported 8692SF without SuperMezz (R or RS series modules)	Maximum number supported 8692SF with SuperMezz (R or RS series modules)	Maximum number supported (E and M modules)
Multiprotocol Label Switching			
MPLS LDP sessions	N/A	200	N/A
MPLS LDP LSPs	N/A	16000	N/A
MPLS RSVP static LSPs	200	200	N/A
Tunnels	2500	2500	N/A
IP Multicast			
DVMRP passive interfaces	1200	1200	1200
DVMRP active interfaces/neighbors	80	80	80
DVMRP routes	2500	2500	2500
PIM instances	on GRT	64	on GRT
PIM active interfaces	500	500 (200 for all VRFs)	500
PIM passive interfaces	1972	1972 (2000 for all VRFs)	1972
PIM neighbors	80	80 (200 for all VRFs)	80
Multicast streams: with SMLT/ without SMLT	500/1500	2000/4000	500/1500
IPX			
IPX interfaces	N/A	N/A	100
IPX RIP routes	N/A	N/A	5000
IPX SAP entries	N/A	N/A	7500
IPv6			
IPv6 interfaces	N/A	250	N/A
IPv6 tunnels	N/A	350	N/A
IPv6 static routes	N/A	2000	N/A
OSPFv3 areas	N/A	5	N/A
OSPFv3 adjacencies	N/A	80	N/A
OSPFv3 routes	N/A	5000	N/A

Table 5
Supported scaling capabilities (cont'd.)

	Maximum number supported 8692SF without SuperMezz (R or RS series modules)	Maximum number supported 8692SF with SuperMezz (R or RS series modules)	Maximum number supported (E and M modules)
Operations, Administration, and Maintenance			
IPFIX	384000 flows per chassis	384000 flows per chassis	N/A
RMON alarms with 4000K memory	2630	2630	2630
RMON events with 250K memory	324	324	324
RMON events with 4000K memory	5206	5206	5206
RMON Ethernet statistics with 250K memory	230	230	230
RMON Ethernet statistics with 4000K memory	4590	4590	4590

The number of hardware forwarding records for M modules is 125 838. 2162 records are used by the system. The record reservation feature reserves 8000 records for traffic types such as ARP, MAC, and so on.

Nortel supports only 25 spanning tree groups (STGs). Although you can configure up to 64 STGs (63 when a Web Switching Module is present), configurations of more than 25 STGs are not supported. If you need to configure more than 25 STGs, contact your Nortel Customer Support representative for more information. The Web Switching Module supports only tagged bridged protocol data units (BPDU) with the default STG value of STG ID 1; this leaves 24 supported STGs.

The following table lists the supported operational modes according to module type.

Table 6
Operational modes versus module type

Operational mode	R and RS module	M module	E module
Default	enabled	enabled	enabled
M	enabled	enabled	disabled
R	enabled	disabled	disabled

The following table lists feature availability according to module and mode type.

Table 7
Modules and feature availability per module

Feature	Module			Comments
	R series	M	E	
sVLAN	No	Yes	Yes	Prestandard
SMLT over 10 GbE	Yes	No	No	Available only with the 10GbE R and RS modules
IPX routing	No	Yes	Yes	
IPv4 ACLs ingress L2 to L4	Yes	N/A	N/A	Ingress filtering (ACT, ACL, ACE)
IPv4 ACLs egress L2 to L4	Yes	N/A	N/A	Egress filtering (ACT, ACL, ACE)
IPv4 ACL pattern matching	Yes	N/A	N/A	Pattern matching for ingress and egress
IPv4 policing L2 to L4	Yes	N/A	N/A	450 policers per LANE (10 1 Gigabit, 1 10 Gigabit), total of 10 800 policers
IPv6 shaping L2 to L4	Yes	N/A	N/A	Per port/per queue shapers 640 queues per LANE (10 1 Gigabit, 1 10 Gigabit) total of 15 360 queues
IPv6 ACLs ingress L2 to L4	Yes	N/A	N/A	Ingress Filtering (ACT, ACL, ACE)
IPv6 ACLs egress L2 to L4	Yes	N/A	N/A	Egress Filtering (ACT, ACL, ACE)
IPv6 ACL pattern matching	Yes	N/A	N/A	Pattern Matching for ingress and egress
Classic filter L2 to L4	No	Yes	Yes	Layer 2 with global filters (limited to 8 per ARU). MAC FDB filters are available only on Classic modules
Classic policing L3 to L4	No	Yes	Yes	

The module types within an operation mode operate whether the chassis is deployed with the same module type or mixed module types. The exception is R mode, which supports only R series modules. You can also independently enable the Enhanced Operational mode.

For additional information about enabling and using Enhanced Operational Mode, see *Nortel Ethernet Routing Switch 8600 Administration* (NN46205-605) .

You can use R series module filters (Access Control Lists, or ACL) in a mixed or R mode chassis, but only for with R series module ports or VLANs that contain R series module ports. In a mixed-mode chassis, ACLs can only be applied to R series module ports and VLANs. You must use Classic filters (source/destination/global) for Classic modules. You can apply an ACL to an VLAN that contains both R series module ports and Classic module ports, but the ACL is only applied to the R series module ports within the VLAN.

Some commands used with filters, including ping-snoop and multimedia filters, apply only to Classic modules and ports, and do not apply to R series module ports. To apply these types of filters to R series module ports, you must use R series module advanced filters.

Optical device guidelines

Use optical devices to enable high bit rate communications and long transmission distances. Use the information in this section to properly use optical devices in your network. For Nortel optical routing system (Coarse Wavelength Division Multiplexing system) information.

Optical device guideline navigation

- "Optical power considerations" (page 44)
- "10 GbE WAN module optical interoperability" (page 45)
- "1000BASE-X and 10GBASE-X reach" (page 45)
- "XFPs and dispersion considerations" (page 45)

Optical power considerations

When connecting the switch to collocated equipment, such as the OPTera Metro 5200, ensure that enough optical attenuation exists to avoid overloading the receivers of each device. Typically, this is approximately 3 to 5 decibels (dB). However, you do not have to attenuate the signal when using the 10GE WAN module in an optically-protected configuration with two OM5200 10G transponders. In such a configuration, use an optical splitter that provides a few dB of loss. Do not attenuate the signal to less than the receiver sensitivity of the OM5200 10G transponder (approximately -11 dBm). Other WAN equipment, such as the Cross Connect DX and the Long Haul 1600G, have transmitters that allow you to change the transmitter power level. By default, they are typically set to -10

dBm, thus requiring no additional receiver attenuation for the 10GE WAN module. For specifications for the 10 GbE modules, see *Nortel Ethernet Routing Switch 8600 Installation — Ethernet Modules* (NN46205-304) .

10 GbE WAN module optical interoperability

Although the 10 GbE WAN module uses a 1310 nanometer (nm) transmitter, it uses a wideband receiver that allows it to interwork with products using 1550 nm 10 Gigabit interfaces. Such products include the Cross Connect DX and the Long Haul 1600G. The Nortel OM5200 10G optical transponder utilizes a 1310 nm client-side transmitter.

1000BASE-X and 10GBASE-X reach

Various SFP (1 Gbit/s), XFP (10 Gbit/s), and GBIC (1 Gbit/s) transceivers can be used to attain different line rates and reaches. The following table shows typical reach attainable with optical devices. To calculate the reach for your particular fiber link, see "Reach and optical link budget" (page 61).

For more information about these devices, including compatible fiber type, see *Nortel Ethernet Routing Switch 8600 Installation — SFP, XFP, GBIC, and OADM Hardware Components* (NN46205-503) .

Table 8
Optical devices and maximum reach

SFP/XFP/GBIC	Maximum reach
1000BASE-SX	Up to 300 m
1000BASE-LRM	Up to 220 m
1000BASE-LX	Up to 10 km
1000BASE-BX	Up to 10 km over one fiber
1000BASE-XD	Up to 40 km
1000BASE-ZX	Up to 70 km
1000BASE-EX	Up to 120 km
10GBASE-LRM	Up to 220 m
10GBASE-SR	Up to 300 m
10GBASE-LR/LW	Up to 10 km
10GBASE-ER/EW	Up to 40 km
10GBASE-ZR/ZW	Up to 80 km

XFPs and dispersion considerations

The optical power budget (that is, attenuation) is not the only factor to consider when you are designing optical fiber links. As the bit rate increases, the system's dispersion tolerance is reduced. As you approach

the 10 Gbit/s limit, dispersion becomes an important consideration in link design. Too much dispersion at high data rates can cause the link bit error rate (BER) to increase to unacceptable limits.

Two important dispersion types that limit the achievable link distance are chromatic dispersion and polarization mode dispersion (PMD). For fibers that run at 10 Gbit/s or higher data rates over long distances, the dispersion must be determined to avoid possible BER increases and/or protection switches. Traditionally, dispersion is not an issue for bit rates of up to 2.5 Gb/s over fiber lengths of less than 500 km. The availability of 10 Gbit/s and 40 Gbit/s devices means that dispersion must now be considered.

Chromatic dispersion

After you have determined the value of the chromatic dispersion of the fiber, ensure that it is within the limits recommended by the International Telecommunications Union (ITU). ITU-T recommendations G.652, G.653, and G.655 specify the maximum chromatic dispersion coefficient. Assuming a zero-dispersion fiber at 1550 nanometers (nm) and an operating wavelength within 1525 to 1575 nm, the maximum allowed chromatic dispersion coefficient of the fiber is 3.5 ps/(nm-km). The total tolerable dispersion over a fiber span at 2.5 Gb/s is 16 000 ps, at 10 Gb/s it is 1000 ps, and at 40 Gb/s it is 60 ps.

Using these parameters, one can estimate the achievable link length. Using a 50 nm-wide optical source at 10 Gbit/s, and assuming that the optical fiber is at the 3.5 ps/(nm-km) limit, the maximum link length is 57 km. To show how link length, dispersion, and spectral width are related, see the following tables.

Table 9
Spectral width and link lengths assuming the maximum of 3.5 ps/(nm-km)

Spectral width (nm)	Maximum link length (km)
1	285
10	28.5
50	5.7

Table 10
Spectral widths and link lengths assuming an average fiber of 1.0 ps/(nm-km)

Spectral width (nm)	Maximum link length (km)
1	1000
10	100
50	20

If your fiber chromatic dispersion is over the limit, you can use chromatic dispersion compensating devices, for example, dispersion compensating optical fiber.

Polarization mode dispersion

Before you put an XFP into service for a long fiber, ensure that the fiber PMD is within the ITU recommendations. The ITU recommends that the total PMD of a fiber link not exceed 10% of the bit period. At 10 Gbit/s, this means that the total PMD of the fiber must not exceed 10 picoseconds (ps). At 40 Gbit/s, the total PMD of the link must not exceed 2.5 ps. For new optical fiber, manufacturers have taken steps to address fiber PMD. However, older, existing fiber plant may have high PMD values. For long optical links over older optical fibers, you should measure the PMD of the fiber proposed to carry 10 Gbit/s.

Table 11
PMD limits

Data rate	Maximum PMD of link (picoseconds)	Maximum PMD coefficient based on a 100 km-long fiber span (ps/sqrt-km)	Maximum PMD coefficient based on a 400 km-long fiber span (ps/sqrtkm)
1 Gbit/s	100	10	5.0
10 Gbit/s	10	1.0	0.5
40 Gbit/s	2.5	0.25	0.125

The dispersion of a fiber can change over time and with temperature change. If you measure fiber dispersion, measure it several times at different temperatures to determine the worst-case value. If you do not consider dispersion in your network design, you may experience an increase in the BER of your optical links.

PMD compensation is a new technology. Until compensating devices are commonly available, if the proposed fiber is over the PMD limit, you may have to use a different optical fiber.

10/100BASE-X and 1000BASE-TX reach

The following tables list maximum transmission distances for 10/100BASE-X and 1000BASE-TX Ethernet cables.

Table 12
10/100BASE-X and 1000BASE-TX maximum cable distances

	10BASE-T	100BASE-TX	100BASE-FX	1000BASE-TX
IEEE standard	802.3 Clause 14	802.3 Clause 21	802.3 Clause 26	802.3 Clause 40
Date rate	10 Mbit/s	100 Mbit/s	100 Mbit/s	1000 Mbit/s

Table 12
10/100BASE-X and 1000BASE-TX maximum cable distances (cont'd.)

	10BASE-T	100BASE-TX	100BASE-FX	1000BASE-TX
Multimode fiber distance	N/A	N/A	412 m (half-duplex) 2 km (full-duplex)	N/A
Cat 5 UTP distance	100 m	100 m	N/A	100 Ω,4pair: 100 m
STP/Coaxial cable distance	500 m	100 m	N/A	

10/100BASE-TX Auto-Negotiation recommendations

Auto-Negotiation lets devices share a link and automatically configures both devices so that they take maximum advantage of their abilities. Auto-Negotiation uses a modified 10BASE-T link integrity test pulse sequence to determine device ability.

The Auto-Negotiation function allows the devices to switch between the various operational modes in an ordered fashion and allows management to select a specific operational mode. The Auto-Negotiation function also provides a Parallel Detection (also called autosensing) function to allow 10BASE-T, 100BASE-TX, and 100BASE-T4 compatible devices to be recognized, even if they do not support Auto-Negotiation. In this case, only the link speed is sensed; not the duplex mode. Nortel recommends the Auto-Negotiation settings as shown in the following table, where A and B are two Ethernet devices.

Table 13
Recommended Auto-Negotiation setting on 10/100BASE-TX ports

Port on A	Port on B	Remarks	Recommendations
Auto-Negotiation enabled	Auto-Negotiation enabled	Ports negotiate on highest supported mode on both sides.	Recommended setting if both ports support Auto-Negotiation mode.
Full-duplex	Full-duplex	Both sides require the same mode.	Recommended setting if full-duplex is required, but Auto-Negotiation is not supported.

Auto-Negotiation cannot detect the identities of neighbors or shut down misconnected ports. These functions are performed by upper-layer protocols.

CANA

The R and RS modules support Custom Auto-Negotiation Advertisement (CANA). Use CANA to control the speed and duplex settings that the R and RS modules advertise during Auto-Negotiation sessions between Ethernet devices. Links can only be established using these advertised settings, rather than at the highest common supported operating mode and data rate.

Use CANA to provide smooth migration from 10/100 Mbit/s to 1000 Mbit/s on host and server connections. Using Auto-Negotiation only, the switch always uses the fastest possible data rates. In scenarios where uplink bandwidth is limited, CANA provides control over negotiated access speeds, and thus improves control over traffic load patterns.

CANA is supported on 10/100/1000 Mbit/s RJ-45 ports only. To use CANA, Auto-Negotiation must be enabled.

CANA is not supported on E modules. It is only supported on R or RS modules.

ATTENTION

If a port belongs to a Multilink Trunking (MLT) group and CANA is configured on the port (that is, an advertisement other than the default is configured), then the same configuration must be applied to all other ports of the MLT group (if they support CANA).

If a 10/100/1000 Mbit/s port that supports CANA is in a MLT group that has 10/100BASE-TX ports, or any other port type that do not support CANA, then CANA should be used only if it does not conflict with MLT abilities.

FEFI and remote fault indication

For information on Far End Fault Indication (FEFI), see “100BASE-FX FEFI recommendations” (page 75). For information on remote fault indication for Gigabit Ethernet, see “Gigabit Ethernet and remote fault indication” (page 76).

Control plane rate limit (CP-Limit)

Control plane rate limit (CP-Limit) controls the amount of multicast control traffic, broadcast control traffic, and exception frames that can be sent to the CPU from a physical port (for example, OSPF hello and RIP updates). It protects the CPU from being flooded by traffic from a single, unstable port. This differs from normal port rate limiting which limits non-control multicast traffic and non-control broadcast traffic on the physical port that would not be sent to the CPU (for example, IP subnet broadcast). The CP-Limit feature is configured on a per-port basis within the chassis.

The CP-Limit default settings are:

- Default state is enabled on all ports
- When creating the IST, CP-Limit is disabled automatically on the IST ports
- Default multicast packets-per-second value is 15000
- Default broadcast packets-per-second value is 10000

If the actual rate of packets-per-second sent from a port exceeds the defined rate, then the port is administratively shut down to protect the CPU from continued bombardment. An SNMP trap and a log file entry are generated indicating the physical port that has been shut down as well as the packet rate causing the shut down. To reactivate the port, you must first administratively disable the port and then reenable the port.

Having CP-Limit disable IST ports in this way could impair network traffic flow, as this is a critical port for SMLT configurations. Nortel recommends that an IST MLT contain at least two physical ports, although this is not a requirement. Nortel also recommends that CP-Limit be disabled on all physical ports that are members of an IST MLT. This is the default configuration. Disabling CP-Limit on IST MLT ports forces another, less critical port to be disabled if the defined CP-Limits are exceeded. In doing so, you preserve network stability should a protection condition (CP-Limit) arise. Please note that, although it is likely that one of the SMLT MLT ports (risers) would be disabled in such a condition, traffic would continue to flow uninterrupted through the remaining SMLT ports.

Extended CP-Limit

The Extended CP-Limit feature goes one step further than CP-Limit by adding the ability to read buffer congestion at the CPU as well as port level congestion on the I/O modules. This feature will protect the CPU from any traffic hitting the CPU by shutting down the ports which are responsible for sending traffic to CPU at a rate greater than desired.

To make use of Extended CP-Limit, configuration must take place at both the chassis and port level. The network administrator must predetermine the number of ports that should be monitored when congestion occurs. Extended CP-Limit can be enabled on all ports in the chassis, but when congestion is detected, Extended CP-Limit will monitor the most highly utilized ports in the chassis. The number of highly utilized ports monitored is configured in the MaxPorts parameter as described below.

When configuring Extended CP-Limit at the chassis level, the following parameters are available:

- MinCongTime (Minimum Congestion Time) sets the minimum time, in milliseconds, the CPU frame buffers can be oversubscribed for before triggering the congestion algorithm.
- MaxPorts (Maximum Ports) sets the total number of ports that need to be analyzed from the may-go-down port list.
- PortCongTime (Port Congestion Time) sets the maximum time, in seconds, a port's bandwidth utilization can exceed the threshold. When this timer is exceeded, the port is disabled - this parameter is only used by SoftDown.
- TrapLevel Sets the manner in which a SNMP trap is sent if a port becomes disabled.
  - — None - no traps are sent (default value)
  - — Normal - sends a single trap if ports are disabled.
  - — Verbose - sends a trap for each port that becomes disabled.

When configuring ext-cp-limit at the port level, the following parameters are available:

- HardDown disables the port immediately once the CPU frame buffers are congested for a certain period of time.
- SoftDown monitors the CPU frame buffer congestion and the port congestion time for a specified time interval - the ports are only disabled if the traffic does not subside after the time has been exceeded. The network administrator can configure the maximum number of SoftDown ports to be monitored.
- CplimitUtilRate defines the percentage of link bandwidth utilization to set as the threshold for the PortCongTime - this parameter is only used by SoftDown.

The following figures detail the flow logic of the HardDown and SoftDown operation of Extended CP-Limit.

Figure 4
Extended CP-Limit HardDown Operation

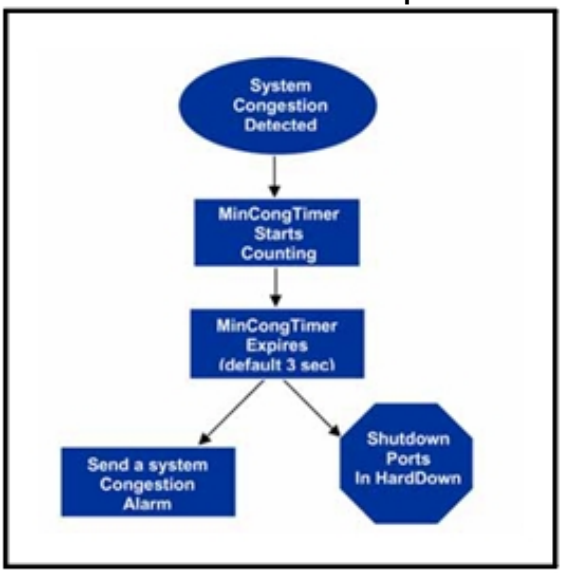
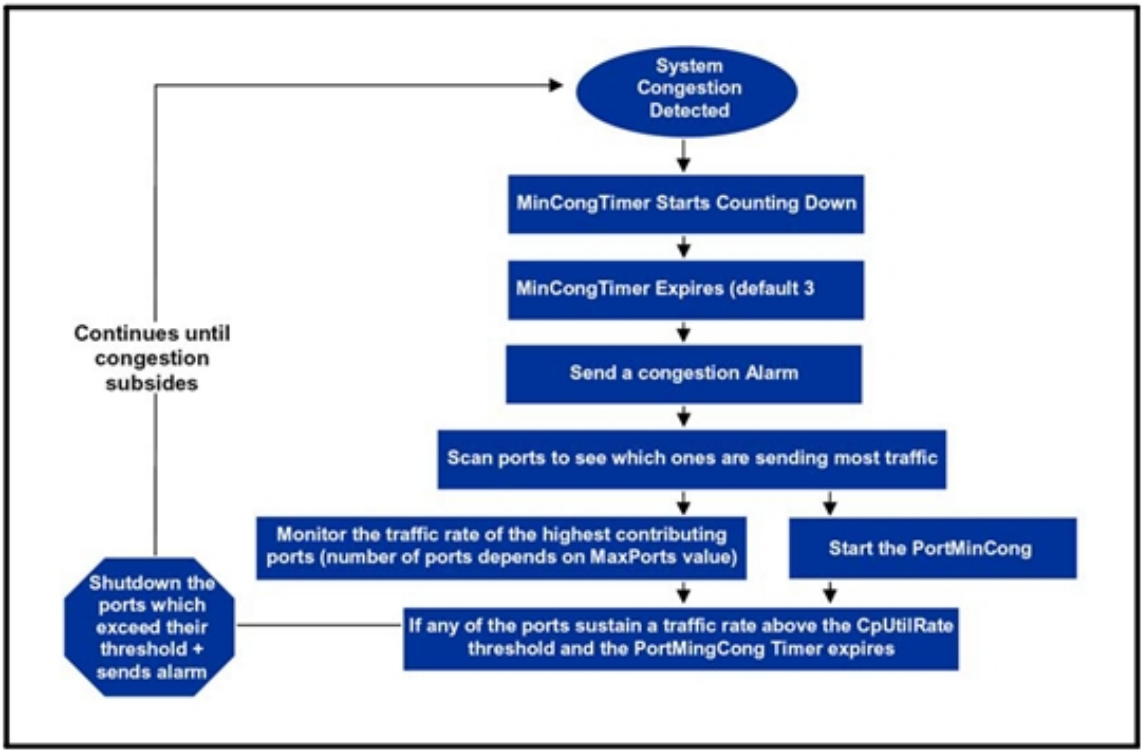


Figure 5
Extended CP-Limit SoftDown Operation



The following table describes the recommended CP-Limit and Extended CP-Limit usage by software release.

Software release	CP-Limit	Extended CP-Limit
3.7.0-3.7.4	Yes (See Note 2)	N/A
3.7.5-3.7.x	Yes (See Note 2)	Yes (HardDown) (See Note 1)
4.0.x	Yes (See Note 2)	N/A
4.1.x-5.1.x	Yes (See Note 2)	Yes (SoftDown) (See Note 2)
Note 1: Loop Protection Note 2: CPU Protection		

For information about using CP-Limit and Extended CP-Limit with SLPP and VLACP, see "SLPP, Loop Detect, and Extended CP-Limit" (page 148).

For more information about CP-Limit and Extended CP-Limit, see *Nortel Ethernet Routing Switch 8600 Administration* (NN46205-605) .

Optical routing design

Use the Nortel optical routing system to maximize bandwidth on a single optical fiber. This section provides optical routing system information that you can use to help design your network.

Navigation

- "Optical routing system components" (page 55)
- "Multiplexer applications" (page 58)
- "Transmission distance" (page 61)
- "DWDM XFPs" (page 68)

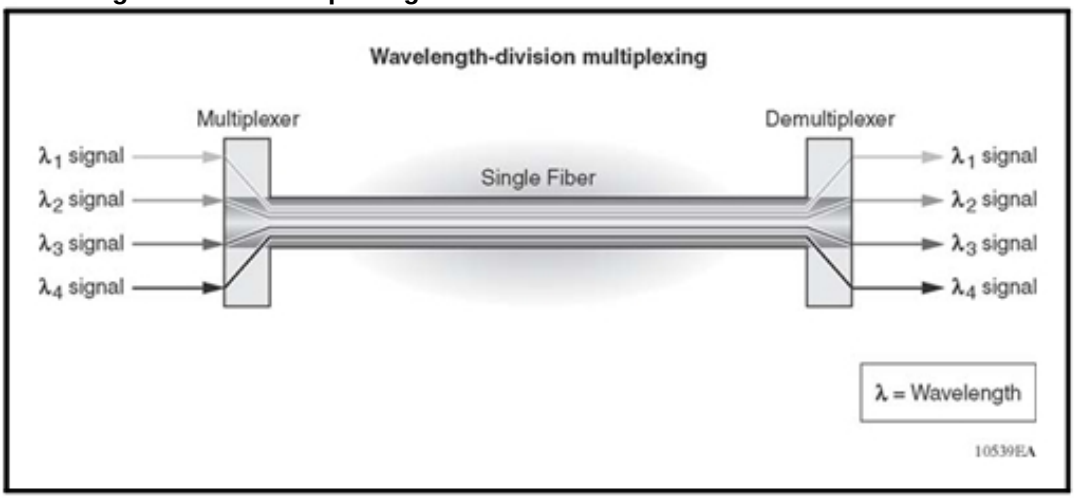
Optical routing system components

The Nortel optical routing system uses coarse wavelength division multiplexing (CWDM) in a grid of eight optical wavelengths. CWDM Gigabit Interface Converters (GBICs) and Small Form Factor Pluggable (SFP) transceivers transmit optical signals from Gigabit Ethernet ports to multiplexers in a passive optical shelf.

Multiplexers combine multiple wavelengths traveling on different fibers onto a single fiber. At the receiver end of the link, demultiplexers separate the wavelengths and route them to different fibers, which terminate at separate CWDM devices. The following figure shows multiplexer and demultiplexer operations.

ATTENTION
For clarity, the following figure shows a single fiber link with signals traveling in one direction only. A duplex connection requires communication in the reverse direction as well.

Figure 6
Wavelength division multiplexing



The Nortel optical routing system supports both ring and point-to-point configurations. The optical routing system includes the following parts:

- CWDM GBICs
- CWDM SFPs
- Optical add/drop multiplexers (OADM)
- Optical multiplexer/demultiplexers (OMUX)
- Optical shelf to house the multiplexers

OADMs drop or add a single wavelength from or to an optical fiber.

The following table describes the parts of the optical routing system and the color matching used. The compatible optical shelf part number is AA1402001-E5.

Table 14
Parts of the optical routing system

Wavelength	GBIC and SFP part numbers	Multiplexer part number		
		OADM	OMUX-4	OMUX-8
1470 nanometers (nm) Gray	AA1419017-E5, up to 120 kilometers (km) GBIC AA1419025-E5, up to 40 km SFP AA1419033-E5, up to 70 km SFP AA1419053-E6, up to 40 km DDI SFP AA1419061-E6, up to 70 km DDI SFP	AA1402002 -E5		AA1402010 -E5

Wavelength	GBIC and SFP part numbers	Multiplexer part number		
		OADM	OMUX-4	OMUX-8
1490 nm Violet	AA1419018-E5, up to 120 km GBIC AA1419026-E5, up to 40 km SFP AA1419034-E5, up to 70 km SFP AA1419054-E6, up to 40 km DDI SFP AA1419062-E6, up to 70 km DDI SFP	AA1402003 -E5	AA1402009 -E5	
1510 nm Blue	AA1419019-E5, up to 120 km GBIC AA1419027-E5, up to 40 km SFP AA1419035-E5, up to 70 km SFP AA1419055-E6, up to 40 km DDI SFP AA1419063-E6, up to 70 km DDI SFP	AA1402004 -E5		
1530 nm Green	AA1419020-E5, up to 120 km GBIC AA1419028-E5, up to 40 km SFP AA1419036-E5, up to 70 km SFP AA1419056-E6, up to 40 km DDI SFP AA1419064-E6, up to 70 km DDI SFP	AA1402005 -E5	AA1402009 -E5	
1550 nm Yellow	AA1419021-E5, up to 120 km GBIC AA1419029-E5, up to 40 km SFP AA1419037-E5, up to 70 km SFP AA1419057-E6, up to 40 km DDI SFP AA1419065-E6, up to 70 km DDI SFP	AA1402006 -E5		
1570 nm Orange	AA1419022-E5, up to 120 km GBIC AA1419030-E5, up to 40 km SFP AA1419038-E5, up to 70 km SFP	AA1402007 -E5	AA1402009 -E5	

Wavelength	GBIC and SFP part numbers	Multiplexer part number		
		OADM	OMUX-4	OMUX-8
	AA1419058-E6, up to 40 km DDI SFP AA1419066-E6, up to 70 km DDI SFP			
1590 nm Red	AA1419023-E5, up to 120 km GBIC AA1419031-E5, up to 40 km SFP AA1419039-E5, up to 70 km SFP AA1419059-E6, up to 40 km DDI SFP AA1419067-E6, up to 70 km DDI SFP	AA1402008 -E5		
1610 nm Brown	AA1419024-E5, up to 120 km GBIC AA1419032-E5, up to 40 km SFP AA1419040-E5, up to 70 km SFP AA1419060-E6, up to 40 km DDI SFP AA1419068-E6, up to 70 km DDI SFP	AA1402011 -E5	AA1402009 -E5	

For more information about multiplexers, SFPs and GBICs, including technical specifications and installation instructions, see *Nortel Ethernet Routing Switch 8600 Installation — SFP, XFP, GBIC, and OADM Hardware Components* (NN46205-320) .

Multiplexer applications

Use OADMs to add and drop wavelengths to and from an optical fiber. Use multiplexers to combine up to eight wavelengths on a single fiber. This section describes common applications for the OADM and OMUX.

Multiplexer application navigation

- “OADM ring” (page 59)
- “Optical multiplexer in a point-to-point application” (page 60)
- “OMUX in a ring” (page 61)

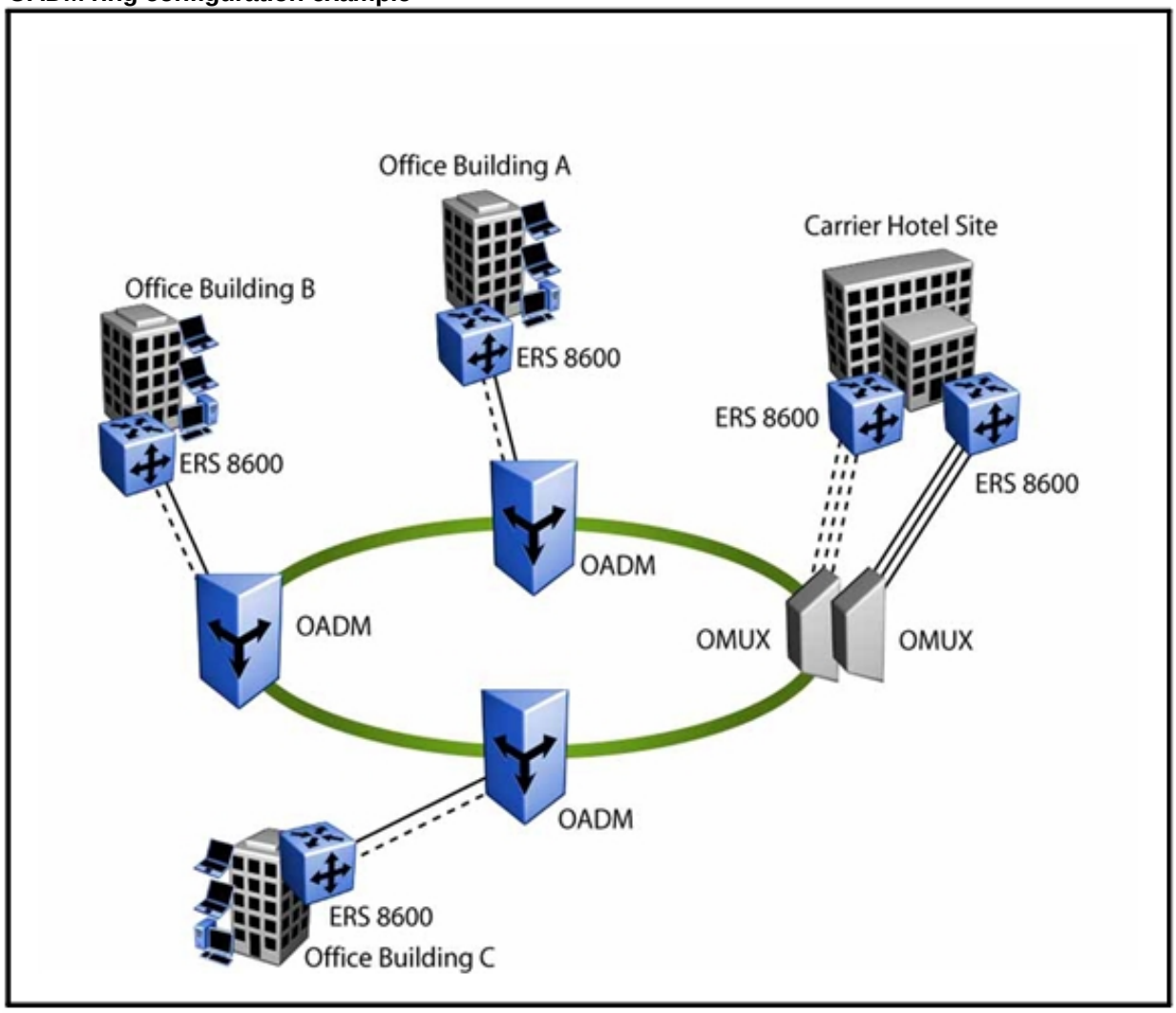
OADM ring

The OADM removes (or adds) a specific wavelength from an optical ring and passes it to (or from) a GBIC or SFP of the same wavelength, leaving all other wavelengths on the ring undisturbed. OADMs are set to one of eight supported wavelengths.

ATTENTION
The wavelength of the OADM and the corresponding GBIC or SFP must match.

The following figure shows an example of two separate fiber paths in a ring configuration traveling in opposite (east/west) directions into the network.

Figure 7
OADM ring configuration example



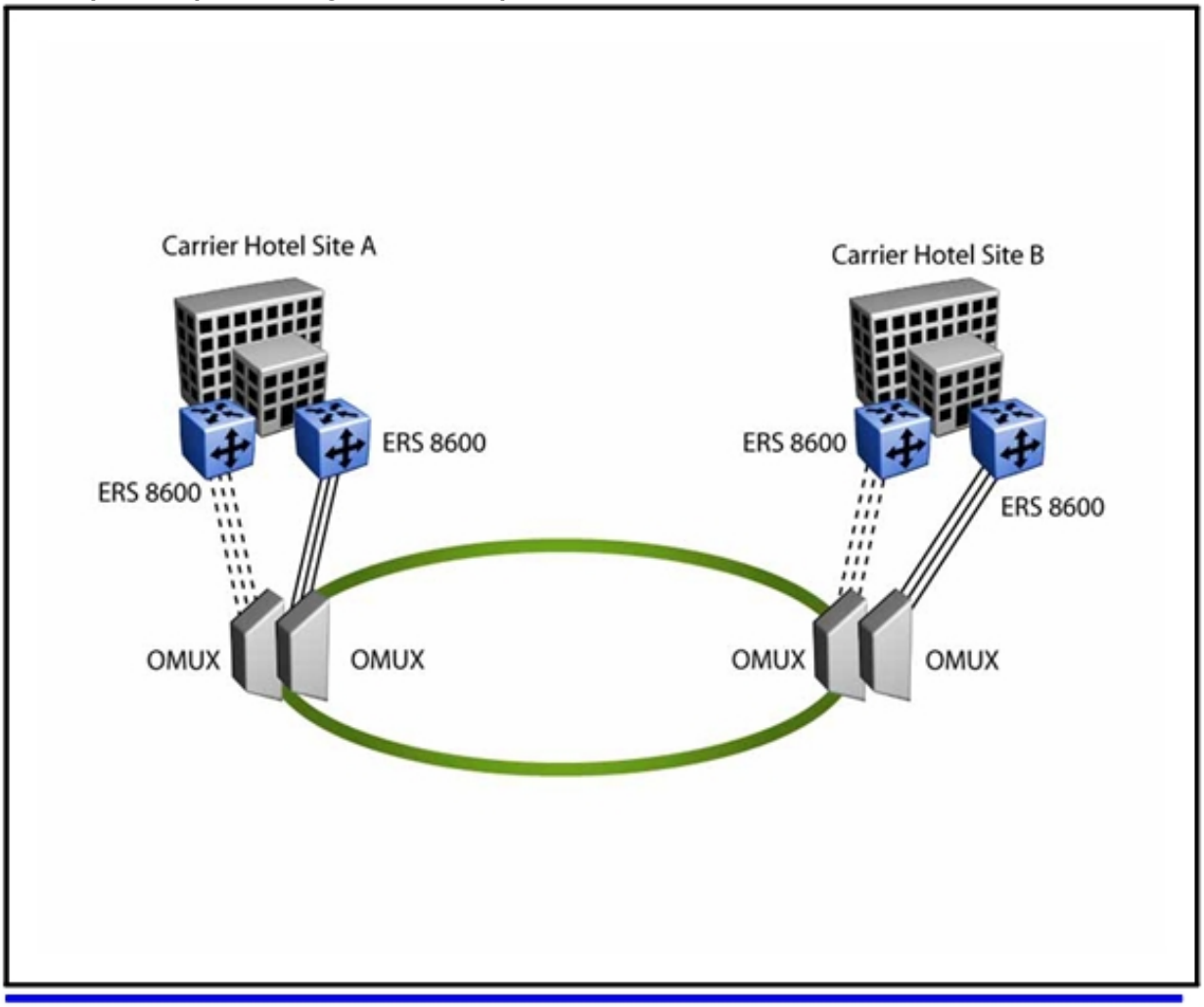
For information about calculating network transmission distance, see "Transmission distance" (page 61).

Optical multiplexer in a point-to-point application

Point-to-Point (PTP) optical networks carry data directly between two end points without branching out to other points or nodes. PTP connections (see the following figure) are made between mux/demuxs at each end. PTP connections transport many gigabits of data from one location to another to support applications, such as the linking of two data centers to become one virtual site, the mirroring of two sites for disaster recovery, or the provision of a large amount of bandwidth between two buildings. The key advantage of a PTP topology is the ability to deliver maximum bandwidth over a minimum amount of fiber.

Each CWDM optical multiplexer/demultiplexer (OMUX) supports one network backbone connection and four or eight connections to GBICs or SFPs. Typically, two OMUXs are installed in a chassis. The OMUX on the left is called the east path, and the OMUX on the right is called the west path.

Figure 8
OMUX point-to-point configuration example



OMUX in a ring

OMUXs are also used as the hub site in OMUX-based ring applications (see Figure 7 "OADM ring configuration example" (page 59)). Two OMUXs are installed in the optical shelf at the central site to create an east and a west fiber path. The east OMUX terminates all the traffic from the east equipment port of each OADM on the ring, and the west OMUX terminates all of the traffic from the west equipment port of each OADM on the ring. In this configuration, the network remains viable even if the fiber is broken at any point on the ring.

Transmission distance

To ensure proper network operation, given your link characteristics, calculate the maximum transmission distance for your fiber link.

Transmission distance navigation

- "Reach and optical link budget" (page 61)
- "Reach calculation examples" (page 62)

Reach and optical link budget

The absorption and scattering of light by molecules in an optical fiber causes the signal to lose intensity. Expect attenuation when you plan an optical network.

Factors that typically affect optical signal strength include:

- optical fiber attenuation (wavelength dependent: typically 0.20 to 0.35 dB/km)
- network devices the signal passes through
- connectors
- repair margin (user-determined)

The loss budget, or optical link budget, is the amount of optical power launched into a system that you can expect to lose through various system mechanisms. By calculating the optical link budget, you can determine the transmission distance (reach) of the link (that is, the amount of usable signal strength for a connection between the point where it originates and the point where it terminates).

ATTENTION
Insertion loss budget values for the optical routing system CWDM OADM and OMUX include connector loss.

Reach calculation examples

The examples in this chapter use the following assumptions and procedure for calculating the maximum transmission distances for networks with CWDM components.

The examples assume the use of the values and information listed in the following table. Use the expected repair margin specified by your organization. For GBIC, SFP, XFP, and multiplexer specifications, see *Nortel Ethernet Routing Switch 8600 Installation — SFP, XFP, GBIC, and OADM Hardware Components* (NN46205-320) . Multiplexer loss values include connector loss.

Attenuation of 0.25 dB/km is used, but the typical attenuation at 1550 nm is about 0.20 dB/km. Be sure to use the appropriate value for your network.

Table 15
Assumptions used in calculating maximum transmission distance

Parameter	Value
Cable	Single mode fiber (SMF)
Repair margin	0 dB
Maximum link budget	30 dB
System margin	3 dB (allowance for miscellaneous network loss)
Fiber attenuation	.25 dB/km
Operating temperature	0 to 40 C (32 to 104 F)
CWDM OADM expected loss	Use OADM specifications
CWDM OMUX expected loss	Use OMUX specifications

To calculate the maximum transmission distance for a proposed network configuration:

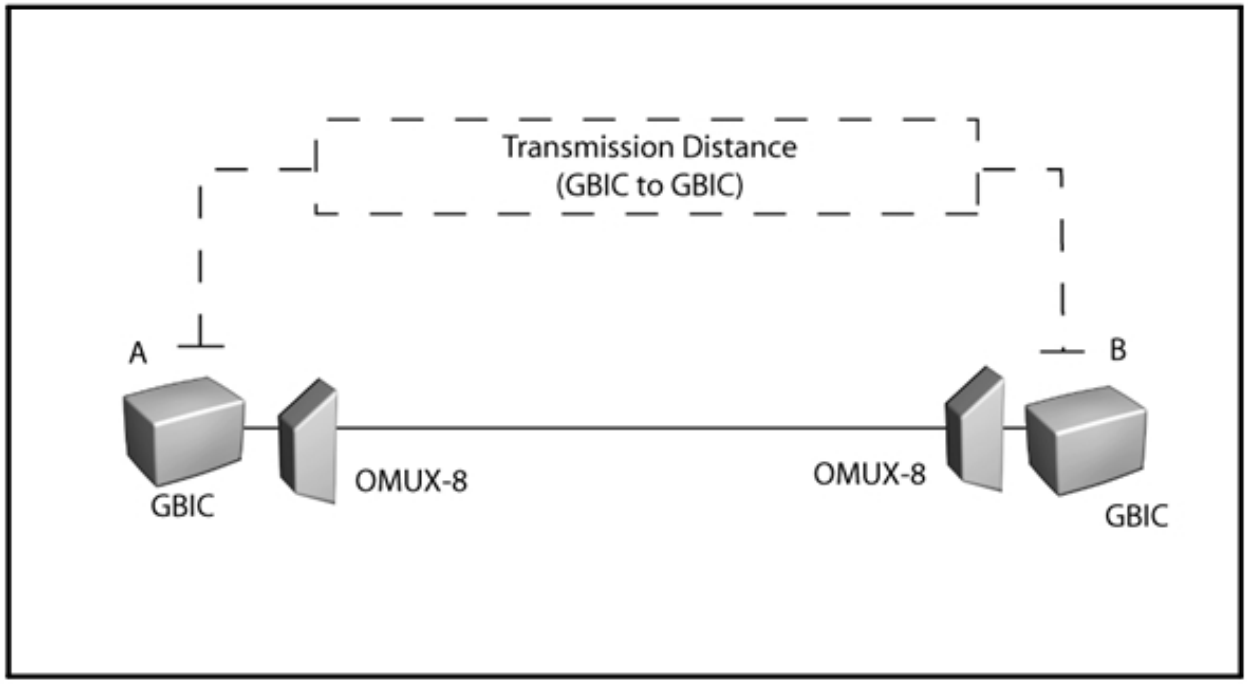
- Identify all points where signal strength is lost.
- Calculate, in dB, the expected loss for each point.
- Find the total passive loss by adding the expected losses together.
- Find the remaining signal strength by subtracting the passive loss and system margin from the total system budget.
- Find the maximum transmission distance by dividing the remaining signal strength by the expected fiber attenuation in dB/km.

Point-to-point reach example

The following factors affect signal strength and determine the point-to-point link budget and the maximum transmission distance for the network shown in the following figure.

- OMUX multiplexer (mux) loss
- OMUX demultiplexer (demux) loss
- Fiber attenuation

Figure 9
Point-to-point network configuration example



The Ethernet switch does not have to be near the OMUX, and the OMUX does not regenerate the signal. Therefore, the maximum transmission distance is from GBIC to GBIC.

The following table shows typical loss values used to calculate the transmission distance for the point-to-point network.

Table 16
Point-to-point signal loss values

Parameter	Value (dB)
Loss budget	30 dB
OMUX-8 mux loss	3.5 dB
OMUX-8 demux loss	4.5 dB

Table 16
Point-to-point signal loss values (cont'd.)

Parameter	Value (dB)
System margin	3.0 dB
Fiber attenuation	.25 dB/km

The equations and calculations used to determine maximum transmission distance for the point-to-point network example are:

Passive loss = mux loss + demux loss
Implied fiber loss = loss budget - passive loss - system margin
Maximum transmission distance = implied fiber loss/attenuation

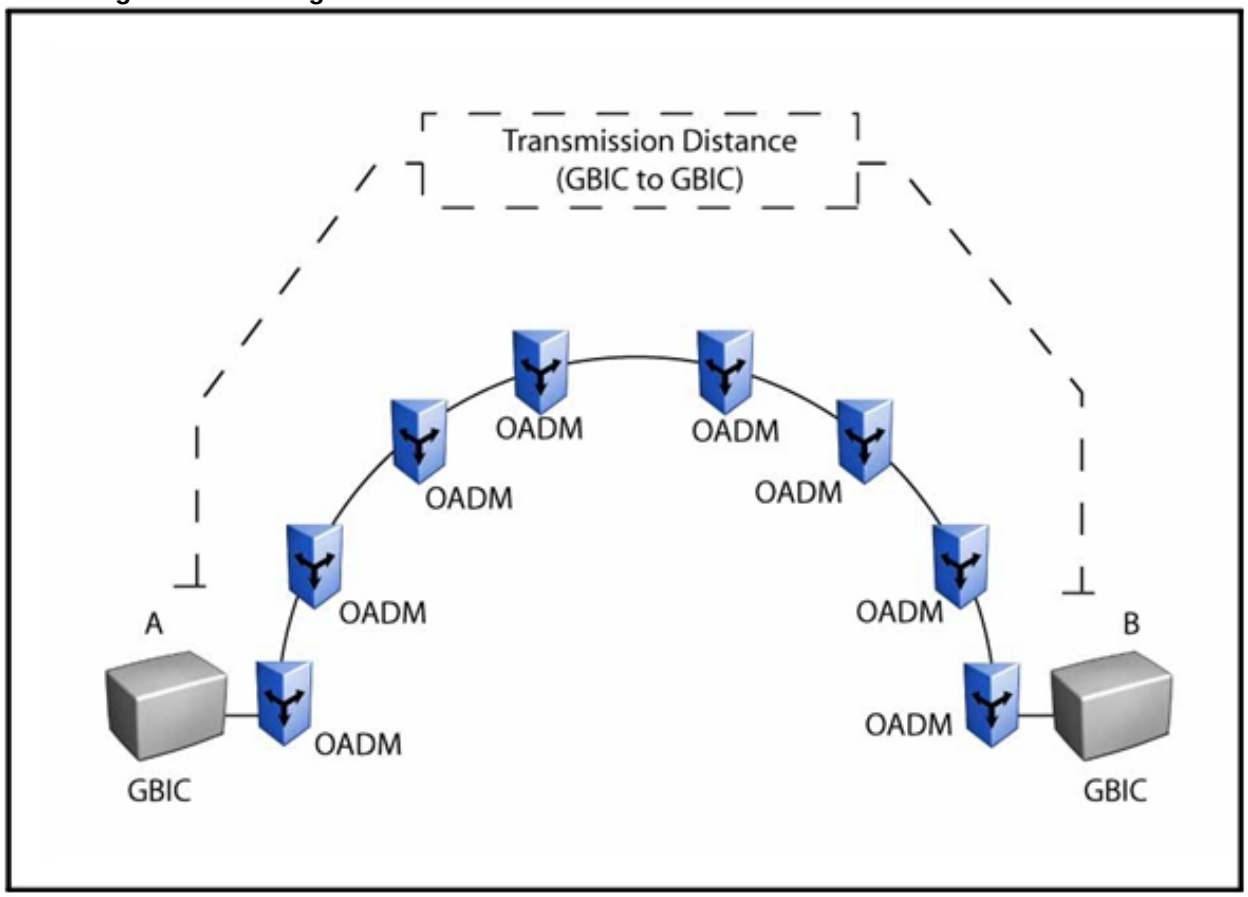
In this case:

Passive loss = 3.5 + 4.5 = 8.0 dB
Implied fiber loss = 30 - 8 - 3 = 19 dB
Maximum reach = (19 dB) / (0.25 dB/km) = 76 km

Mesh ring reach example

The transmission distance calculation for the mesh ring configuration shown in the following figure is similar to that of the point-to-point configuration, with some additional loss generated in the passthrough of intermediate OADM nodes.

Figure 10
Mesh ring network configuration



As the signal passes from point A to point B (the most remote points in the mesh ring network example), the signal loses intensity in the fiber optic cable, and in each connection between the individual OADMs and GBICs.

The following factors determine the mesh ring link budget and the transmission distance for the network:

- OADM insertion loss for Add port
- OADM insertion loss for Drop port
- OADM insertion loss for Through port at intermediate nodes
- Fiber attenuation of 0.25 dB/km

The maximum transmission distance is from GBIC to GBIC.

The number of OADMs that can be supported is based on loss budget calculations.

The following table shows the typical loss values used to calculate the transmission distance for the mesh ring network example.

Table 17
Mesh ring signal loss values

Parameter	Value
Loss budget	30 dB
OADM insertion loss for Add port	1.9 dB
OADM insertion loss for Through port	2.0 dB
OADM insertion loss for Drop port	2.3 dB
System margin	3.0 dB
Fiber attenuation	.25 dB/km

The equations and calculations used to determine the maximum transmission distance for this network example are:

Passthrough nodes = nodes - 2
Passive loss = OADM add + OADM drop + (passthrough nodes*OADM passthrough loss)
Implied fiber loss = loss budget - passive loss - system margin
Maximum transmission distance = implied fiber loss/attenuation

In this case:

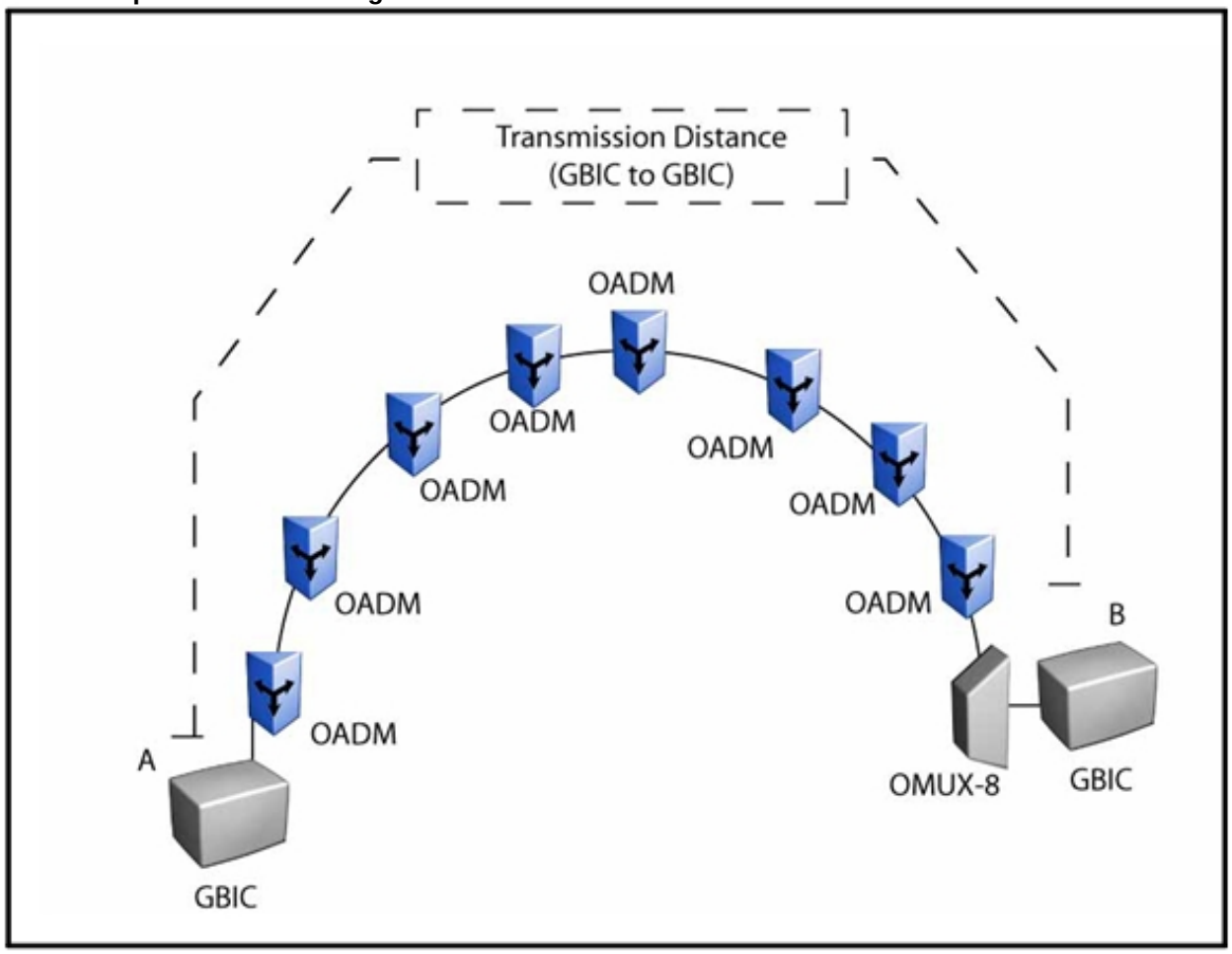
Passthrough nodes = 8 - 2 = 6 nodes
Passive loss = 1.9 + 2.3 + (6*2.0)= 16.2 dB
Implied fiber loss = 30 - 16.2 - 3 = 10.8 dB
Maximum reach = (10.8 dB) / (0.25 dB/km) = 43.2 km

Hub-and-spoke reach example

Hub-and-spoke topologies are complex. The characteristics of all components designed into the network must be considered in calculating the transmission distance. The following factors determine the maximum transmission distance for the configuration shown in the following figure.

- OADM insertion loss for Add port
- OADM insertion loss for Drop port
- OADM insertion loss for Through port for intermediate nodes
- Fiber attenuation of 0.25 dB/km

Figure 11
Hub and spoke network configuration



As the signal passes from point A to point B (the most remote points), it loses intensity in the fiber optic cable, and in each connection between the individual OADMs, the OMUX-8, and the GBICs. The number of OADMs that can be supported is based on the loss budget calculations.

The following table shows typical loss values used to calculate the transmission distance for the hub and spoke network.

Table 18
Hub and spoke signal loss values

Parameter	Value
Loss budget	30 dB
OADM insertion loss for Add port	1.9 dB
OADM insertion loss for Through port	2.0 dB

Table 18
Hub and spoke signal loss values (cont'd.)

Parameter	Value
OMUX-8 demux loss	4.5 dB
System margin	3.0 dB
Fiber attenuation	.25 dB/km

The equations and calculations used to determine maximum transmission distance for the network example are:

Passthrough nodes = number of OADMs between first OADM and OMUX
Passive loss = OADM add + OMUX-8 demux+ (passthrough nodes*OADM passthrough loss)
Implied fiber loss = loss budget - passive loss - system margin
Maximum transmission distance = implied fiber loss/attenuation

In this case:

Passthrough nodes = 7 nodes
Passive loss = 1.9 + 4.5 + (67*2.0)= 20.4 dB
Implied fiber loss = 30 - 20.4 - 3 = 6.6 dB
Maximum reach = (6.6 dB) / (0.25 dB/km) = 26.4 km

DWDM XFPs

The Ethernet Routing Switch 8600 provides support for DWDM XFP devices on all 10 Gigabit ports for R/RS modules (8683XLR, 8683XZR, 8612XLRS, 8634XGRS). The Ethernet Routing Switch 8600 can support 10 Gigabit, frequency multiplexed, direct connections to Nortel CPL rings. The 8006, 8010 and 8010co chassis support the required cooling of the DWDM XFP devices in all XFP ports in all slots.

As shown in the following table, Release 5.1 supports 20 DWDM XFPs with different Lambdas. These are C band wavelengths, with 100 GHz spacing with both ITU frequencies (THz) and ITU wavelenghts (nm) specified.

Table 19
Supported DWDM XFPs

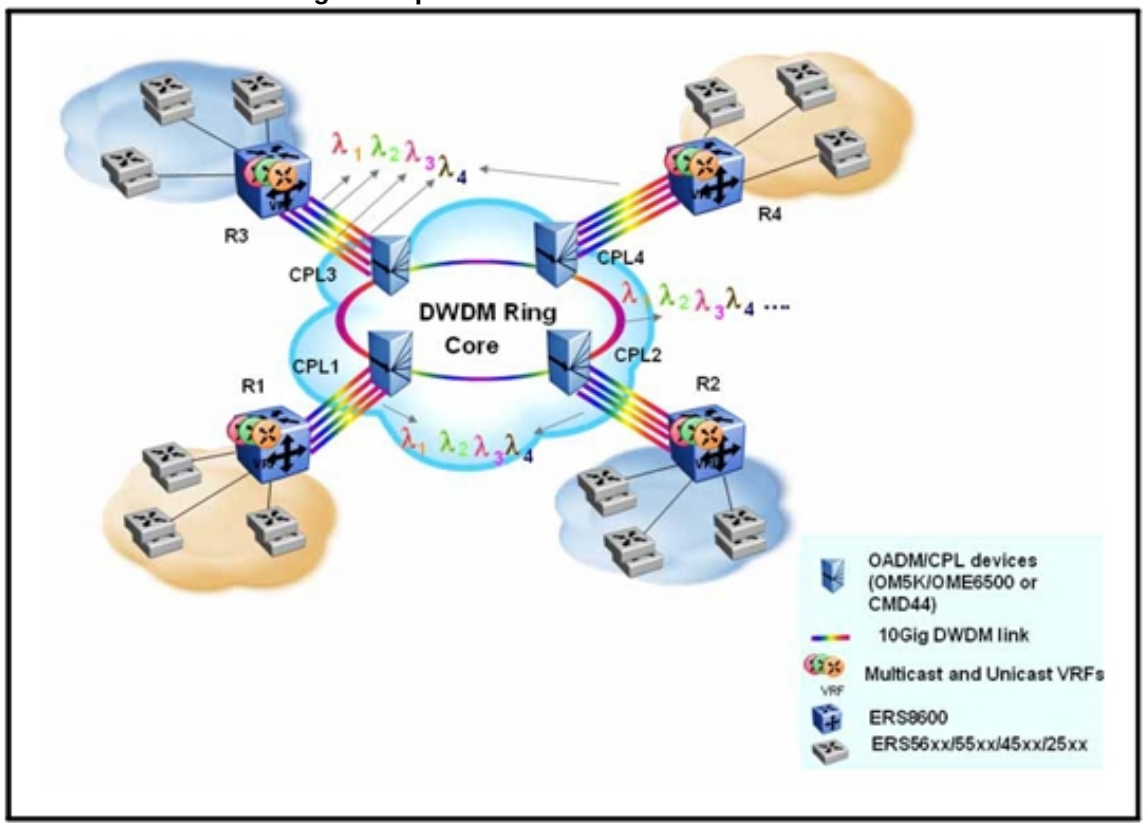
Product number	Centre wavelength (nm)	Centre wavelength (THz)
NTK587AEE5	1530.33	195.90
NTK587AGE5	1531.12	195.80
NTK587AJE5	1531.90	195.70
NTK587ALE5	1532.68	195.60

Table 19
Supported DWDM XFPs (cont'd.)

Product number	Centre wavelength (nm)	Centre wavelength (THz)
NTK587ANE5	1533.47	195.50
NTK587AQE5	1534.25	195.40
NTK587ASE5	1535.04	195.30
NTK587AUE5	1535.82	195.20
NTK587AWE5	1536.61	195.10
NTK587AYE5	1537.4	195.0
NTK587BAE5	1538.19	194.9
NTK587BCE5	1538.98	194.8
NTK587BEE5	1539.77	194.7
NTK587BGE5	1540.56	194.6
NTK587BJE5	1541.35	194.5
NTK587BLE5	1542.14	194.4
NTK587BNE5	1542.94	194.3
NTK587BQE5	1543.73	194.2
NTK587BSE5	1544.53	194.1
NTK587BUE5	1545.32	194.0

The following figure shows a sample network topology using DWDM XFPs in a large enterprise.

Figure 12
DWDM XFPs for MSO/large enterprise



Software considerations

The software you install, in conjunction with the hardware present in the chassis and the operation mode, determine the features available on the switch. Use this section to help you determine which modes to use.

Navigation

- “Operational modes” (page 71)

Operational modes

An Ethernet Routing Switch 8600 can run in the following hardware operating modes:

- Default mode (32 000 table entries) supports up to 32 000 hardware records. This mode supports all modules.
- Mmode(128000tableentries)supportsupto128000hardware records. This mode supports M and R series modules only. If an E module is installed in a chassis with M mode enabled, the E module is disabled. This protects the system forwarding database from inconsistencies.
- Rmodesupportsupto256000IProutes,64000MACentries,and 32000 Address Routing Protocol (ARP) entries. This mode supports R and RS modules only. With R mode enabled, any E or M modules installed in the chassis are disabled.

M mode allows increased record scalability. For M modules, Nortel strongly recommends that you use 8691 or 8692 SF/CPUs (otherwise, the chassis operates in normal mode). Based on the internal hardware architecture, when using the 10GE modules, Nortel further recommends that you employ two 8692 SF/CPUs for traffic balancing and redundancy.

To enable certain features such as Feedback Output Queueing (FOQ), MultiLink Trunking (MLT), and Equal Cost Multi-Path (ECMP), you must enable R mode. With Software Release 4.1 and later, R and RS modules support up to 128 MLT groups and up to eight ECMP routing paths.

The switch can additionally operate in the following modes:

- Enhanced Operational Mode increases the maximum number of Virtual Local Area Networks (VLANs) when using MultiLink Trunking (MLT) (1980 or 1972 VLANs) and Split MultiLink Trunking (SMLT) (989 VLANs).

  Release 5.0 and later supports up to 4000 VLANs with a default of 1972.

- VLAN optimization mode supports all modules, except the 8648TX and 8648TXE modules. VLAN optimization mode is not applicable to R mode.

The following table lists features supported per operation mode. In this table, *EN* denotes enabled, and *DIS* denotes disabled.

Table 20
Features supported per operation mode

Chassis configuration	Operation mode	Module type		
		R	**M**	**E**
All same type modules	Default	—	—	EN
	M	—	EN	—
	R	EN	—	—
Mixed module types	Default (32000 records)	EN	EN	EN
	M (128000 records)	EN	EN	DIS
	R (256000 records)	EN	DIS	DIS

Enhanced operational mode

When enhanced operational mode is enabled, only E modules, M modules, and R series modules are initialized. Enhanced Operational Mode increases the maximum number of VLANs when using MultiLink Trunking (MLT) (1980 or 1972 VLANs) and Split MultiLink Trunking (SMLT) (989 VLANs). VLAN scaling is reduced if you use multicast MAC filters.

The boot mode is determined by the modules in the chassis and whether the enhanced operational mode is enabled.

If enhanced operational mode is disabled, the system starts in nonenhanced operational mode. The enabling or disabling of, and the hardware operating mode of the module, depends on the module configuration in the chassis.

When you insert a module into a running chassis, the enhanced operational mode status determines the initialization mode of the module.

Enhanced operational mode configuration is not relevant in R-mode. The system always operates in enhanced mode.

Redundant network design

Provide redundancy to eliminate a single point of failure in your network. This section provides guidelines that help you design redundant networks.

Navigation

Physical layer redundancy

Provide physical layer redundancy to ensure that a faulty link does not cause a service interruption. You can also configure the switch to detect link failures.

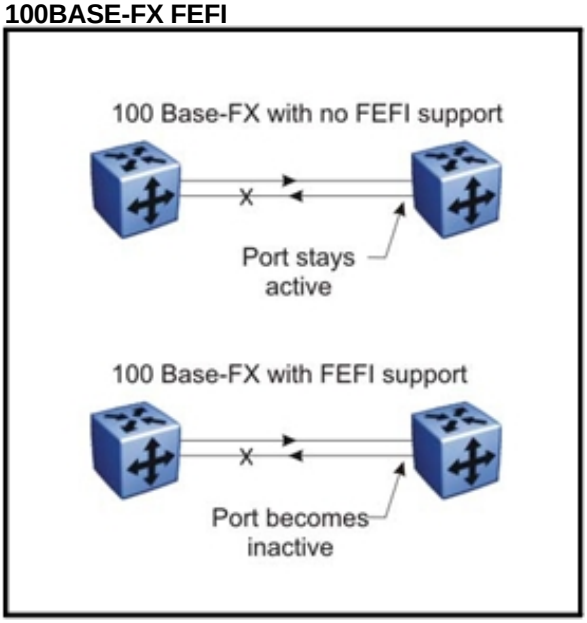
Physical layer redundancy navigation

100BASE-FX FEFI recommendations

The Ethernet Routing Switch 8600 supports Far End Fault Indication (FEFI). FEFI ensures that link failures are reported to the switch. FEFI is enabled when the Auto-Negotiation function is enabled. However, not all 100BASE-FX drivers support FEFI. Without FEFI support, if one of two unidirectional fibers forming the connection between the two switches fails, the transmitting side cannot determine that the link is broken in one

direction (see Figure 13 "100BASE-FX FEFI" (page 76)). This leads to network connectivity problems because the transmitting switch keeps the link active as it still receives signals from the far end. However, the outgoing packets are dropped because of the failure.

Figure 13
100BASE-FX FEFI



With Nortel-to-Nortel connections, to avoid loss of connectivity for devices that do not support FEFI, you can use VLACP as an alternative failure detection method. For more information, see “End-to-end fault detection and VLACP” (page 77).

Gigabit Ethernet and remote fault indication

The 802.3z Gigabit Ethernet standard defines remote fault indication (RFI) as part of the Auto-Negotiation function. RFI provides a means for the stations on both ends of a fiber pair to be informed when a problem occurs on one of the fibers. Because RFI is part of the Auto-Negotiation function, if Auto-Negotiation is disabled, RFI is automatically disabled. Therefore, Nortel recommends that Auto-Negotiation be enabled on Gigabit Ethernet links when Auto-Negotiation is supported by the devices on both ends of a fiber link.

For information about Auto-Negotiation for 10 and 100 Mbit/s links, see “10/100BASE-TX Auto-Negotiation recommendations” (page 48).

SFFD recommendations

The Ethernet switching devices listed in the following table do not support Auto-Negotiation on fiber-based Gigabit Ethernet ports. These devices are unable to participate in remote fault indication (RFI), which is a part of the Auto-Negotiation specification. Without RFI, and in the event of a

single fiber strand break, one of the two devices may not detect a fault, and continues to transmit data even though the far-end device does not receive it.

Table 21
Ethernet switching devices that do not support Auto-Negotiation

Switch name / Part number	Port or MDA type / Part number
Ethernet Switch 470-48T (AL2012x34) Ethernet Switch 470-24T (AL2012x37)	SX GBIC (AA1419001)
	LX GBIC (AA1419002)
	XD GBIC (AA1419003)
	ZX GBIC (AA1419004)
Ethernet Switch 460-24T-PWR (AL20012x20)	2-port SFP GBIC MDA (AL2033016)
OM1200 (AL2001x19)	2-port SFP GBIC MDA (AL2033016)
OM1400 (AL2001x22)	2-port SFP GBIC MDA (AL2033016)
OM1450 (AL2001x21)	2-port SFP GBIC MDA (AL2033016)

If you must connect the switch to a device that does not support Auto-Negotiation, you can use Single-fiber Fault Detection (SFFD). SFFD can detect single fiber faults and bring down faulty links immediately. If the port is part of a multilink trunk (MLT), traffic fails over to other links in the MLT group. Once the fault is corrected, SFFD brings the link up within 12 seconds. For SFFD to work properly, both ends of the fiber connection must have SFFD enabled and Auto-Negotiation disabled.

A better alternative to SFFD is VLACP (see "End-to-end fault detection and VLACP" (page 77)).

End-to-end fault detection and VLACP

A major limitation of the RFI and FEFI functions is that they terminate at the next Ethernet hop. Therefore, failures cannot be determined on an end-to-end basis over multiple hops.

To mitigate this limitation, Nortel has developed a feature called Virtual LACP (VLACP), that provides an end-to-end failure detection mechanism. With VLACP, far-end failures can be detected. This allows MLT to properly failover when end-to-end connectivity is not guaranteed for certain links in an aggregation group.

VLACP allows you to switch traffic around entire network devices before Layer 3 protocols detect a network failure, thus minimizing network outages.

VLACP operation

Virtual Link Aggregation Control Protocol (VLACP) is an extension to LACP used for end-to-end failure detection. VLACP is not a link aggregation protocol. It is a mechanism to periodically check the end-to-end health of a point-to-point connection. VLACP uses the Hello mechanism of LACP to periodically send Hello packets to ensure an end-to-end communication. When Hello packets are not received, VLACP transitions to a failure state, which indicates a service provider failure and that the port is disabled.

The VLACP only works for port-to-port communications where there is a guarantee for a logical port-to-port match through the service provider. VLACP does not work for port-to-multiport communications where there is no guarantee for a point-to-point match through the service provider. You can configure VLACP on a port.

VLACP can also be used with MLT to complement its capabilities and provide quick failure detection. VLACP is recommended for all SMLT access links when the links are configured as MLT to ensure both end devices are able to communicate. By using VLACP over SLT, enhanced failure detection is extended beyond the limits of the number of SMLT or LACP instances that can be created on a Nortel switch.

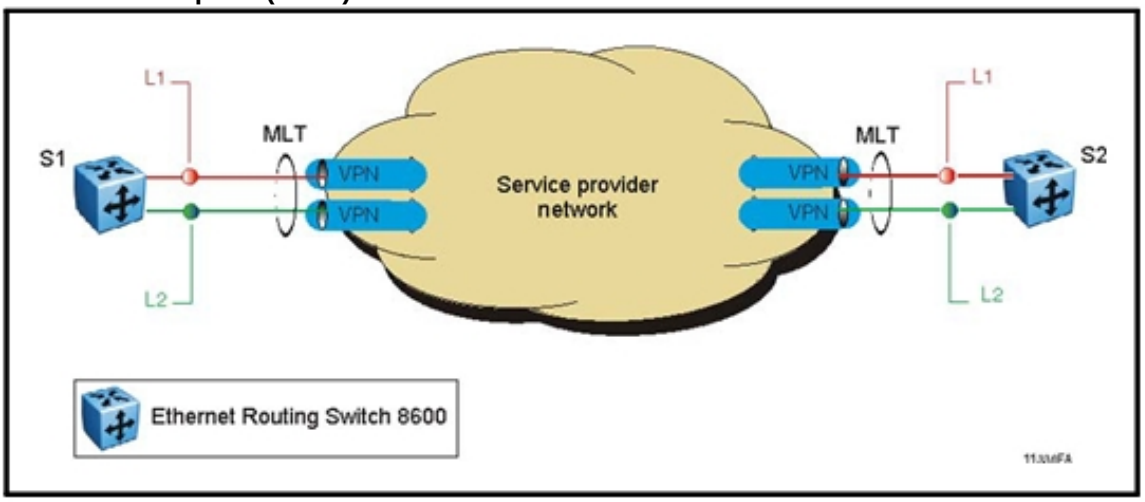
VLACP trap messages are sent to the management stations if the VLACP state changes. If the failure is local, the only traps that are generated are port linkdown or port linkup.

The Ethernet cannot detect end-to-end failures. Functions such as remote fault indication or far-end fault indication extend the Ethernet to detect remove link failures. A major limitation of these functions is that they terminate at the next Ethernet hop. They cannot determine failures on an end-to-end basis.

For example, in Figure 14 "Problem description (1 of 2)" (page 79) when the Enterprise networks connect the aggregated Ethernet trunk groups through a service provider network connection (for example, through a VPN), far-end failures cannot be signaled with Ethernet-based functions that operate end-to-end through the service provider network. The MultiLink trunk (between Enterprise switches S1 and S2) extends through the Service Provider (SP) network.

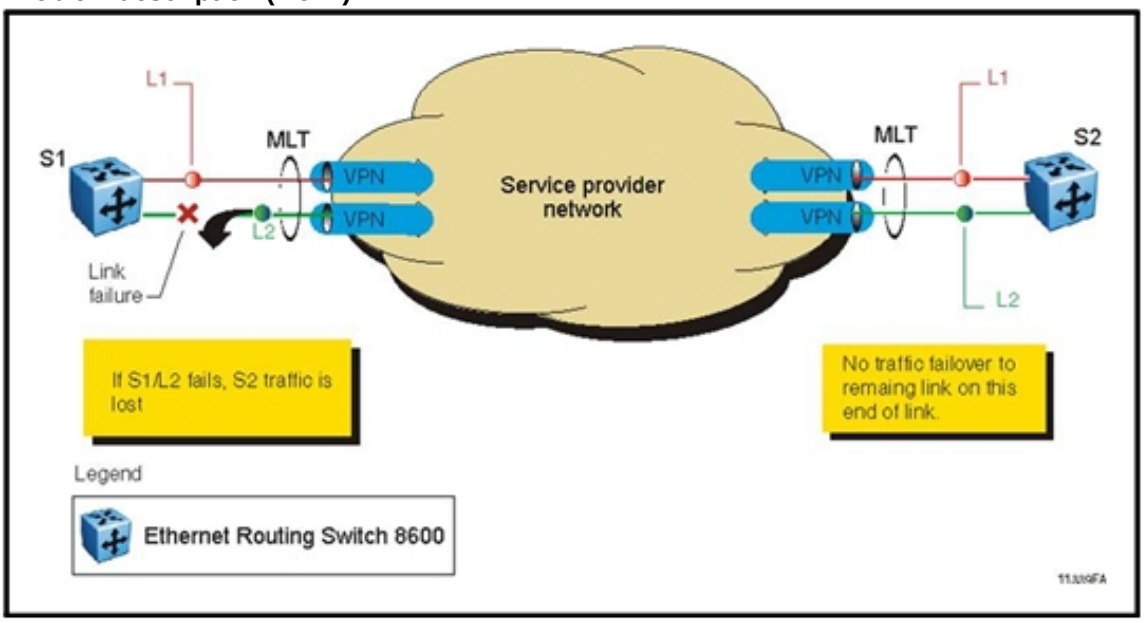
Figure 14 "Problem description (1 of 2)" (page 79) shows a MLT running with VLACP. VLACP can operate end-to-end, but can be used in a point-to-point link.

Figure 14
Problem description (1 of 2)



In the following figure, if the L2 link on S1 (S1/L2) fails, the link-down failure is not propagated over the SP network to S2 and S2 continues to send traffic over the failed S2/ L2 link.

Figure 15
Problem description (2 of 2)



Use VLACP to detect far-end failures, which allows MLT to failover when end-to-end connectivity is not guaranteed for links in an aggregation group. VLACP prevents the failure scenario.

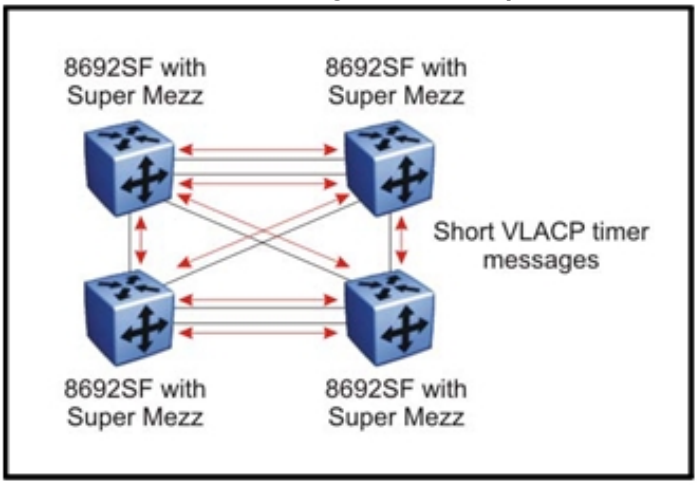
When used in conjunction with SMLT, VLACP allows you to switch traffic around entire network devices before Layer 3 protocols detect a network failure, thus minimizing network outages.

VLACP sub-100 ms convergence using SuperMezz

With Software Release 4.1 and later, the Ethernet Routing Switch 8600 can provide sub-100 millisecond failover using short timers and a new module called the Enterprise Enhanced CPU Daughter Card (SuperMezz). SuperMezz also supports IPv6. The target scenario, as shown in Figure 16 "VLACP sub 100ms convergence with Supermezz" (page 80), is a core network of at least two Ethernet Routing Switch 8600s (this feature works only between at least two Ethernet Routing Switch 8600s equipped with SuperMezz).

The Ethernet Routing Switch 8600 supports sub 100msec failover, but not as a best practice general recommendation. This functionality is only supported between two Ethernet Routing Switch 8600 switches, generally across the core of a square of a full-mesh multiple cluster design. As an environment is scaled, sub 100msec failover may not be stable. Therefore, if you enable this feature, minimize the number of links running sub 100msec operation. Upon implementing sub 100msec links or timers, if any VLACP instability is seen, increase the timers.

Figure 16
VLACP sub 100ms convergence with Supermezz



VLACP recommendations and considerations

Nortel recommends the following:

- The best practice standard settings for VLACP are a short timer of no less than 500 milliseconds and a time-out scale of 5. Both faster timers and lower time-out scales are supported, but if any VLACP flapping occurs, increase the short timer and the time-out scale to their recommended values: 5 and 500, respectively.
- Do not use VLACP on any configured LACP MLTs because LACP provides the same functionality as VLACP for link failure. VLACP and LACP running on the same link is not supported.

- Although the software configuration supports VLACP short timers of less than 30 ms, using values less than 30 ms is not supported in practice. The shortest (fastest) supported VLACP timer is 30 ms with a timeout of 3, which is used to achieve sub-100 ms failover (see “VLACP sub-100 ms convergence using SuperMezz” (page 80)). 30 ms timers are not supported in High Availability (HA) mode, and may not be stable in scaled networks. At this time, 30 ms timers are only supported between two Ethernet Routing Switch 8600 switches using SuperMezz. VLACP timer values below 500 ms may also require the SuperMezz option, particularly in scaled environments or when very short (fast) timer values are used.
- Interswitch trunk (IST) links do not support VLACP with short timers. Use only long timers. For IST MLTs, Nortel recommends that you do not set the VLACP long periodic timer to less than 30 seconds.
- If you plan to use a Layer 3 core with Equal Cost Multipath Protocol (ECMP), do not configure VLACP timers to less than 100 ms. This recommendation assumes a combination of basic Layer 2 and Layer 3 with OSPF. If you have more complex configurations, higher timer values may be required.
- When a VLACP-enabled port does not receive a VLACPDU, it should enter the disabled state. There are occasions when a VLACP-enabled port does not receive a VLACPDU but remains in the forwarding state. To avoid this situation, ensure that the VLACP configuration at the port level is consistent - both sides of the point-to-point connection should be either enabled or disabled.
- The fast periodic timer value of 200 milliseconds (ms) is not supported unless you are using the SuperMezz module. The minimum supported fast periodic timer value without the SuperMezz is 400 ms.
- VLACP is configured on a per port basis. The port can be either an individual port or a MLT member. VLACPDUs are sent periodically on each port where VLACP is enabled. This allows the exchange of VLACPDUs from an end-to-end perspective. If VLACPDUs are not received on a particular link, that link will be taken down after the expiry timeout occurs (timeout scale x periodic time). This implies that unless VLACP is enabled on the IST peer, the ports will stay in a disabled state. When VLACP is enabled at the IST peer, the VLACPDU is received and the ports will be re-enabled. This behavior can be replicated despite the IST connectivity between the end-to-end peers. When you enable VLACP on the IST ports at one end of the IST, the ports are taken down along with the IST. However, the IST at the other end will stay active until the expiry timeout occurs on the other end. As soon you enable VLACP at the other end, the VLACPDU is received by the peer and the ports are brought up at the software level.

Platform redundancy

Provide platform layer redundancy to ensure that faulty hardware does not cause a service interruption.

Nortel recommends that you use the following mechanisms to achieve device-level redundancy:

- Redundant power supplies

  Employ N + 1 power supply redundancy, where N is the number of required power supplies to power the chassis and its modules. Connect the power supplies to an additional power supply line to protect against supply problems.

  To provide additional redundancy, you can use the 8005DI AC power supply, which is a dual AC input, 1170/1492 watts (W) AC-DC power supply. On the 8005DI AC power supply, the two AC input sources can be out of synchronization with each other, having a different voltage, frequency, phase rotation, and phase angle as long as the power characteristics for each separate input AC source remain within the range of the manufacturer's specifications.

  The 8000 Series switch has two slots for fan trays or cooling modules, each with eight individual fans. Sensors are used to monitor board health.

- Input/output (I/O) port redundancy

  You can protect I/O ports using a link aggregation mechanism. MLT, which is compatible with 802.3ad static(Link Access Control Protocol [LACP] disabled), provides you with a load sharing and failover mechanism to protect against module, port, fiber or complete link failures.

- Switch fabric redundancy

  Nortel recommends that you use two SF/CPUs to protect against switch fabric failures. The two SF/CPUs load share and provide backup for each other. Using the 8006 or 8010 chassis, full switching capacity is available with both SF/CPU modules.With two SF/CPUs, you can use High Availability mode. For more information about High Availability (HA) mode, see "High Availability mode" (page 83).

- SF/CPU redundancy

  The CPU is the control plane of the switch. It controls all learning, calculates routes, and maintains port states. If the last SF/CPU in a system fails, the switch resets the I/O cards after a heartbeat timeout period of 3 seconds.

  To protect against CPU failures, Nortel has developed two different types of control plane (CPU) protection:

  — Warm Standby mode

In this mode, the Standby CPU is ready and the system image is loaded.

— High Availability mode (Hot Standby)

- Configuration and image redundancy

  You can define a primary, secondary, and tertiary configuration and system image file paths. This protects against system flash failures. For example, the primary path can point to system flash memory, the secondary path to the PCMCIA card, and the tertiary path to a network device.

  Both SF/CPU modules are identical and support flash and Personal Computer Memory Card International Association (PCMCIA) storage. If you enable the system flag called **`save to standby`**, it ensures that configuration changes are always saved to both CPUs.

When you use SMLT, Nortel recommends that you use VLACP to avoid packet forwarding to a failed switch that cannot process them.

High Availability mode

High Availability (HA) mode activates two CPUs simultaneously. These CPUs exchange topology data so that, if a failure occurs, either CPU can take precedence (with current topology data) very quickly.

In HA mode, the two CPUs are active and exchange topology data through an internal dedicated bus. This allows for a complete separation of traffic. To guarantee total security, users cannot access this bus.

In HA mode, also called Hot Standby, the two CPUs are synchronized. This means the CPUs are compatible and configured in the same mode. In nonHA mode, also called Warm Standby, the two CPUs are not synchronized. Either the CPUs are incompatible, or one of them is configured in a mode that it cannot support. Synchronization also applies to software parameters.

Depending on the protocols and data exchanged (Layer 2, Layer 3, or platform), the CPUs perform different tasks. This ensures that if a failure occurs, the backup CPU can take precedence with the most recent topology data.

Layer 2 (L2) redundancy supports the synchronization of VLAN and QoS software parameters. Layer 3 redundancy, which is an extension to and includes the Layer 2 redundancy software feature, supports the synchronization of VLAN and Quality of Service (QoS) software parameters, static and default route records, ARP entries, and LAN virtual interfaces. Specifically, Layer 3 (L3) redundancy passes table information and Layer 3 protocol-specific control packets to the Standby CPU. When

using L2/L3 redundancy, the bootconfig file is saved to both the Master and the Standby CPUs, and the Standby CPU is reset automatically. You must manually reset the Master CPU.

The following tables lists feature support and synchronization information for HA in specified software release versions.

Table 22
Feature support for HA in specified software release versions

Release/ Feature	3.5.0	3.7.0	4.0.0	4.1.0	5.0	5.1
Modules	Classic	Classic	Classic and R	Classic and R	Classic, R, and RS	Classic, R, and RS
Platform	Yes	Yes	Yes	Yes	Yes	Yes
Layer 2	Yes	Yes	Yes (3.5 based)	Yes	Yes	Yes
Layer 3	Yes (Static/ ARP)	Yes (3.5 + RIP, OSPF, VRRP, Filters, Route Policies). No BGP	N, 3.5 based	Yes as in 3.7.0+. ACE/ACLs. No BGP	Yes as in 4.1.0. Partial-HA for BGP	Yes as in 4.1.0. Partial-HA for BGP and BFD (See Note 1)
Multicast	No	No	No	No	Yes, partial HA for DVMRP and PIM. No PGM.	Yes, partial HA for DVMRP, PIM, and virtualization. No PGM. (See Note 1)
(See Note 1)IPv6	NA	NA	NA	Yes, Restart	Yes, Restart	Yes, Restart
Security	Yes	Yes	Yes (3.5 based)	Yes	Yes	Yes

Table 22
Feature support for HA in specified software release versions (cont'd.)

Release/ Feature	3.5.0	3.7.0	4.0.0	4.1.0	5.0	5.1
ATM, POS, WSM, SAM, SDM modules	No	No	No	No	No	No
Note 1: HA-CPU supports the following in Warm Standby mode. After failover, these protocols are restarted: • DVMRP, PIM-SM, PIM-SSM • BGP • MPLS • BFD						

Table 23
Synchronization capabilities in HA mode

Synchronization of:	Release 3.5	Release 3.7	Release 4.0	Release 4.1	Release 5.0	Release 5.1
Layer 1						
Port configuration parameters	Yes	Yes	Yes	Yes	Yes	Yes
Layer 2						
VLAN parameters	Yes	Yes	Yes	Yes	Yes	Yes
STP parameters	Yes	Yes	Yes	Yes	Yes	Yes
RSTP/MSTP parameters	N/A	N/A	N/A	Yes	Yes	Yes
SMLT parameters	Yes	Yes	Yes	Yes	Yes	Yes
QoS parameters	Yes	Yes	Yes	Yes	Yes	Yes
Layer 3						
Virtual IP (VLANs)	Yes	Yes	Yes	Yes	Yes	Yes
ARP entries	Yes	Yes	Yes	Yes	Yes	Yes
Static and default routes	Yes	Yes	Yes	Yes	Yes	Yes

Table 23
Synchronization capabilities in HA mode (cont'd.)

Synchronization of:	Release 3.5	Release 3.7	Release 4.0	Release 4.1	Release 5.0	Release 5.1
VRRP	No	Yes	No	Yes	Yes	Yes
RIP	No	Yes	No	Yes	Yes	Yes
OSPF	No	Yes	No	Yes	Yes	Yes
Layer 3 Filters; ACE/ACLs	No	Yes	No	Yes	Yes	Yes
BGP	No	No	No	No	Yes	Yes
DVMRP/PIM	No	No	No	No	Yes	Yes
IGMP, PIM-SM , and PIM-SSM virtualization	No	No	No	No	No	Yes
BFD	No	No	No	No	No	Yes

For more information about configuring HA, see *Nortel Ethernet Routing Switch 8600 Administration* (NN46205-605) .

L3 redundancy (HA-CPU) limitations and considerations

This section describes the limitations and considerations of the L3 redundancy (High Availability) feature.

In HA mode, you cannot configure protocols that are not supported by HA. If HA is enabled on an existing system, a protocol that is not supported by HA is disabled and all configuration information associated with that protocol is removed.

L3 redundancy (HA-CPU) is not compatible with the Packet Capture (PCAP) Tool. Be sure to disable HA-CPU prior to using PCAP.

A reboot is necessary to make HA-CPU mode active.

For information about configuring ARP, IP static routes, and IP dynamic routing protocols (OSPF and RIP), see *Nortel Ethernet Routing Switch 8600 Configuration — IP Routing* (NN46205-523) and *Nortel Ethernet Routing Switch 8600 Configuration — OSPF and RIP* (NN46205-522) .

HA does not currently support the following protocols:

- PGM
- IPX RIP/SAP
- Web Switch Module (WSM)
- 8683 POS module

- 8660 SDM module
- 8672 ATM module

If you want to use High Availability (HA) mode, verify that the link speed and duplex mode for the CPU module are 100 Mbit/s and full-duplex. If the link is not configured in 100 Mbit/s and full-duplex mode, either you cannot synchronize the two CPUs, or the synchronization may take a long time. Error messages can appear on the console.

In HA mode, Nortel recommends that you do not configure the OSPF hello timers for less than one second, and the dead router interval for less for than 15 seconds.

HA mode and short timers

Prior to Release 5.0, protocols that used short timers could bounce (restart) during HA failover. These protocols include VLACP, LACP, VRRP, OSPF, and STP. Release 5.0 introduces enhancements that support fast failover for configurations that use short timers. In Release 5.0:

- all HA configurations that use R or RS modules in R mode, along with the 8692 SF/CPU with SuperMezz, supports protocols with short timers. Fast failover under these conditions is supported.
- all HA configurations that use E or M modules, along with the 8692 SF/CPU with SuperMezz, may restart the protocols upon failover.
- all HA configurations that use E, M, R or RS modules in default, M, or R mode, along with the 8692 SF/CPU without SuperMezz, supports protocols with short timers. Fast failover under these conditions is supported.
- any HA configuration that uses the 8691 SF/CPU may restart the protocols upon failover.

Link redundancy

Provide link layer redundancy to ensure that a faulty link does not cause a service interruption. The sections that follow explain design options that you can use to achieve link redundancy. These mechanisms provide alternate data paths in case of a link failure.

Link redundancy navigation

- “Multilink Trunking” (page 88)
- “802.3ad-based link aggregation” (page 91)

- “Bidirectional Forwarding Detection” (page 95)
- “Multihoming” (page 96)

Multilink Trunking

Multilink trunking is used to provide link layer redundancy. You can use Multilink Trunking (MLT) to provide alternate paths around failed links. When you configure MLT links, consider the following information:

- Software Release 4.1 and later supports 128 MLT aggregation groups with up to 8 ports (R or RS series modules and R mode required).
- You can create up to 32 MLT groups for non R and RS modules.
- Up to eight same-type ports can belong to a single MLT group (Default, E, M mode). Same port type means that the ports operate on the same physical media, at the same speed, and in the same duplex mode.
- For Software Releases 4.0.x and 4.1, MLT configuration for the 8648GTR module is allowed only with other 8648GTR ports. No other configuration option is supported. For Release 4.1.1 and later, MLT ports can run between an 8648GTR and other module types. MLT ports must run at the same speed with the same interface type, even if using different I/O module types.

MLT navigation

- “MLT/LACP groups and port speed” (page 88)
- “Switch-to-switch MLT link recommendations” (page 89)
- “Brouter ports and MLT” (page 89)
- “MLT and spanning tree protocols” (page 90)
- “MLT protection against split VLANs” (page 91)

MLT/LACP groups and port speed

Ensure that all ports that belong to the same MLT/LACP group use the same port speed, for example, 1 Gbit/s, even if Auto-Negotiation is used. The software does not enforce this requirement. Nortel recommends that you use CANA to ensure proper speed negotiation in mixed-port type scenarios.

To maintain LAG stability during failover, use CANA: configure the advertised speed to be the same for all LACP links. For 10/100/1000 ports, ensure that CANA uses one particular setting, for example, 1000-full or 100-full. Otherwise, a remote device could restart Auto-Negotiation and the link could use a different capability.

It is important that each port uses only one speed and duplex mode. This way, all links in Up state are guaranteed to have the same capabilities. If Auto-Negotiation and CANA are not used, the same speed and duplex mode settings should be used on all ports of the MLT. This is true for both 10/100/1000 modules, and for 10/100 Classic modules that do not support CANA.

Switch-to-switch MLT link recommendations

Nortel recommends that physical connections in switch-to-switch MLT and link aggregation links be connected in a specific order. To connect an MLT link between two switches, connect the lower number port on one switch with the lower number port on the other switch. For example, to establish an MLT switch-to-switch link between ports 2/8 and 3/1 on switch A with ports 7/4 and 8/1 on switch B, do the following:

- Connect port 2/8 on switch A to port 7/4 on switch B
- Connect port 3/1 on switch A to port 8/1 on switch B

Brouter ports and MLT

In the Ethernet Routing Switch 8600, brouter ports do not support MLT. Thus, you cannot use brouter ports to connect two switches with a MLT. An alternative is to use a VLAN. This configuration option provides a routed VLAN with a single logical port (MLT).

To prevent bridging loops of bridge protocol data units (BPDUs) when you configure this VLAN:

1. Create a new Spanning Tree Group (STGx) for the two switches (switch A and switch B).
2. Add all the ports you would use in the MLT to STGx.
3. Enable the Spanning Tree Protocol (STP) for STGx.
4. On each of the ports in STGx, disable the STP. By disabling STP per port, you ensure that all BPDUs are discarded at the ingress port, preventing bridging loops.
5. Create a VLAN on switch A and switch B (VLAN AB) using STGx. Do not add any other VLANs to STGx; this action could potentially create a loop.
6. Add an IP address to both switches in VLAN AB.

MLT and spanning tree protocols

When you combine MLTs and STGs, the Spanning Tree Protocol treats multilink trunks as another link, which can be blocked. If two MLT groups connect two devices and belong to the same STG, the Spanning Tree Protocol blocks one of the MLT groups to prevent looping.

To calculate path cost defaults, the 8000 Series switch uses the following STP formulas (based on the 802.1D standard):

- Bridge Path_Cost = 1000/Attached_LAN_speed_in_Mbit/s
- MLT Path_Cost = 1000/(Sum of LAN_speed_in_Mbit/s of all Active MLT ports)

The bridge port and MLT path cost defaults for both a single 1000 Mbit/s link and an aggregate 4000 Mbit/s link is 1. Because the root selection algorithm chooses the link with the lowest port ID as its root port (ignoring the aggregate rate of the links), Nortel recommends that the following methods be used when you define path costs:

- Use lower port numbers for multilink trunks so that the multilink trunks with the most active links gets the lowest port ID.
- Modify the default path cost so that nonMLT ports, or the MLT with the least active links, has a higher value than the MLT link with the most active ports.

With the implementation of 802.1w (Rapid Spanning Tree Protocol—RSTP) and 802.1s (Multiple Spanning Tree Protocol—MSTP), a new path cost calculation method is implemented. The following table describes the new path costs associated with each interface type:

Table 24
New path cost for RSTP or MSTP mode

Link speed	Recommended path cost
Less than or equal 100 Kbit/s	200000000
1 Mbit/s	20000000
10 Mbit/s	2 000000
100 Mbit/s	200000
1 Gbit/s	20000
10 Gbit/s	2000
100 Gbit/s	200
1 Tbit/s	20
10 Tbit/s	2

MLT protection against split VLANs

When you create distributed VLANs, consider link redundancy. In a link failure, split subnets or separated VLANs disrupt packet forwarding.

The split subnet problem can occur when a VLAN carrying traffic is extended across multiple switches, and a link between the switches fails or is blocked by STP. The result is a broadcast domain that is divided into two noncontiguous parts. This problem can cause failure modes that higher level protocols cannot recover.

To avoid this problem, protect your single-point-of-failure links with an MLT backup path. Configure your spanning tree networks so that blocked ports do not divide your VLANs into two noncontiguous parts. Set up your VLANs so that device failures do not lead to the split subnet VLAN problem. Analyze your network designs for such failure modes.

802.3ad-based link aggregation

Link aggregation provides link layer redundancy. Use IEEE 802.3ad-based link aggregation (IEEE 802.3 2002 clause 43) to aggregate one or more links together to form Link Aggregation Groups (LAG) to allow a MAC client to treat the LAG as if it were a single link. Using link aggregation increases aggregate throughput of the interconnection between devices and provides link redundancy. LACP can dynamically add or remove LAG ports, depending on their availability and states.

Although IEEE 802.3ad-based link aggregation and MLT provide similar services, MLT is statically defined. By contrast, IEEE 802.3ad-based link aggregation is dynamic and provides additional functionality.

802.3ad-based link aggregation navigation

- “LACP and MLT” (page 91)
- “LACP and SMLT: Interoperability with servers (and potentially third party switches)” (page 92)
- “LACP and spanning tree interaction” (page 93)
- “LACP and Minimum Link” (page 93)
- “Link aggregation group rules” (page 94)

LACP and MLT

When you configure standards-based link aggregation, you must enable the aggregatable parameter. After you enable the aggregatable parameter, the LACP aggregator is one-to-one mapped to the specified MLT.

A newly-created MLT/LAG adopts the VLAN membership of its member ports when the first port is attached to the aggregator associated with this LAG. When a port is detached from an aggregator, the port is deleted from the associated LAG port member list. When the last port member is deleted from the LAG, the LAG is deleted from all VLANs and STGs.

After the MLT is configured as aggregatable, you cannot add or delete ports or VLANs manually.

To enable tagging on ports belonging to a LAG, first disable LACP on the port, enable tagging on the port, and then enable LACP.

LACP and SMLT: Interoperability with servers (and potentially third party switches)

To better serve interoperability with servers (and potentially certain third party switches) in SMLT designs, the Ethernet Routing Switch 8600 provides a system ID configuration option for Split MultiLink Trunk (SMLT).

Prior to this enhancement, if the SMLT Core Aggregation Switches were unable to negotiate the system ID (for example, if the inter-switch trunk [IST] or one of the aggregate switches failed), the Ethernet Routing Switch 8600 SMLT/LACP implementation modified the Link Aggregation Control Protocol (LACP) system ID to aa:aa:aa:aa:aa:xx (where xx is the LACP key). When SMLT-attached servers (and certain third party wiring closet switches) received this new system ID, in some cases, ports moved to a different link aggregation group (LAG) resulting in data loss.

To avoid this issue, the Ethernet Routing Switch 8600 provides an option to configure a static system ID that is always used by the SMLT Core Aggregation Switches. In this way, the same LACP key is always used, regardless of the state of the SMLT Core Aggregation Switch neighbor (or the IST link). Therefore no change in LAGs occur on the attached device, be it a server or a third party switch. This situation (and therefore this advanced configuration option) does not affect Nortel edge switches used in SMLT configurations.

The actor system priority of LACP_DEFAULT_SYS_PRIO, the actor system ID configured by the user, and an actor key equal to the SMLT-ID or SLT-ID are sent to the wiring closet switch. Nortel recommends that you configure the system ID to be the base MAC address of one of the aggregate switches along with its SMLT-ID. You must ensure that the same value for system ID is configured on both of the SMLT Core Aggregation Switches.

To configure the system ID, use the following CLI command:

```
ERS8610:5# config lacp smlt-sys-id <MAC Address>
```

OR

Use the following NNCLI command:

```
ERS8610:5(config)# lacp smlt-sys-id <MAC Address>
```

For more information about SMLT, see *Switch Clustering using Split Multi-Link Trunking (SMLT) with ERS 8600, 8300, 5500 and 1600 Series Technical Configuration Guide* (NN48500-518).

LACP and spanning tree interaction

The operation of LACP module is only affected by the physical link state or its LACP peer status. When a link goes up and down, the LACP module is notified. The STP forwarding state does not affect the operation of the LACP module. LACP data units (LACPDU) can be sent even if the port is in STP blocking state.

Unlike legacy MLT, configuration changes (such as speed, duplex mode, and so on) made to a LAG member port are not applied to all the member ports of the MLT. Instead, the changed port is taken out of the LAG, and the corresponding aggregator and user is alerted.

In contrast to MLT, IEEE 802.3ad-based link aggregation does not require BPDUs to be replicated over all ports in the trunk group. Therefore, use the CLI command**`config stg <stg> ntstg disable`** to disable the parameter on the STG for LACP-based link aggregation.

In the NNCLI, the command is **`no spanning-tree stp <1-64> ntstp`**.

This parameter applies to all trunk groups that are members of this STG. This parameter is necessary when interworking with devices that only send BPDUs out one port of the LAG.

LACP and Minimum Link

The Minimum Link parameter defines the minimum number of active links required for a LAG to remain in the forwarding state. Use the Minimum-Link (MinLink) feature so that when the number of active links in a LAG is less than the MinLink parameter, the entire LAG is declared down. Prior to MinLink support, a LAG was always declared up if one physical link of the LAG was up.

Configure MinLink for each LAG; each LAG can have a different value, if required. The number of minimum links configured for an end of a LAG is independent of the other end; a different value can be configured for each end of a LAG. The default MinLink value is 1, with a range of 1 to 8.

If the number of active links in the LAG becomes less than the MinLink setting, the Ethernet Routing Switch 8600 marks the LAG as down, and informs the remote end of the LAG state by using a Link Aggregation Protocol Data Unit (LACPDU). The switch continues to send LACPDUs to neighbors on each available link based on the configured timers. When the number of active links in the LAG is greater than or equal to the MinLink parameter, LACP informs the remote end, and the LAG transitions to the forwarding (up) state.

The maximum number of active links in a LAG is 8; however, it is possible to configure up to 16 links in a LAG. The eight inactive links are in Standby mode. If a link goes down, Standby links take precedence over MinLink. When an active link is disabled, the standby link with the lowest port number immediately becomes active. MinLink operates after the Standby processes finish.

To use MinLink on the Ethernet Routing Switch 8600, you require:

- 8691 or 8692 SF/CPU modules
- Software Release 4.1.5.9, or 4.1.8.0 or later

The SuperMezz daughtercard is not required.

On standard MLT links, you must enable LACP to enable MinLink.

You cannot enable MinLink on Split MultiLink Trunking (SMLT) links because the minimum number of links with SMLT can only be set to 1.

Link aggregation group rules

Link aggregation is compatible with the Spanning Tree Protocol (STP/RSTP/MSTP). Link aggregation groups operate under the following rules:

- All ports in a link aggregation group must operate in full-duplex mode.
- All ports in a link aggregation group must use the same data rate.
- All ports in a link aggregation group must be in the same VLANs.
- Link aggregation groups must be in the same STP groups.
- If the ntstg parameter is false, STP BPDU transmit on only one link.
- Ports in a link aggregation group can exist on different modules.
- Link aggregation groups are formed using LACP.
- Amaximumof32linkaggregationgroupsaresupportedinnonRmode (both Classic and R series modules).

- A maximum of 128 link aggregation groups are supported in Rmode (R or RS series modules only).
- A maximum of eight active links are supported per LAG.

For LACP fundamentals and configuration procedures, see *Nortel Ethernet Routing Switch 8600 Configuration — Link Aggregation, MLT, and SMLT* (NN46205-518) .

Bidirectional Forwarding Detection

The Ethernet Routing Switch 8600 supports Bidirectional Forwarding Detection (BFD). BFD is a simple Hello protocol used between two peers. In BFD, each peer system periodically transmits BFD packets to each other. If one of the systems does not receive a BFD packet after a certain period of time, the system assumes that the link or other system is down.

BFD provides low-overhead, short-duration failure detection between two systems. BFD also provides a single mechanism for connectivity detection over any media, at any protocol layer.

Because BFD sends rapid failure detection notifications to the routing protocols that run on the local system, which initiates routing table recalculations, BFD helps reduce network convergence time.

BFD supports IPv4 single-hop detection for static routes, OSPF, and BGP. The Ethernet Routing Switch 8600 BFD implementation complies with IETF drafts draft-ietf-bfd-base-06 and draft-ietf-bfd-v4v6-1hop-06.

Operation

The Ethernet Routing Switch 8600 uses one BFD session for all protocols with the same destination. For example, if a network runs OSPF and BGP across the same link with the same peer, only one BFD session is established, and BFD shares session information with both routing protocols.

ATTENTION
BFD requires the 8692 SF/CPU. Although SuperMezz is not mandatory, Nortel recommends that you use it if you use BFD.

You can enable BFD over data paths with specified OSPF neighbors, BGP neighbors, and static routing next-hop addresses.

The Ethernet Routing Switch 8600 supports BFD asynchronous mode, which sends BFD control packets between two systems to activate and maintain BFD neighbor sessions. To reach an agreement with its neighbor about how rapidly failure detection occurs, each system estimates how quickly it can send and receive BFD packets.

A session begins with the periodic, slow transmission of BFD Control packets. When bidirectional communication is achieved, the BFD session comes up. The switch only declares a path as operational when two-way communication is established between systems.

After the session is up, the transmission rate of Control packets can increase to achieve detection time requirements. If Control packets are not received within the calculated detection time, the session is declared down. After a session is down, Control packet transmission returns to the slow rate.

If a session is declared down, it cannot come back up until the remote end signals that it is down (three-way handshake). A session can be kept administratively down by configuring the state of AdminDown.

BFD restrictions

The Ethernet Routing Switch 8600 supports up to 256 BFD sessions, however, the number of BFD sessions plus the number of VLACP sessions cannot exceed 256.

The Ethernet Routing Switch 8600 does not support the following IETF BFD options:

- Echo packets
- BFD over IPv6
- Demand mode
- authentication

The Ethernet Routing Switch 8600 does not support:

- BFD on a VRRP virtual interface
- High Availability (HA) for BFD

The Ethernet Routing Switch 8600 supports partial HA for BFD.

The Ethernet Routing Switch 8600 also supports the modification of transmit and receive intervals during an active BFD session.

Multihoming

Multihoming enables the Ethernet Routing Switch 8600 to support clients or servers that have multiple IP addresses associated with a single MAC address.

Multihomed hosts can be connected to port-based, policy-based, and IP subnet-based VLANs.

The IP addresses that you associate with a single MAC address on a host must be located in the same IP subnet. The Ethernet Routing Switch 8600 supports multihomed hosts with up to 16 IP addresses per MAC address.

For more information about multihoming, see *Nortel Ethernet Routing Switch 8600 Configuration — VLANs and Spanning Tree* (NN46205-517) .

Network redundancy

Provide network redundancy so that a faulty switch does not interrupt service. You can configure mechanisms that direct traffic around a malfunctioning switch. The sections that follow describe designs you can follow to achieve network redundancy.

Network redundancy navigation

- "Modular network design for redundant networks" (page 97)
- "Network edge redundancy" (page 101)
- "Split Multi-Link Trunking" (page 102)
- "Routed SMLT" (page 115)
- "Switch clustering topologies and interoperability with other products" (page 120)

Modular network design for redundant networks

Network designs normally depend on the physical layout and the fiber and copper cable layout of the area. When designing networks, Nortel recommends that you use a modular approach. Break the design into different sections, which can then be replicated as needed using a recursive model. You must consider several functional layers or tiers. To define the functional tiers, consider campus architectures separately from data center architectures.

Campus architecture

A three-tier campus architecture consists of an edge layer, a distribution layer, and a core layer.

- **Edge layer**: The edge layer provides direct connections to end user devices. These are normally the wiring closet switches that connect devices such as PCs, IP phones, and printers.
- **Distribution layer**: The distribution layer provides connections to the edge layer wiring closets in a three-tier architecture. This layer connects the wiring closets to the core.
- **Core layer**: The core layer is the center of the network. In a three-tier architecture, all distribution layer switches terminate in the core. In a

two-tier architecture, the edge layer terminates directly in the core, and no distribution layer is required.

ATTENTION
Nortel recommends that you do not directly connect servers and clients in core switches. If one IST switch fails, connectivity to the server is lost.

Data center architecture

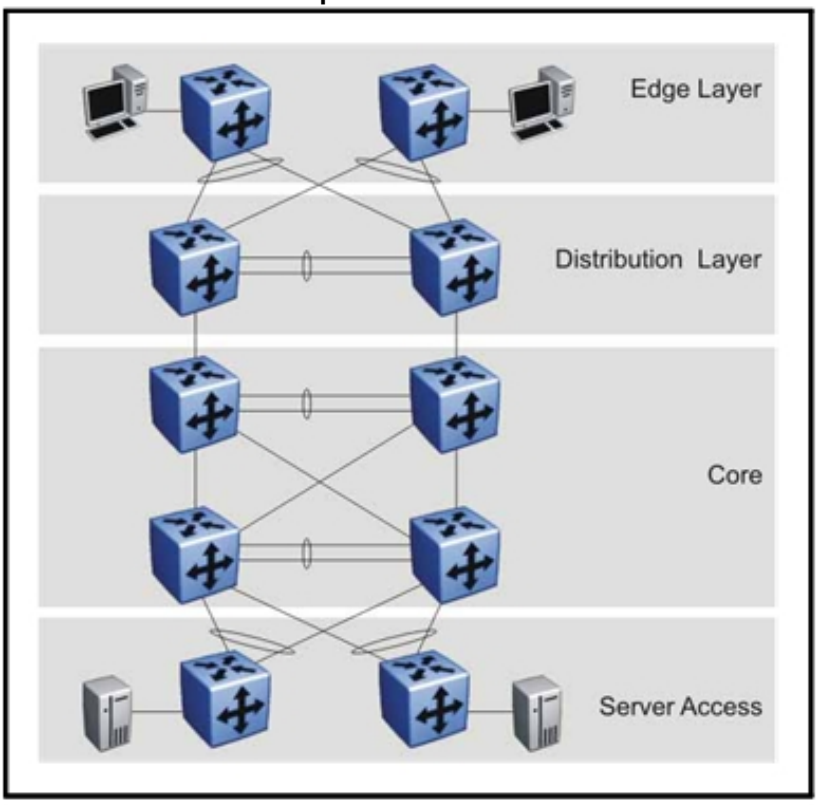
The tiered network architecture also applies to a data center architecture. In this case, the core and distribution layers provide similar functions to those in a campus architecture, while the edge layer is replaced by the server access layer:

- **Server Access layer**: The server access layer provides direct connections to servers.
- **Distribution layer**: The distribution layer provides connections to the server access layer in a three-tier architecture.
- **Core layer**: The core layer is the center of the network. In a three-tier architecture, all distribution layer switches terminate in the core. In a two-tier architecture, the server access layer terminates directly in the core, and no distribution layer is required.

Example network layouts

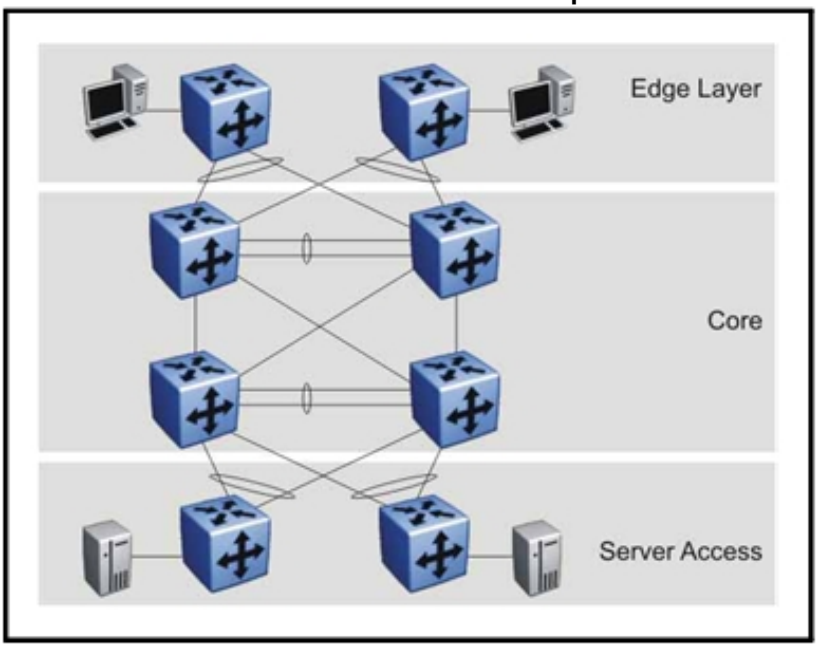
The following figure shows a three-tiered campus architecture with edge, distribution, and core layers. In addition, a server access layer is directly connected to the core, representing a two-layer data center architecture.

Figure 17
Three-tiered architecture plus data center



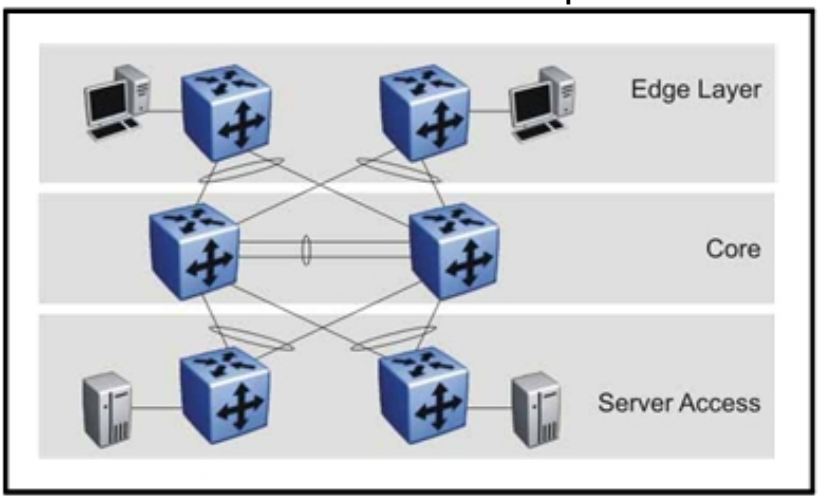
In many cases, you can remove the distribution layer from the campus network layout. This maintains functionality, but decreases cost, complexity, and network latency. The following figure shows a two-tiered architecture where the edge layer is connected directly into the core.

Figure 18
Two-tiered architecture with four-switch core plus data center



The following figure shows a two-tiered architecture with a two-switch core.

Figure 19
Two-tiered architecture with two-switch core plus data center

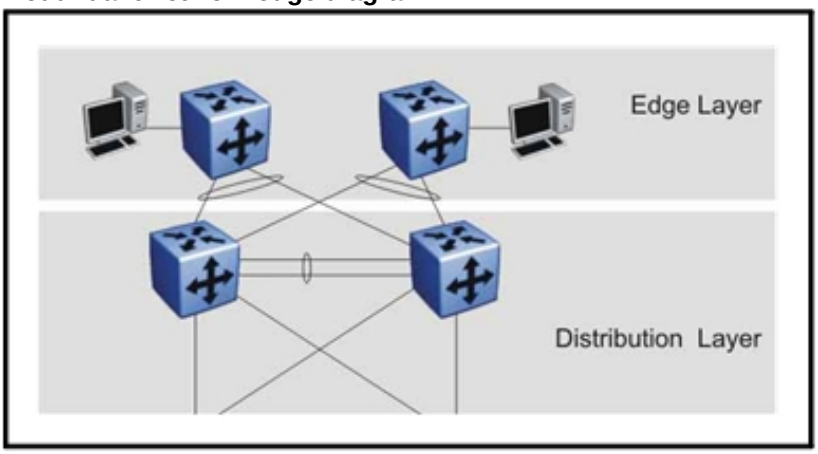


For specific design and configuration parameters, see *Converged Campus Technical Solutions Guide* (NN48500-516) and *Switch Clustering using Split-Multilink Trunking (SMLT) Technical Configuration Guide* (NN48500-518) .

Network edge redundancy

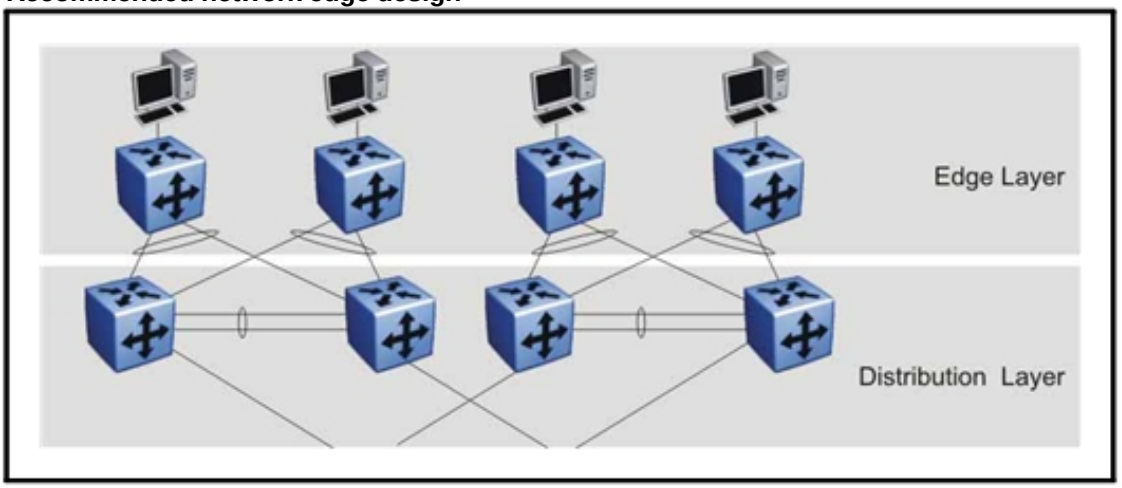
Provide network edge redundancy. The following figure depicts an distribution switch pair distributing riser links to wiring closets. If one edge layer switch fails, the other can maintain user services.

Figure 20
Redundant network edge diagram



Nortel recommends the network edge design shown in Figure 21 "Recommended network edge design" (page 101). This setup is simple to implement and maintain, yet still provides redundancy if one of the edge or distribution layer switches fails.

Figure 21
Recommended network edge design



Split Multi-Link Trunking

A split multilink trunk is a multilink trunk with one end split (shared) between two aggregation switches. Using Single Link Trunking (SLT), you can configure a split multilink trunk using a single port. This permits the scaling of the number of split multilink trunks on a switch to the maximum number of available ports.

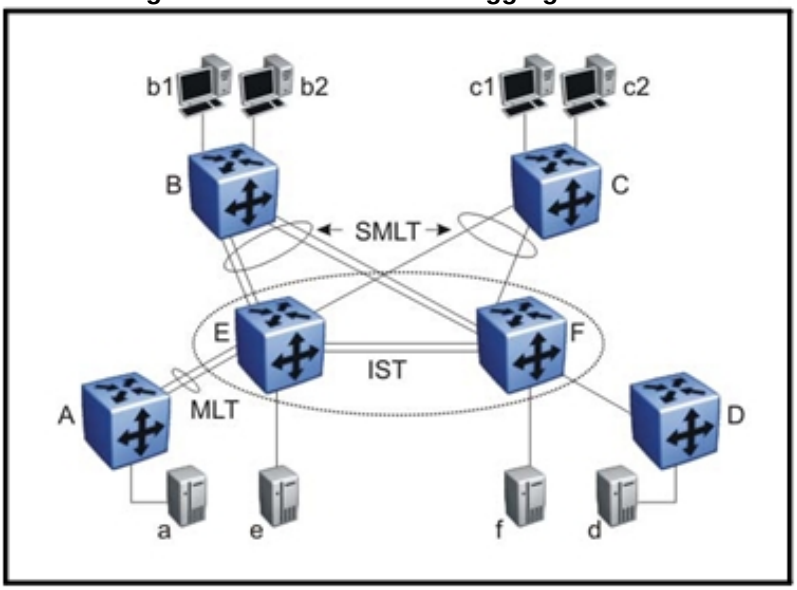
For configuration procedures for the Nortel Split Multi-Link Trunking feature for the Ethernet Routing Switch 8600, see *Switch Clustering using Split-Multilink Trunking (SMLT) Technical Configuration Guide* (NN48500-518) or *Switch Clustering (SMLT/SLT) Configuration Tool* (NN48500-536) .

SMLT navigation

SMLT redundancy

The following figure shows an SMLT configuration that contains a pair of Ethernet Routing Switch acting as aggregation switches (E and F). Four separate wiring closet switches are shown, labeled A, B, C, and D (MLT-compatible devices).

Figure 22
SMLT configuration with switches as aggregation switches



B and C are connected to the aggregation switches through multilink trunks that are split between the two aggregation switches. For example, SMLT client switch B can use two parallel links for its connection to E, and two additional parallel links for its connection to F. This provides redundancy.

The SMLT client switch C may have only a single link to both E and F. Switch A is configured for MLT, but the MLT terminates on only one switch in the network core. Switch D has a single connection to the core. Although you could configure both switch A and switch D to terminate across both of the aggregation switches using SMLT, neither switch would benefit from SMLT in this network configuration.

The SMLT client switches are dual-homed to the two aggregation switches, yet they require no knowledge of whether they are connected to a single switch or to two switches. SMLT intelligence is required only on the aggregation switches. Logically, they appear as a single switch to the edge switches. Therefore, the SMLT client switches only require an MLT configuration. The connection between the SMLT aggregation switches and the SMLT client switches are called the SMLT links. The client switch can use any proprietary link aggregation protocol, such as MLT or EtherChannel, in addition to standards-based LACP.

Figure 22 "SMLT configuration with switches as aggregation switches" (page 103) also includes end stations connected to each of the switches. End stations a, b1, b2, c1, c2, and d are typically hosts, while e and f may be hosts, servers, or routers. SMLT client switches B and C can use any

method to determine which multilink trunk link to use to forward a packet, so long as the same link is used for a given Source/Destination address (SA/DA) pair (regardless of whether or not the DA is known by B or C).

ATM, Packet over SONET (POS), and Ethernet interfaces are supported as operational SMLT links.

Aggregation switches always send traffic directly to an SMLT client switch. They only use the interswitch trunk for traffic that they cannot forward in another, more direct way.

SMLT and VLACP

VLACP is recommended for all SMLT access links when the links are configured as MLT to ensure both end devices are able to communicate. By using VLACP over SLT, enhanced failure detection is extended beyond the limits of the number of SMLT or LACP instances that can be created on a Nortel switch.

For more information about VLACP, see “End-to-end fault detection and VLACP” (page 77).

SMLT and loop prevention

Split MultiLink Trunking (SMLT) based network designs form physical loops for redundancy that logically do not function as a loop. Under certain adverse conditions, incorrect configurations or cabling, loops can form.

The two solutions to detect loops are Loop Detect and Simple Loop Prevention Protocol (SLPP). Loop Detect and SLPP detect a loop and automatically stop the loop. Both solutions determine on which port the loop is occurring and shuts down that port.

For more information, see “SLPP, Loop Detect, and Extended CP-Limit” (page 148).

Interswitch Trunking recommendations

Figure 22 "SMLT configuration with switches as aggregation switches" (page 103) shows that SMLT requires only two SMLT-capable aggregation switches connected by an interswitch trunk. The aggregation switches use the interswitch trunk to:

- Confirm that each switch is alive and to exchange MAC address information. Thus, the link must be reliable and must not exhibit a single point of failure in itself.
- Forward flooded packets or packets destined for nonSMLT connected switches, or for servers physically connected to the other aggregation switch.

The amount of traffic from a single SMLT wiring-closet switch that requires forwarding across the interswitch trunk is usually small. However, if the aggregation switches terminate connections to single-homed devices, or if uplink SMLT failures occur, the interswitch trunk traffic volume may be significant. To ensure that no single point of failure exists in the interswitch trunk, Nortel recommends that the interswitch trunk be a multi-gigabit multilink trunk with connections across different modules on both aggregation switches.

The Interswitch Trunking (IST) session is established between the peering SMLT aggregation switches. The basis for this connection is a common VLAN and the knowledge about the peer IP addressing for the common VLAN. Nortel recommends that you use an independent VLAN for this IST peer session. You can do so only by including the interswitch trunk ports in the VLAN because only the interswitch trunk port is a member of the interswitch trunk VLAN.

Nortel recommends that you not enable any dynamic routing protocols on the IST VLAN. The IST VLAN's purpose is to support adjacent switches; the IST should not be used as a next-hop route for nonIST traffic or routing traffic. One exception to this rule is the case of multicast traffic with PIM-SM. In this case, you must enable PIM-SM on the IST VLAN.

Nortel also recommends that you use low slot number ports for the IST, for example ports 1/1 and 2/1, because the low number slots boot up first.

Nortel recommends that you use an independent Virtual Local Area Network (VLAN) for the IST peer session. To avoid the dropping of IST control traffic, Nortel recommends that you use a nonblocking port for the IST—for example, any R series module Gigabit Ethernet port.

Nortel recommends that an interswitch multilink trunk contain at least two physical ports, although this is not a requirement.

Nortel recommends that CP-Limit be disabled on all physical ports that are members of an IST multilink trunk. Disabling CP-Limit on IST MLT ports forces another, less-critical port to be disabled if the defined CP-Limit is exceeded. By doing this, you preserve network stability if a protection condition arises. Although it is likely that one SMLT MLT port (riser) is disabled in such a condition, traffic continues to flow through the remaining SMLT ports.

SMLT and client/server applications

Do not use unbalanced client-server configuration, where core switches have directly-connected servers and/or clients. This is not recommended because a loss of one of the IST pair switches causes connectivity to the server to be lost.

SMLT ID recommendations

SMLT links on both aggregation switches share an SMLT link ID called SmltId. The SmltId identifies all members of a split multilink trunk group. Therefore, you must terminate both sides of each SMLT having the same SmltId at the same SMLT client switch. For the exceptions to this rule, see Figure 23 "SMLT square configuration" (page 113) and Figure 24 "SMLT full-mesh configuration" (page 113).

The SMLT IDs can be, but are not required to be, identical to the MLT IDs. SmltId ranges are:

- 1 to 32 for MLT-based SMLTs in non-R-mode enabled chassis
- 1 to 128 for MLT-based SMLTs in R-mode enabled chassis
- 1 to 512 for SLTs

ATTENTION
Nortel recommends to use SLT IDs of 129 to 512 and that you reserve the lower number IDs of 1 to 128 for SMLT only.

Single Link Trunking (SLT)

Use Single Link Trunking (SLT) to configure a split multilink trunk that uses a single port. A single-port split multilink trunk behaves like an MLT-based split multilink trunk and can coexist with split multilink trunks in the same system. However, an SMLT ID can belong to either an MLT-SMLT or to an SLT per chassis. Use SLT to scale the number of split multilink trunks on a switch to the maximum number of available ports.

On the SMLT aggregation switch pair, split multilink trunks can exist in the following combinations:

- MLT-based split multilink trunks and MLT-based split multilink trunks
- MLT-based split multilink trunks and SLTs
- SLTs and SLTs

SLT configuration rules include:

- The dual-homed device that connects the aggregation switches must support MLT.
- SLT is supported on Ethernet, POS, and ATM ports.
- Assign SMLT IDs of 129 to 512 to SLTs and reserve the lower number IDs of 1 to 128 for SMLT only.
- SLT ports can be designated access or trunk (that is, IEEE 802.1Q tagged or untagged), and changing the type does not affect their behavior.

- You cannot change an SLT into an MLT-based SMLT by adding more ports. You must delete the SLT and then reconfigure the port as SMLT/MLT.
- You cannot change an MLT-based SMLT into an SLT by deleting all ports but one. You must first remove the SMLT/MLT and then reconfigure the port as SLT.
- AportcannotbeconfiguredasMLT-basedSMLTandasSLTatthe same time.

For information about configuring SLT, see *Nortel Ethernet Routing Switch 8600 Configuration — Link Aggregation, MLT, and SMLT* (NN46205-518) .

SMLT and Layer 2 traffic load sharing

On the edge switch, SMLT achieves load sharing by using the MLT path selection algorithm (for a description of the algorithm, see *Nortel Ethernet Routing Switch 8600 Configuration — Link Aggregation, MLT, and SMLT* (NN46205-518) . Usually, the algorithm operates on a source/destination MAC address basis or a source/destination IP address basis.

On the aggregation switch, SMLT achieves load sharing by sending all traffic destined for the SMLT client switch directly to the SMLT client, and not over the IST trunk. The IST trunk is never used to cross traffic to and from an SMLT dual-homed wiring closet. Traffic received on the IST by an aggregation switch is not forwarded to SMLT links (the other aggregation switch does this), thus eliminating the possibility of a network loop.

SMLT and Layer 3 traffic Redundant Default Gateway: VRRP

On SMLT aggregation switches, you can route VLANs that are part of an SMLT network. Routing VLANs enables the SMLT edge network to connect to other Layer 3 networks. VRRP, which provides redundant default gateway configurations, additionally has BackupMaster capability. BackupMaster improves the Layer 3 capabilities of VRRP operating in conjunction with SMLT. Nortel recommends that you use a VRRP BackupMaster configuration with any SMLT configuration that has an existing VRRP configuration.

A better alternative than SMLT with VRRP BackupMaster is to use RSMLT L2 Edge. For Release 5.0 and later, Nortel recommends that you use RSMLT L2 Edge configuration, rather than SMLT with VRRP BackupMaster, for those products that support RSMLT L2 Edge. RSMLT L2 Edge provides:

- Greater scalability—VRRP scales to 255 instances, while RSMLT scales to the maximum number of VLANs.
- Simpler configuration—Simply enable RSMLT on a VLAN; VRRP requires virtual IP configuration, along with other parameters.

For connections in pure Layer 3 configurations (using a static or dynamic routing protocol), a Layer 3 RSMLT configuration is recommended over SMLT with VRRP. In these instances, an RSMLT configuration provides faster failover than one with VRRP because the connection is a Layer 3 connection, not just a Layer 2 connection for default gateway redundancy.

ATTENTION

In an SMLT-VRRP environment that has VRRP critical IP configured within both IST core switches, routing between directly connected subnets ceases to work when connections from each of the switches to the exit router (the critical IP) fail. Nortel recommends that you do not configure VRRP critical IPs within SMLT or R-SMLT environments because SMLT operation automatically provides the same level of redundancy.

As well, do not use VRRP BackupMaster and critical IP at the same time. Use one or the other. Do not use VRRP in RSMLT environments.

Typically, only the VRRP Master forwards traffic for a given subnet. If you use BackupMaster on the SMLT aggregation switch, and it has a destination routing table entry, then the Backup VRRP switch also routes traffic. The VRRP BackupMaster uses the VRRP standardized backup switch state machine. Thus, VRRP BackupMaster is compatible with standard VRRP. This capability is provided to prevent the traffic from edge switches from unnecessarily utilizing the IST to deliver frames destined for a default gateway. In a traditional VRRP implementation, this operates only on one of the aggregation switches.

The BackupMaster switch routes all traffic received on the BackupMaster IP interface according to the switch routing table. The BackupMaster switch does not Layer 2-switch the traffic to the VRRP Master.

You must ensure that both SMLT aggregation switches can reach the same destinations by using a routing protocol. Therefore, Nortel recommends that, for routing purposes, you configure per-VLAN IP addresses on both SMLT aggregation switches. Nortel further recommends that you introduce an additional subnet on the IST that has a shortest-route-path to avoid issuing Internet Control Message Protocol (ICMP) redirect messages on the VRRP subnets. (To reach the destination, ICMP redirect messages are issued if the router sends a packet back out through the same subnet on which it is received).

SMLT failure and recovery

Traffic can cease if an SMLT link is lost. If a link is lost, the SMLT client switch detects the loss and sends traffic on the other SMLT links, as it does with standard MLT. If the link is not the only one between the SMLT client and the aggregation switches in question, the aggregation switch also uses standard MLT detection and rerouting to move traffic to the remaining links. However, if the link is the only route to the aggregation

switch, the switch informs the other aggregation switch of the SMLT trunk failure. The other aggregation switch then treats the SMLT trunk as a regular multilink trunk. In this case, the MLT port type changes from splitMLT to normalMLT. If the link is reestablished, the aggregation switches detect it and move the trunk back to regular SMLT operations. The operation mode changes from normalMLT back to splitMLT.

Traffic can also cease if an aggregation switch fails. If an aggregation switch fails, the SMLT client switch detects the failure and sends traffic out on other SMLT links, as in standard MLT. The operational aggregation switch detects the loss of the partner IST. The SMLT trunks are modified to regular MLT trunks, and the operation mode is changed to normalMLT. If the partner switch IST returns, the operational aggregation switch detects it. The IST again becomes active, and after full connectivity is reestablished, the trunks are moved back to regular SMLT.

If an IST link fails, the SMLT client switches do not detect a failure and continue to communicate as usual. Normally, more than one link in the IST is available (the interswitch trunk is itself a distributed MLT). Thus, IST traffic resumes over the remaining links in the IST.

Finally, if all IST links are lost between an aggregation switch pair, the aggregation switches cannot communicate with each other. Both switches assume that the other switch has failed. Generally, a complete IST link failure causes no ill effects in a network if all SMLT client switches are dual-homed to the SMLT aggregation switches. However, traffic that comes from single attached switches or devices no longer predictably reaches the destination. IP forwarding may cease because both switches try to become the VRRP Master. Because the wiring closets switches do not know about the interswitch trunk failure, the network provides intermittent connectivity for devices that are attached to only one aggregation switch. Data forwarding, while functional, may not be optimal because the aggregation switches may not learn all MAC addresses, and the aggregation switches can flood traffic that would not normally be flooded.

SMLT and IEEE 802.3ad interaction

The Ethernet Routing Switch 8600 switch fully supports the IEEE 802.3ad Link Aggregation Control Protocol (LACP) on MLT and distributed MLT links, and on a pair of SMLT switches. Note the following information:

- MLT peer and SMLT client devices can be network switches or any type of server/workstation that supports link bundling through IEEE 802.3ad.
- Single-link and multilink SMLT solutions support dual-homed connectivity for more than 350 attached devices, thus allowing you to build dual-homed server farm solutions.

Only dual-homed devices benefit from LACP and SMLT interactivity.

SMLT/IEEE link aggregation supports all known SMLT scenarios where an IEEE 802.3ad SMLT pair can be connected to SMLT clients, or where two IEEE 802.3ad SMLT pairs can be connected to each other in a square or full-mesh topology.

Known SMLT/LACP failure scenarios include:

- Wrong ports connected
- Mismatched SMLT IDs assigned to SMLT client

  SMLT switches detect inconsistent SMLT IDs. In this case, the SMLT aggregation switch that has the lowest IP address does not allow the SMLT port to become a member of the aggregation group.
- SMLT client switch has LACP disabled

  SMLT aggregation switches detect that aggregation is disabled on the SMLT client, thus no automatic link aggregation is established until the configuration is resolved.
- Single CPU failure

  In this case, LACP on other switches detects the remote failure, and all links connected to the failed system are removed from the link aggregation group. This process allows failure recovery to a different network path.

SMLT and LACP System ID

Since Release 4.1.1, an administrator can configure the LACP SMLT System ID used by SMLT core aggregation switches. Prior to Release 4.1.1, if the SMLT core aggregation switches did not know and were unable to negotiate the LACP system ID, data could be lost. Nortel recommends that you configure the LACP SMLT system ID to be the base MAC address of one of the aggregate switches, and that you include the SMLT-ID. Ensure that the same System ID is configured on both of the SMLT core aggregation switches.

An explanation of the importance of configuring the System ID is as follows.

The LACP System ID is the base MAC address of the switch, which is carried in Link Aggregation Control Protocol Data Units (LACPDU). When two links interconnect two switches that run LACP, each switch knows that both links connect to the same remote device because the LACPDUs originate from the same System ID. If the links are enabled for aggregation using the same key, then LACP can dynamically aggregate them into a LAG (MLT).

When SMLT is used between the two switches, they act as one logical switch. Both aggregation switches must use the same LACP System ID over the SMLT links so that the edge switch sees one logical LACP peer, and can aggregate uplinks towards the SMLT aggregation switches. This process automatically occurs over the IST connection, where the base MAC address of one of the SMLT aggregation switches is chosen and used by both SMLT aggregation switches.

However, if the switch that owns that Base MAC address reboots, the IST goes down, and the other switch reverts to using its own Base MAC address as the LACP System ID. This action causes all edge switches that run LACP to think that their links are connected to a different switch. The edge switches stop forwarding traffic on their remaining uplinks until the aggregation can reform (which can take several seconds). Additionally, when the rebooted switch comes back on line, the same actions occur, thus disrupting traffic twice.

The solution to this problem is to statically configure the same SMLT System ID MAC address on both aggregation switches.

For more information about configuring the LACP SMLT system ID, see *Nortel Ethernet Routing Switch 8600 Configuration — Link Aggregation, MLT, and SMLT* (NN46205-518) .

SMLT and Spanning Tree Protocol

When you configure an SMLT interswitch trunk, Spanning Tree Protocol is disabled on all ports that belong to the interswitch trunk. As of Release 3.3, you cannot have an interswitch trunk link with STP enabled, even if the interswitch trunk link is tagged and belongs to other STGs.

Connecting a VLAN to both SMLT aggregation switches with nonSMLT link introduces a loop and is not a supported configuration. Ensure that the connections from the SMLT aggregation switch pair are SMLT links or make the connection through routed VLANs.

SMLT scalability

To determine the maximum number of VLANs supported per device on an MLT/SMLT, use the following formulas.

To calculate the total number of VLANs that you can configure with SMLT/IST without Enhanced Operational mode, use the following formula (if you are using R series modules, replace 1980 with 1972):

(2 * number of VLANs on regular ports) + (16 * number of VLANs of SMLT/MLT ports) = 1980

To calculate the total number of VLANs that you can configure with SMLT/IST with Enhanced Operational mode, use the following formula (if you are using R series modules, replace 1980 with 1972):

(number of VLANs on regular ports or MLT ports) + (2 * number of VLANs on SMLT ports) = 1980

ATTENTION
Enable Enhanced Operational mode in chassis with E and M modules only. Do not enable Enhanced Operational mode in a mixed chassis that contains R or RS modules with E or M modules.

If you are operating the system in R-mode, the available VLANs in an SMLT setup are based on the following:

- If **`config sys set max-vlan-resource-reservation enable`** is enabled, then 2042 VLANs are available for SMLT.
- If **`config sys set multicast-resource-reservation <value>`** is configured (range of value: 64-4084), then the number of available VLANs on the SMLT switch is calculated as the configured value divided by 2 (VLANs available = <value>/2)

In this case the number of available VLANs on SMLT switch is calculated by using the configured value and divided by 2 (value/2).

A maximum of one IST MLT can exist per switch. With R and RS modules, you can have a total of 127 MLT/SMLT groups (128 MLT groups minus 1 MLT group for the IST). For E and M Modules, the maximum is 31 MLT/SMLT groups with 1 IST.

SMLT IDs can be either MLT- or port-based. The maximum value for the Port/SMLT ID is 512, but in practice, this is limited by the number of available ports on the switch.

Port/SMLT IDs allow only one port per switch to be a member of an SMLT ID; MLT/SMLT allows up to eight ports to be members of an SMLT ID per switch.

When you use SMLT, the total number of supported MAC addresses (if all records are available for MAC address learning) is 64 000 for M, R, and RS modules.

For more information about SMLT scalability and multicast routing, see “Multicast network design” (page 193).

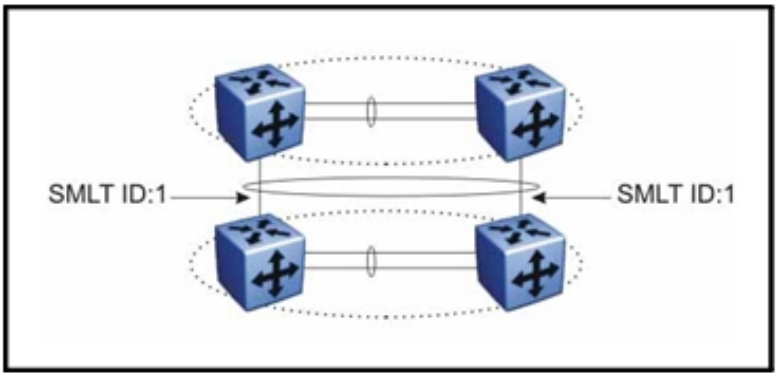
For more information about VLAN scalability, see *Nortel Ethernet Routing Switch 8600 Configuration — VLANs and Spanning Tree* (NN46205-517) .

SMLT topologies

Several common network topologies are used in SMLT networks. These include the SMLT triangle, the SMLT square, and the SMLT full-mesh.

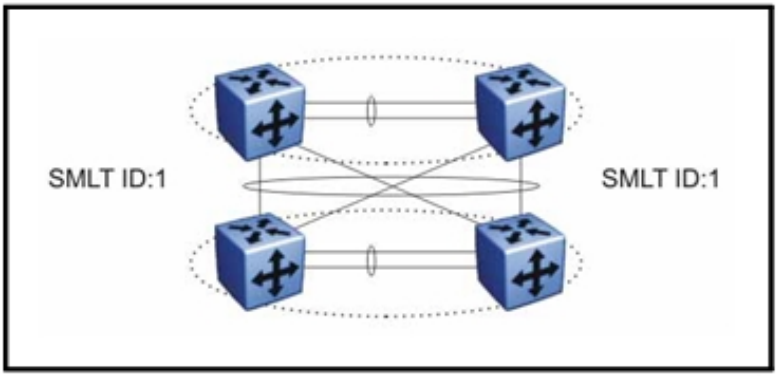
A triangle design is an SMLT configuration in which you connect edge switches or SMLT clients to two aggregation switches. You connect the aggregation switches together with an interswitch trunk that carries all the SMLTs configured on the switches. Each switch pair can have up to 31 SMLT client switch connections, and up to 512 SLT connections. When you use the square design (Figure 23 "SMLT square configuration" (page 113)), keep in mind that all links facing each other (denoted by the MLT ring on an aggregation pair) must use the same SMLT IDs.

Figure 23
SMLT square configuration



You can configure an SMLT full-mesh configuration as shown in Figure 24 "SMLT full-mesh configuration" (page 113). In this configuration, all SMLT ports must use the same SmltId (denoted by the MLT ring). Because the full-mesh configuration requires MLT-based SMLT, you cannot configure SLT in a full-mesh. In the following figure, the vertical and diagonal links emanating from any switch are part of an MLT.

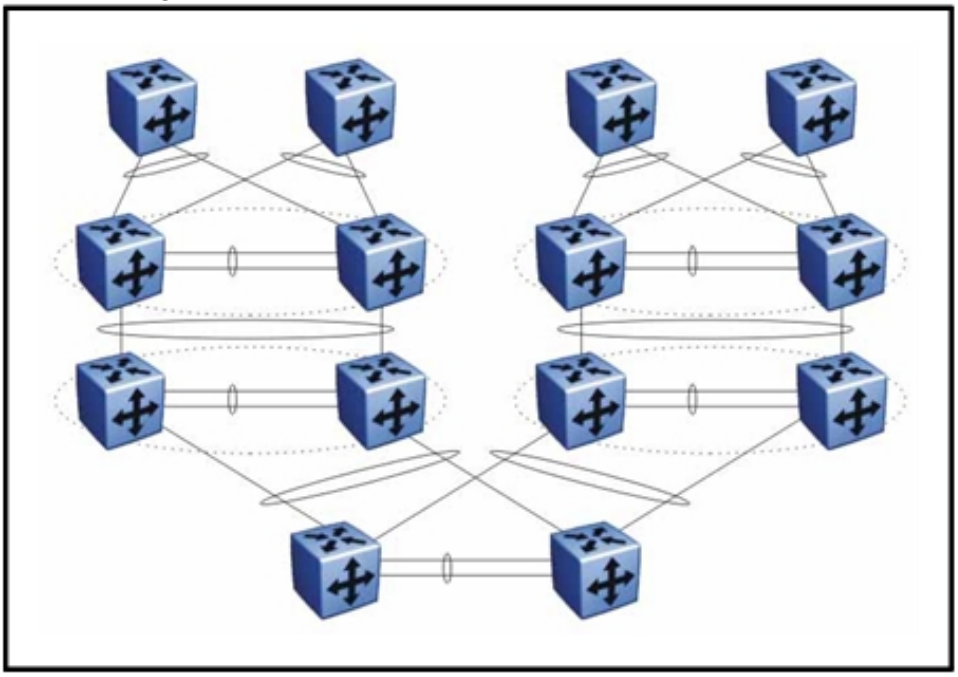
Figure 24
SMLT full-mesh configuration



R series modules, in Release 4.1 and later, and RS modules, in Release 5.0 and later, support up to 128 MLT groups of 8 ports. Within the network core, you can configure SMLT groups as shown in the following figure. Both sides of the links are configured for SMLT. No state information passes across the MLT link; both ends believe that the other is a single switch. The result is that no loop is introduced into the network. Any of the core switches or any of the connecting links between them may fail, but the network recovers rapidly.

You can scale SMLT groups to achieve hierarchical network designs by connecting SMLT groups together. This allows redundant loop-free Layer 2 domains that fully use all network links without using an additional protocol.

Figure 25
SMLT scaling



For more information about the SMLT triangle, square, and full-mesh designs, see *Nortel Ethernet Routing Switch 8600 Configuration — Link Aggregation, MLT, and SMLT* (NN46205-518) .

For more information about SMLT, see the Internet Draft draft-lapuh-network-smlt-06.txt available at www.ietf.org.

SMLT full-mesh recommendations with OSPF

In a full-mesh SMLT configuration between two clusters running OSPF (typically an RSMLT configuration), Nortel highly recommends that you place the MLT ports that form the square leg of the mesh (rather than

the cross connect) on lower numbered slots/ports. This is because CP-generated traffic is always sent out on the lower numbered MLT ports when active. This configuration will keep some OSPF adjacencies up in case the IST on one cluster fails. Without such a configuration, a booted switch in the scenario where the IST is also down can lose complete OSPF adjacency to both switches in the other cluster and therefore become isolated.

Routed SMLT

In many cases, core network convergence time depends on the length of time a routing protocol requires to successfully converge. Depending on the specific routing protocol, this convergence time can cause network interruptions ranging from seconds to minutes.

Routed Split Multilink Trunking (RSMLT) allows rapid failover for core topologies by providing an active-active router concept to core SMLT networks. RSMLT is supported on SMLT triangles, squares, and SMLT full-mesh topologies that have routing enabled on the core VLANs. RSMLT provides redundancy as well: if a core router fails, RSMLT provides packet forwarding, which eliminates dropped packets during convergence.

Routing protocols used to provide convergence can be any of the following: IP unicast static routes, RIPv1, RIPv2, OSPF, or BGP.

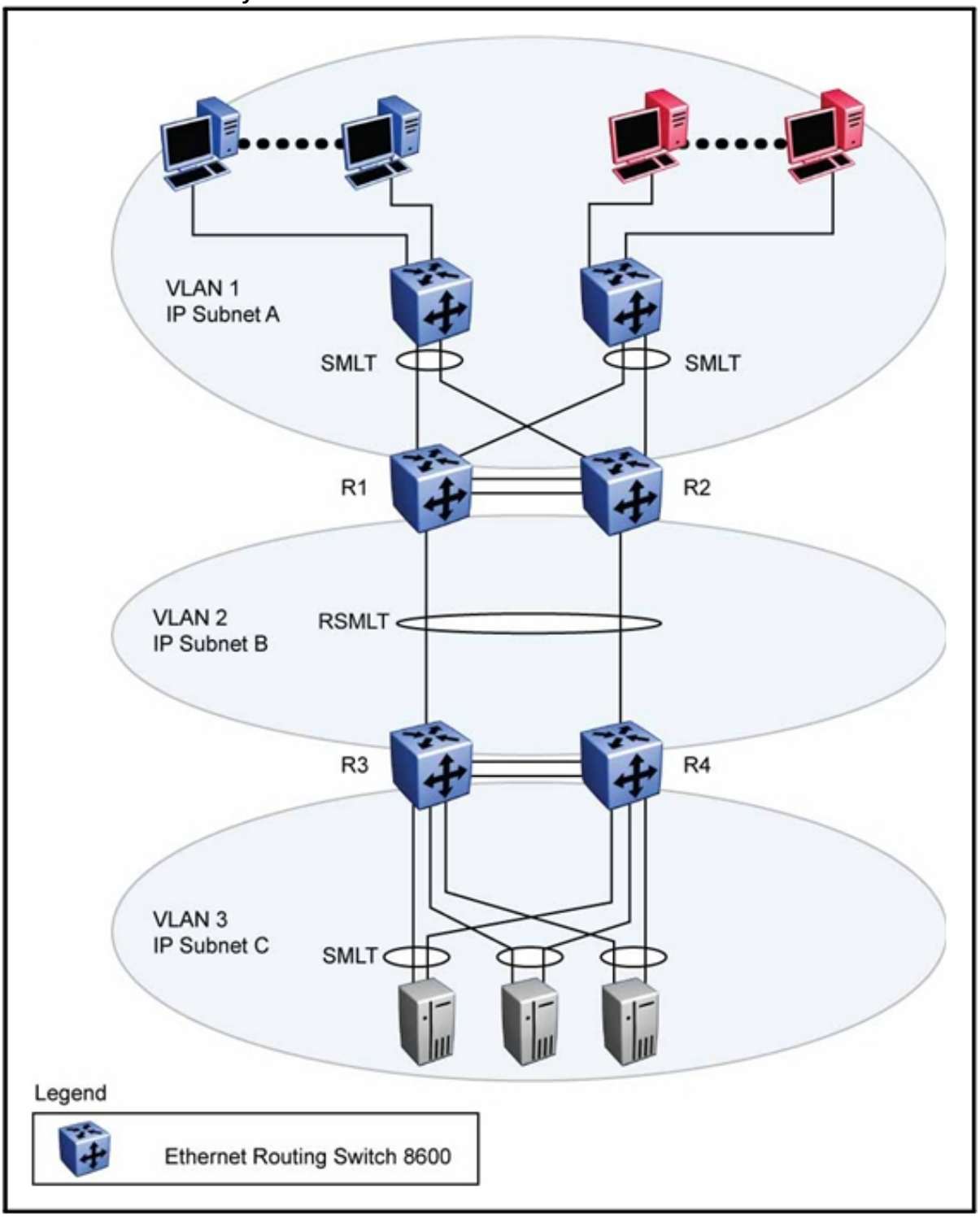
RSMLT navigation

- "SMLT and RSMLT operation " (page 115)
- "RSMLT router failure and recovery" (page 117)
- "RSMLT guidelines" (page 118)
- "RSMLT timer tuning" (page 118)
- "Example: RSMLT redundant network with bridged and routed edge VLANs" (page 119)
- "Example: RSMLT network with static routes at the access layer" (page 120)

SMLT and RSMLT operation

The following figure shows a typical redundant network with user aggregation, core, and server access layers. To minimize the creation of many IP subnets, one VLAN (VLAN 1, IP subnet A) spans all wiring closets. SMLT provides loop prevention and enables all links to forward to VLAN 1, IP Subnet A.

Figure 26
SMLT and RSMLT in Layer 3 environments



The aggregation layer switches are routing-enabled and provide active-active default gateway functions through RSMLT. Routers R1 and R2 forward traffic for IP subnet A. RSMLT provides both router failover and link failover. For example, if the SMLT link in between R2 and R4 are broken, the traffic fails over to R1.

For IP subnet A, VRRP Backup-Master can provide the same functions as RSMLT, as long as an additional router is not connected to IP subnet A.

RSMLT provides superior router redundancy in core networks (for example, IP subnet B) in which OSPF is used. Routers R1 and R2 provide router backup for each other—not only for the edge IP subnet A but also for the core IP subnet B. Similarly, routers R3 and R4 provide router redundancy for IP subnet C and also for core IP subnet B.

RSMLT router failure and recovery

This section describes the failure and recovery of router R1 in Figure 26 "SMLT and RSMLT in Layer 3 environments" (page 116).

R3 and R4 both use both R1 as their next-hop to reach IP subnet A. Even though R4 sends packets to R2, these packets are routed directly to subnet A at R2. R3 sends its packets towards R1; these packets are also sent directly to subnet A. When R1 fails, with the help of SMLT, all packets are directed to R2. R2 provides routing for R2 and R1.

After OSPF converges, R3 and R4 change their next-hop to R2 to reach IP subnet A. The network administrator can set the hold-up timer (that is, for the amount of time R2 routes for R1 in the event of failure) to a time period greater than the routing protocol convergence or to indefinite (that is, the pair always routes for each other). Nortel recommends that you set the hold up and hold down timer to 1.5 times the convergence time of the network.

In an application where RSMLT is used at the edge instead of VRRP, Nortel recommends that you set the hold-up timer value to indefinite.

When R1 reboots after a failure, it first becomes active as a VLAN bridge. Using the bridging forwarding table, packets destined to R1 are switched to R2 for as long as the hold-down timer is configured. These packets are routed at R2 for R1. Like VRRP, to converge routing tables, the hold-down timer value needs to be greater than the one required by the routing protocol.

When the hold-down time expires and the routing tables have converged, R1 starts routing packets for itself and also for R2. Therefore, it does not matter which one of the two routers is used as the next-hop from R3 and R4 to reach IP subnet A.

If single-homed IP subnets are configured on R1 or R2, Nortel recommends that you add another routed VLAN to the ISTs. As a traversal VLAN/subnet, this additional routed VLAN needs lower routing protocol metrics to avoid unnecessary ICMP redirect generation messages. This recommendation also applies to VRRP implementations.

RSMLT guidelines

Because RSMLT is based on SMLT, all SMLT configuration rules apply. In addition, RSMLT is enabled on the SMLT aggregation switches on a per-VLAN basis. The VLAN must be a member of SMLT links and the IST trunk.

The VLAN also must be routable (IP address configured). On all four routers in a square or full-mesh topology, an Interior Routing Protocol, such as OSPF, must be configured, although the protocol is independent from RSMLT.

You can use any routing protocol, including static routes, with RSMLT.

RSMLT pair switches provide backup for each other. As long as one of the two routers in an IST pair is active, traffic forwarding is available for both next-hops.

For design examples using RSMLT, see the following sections and "RSMLT redundant network with bridged and routed VLANs in the core" (page 290).

RSMLT timer tuning

RSMLT enables RSMLT peer switches to act as a router for its peer (by MAC address), which doubles router capacity and enables fast failover in the event of a peer switch failure. RSMLT provides hold-up and hold-down timer parameters to aid these functions.

The hold-up timer defines the length of time the RSMLT-peer switch routes for its peer after a peer switch failure. Configure the hold-up timer to at least 1.5 times greater than the routing protocol convergence time.

The RSMLT hold-down timer defines the length of time that the recovering/rebooting switch remains in a nonLayer 3 forwarding mode for MAC address of its peer. Configure the hold-down timer to at least 1.5 times greater than the routing protocol convergence time. The

configuration of the hold-down timer allows RIP, OSPF or BGP some time to build up the routing table before Layer 3 forwarding for the peer router MAC address begins again.

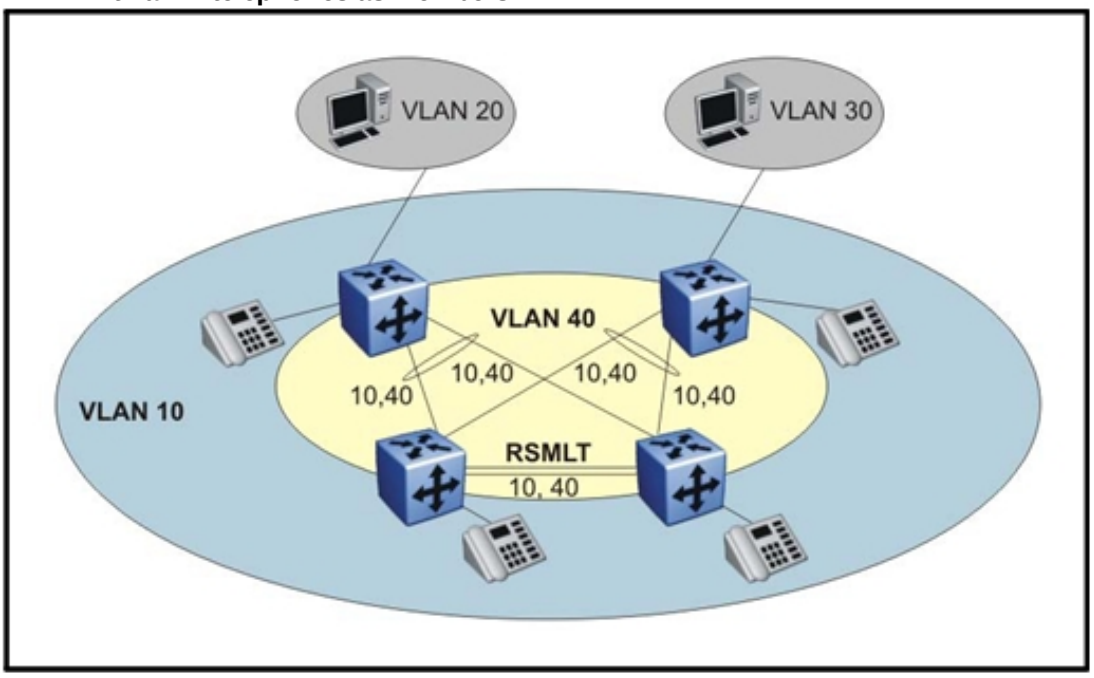
ATTENTION
If you use a Layer 3 SMLT client switch without a routing protocol, configure two static routes to point to both RSMLT switches or configure one static route. Set the RSMLT hold-up timer to 9999 (infinity). Nortel also recommends that you set the RSMLT hold-up timer to 9999 (infinity) for RSMLT Edge (Layer 2 RSMLT).

Example: RSMLT redundant network with bridged and routed edge VLANs

Many Enterprise networks require the support of VLANs that span multiple wiring closets as in, for example, a Voice over IP (VoIP) VLAN. VLANs are often local to wiring closets and routed towards the core. The following figure shows VLAN-10, which has all IP phones as members and resides everywhere, while at the same time VLANs 20 and 30 are user VLANs that are routed through VLAN-40.

A combination of SMLT and RSMLT provide sub-second failover for all VLANs bridged or routed. VLAN-40 is RSMLT enabled that provides for the required redundancy. You can use any unicast routing protocols—such as RIP, OSPF, or BGP—and routing convergence times do not impact the network convergence time provided by RSMLT.

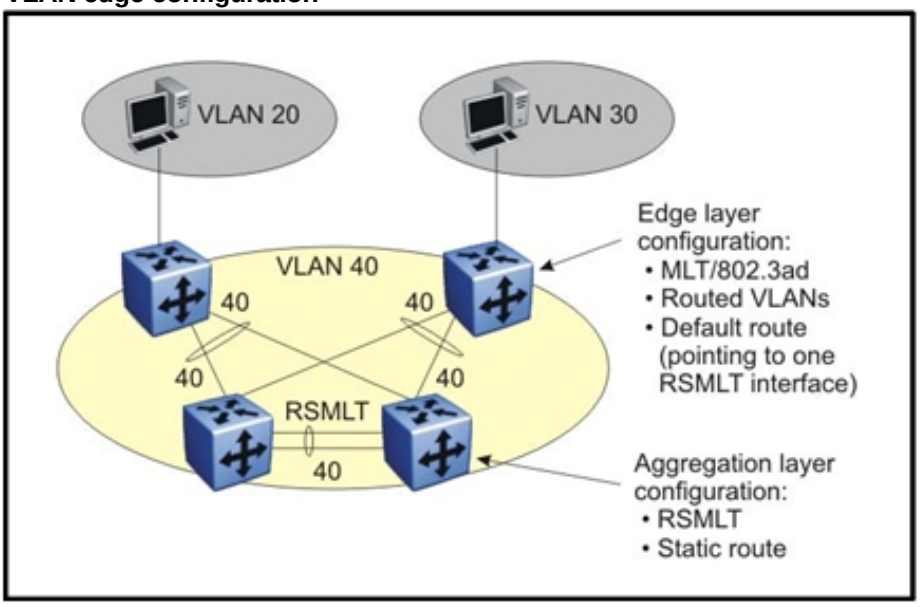
Figure 27
VLAN with all IP telephones as members



Example: RSMLT network with static routes at the access layer

You can use default routes that point towards the RSMLT IP interfaces of the aggregation layer to achieve a very robust redundant edge design, as shown in the following figure. As well, you can install a static route towards the edge.

Figure 28
VLAN edge configuration



Switch clustering topologies and interoperability with other products

When the Ethernet Routing Switch 8600 is used with other Ethernet Routing Switch products, the switch clustering bridging, unicast routing, and multicast routing configurations vary with switch type. Nortel recommends that you use the supported topologies and features when you perform inter-product switch clustering. For more information, see *Switch Clustering (SMLT/SLT/RSMLT/MSMLT) Supported Topologies and Interoperability with ERS 8600 / 5500 / 8300 / 1600* (NN48500-555) , available on the Nortel Technical Support Web site.

For specific design and configuration parameters, see *Converged Campus Technical Solutions Guide* (NN48500-516) and *Switch Clustering using Split-Multilink Trunking (SMLT) Technical Configuration Guide* (NN48500-518) .

sVLANs

Stacked VLANs (sVLAN) allow the 8000 Series switch to use a single VLAN to support customers who have multiple VLANs. The core service provider network carries sVLAN (802.1Q-in-Q) packets of multiple customers while maintaining customer configurations.

Navigation

Overview

sVLANs allow packets to have multiple tags, or stacked tags, so that service providers can transparently bridge tagged or untagged customer traffic through a core network.

The current provider bridging project in IEEE standard 802.1ad:

- provisions multiple Virtual Bridged LANs by using the common LAN equipment of a single organization
- uses a common infrastructure of Bridges and LANs to offer independent organizations the equivalent of separate LANs, Bridged LANs, or Virtual Bridged LANs
- allows tunneling of 802.1q tagged or untagged traffic through service provider core networks, which allows overlapping VLAN configurations
- offers improved VLAN scalability by summarizing VLANs into core VLANs
- offers improved VLAN scalability by using a layered architecture
- offers loop detection mechanisms

For conceptual and configuration sVLAN information, see *Nortel Ethernet Routing Switch 8600 Configuration — VLANs and Spanning Tree* (NN46205-517) .

ATTENTION
The sVLAN feature is available only with Classic modules. It is not available with R series modules.

sVLAN recommendations

This section describes sVLAN design information and guidelines.

When you design a multilevel sVLAN hierarchy, keep the physical layout of the hierarchy consistent with a logical layout based on the default Ether-type values for each sVLAN level. For example, if the sVLAN network consists of only one level, use default sVLAN level 1, which maps to Ether-type 8020. This eliminates any confusion or complexities in the engineering and support of the network.

Enable loop detection on all User-to-Network Interface (UNI) customer ports and on Split Multilink Trunking (SMLT) links. Do not use loop detection on interswitch trunks.

Because the sVLAN feature is based on regular VLAN bridging, all MAC addresses of an sVLAN are viewable by all provider switches that have this sVLAN provisioned.

Classic modules that have multiple physical ports share a common OctaPID. All ports on the same OctaPID must be configured either as normal ports or as UNI or Network-to-Network Interface (NNI) ports. For example, if port 1 on an 8648TX module is configured as a UNI port, then the remaining ports on that OctaPID (ports 2 to 8) must be configured either as UNI ports or NNI ports—they cannot be configured as normal tagged ports.

sVLAN MAC address learning

Duplicate MAC addresses with multiple levels of sVLANs can lead to connectivity problems.

Independent VLAN learning is only applicable within the VLAN context of the sVLAN first level. This means that a switch can apply a MAC address to a VLAN/sVLAN to maintain duplicate MAC addresses only as long as the addresses are in separate VLANs.

When multiple sVLAN levels are used, sVLANs are aggregated into levels. This process can introduce duplicate MAC addresses; they are learned on different ports. Duplicate MAC addresses result in a flapping MAC address from the provider NNI port to another provider NNI port, or from a customer UNI port.

Duplicate MAC addresses can be very common for control traffic such as VRRP. VRRP source MAC addresses are defined by Internet Engineering Task Force RFCs and therefore are used by many customers.

To overcome such issues, Nortel recommends that you connect routers to UNI ports. This limits the number of MAC addresses and reduces the potential of duplicate MAC addresses.

Management sVLAN

Normal VLANs are currently not supported on sVLAN NNI links. To transport regular VLANs in an sVLAN network, Nortel recommends that you use separate links between the core devices.

For management purposes, Nortel recommends that you define a management sVLAN and connect the external Ethernet management ports to the management sVLAN UNI ports. The management station must also be a member of this sVLAN or have a routing connection to it.

sVLAN restrictions

Note the following sVLAN restrictions:

- For the 8648 and 8632 modules, the eight 10/100 ports that share an OctaPID must run in the same mode—either normal or sVLAN UNI/NNI.
- For 8616 modules, the two gigabit ports that share an OctaPID must run in the same mode: either normal or sVLAN UNI/NNI
- The 8672 and 8684 modules do not support sVLAN.
- sVLAN NNI ports do not support normal VLANs
- Routing is not supported on sVLANs.
- IP filters are not supported on sVLANs.
- QoS can be applied through sVLAN QoS only (no filter support).
- sVLAN switches cannot be managed in-band. Nortel recommends an out-of-band network for management. Connect the Management port to a separate Management sVLAN, and bridge the management port to the Network Management System segment.

ATM guidelines

You can use an Asynchronous Transfer Mode (ATM) module to connect Ethernet networks to Wide Area Networks (WAN). The Ethernet Routing Switch 8672ATM module supports many configuration options for your ATM networks. This section highlights some general design factors and techniques you need to be aware of when you configure an Ethernet Routing Switch 8672ATM module.

For more information about the ATM modules, see *Nortel Ethernet Routing Switch 8600 Configuration — 8672ATME and 8672ATMM Modules* (NN46205-511) .

Navigation

ATM scalability

The maximum number of Ethernet Routing Switch 8672ATM modules supported per chassis:

- In a 10-slot chassis, 6 modules
- In a 6-slot chassis, 3 modules
- In a 3-slot chassis, 1 module

The maximum supported number of Emulated Local Area Networks (ELAN), Permanent Virtual Circuits (PVC), and Virtual Local Area Networks (VLAN) are:

- 256 RFC1483 bridged/routed ELANs per media dependent adapter (MDA)
- 500 RFC1483 bridged/routed ELANs per switch

  (you can configure 12 more RFC1483 bridged ELANs per switch)
- 64 PVCs per RFC1483 bridged ELAN
- 1 PVC per RFC1483 routed ELAN

ATM performance

Because ATM uses a fixed cell size, the Ethernet Routing Switch 8672ATM interface exhibits throughput of less than 50% of link bandwidth when it processes a continuous stream of small packets (less than 512 bytes). However, in a real network scenario, the Ethernet Routing Switch 8672ATM interface throughput is generally close to line rate. Nortel tested this scenario by simultaneously sending multiple packet sizes over the ATM link. The observed throughput was:

- For OC-3 bridged, the throughput is 125.9 Mbit/s
- For OC-3 routed, the throughput is 126.9 Mbit/s
- For OC-12 bridged, the throughput is 520.6 Mbit/s
- For OC-12 routed, the throughput is 507.4 Mbit/s

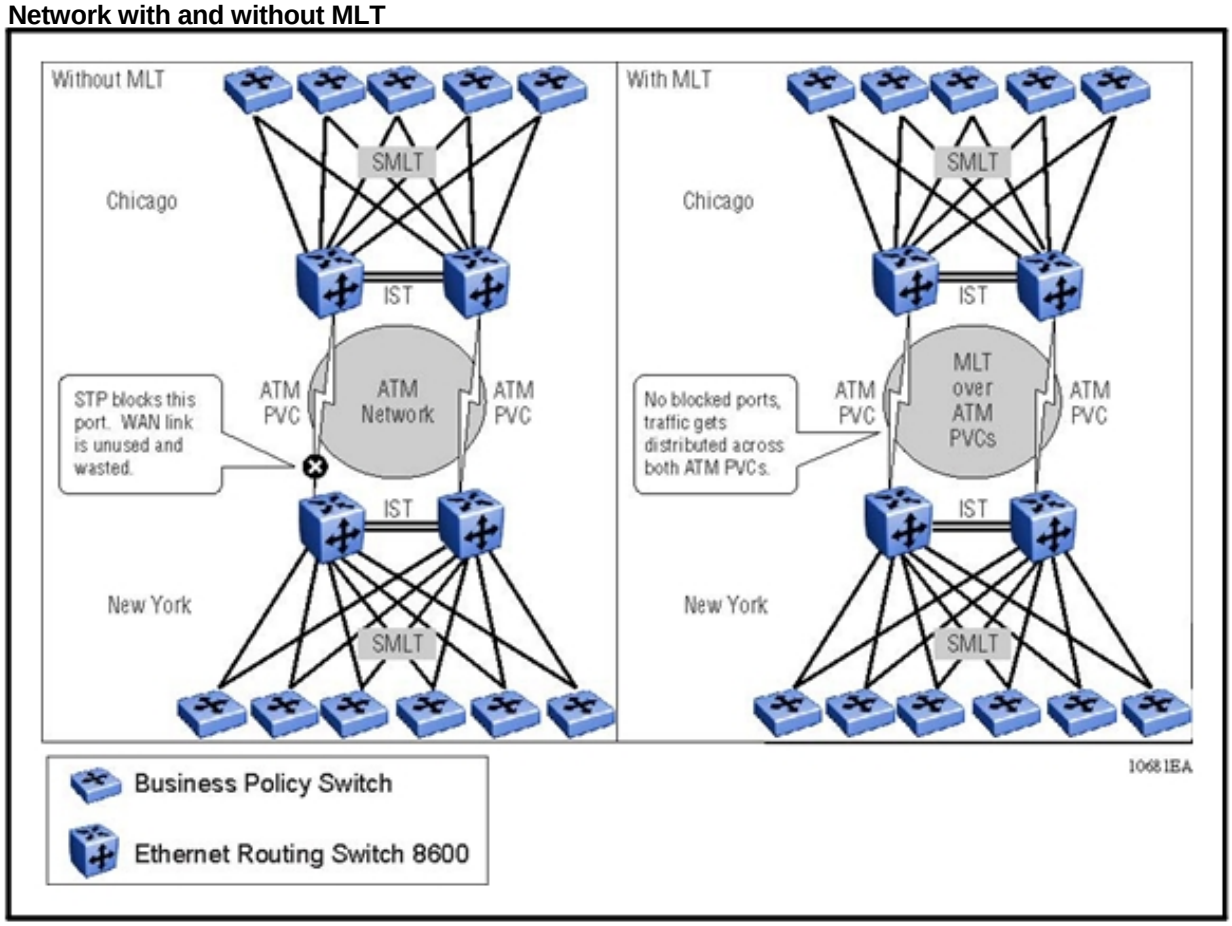
Tests involved simultaneously sending 64, 128, 512, 1024, 1280, and 1518 bytes of traffic from a 1 Gbit/s Smartbit port over an ATM link in both directions. The Smartflow application ensures that large packet sizes contribute much more to the link bandwidth than small packet sizes (Smartflow simulates a real network scenario).

ATM resiliency

When used with MultiLink Trunking (MLT) and Split MultiLink Trunking (SMLT), the Ethernet Routing Switch 8672ATM module provides resiliency. By using MLT, Spanning Tree Protocol (STP) for the WAN links is not required; the ATM link is not blocked by STP. MLT removes the need to use Spanning Tree and utilizes the trunk group in a load-sharing manner based on the MAC address hash algorithm.

SMLT introduces greater redundancy by allowing an MLT to terminate at two separate colocated switches (which acts as one switch). This protects against switch failure at the core. The Ethernet Routing Switch 8672ATM module supports MLT and SMLT, with the restriction that the interswitch trunk (IST) link between two switches cannot be ATM. The following figure compares a network that uses STP and a network that uses SMLT across ATM WAN links.

Figure 29
Network with and without MLT

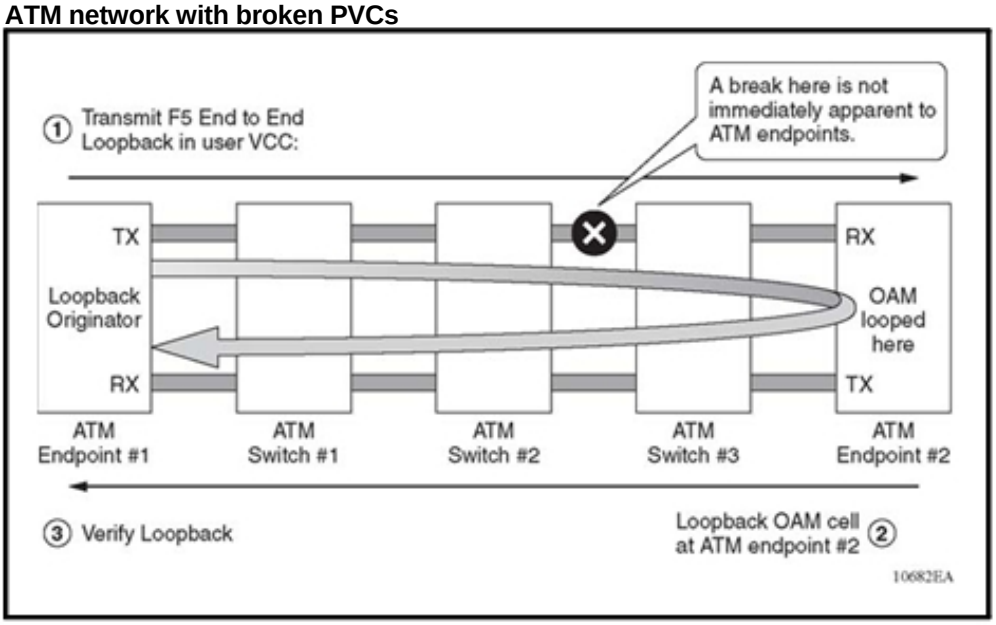


A separate PVC is required for each VLAN that is carried on a trunk.

F5 OAM loopback request/reply

The F5 Operations and Maintenance (OAM) Loopback Request/reply feature detects downstream failures within an ATM network. Without this feature, no notification of a broken PVC within the ATM network is sent. Packets are still sent down the link and are lost. The F5 OAM feature uses an ATM loopback cell that is sent end-to-end to determine the health of the link. If a configured number of loopback cells is not returned by the far end device, the PVC is declared down.

Figure 30
ATM network with broken PVCs



To use the F5 OAM feature, all PVCs on the port must have the F5 OAM loopback feature enabled. To propagate the failure up to higher layer protocols, such as Open Shortest Path First (OSPF) and MLT, all PVCs must fail the loopback test. Therefore, the feature may work well in point-to-point scenarios but not for other topologies. For example, if a port has multiple PVCs that terminate at different locations (for example, hub and spoke), failure of a single PVC does not propagate up to higher levels, and features such as MLT do not work. Traffic is still sent out of this PVC and is lost until a bridging or routing timer expires.

By default, the F5 OAM Loopback Request feature is disabled.

ATM considerations

Keep in mind the feature considerations described in the following sections when you configure an Ethernet Routing Switch 8672ATM module.

ATM considerations navigation

- "ATM and MLT" (page 129)
- "ATM and 802.1q tags" (page 129)
- "ATM and DiffServ" (page 129)
- "ATM and IP multicast" (page 129)
- "ATM traffic shaping" (page 130)
- "ATM and ingress port mirroring" (page 130)

ATM and MLT

If you add an ATM port to an MLT VLAN, it can belong only to that MLT VLAN, and to no other 802.1q-tagged VLANs.

ATM and 802.1q tags

802.1q trunk configuration over ATM links is not supported because the 802.1q tag is removed at the ATM egress interface. As a result, multiple Ethernet VLANs can not be carried across a single ATM PVC. However, multiple Ethernet VLANs can be carried across an ATM link by using individual PVCs mapped to individual VLANs.

ATM and DiffServ

The Ethernet Routing Switch 8672ATM module can function as a DiffServ core port for the Ethernet Routing Switch 8600. DiffServ CodePoint (DSCP) marking by Ethernet ports behaving as DiffServ access ports are preserved over ATM links.

The DSCP does not map directly to ATM Class of Service. To map Quality of Service (QoS) levels to ATM Class of Service levels, assign a QoS level to the Ethernet VLAN and configure a variable bit rate (VBR) Class of Service for a PVC in that VLAN.

ATM and IP multicast

If you configure a network in point-to-multipoint mode (hub to spoke mode) and connect a central Ethernet Routing Switch 8600 to several switches with PVCs on the same port in the same VLAN, multicast traffic is flooded on these PVCs if just one of them has a member of an active multicast group. Internet Group Membership Protocol (IGMP) does not distinguish between different PVCs on the same VLAN when they are configured on the same port.

Nortel recommends that you do not use IP multicast over ATM in point-to-multipoint mode. However, if such a configuration is essential to your network requirements, you must ensure that traffic that floods all PVCs is required on these PVCs. Ensure that the traffic does not use a high amount of bandwidth that can lead to loss of performance or an application malfunction, as is the case with television or streaming applications.

If IP multicast over ATM is an essential requirement, use PVCs on different VLANs that connect to the central switch and use routing between these PVCs. If PVCs must be on the same VLAN, use different ports for these PVCs so that IP multicast traffic flows only on the ports/PVCs with receivers.

Some implications of configuring IP multicast with ATM PVCs include the following:

- Multicast data sent from a PVC on a port is not received by another PVC on the same port on the same VLAN.
- Multicast data sent from a PVC on a port is multicast to all other PVCs in the same VLAN on different ports if they have multicast receivers.
- Multicast data sent from a PVC on a port is multicast to all other VLANs on the port and to other ports with Multicast routing enabled if they have multicast receivers.

When you use the IGMP Fast Leave feature for PVCs on the same port (or VLAN) that are flooded with traffic for a given group, if one member leaves the group, all traffic for this group stops on all PVCs on the port or VLAN.

ATM traffic shaping

When connecting to a service provider ATM network, shape your egress flows so as not to exceed the traffic contract negotiated with the service provider. Cells that do not meet the traffic contract are either discarded immediately or tagged for discard if congestion is encountered further downstream.

The Ethernet Routing Switch 8672 ATM module supports shaping on a per-Virtual Circuit (VC) basis. For each PVC, you can configure the Peak Cell Rate (PCR), Sustained Cell Rate (SCR), and Maximum Burst Size (MBS) parameters. For Variable Bit Rate (VBR) service, a channel can burst at the PCR for MBS cells. If the MBS is exhausted, the channel reduces to the SCR while credits are accumulated to support another burst. The minimum PCR or SCR for a channel is 86 cells/second (cells/s) or 36.67 kbit/s. The maximum shaping rate per PVC is half of the link rate (353207 cells/s for OC-3/STM-1 and 733 490 cells/sec for OC-12/STM-4).

ATM and ingress port mirroring

The 8672 ATM module supports port mirroring. The module removes the SONET framing and the ATM header from the ingress packets. The appropriate 802.3 header, including the 802.1q tag, is added to the packet. The frame is buffered, queued, and segmented, and then forwarded to the switch fabric. Mirroring is performed by replicating each cell as it is delivered to the switch fabric and monitoring port.

For more information on port mirroring, see *Passport 8600 Technical Configuration Guide For Remote Port Mirroring*.

ATM applications

Use this section when you consider designs that use the ATM module.

ATM applications navigation

ATM WAN connectivity and OE/ATM interworking

In a typical enterprise environment, WAN connectivity can be achieved by several means:

- point-to-point leased line
- Frame Relay
- Packet Over SONET (POS) or ATM

You can use the Ethernet Routing Switch 8672ATM module to provide WAN connectivity for sites that have ready access to an ATM network. ATM is more economical than a leased line, and provides higher bandwidth than Frame Relay.

In a Carrier environment, you can use the Ethernet Routing Switch 8672ATM module as an interworking point between new Optical Ethernet architectures and existing revenue-generating ATM networks. ATM service can provide a bridge until a full gigabit Ethernet core with MultiProtocol Label Switching is realized. In addition, you can use the Ethernet Routing Switch 8672ATM module to bring services into an aggregation Point-of-Presence (PoP). Access sites that are traditionally served by ATM, or are not serviced by dark fiber, can still use ATM to reach the aggregation site, which may have already migrated to a gigabit Ethernet/dark fiber access architecture.

Point-to-point WAN connectivity

Point-to-point is the simplest and most common application for the Ethernet Routing Switch 8672ATM module. Enterprise network sites can be interconnected over a service provider ATM network. The use of the Spanning Tree Protocol (STP) can potentially require that one of the expensive WAN links be in a blocked state, which wastes valuable resources. To avoid such a situation, Nortel recommends that you instead use MLT or SMLT, rather than STP, to provide Layer 2 redundancy.

For efficiency, SMLT provides a load-sharing mechanism across the two WAN links. Across high-priority sites, running the two links across different service providers adds another layer of resiliency and reduces dependency on any one network.

For routed backbones, OSPF and Equal Cost MultiPath (ECMP) can loadshare across multiple ATM links (assuming equal metrics) and provide the required resiliency.

Optical Ethernet and ATM interworking

Service providers are faced with the challenge of interworking existing ATM networks with new Optical Ethernet (OE) networks. Today, the majority of data network revenue still comes from Frame Relay and ATM networks. Optical Ethernet is a new architecture that allows seamless Layer 2 Ethernet connectivity for Enterprise users across both MANs and WANs. The Ethernet Routing Switch 8600, with the Ethernet Routing Switch 8672ATM module, can be viewed as a way to extend the reach of ATM networks and services into an Optical Ethernet arena, and vice versa.

In the following figure, the Ethernet Routing Switch 8672ATM module is used to bring remote sites into an aggregation PoP using public ATM service.

An ATM connection is useful for sites that may not be reachable via gigabit Ethernet due to lack of dark fiber, but are readily serviced by public ATM. In this scenario, use high speed gigabit Ethernet uplinks, rather than ATM links, to connect the Ethernet Routing Switch 8600 to the core. ATM to ATM switching through an Ethernet Routing Switch 8600 is not recommended because the Ethernet Routing Switch 8672ATM module is a User-to-Network Interface (UNI) device, and not a Network to Network Interface (NNI) device.

Figure 31
Bringing remote sites into an aggregation PoP

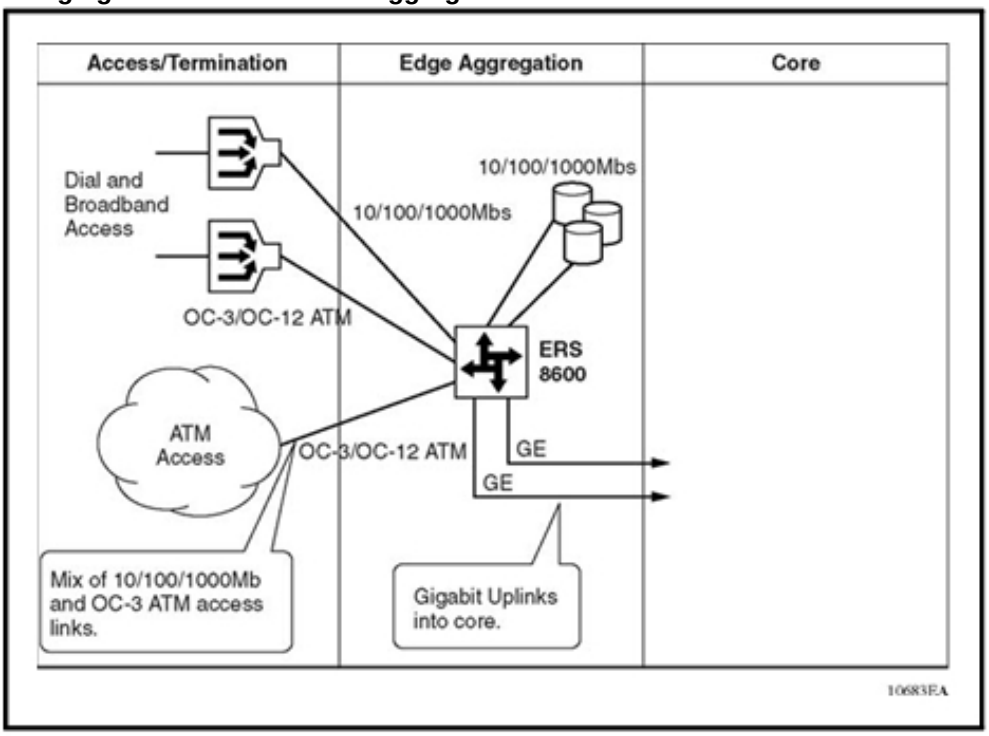
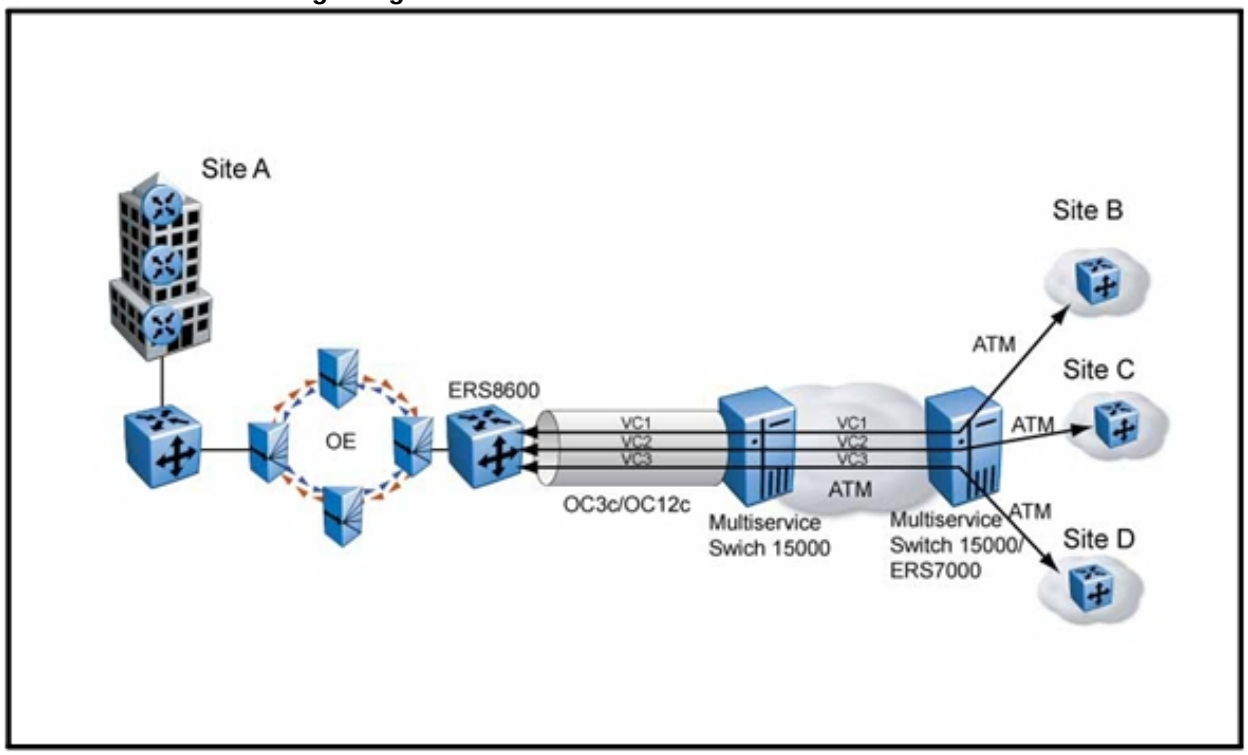


Figure 32 "OE and ATM interworking using home-run PVCs" (page 134), Figure 33 "OE and ATM interworking using RFC 1483 bridge termination" (page 135), and Figure 34 "OE and ATM interworking using RFC 1483 bridge termination with cVRs" (page 136) provide a detailed view about how the Ethernet Routing Switch 8600 can facilitate the interworking of OE and ATM networks based on the Multiservice Switch 15000.

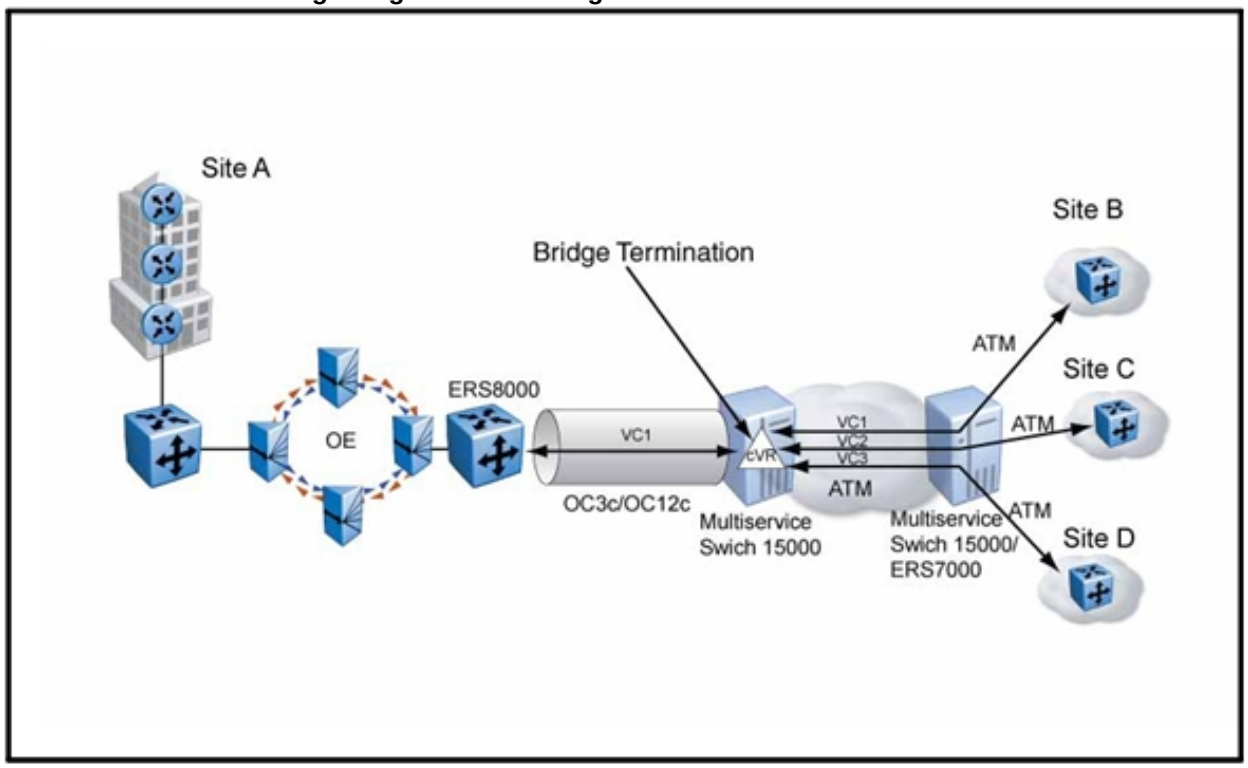
Figure 32
OE and ATM interworking using home-run PVCs



In Figure 32 "OE and ATM interworking using home-run PVCs" (page 134), ATM PVCs are mapped one to one from each remote site through the ATM network (via Multiservice Switch 15000), and terminate on the Ethernet Routing Switch 8672ATM module. You can bridge from each remote ATM site to anywhere on the OE ring, but remote ATM sites cannot bridge to each other. This should not be an issue because most remote sites need to access a data center on the OE core.

The solution shown in Figure 32 "OE and ATM interworking using home-run PVCs" (page 134) works well for a customer that has a number of key sites on the OE ring, but who needs to bring in some remote sites that do not have direct access to the ring. The ATM network can span a wide area (across MANs) or can be a local ATM access network within a metropolitan area. For large numbers of remote sites, consider "ATM scalability" (page 125) restrictions. A one-to-one mapping of PVCs to remote sites means that for any single ELAN can support up to 64 remote sites (PVCs).

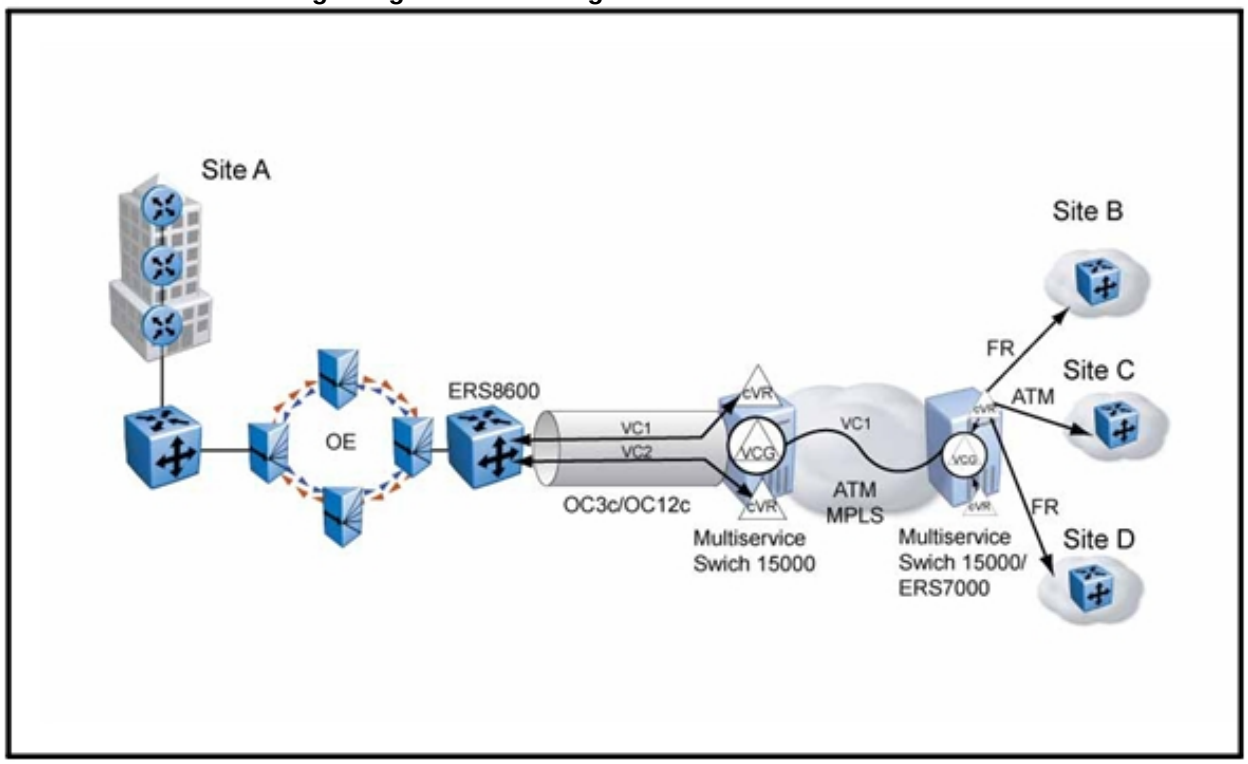
Figure 33
OE and ATM interworking using RFC 1483 bridge termination



In Figure 33 "OE and ATM interworking using RFC 1483 bridge termination" (page 135), routed PVCs in the ATM network bring the remote sites into a customer Virtual Router (cVR) on Multiservice Switch 15000 at the interworking point. A single bridged PVC per customer is used to interconnect the Ethernet Routing Switch 8600 to the Multiservice Switch 15000. Different customers home to different cVRs at the interworking point, with a different PVC between Ethernet Routing Switch 8600 and Multiservice Switch 15000. Bridged frames are terminated on the Multiservice Switch 15000 and routed over ATM PVCs to their destination. This solution reduces the number of PVCs between the Ethernet Routing Switch 8600 and Multiservice Switch 15000, and is a hybrid routed/bridged solution (no bridging end-to-end).

If necessary, you can use multiple ATM ports between the Ethernet Routing Switch 8600 and the Multiservice Switch 15000. The multiple ports must be located on different MDAs because ports on an MDA share the same forwarding engine. Use egress shaping on the Multiservice Switch 15000 because the Ethernet Routing Switch 8672ATM module does not support policing. Without shaping, a burst of traffic from the Multiservice Switch 15000 may overwhelm the capabilities of the Ethernet Routing Switch 8672ATM module, leading to lost cells and poor throughput.

Figure 34
OE and ATM interworking using RFC 1483 bridge termination with cVRs



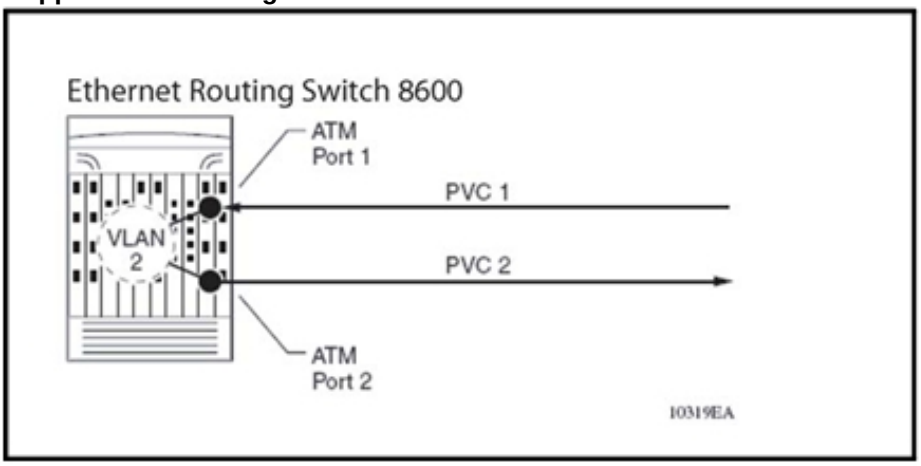
The solution shown in Figure 34 "OE and ATM interworking using RFC 1483 bridge termination with cVRs" (page 136) is similar to the routed/bridged model shown in Figure 33 "OE and ATM interworking using RFC 1483 bridge termination" (page 135). However, it uses cVRs at each customer PoP to provide a variety of connectivity options to the customer site, including Frame Relay. Multiple customers can be supported at each site by using multiple cVRs. In this example, the cVRs are aggregated with a virtual connection gateway (VCG) into the ATM core. This simplifies engineering, but the cVRs must share the bandwidth of a single VC.

Another option is to provide premium service by using an individual mesh for each set of cVRs, which provides a reserved VC for each cVR.

Transparent LAN services

The following figure shows a supported transparent LAN services (TLS) configuration.

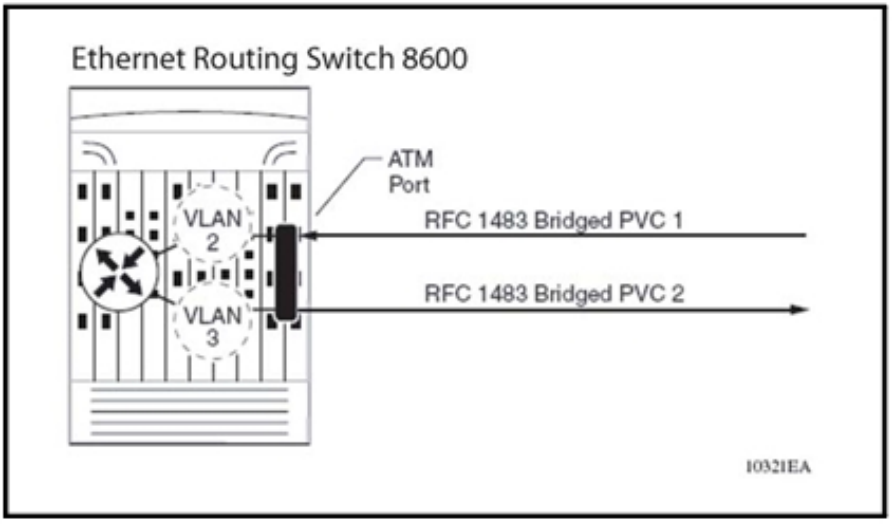
Figure 35
Supported TLS configuration



The Ethernet Routing Switch 8672ATM module does not allow bridging between PVCs in the same VLAN on the same ATM port. This is generally not an issue if the application requires many remote sites to access a central resource, rather than exchanging traffic between different remote sites. Not allowing bridging can enhance security because you can place multiple customers in the same VLAN on the same ATM port and they cannot see each others' traffic, including broadcast traffic.

Alternatively, you can place two bridged PVCs in different VLANs on the same ATM port, and route between them using routing on the Ethernet Routing Switch 8600 (see Figure 36 "PVCs in different VLANs on the same ATM port" (page 137)).

Figure 36
PVCs in different VLANs on the same ATM port



Video over DSL over ATM

Nortel recommends that you do not configure the Ethernet Routing Switch 8672ATM module to support Video over Digital Subscriber Line (DSL) over ATM applications. Asymmetric Digital Subscriber Line (ADSL) and Digital Subscriber Line Access Multiplexer (DSLAM) devices have the following limitations:

- ADSL has bandwidth limitations which make it difficult to support two video channels for each subscriber
- DSLAM devices do not support IGMP Snoop
- DSLAM devices do not support customer-based security features

Security issues relating to point-to-multipoint bridged ATM PVC mode arise because multicast traffic is treated as broadcast traffic for that VLAN on the ATM port. This characteristic limits the use of point-to-multipoint bridged ATM PVC mode to a single customer or subscriber per Ethernet VLAN on the same ATM port.

The limitation of one customer per Ethernet VLAN limits the total number of subscribers to an Ethernet Routing Switch 8600 to 500 because this is the maximum number of routable VLANs currently supported. An Ethernet Routing Switch 8600 supports 1972 bridged VLANs. The number of ATM PVCs available on the Ethernet Routing Switch 8672ATM module is a secondary issue with respect to video over DSL applications.

ATM and voice traffic recommendations

The use of ATM to carry voice traffic has significant challenges. Note the following Ethernet Routing Switch 8600 and voice application characteristics:

- The Ethernet Routing Switch 8600 architecture is optimized for frame-based applications. The introduction of ATM cell-based interfaces, such as OC-3, OC-12, or DS3, in such a system is a challenge in itself.
- The conversion of Ethernet frames to fixed-size ATM cells involves considerable overhead. This overhead can have a significant performance impact, especially for small Ethernet frame sizes (less than 512 bytes).
- The inherent characteristics of voice traffic include:
  - — The average Ethernet frame size is 120 Bytes.
  - — The bursty nature of voice traffic is not efficient for ATM systems.
- Voice applications have very small delay and jitter tolerance.

When using the Ethernet Routing Switch 8672ATM module for voice applications, Nortel recommends that you:

- Under-provision the Ethernet Routing Switch 8672ATM link bandwidth when using it for voice over IP (VoIP) or ATM applications (that is, provision your network in such a way that ATM link is not oversubscribed and is not running at line rate).
- Ensure that a good mix of small and large packets traverse the ATM link by designating only 20% of the link bandwidth for voice traffic.
- Use separate VLANs for voice and data traffic.
- Assign a higher QoS level to voice traffic VLANs.
- Map voice traffic VLANs to VBR PVCs and choose the SCR carefully. By doing this, the Sustained Cell Rate is guaranteed for the voice traffic.

ATM segmentation and reassembly incurs some additional delay when a packet is transferred through a switch, ranging from 65 microseconds to 232 microseconds for very large packets. However, the total latency introduced is well within the required tolerance (milliseconds) for time-sensitive applications such as VoIP. Testing was performed with a single flow of traffic between back-to-back Ethernet Routing Switch 8600s connected with gigabit Ethernet trunks, and then ATM trunks, with traffic at 50% or below to avoid congestion.

Layer 2 loop prevention

To use bandwidth and network resources efficiently, prevent layer 2 data loops. Use the information in this section to help you use loop prevention mechanisms.

Navigation

Spanning tree

Spanning Tree prevents loops in switched networks. The Ethernet Routing Switch 8600 supports several spanning tree protocols and implementations. These include the Spanning Tree Protocol (STP), Per-VLAN Spanning Tree Plus (PVST+), Rapid Spanning Tree Protocol (RSTP), and Multiple Spanning Tree Protocol (MSTP). This section describes some issues to consider when you configure spanning tree.

For more information about spanning tree protocols, see *Nortel Ethernet Routing Switch 8600 Configuration — VLANs and Spanning Tree* (NN46205-517) .

Spanning tree navigation

Spanning Tree Protocol

Use Spanning Tree Protocol (STP) to prevent loops in your network. This section provides some STP guidelines.

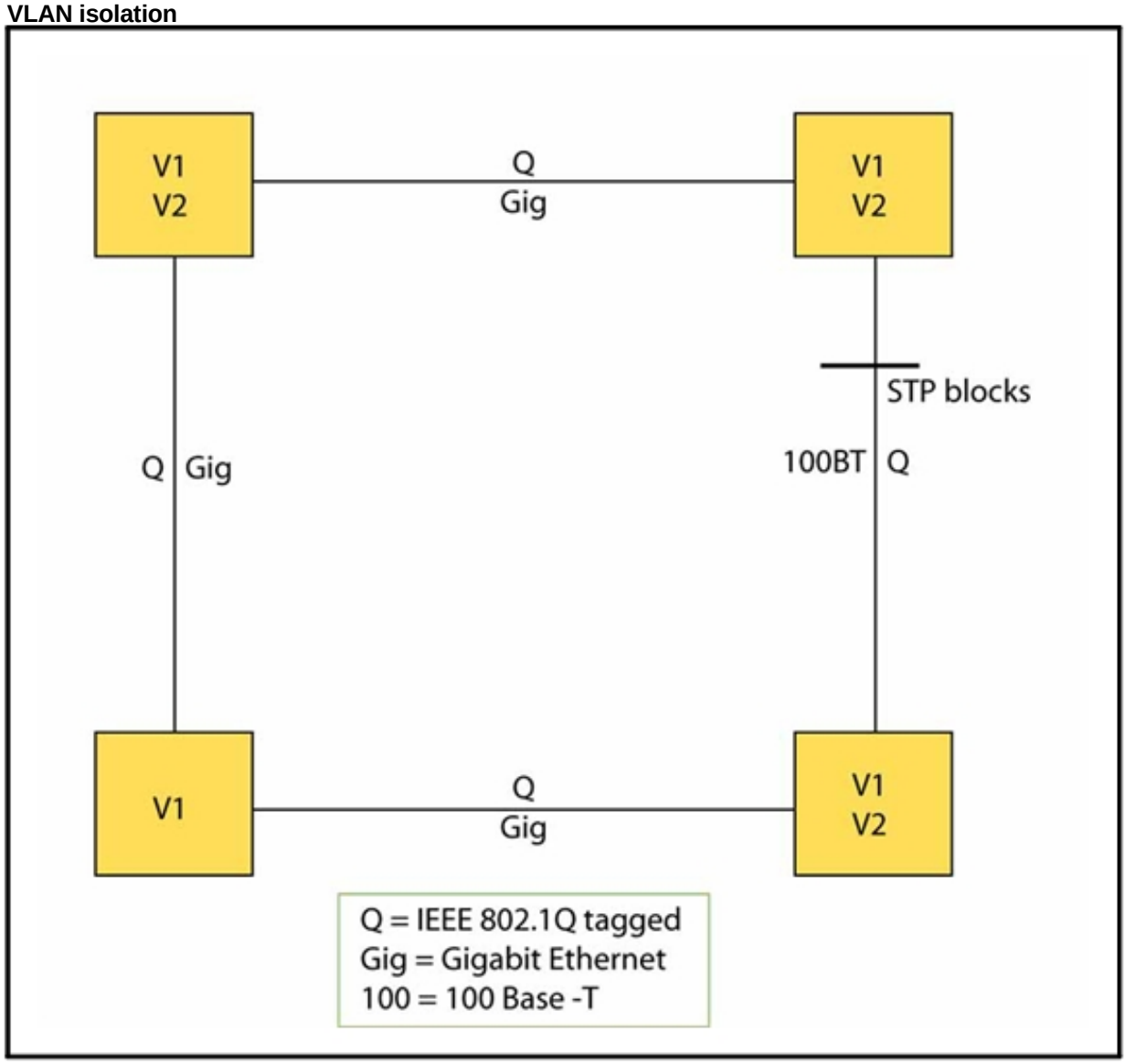
STP and BPDU forwarding

You can enable or disable STP at the port or at the spanning tree group (STG) level. If you disable the protocol at the STG level, Bridge Protocol Data Units (BPDU) received on one port in the STG are flooded to all ports of this STG regardless of whether the STG is disabled or enabled on a per port basis. When you disable STP at the port level and STG is enabled globally, the BPDUs received on this port are discarded by the CPU.

Spanning Tree and protection against isolated VLANs

Virtual Local Area Network (VLAN) isolation disrupts packet forwarding. The problem is shown in the following figure. Four devices are connected by two VLANs (V1 and V2) and both VLANs are in the same STG. V2 includes three of the four devices, whereas V1 includes all four devices. When the Spanning Tree Protocol detects a loop, it blocks the link with the highest link cost. In this case, the 100 Mbit/s link is blocked, which isolates a device in V2. To avoid this problem, either configure V2 on all four devices or use a different STG for each VLAN.

Figure 37
VLAN isolation

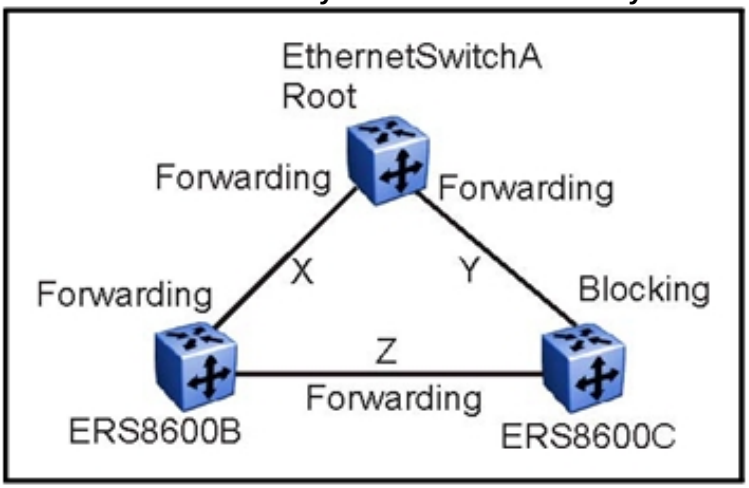


Multiple STG interoperability with single STG devices

Nortel provides multiple spanning tree group (STG) interoperability with single STG devices. When you connect the Ethernet Routing Switch 8600 with Layer 2 switches, be aware of the differences in STG support between the two types of devices. Some switches support only one STG, whereas the Ethernet Routing Switch 8600 supports 25 STGs.

In the following figure, all three devices are members of STG1 and VLAN1. Link Y is in a blocking state to prevent a loop, and links X and Z are in a forwarding state. With this configuration, congestion on link X is possible because it is the only link that forwards traffic between EthernetSwitchA and ERS8600C.

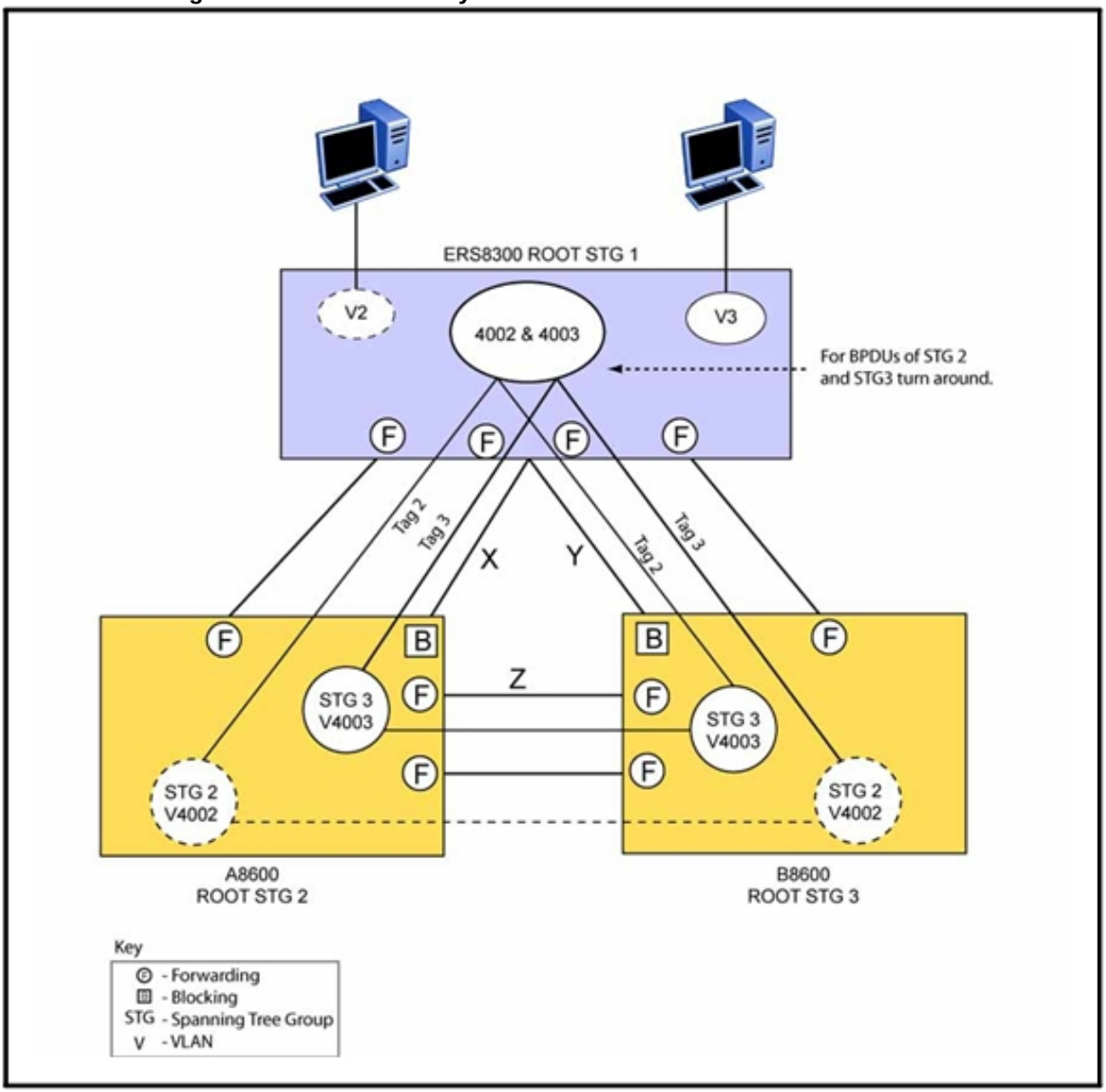
Figure 38
One STG between two Layer 3 devices and one Layer 2 device



To provide load sharing over links X and Y, create a configuration with multiple STGs that are transparent to the Layer 2 device and that divide the traffic over different VLANs. To ensure that the multiple STGs are transparent to the Layer 2 switch, the BPDUs for the two new STGs (STG2 and STG3) must be treated by the Ethernet Switch as regular traffic, not as BPDUs.

In the configuration in Figure 39 "Alternative configuration for STG and Layer 2 devices" (page 145), the BPDUs generated by the two STGs (STG2 and STG3) are forwarded by the Ethernet Switch 8100. To create this configuration, you must configure STGs on the two Ethernet Routing Switch 8600s, assign specific MAC addresses to the BPDUs created by the two new STGs, create VLANs 4002 and 4003 on the Layer 2 device, and create two new VLANs (VLAN 2 and VLAN 3) on all three devices.

Figure 39
Alternative configuration for STG and Layer 2 devices



When you create STG2 and STG3, you must specify the source MAC addresses of the BPDUs generated by the STGs. With these MAC addresses, the Layer 2 switch does not process the STG2 and STG3 BPDUs as BPDUs, but forwards them as regular traffic.

To change the MAC address, you must create the STGs and assign the MAC addresses as you create these STGs. You can change the MAC address by using the CLI command **`config stg <stgid> create [vlan <value>] [mac <value>]`**.

In the NNCLI, the command is **`spanning-tree stp <1-64> create`**.

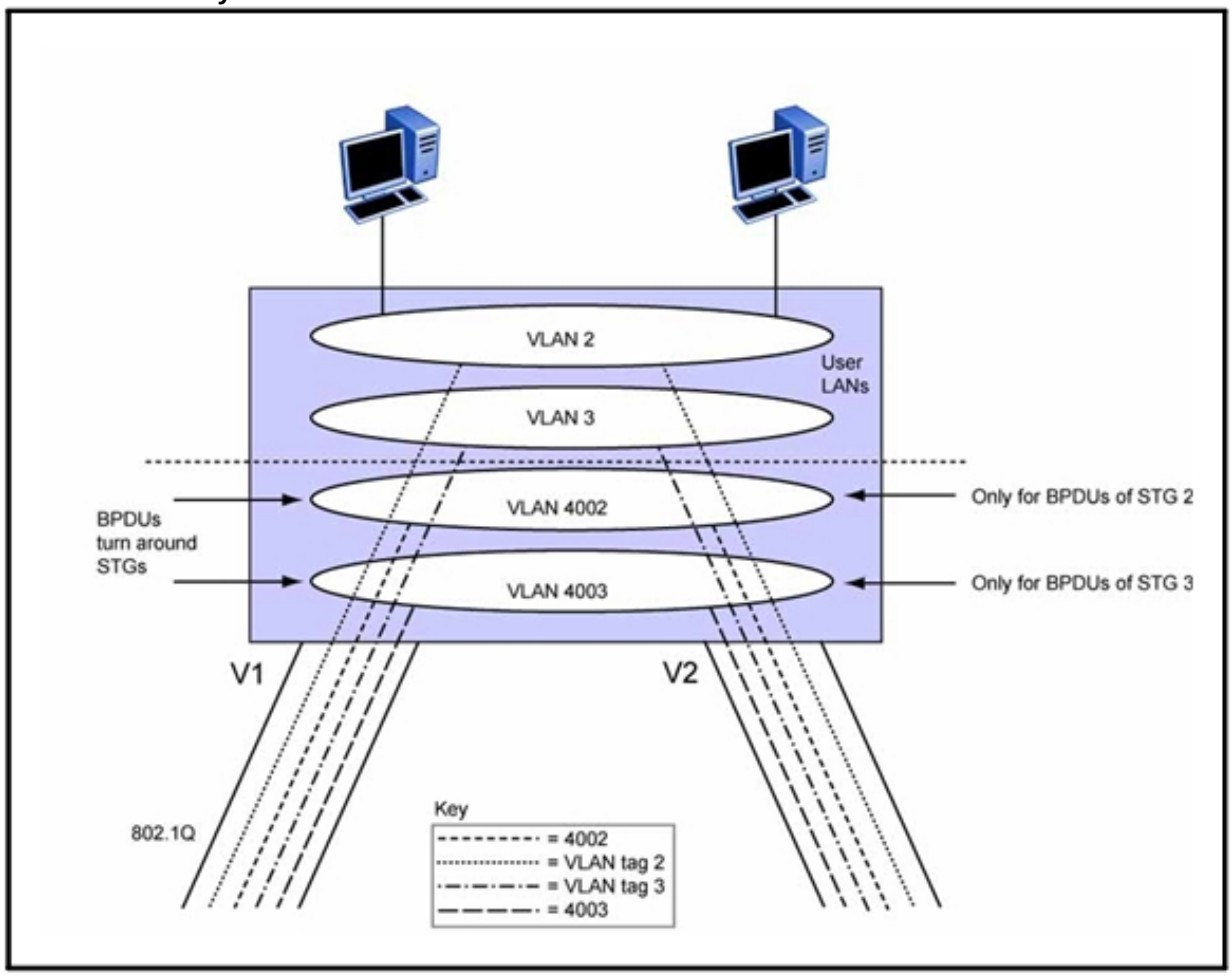
On the Ethernet Routing Switch 8600s (A8600 and B8600), configure A8600 as the root of STG2 and B8600 as the root of STG3. On the Ethernet Switch 8100 (Layer 2), configure A8600 as the root of STG1. Configure a switch to be the root of an STG by giving it the lowest root bridge priority.

Configure the four VLANs on the Layer 2 switch to include the tagged ports connected to the Ethernet Routing Switch 8600. To ensure that the BPDUs from STG2 and STG3 are seen by the Layer 2 switch as traffic for the two VLANs, and not as BPDUs, give two of the VLANs the IDs 4002 and 4003. Figure 40 "VLANs on the Layer 2 switch" (page 147) illustrates the four VLANs configured on the Ethernet Switch 8100 and the traffic associated with each VLAN.

After you configure the Ethernet Switch 8100, configure VLAN 2 and VLAN 3 on the Ethernet Routing Switch 8600s.

The IDs of these two VLANs are important because they must have the same ID as the BPDUs generated from them. The BPDUs generated from these VLANs is tagged with a TaggedBpduVlanId that is derived by adding 4000 to the STG ID number. For example, for STG3 the TaggedBpduVlanId is 4003.

Figure 40
VLANs on the Layer 2 switch



Per-VLAN Spanning Tree Plus

PVST+ is the Cisco-proprietary spanning tree mechanism that uses a spanning tree instance per VLAN. PVST+ is an extension of the Cisco PVST with support for the IEEE 802.1Q standard. PVST+ is the default spanning tree protocol for Cisco switches and uses a separate spanning tree instance for each configured VLAN. In addition, PVST+ supports IEEE 802.1Q STP for support across IEEE 802.1Q regions.

For more information about PVST+, see *Nortel Ethernet Routing Switch 8600 Configuration — VLANs and Spanning Tree* (NN46205-517) .

MSTP and RSTP considerations

The Spanning Tree Protocol provides loop protection and recovery, but it is slow to respond to a topology change in the network (for example, a dysfunctional link in a network). The Rapid Spanning Tree protocol (RSTP or IEEE 802.1w) reduces the recovery time after a network failure.

It also maintains a backward compatibility with IEEE 802.1D. Typically, the recovery time of RSTP is less than 1 second. RSTP also reduces the amount of flooding in the network by enhancing the way that Topology Change Notification (TCN) packets are generated.

Use to configure multiple instances of RSTP on the same switch. Each RSTP instance can include one or more VLANs. The operation of the MSTP is similar to the current Nortel proprietary MSTP, except that the Nortel version has faster recovery time.

In MSTP mode, eight instances of RSTP can be supported simultaneously for the Ethernet Switch 460/470 or Ethernet Routing Switch 1600. Instance 0 or Common and Internal Spanning Tree (CIST) is the default group, which includes default VLAN 1. Instances 1 to 7 are called Multiple Spanning Tree Instances (MSTI) 1 to 7. You can configure up to 64 instances, of which only 25 can be active at one time.

RSTP provides a new parameter called ForceVersion for backward compatibility with legacy STP. You can configure a port in either STP-compatible mode or RSTP mode:

- An STP-compatible port transmits and receives only STP BPDUs. Any RSTP BPDU that the port receives in this mode is discarded.
- An RSTP port transmits and receives only RSTP BPDU. If an RSTP port receives an STP BPDU, it becomes an STP port. User intervention is required to bring this port back to RSTP mode. This process is called Port Protocol Migration.

You must be aware of the following recommendations before implementing 802.1w or 802.1s:

- 25 STP groups are supported.
- Configuration files are not compatible between regular STP and 802.1w/s modes. A special bootconfig flag identifies the mode. The default mode is 802.1D. If you choose 802.1w or 802.1s, new configuration files cannot be loaded if the flag is changed back to regular STP.
- For best interoperability results, contact your Nortel representative.

SLPP, Loop Detect, and Extended CP-Limit

Split MultiLink Trunking (SMLT) based network designs form physical loops for redundancy that logically do not function as a loop. Under certain adverse conditions, incorrect configurations or cabling, loops can form.

The two solutions to detect loops are Loop Detect and Simple Loop Prevention Protocol (SLPP). Loop Detect and SLPP detect a loop and automatically stop the loop. Both solutions determine on which port the loop is occurring and shuts down that port.

Control packet rate limit (CP-Limit) controls the amount of multicast and broadcast traffic sent to the SF/CPU from a physical port. CP-Limit protects the SF/CPU from being flooded with traffic from a single, unstable port. The CP-Limit functionality only protects the switch from broadcast and control traffic with a QoS value of 7.

Do not use only the CP-Limit for loop prevention. Nortel recommends the following loop prevention and recovery features in order of preference:

- SLPP
- Extended CP-Limit (Ext-CP-Limit) HardDown
- Loop Detect with ARP-Detect activated, when available

For information about configuring CP-Limit and SLPP, see *Nortel Ethernet Routing Switch 8600 Administration* (NN46205-605) . For more information about loop detection, see *Nortel Ethernet Routing Switch 8600 Configuration — VLANs and Spanning Tree* (NN46205-517) .

Simple Loop Prevention Protocol (SLPP)

Beginning with Software Release 4.1, Nortel recommends that you use Simple Loop Prevention Protocol (SLPP) to protect the network against Layer 2 loops. When you configure and enable SLPP, the switch sends a test packet to the VLAN. A loop is detected if the switch or if a peer aggregation switch on the same VLAN receives the original packet. If a loop is detected, the switch disables the port. To enable the port requires manual intervention. As an alternative, you can use port auto-enable to re-enable the port after a predefined interval.

SLPP is used to prevent loops in an SMLT network, but also works with other configurations, including Spanning Tree networks.

Loops can be introduced into the network in many ways. One way is through the loss of a multilink trunk configuration caused by user error or malfunction. This scenario does not introduce a broadcast storm, but because all MAC addresses are learned through the looping ports, Layer 2 MAC learning is significantly impacted. Spanning Tree cannot always detect such a configuration issue, whereas SLPP reacts and disables the malfunctioning links, minimizing the impact on the network.

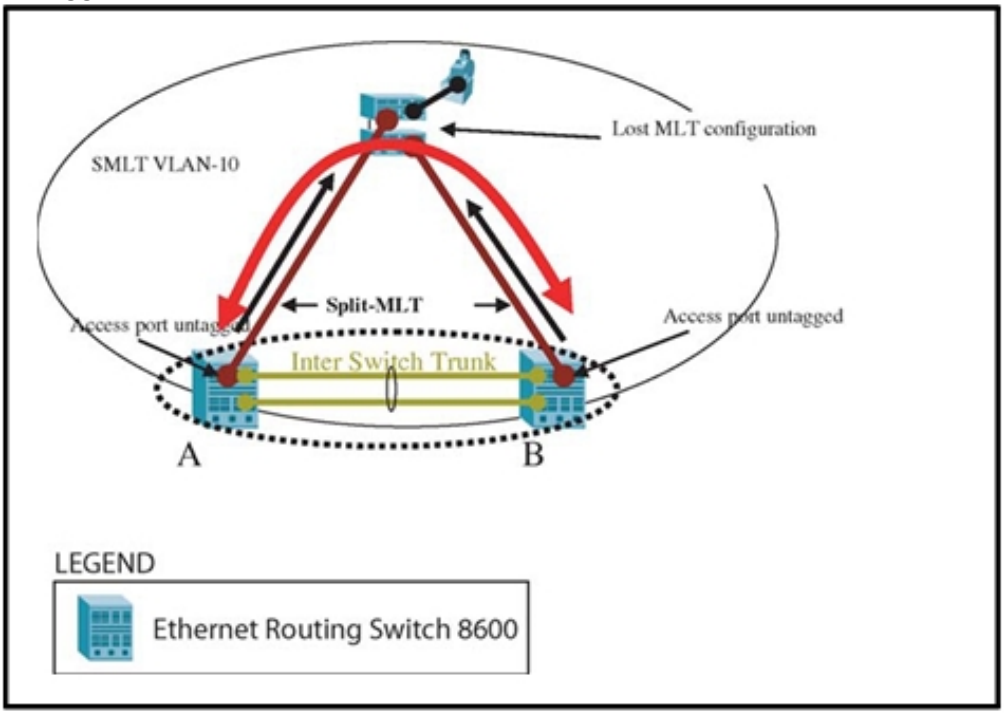
In addition to using SLPP for loop prevention, you can use the extended CP-Limit softdown feature to protect the SF/CPU against Denial of Service (DOS) attacks where required. The extended CP-Limit harddown option should only be used as a loop prevention mechanism in Software Release 3.7.x.

SLPP and SMLT examples

The following configurations show how to configure SLPP so that it detects VLAN-based network loops for untagged and tagged IEEE 802.1Q VLAN link configurations.

The following figure shows the network configuration. A and B exchange untagged packets over the SMLT.

Figure 41
Untagged SMLT links

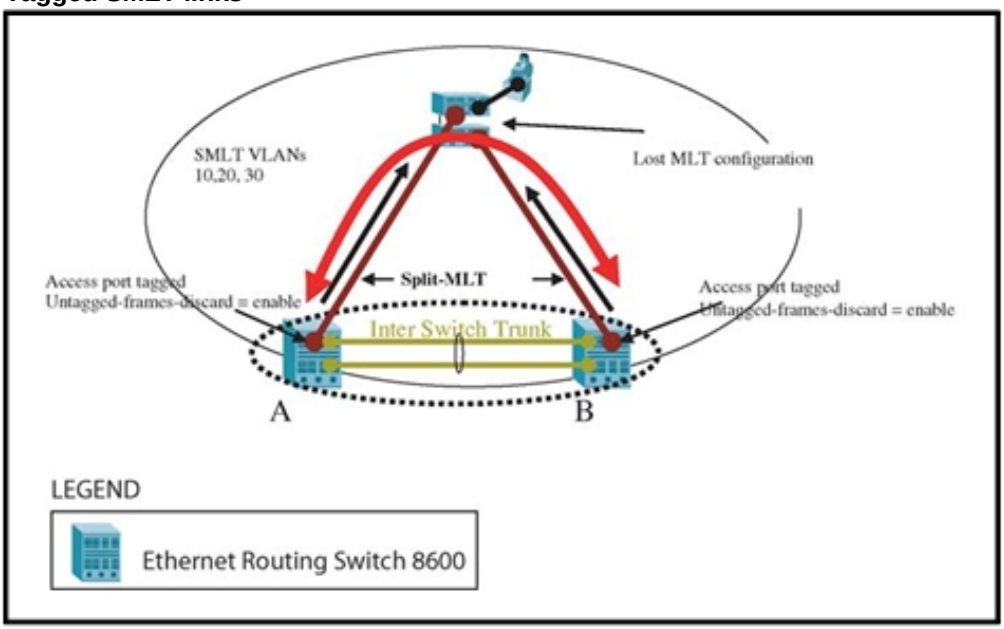


For the network shown in Figure 41 "Untagged SMLT links" (page 150), the configuration consists of the following:

- SLPP-Tx is enabled on SMLT VLAN-10.
- On switches A and B, SLPP-Rx is enabled on untagged access SMLT links.
- On switch A, the SLPP-Rx threshold is set to 5.
- In case of a network failure, to avoid edge isolation, the SLPP rx-threshold is set to 50 on SMLT switch B.

This configuration detects loops and avoids edge isolation. For tagged data, consider the following configuration:

Figure 42
Tagged SMLT links



The configuration is changed to:

- SLPP-Tx is enabled on SMLT VLANs 10, 20, and 30. A loop in any of these VLANs triggers an event and resolves the loop.
- On switches A and B, SLPP-Rx is enabled on tagged SMLT access links.
- On switch A, the SLPP Rx threshold is set to 5.
- On SMLT switch B, the SLPP Rx threshold is set to 50 to avoid edge isolation in case of a network failure.

In this scenario, Nortel recommends that you enable the untagged-frame s-discard parameter on the SMLT uplink ports.

SLPP configuration considerations and recommendations

SLPP uses a per-VLAN hello packet mechanism to detect network loops. Sending hello packets on a per-VLAN basis allows SLPP to detect VLAN-based network loops for untagged and tagged IEEE 802.1Q VLAN link configurations. The network administrator decides which VLANs to which a switch should send SLPP hello packets. The packets are replicated out of all ports that are members of the SLPP-enabled VLAN.

Use the information in this section to understand the considerations and recommendations when configuring SLPP in an SMLT network.

- You must enable SLPP packet receive on each port to detect a loop.
- Vary the SLPP packet receive threshold between the two core SMLT switches so that if a loop is detected, the access ports on both switches do not go down, and SMLT client isolation is avoided.
- SLPP test packets (SLPP-PDU) are forwarded for each VLAN.
- SLPP-PDUs are automatically forwarded VLAN ports configured for SLPP.
- The SLPP-PDU destination MAC address is the switch MAC address (with the multicast bit set) and the source MAC address is the switch MAC address.
- The SLPP-PDU is sent out as a multicast packet and is constrained to the VLAN on which it is sent.
- If an MLT port receives an SLPP-PDU the port goes down.
- The SLPP-PDU can be received by the originating CP or the peer SMLT CP. All other switches treat the SLPP-PDU as a normal multicast packet, and forward it to the VLAN.
- SLPP-PDU transmission and reception only operates on ports for which STP is in a forwarding state (if STP is enabled on one switch in the path).
- SLPP is port-based, so a port is disabled if it receives SLPP-PDU on one or more VLANs on a tagged port. For example, if the SLPP packet receive threshold is set to 5, a port is shut down if it receives 5 SLPP-PDU from one or more VLANs on a tagged port.
- The switch does not act on any other SLPP packet but those that it transmits.
- Enable SLPP-Rx only on SMLT edge ports, and never on core ports. Do not enable SLPP-Rx on SMLT IST ports or SMLT square or full-mesh core ports.
- In an SMLT Cluster, Nortel recommends an SLPP Packet-RX Threshold of 5 on the primary switch and 50 on the secondary switch .
- The administrator can tune network failure behavior by choosing how many SLPP packets must be received before a switch takes action.
- SLPP-Tx operationally disables ports that receive their own SLPP packet.

The following table provides the Nortel recommended SLPP values.

Table 25
SLPP recommended values

	Setting
Enable SLPP	
Access SMLT	Yes
Access SLT	Yes
Core SMLT	No
IST	No
Primary switch	
Packet Rx threshold	5
Transmission interval	500 milliseconds (ms) (default)
Ethertype	Default
Secondary switch	
Packet Rx threshold	50
Transmission interval	500 ms (default)
Ethertype	Default

Extended CP-Limit

The Extended CP-Limit function protects the SF/CPU by shutting down ports that send traffic to the SF/CPU at a rate greater than desired through one or more ports. You can configure the Extended CP-Limit functionality to prevent overwhelming the switch with high traffic. To use the Extended CP-Limit functionality, configure CP-Limit at the chassis and port levels.

ATTENTION
The Extended CP-Limit feature differs from the rate-limit feature by monitoring only packets that are sent to the SF/CPU (control plane), instead of all packets that are forwarded through the switch (data plane).

The set of ports to check for a high rate of traffic must be predetermined, and configured as either SoftDown or HardDown.

HardDown ports are disabled immediately after the SF/CPU is congested for a certain period of time.

SoftDown ports are monitored for a specified time interval, and are only disabled if the traffic does not subside. The user configures the maximum number of monitored SoftDown ports.

To enable this functionality and set its general parameters, configuration must take place at the chassis level first. After you enable this functionality at the chassis level, configure each port individually to make use of it.

The following table provides the Nortel recommended Extended CP-Limit values.

Table 26
Extended CP-Limit recommended values

Setting	Value
SoftDown - use with 4.1	
Maximum ports	5
Minimum congestion time	3 seconds (default)
Port congestion time	5 seconds (default)
CP-Limit utilization rate	Dependent on network traffic
HardDown - use with 3.7	
Maximum ports	5
Minimum congestion time	P=4000ms S=70000ms T=140000ms Q=210000ms
Port congestion time	P=4seconds S=70seconds T=140seconds Q=210seconds
Legend: Primary (P) - primary target for convergence, Secondary (S) - secondary target for convergence, Tertiary (T) - third target for convergence, Quarternary (Q) - fourth target for convergence Nortel does not recommend the Ext CP-Limit HardDown option for software Release 4.1 or later. Only use this option if SLPP is not available.	

Loop Detect

The Loop Detection feature is used at the edge of a network to prevent loops. It detects whether the same MAC address appears on different ports. This feature can disable a VLAN or a port. The Loop Detection feature can also disable a group of ports if it detects the same MAC address on two different ports five times in a configurable amount of time.

On a individual port basis, the Loop Detection feature detects MAC addresses that are looping from one port to other ports. After a loop is detected, the port on which the MAC addresses were learned is disabled. Additionally, if a MAC address is found to loop, the MAC address is disabled for that VLAN.

ARP Detect

The ARP-Detect feature is an enhancement over Loop Detect to account for ARP packets on IP configured interfaces. For network loops involving ARP frames on routed interfaces, Loop-Detect does not detect the network

loop condition due to how ARP frames are copied to the SF/CPU . Use ARP-Detect on Layer 3 interfaces. The ARP-Detect feature supports only the vlan-block and port-down options.

VLACP

Although VLACP has already been discussed previously in this document, it is important to discuss this feature in the context of Loop Prevention and CPU protection of Switch Cluster networks. This feature provides an end-to-end failure detection mechanism which will help to prevent potential problems caused by misconfigurations in a Switch Cluster design.

VLACP is configured on a per port basis and traffic will only be forwarded across the uplinks when VLACP is up and running correctly. The ports on each end of the link must be configured for VLACP. If one end of the link does not receive the VLACP PDUs, it will logically disable that port and no traffic will pass. This insures that even if there is link on the port at the other end, if it is not processing VLACP PDU's correctly, no traffic will be sent. This alleviates potential black hole situations by only sending traffic to ports that are functioning properly.

Loop prevention recommendations

Depending upon code release usage, select the set of features listed in Table 27 "Loop prevention by release" (page 155). For best loop prevention, Nortel Global Network Product Support recommends that you upgrade to release 4.1.1 or greater and use SLPP.

Table 27
Loop prevention by release

Software release	CP-Limit	Loop detect	Ext-CP-Limit	SLPP
3.7.0 - 3.7.4	Yes (see Note 2)	Yes (see Note 1)	N/A	N/A
3.7.5 - 3.7.x	Yes (see Note 2)	Yes (see Notes 1 and 5)	Yes (hard down) (see Notes 2 and 4)	N/A
4.0.x	Yes (see Note 2)	Yes (see Note 1)	N/A	N/A
4.1.x and on	Yes (see Note 2)	No	Yes (soft down) (see Notes 2 and 4)	Yes (see Note 3)

Note 1: Do not enable on IST links and do not use the VLAN down option for SMLT configurations.

Note 2: SF/CPU protection mechanism; do not enable on IST links.

Note 3: Do not enable SLPP on IST links.

Note 4: With Release 4.1.1.0 and later, Nortel recommends that you use the Soft Down option verses Hard Down.

Software release	CP-Limit	Loop detect	Ext-CP-Limit	SLPP
Note 5: For this configuration, always set ARP-detect option to activated as well.				

The following table provides the Nortel recommended CP-Limit values.

Table 28
CP-Limit recommended values

	CP-Limit Values	
	Broadcast	**Multicast**
Aggressive		
Access SMLT/SLT	1000	1000
Server	2500	2500
Core SMLT	7500	7500
Moderate		
Access SMLT/SLT	2500	2500
Server	5000	5000
Core SMLT	9000	9000
Relaxed		
Access SMLT/SLT	4000	4000
Server	7000	7000
Core SMLT	10000	10000

SF/CPU protection and loop prevention compatibility

Nortel recommends several best-practice methods for loop prevention, especially in any Ethernet Routing Switch 8600 Switch cluster environment. For more information about loop detection and compatibility for each software release, see *Converged Campus Technical Solution Guide — Enterprise Solution Engineering* (NN48500-516) .

Layer 3 network design

This section describes some Layer design considerations that you need to be aware of to properly design an efficient and robust network.

Navigation

VRF Lite

Prior to release 5.0, the Ethernet Routing Switch 8600 supported a single routing domain. Release 5.0 and later supports the Virtual Router Forwarding (VRF) Lite feature, which supports many virtual routers, each with its own routing domain. VRF Lite virtualizes the routing tables to form independent routing domains, which eliminates the need for multiple physical routers.

To use VRF Lite, you must use R or RS modules. E and M modules do not support VRF Lite. The chassis can be either in mixed or R mode. VRF Lite also requires the SuperMezz CPU-Daughter card and the Premier Software License.

VRF Lite fully supports the High Availability feature. Dynamic tables built by VRF Lite are synchronized. If failover occurs when HA is enabled, VRF Lite does not experience an interruption.

For more information about VRF Lite, see *Nortel Ethernet Routing Switch 8600 Configuration — IP Routing* (NN46205-523) .

VRF Lite route redistribution

Using VRF Lite, the Ethernet Routing Switch 8600 can function as many routers; each Virtual Router and Forwarder (VRF) autonomous routing engine works independently. Normally, no route leak occurs between different VRFs. Sometimes users may have to redistribute OSPF or RIP routes from one VRF to another. The route redistribution option facilitates the redistribution or routes.

If you enable route redistribution between two VRFs, ensure that the IP addresses do not overlap. The software does not enforce this requirement.

This feature is available in R mode. This feature operates on R and RS modules (only) in a mixed-mode chassis. Even though Classic modules are supported on the chassis, route redistribution from other VRFs to the base router (default VRF) will not work on Classic modules.

VRF Lite capability and functionality

On any VRF instance, VRF Lite supports the following protocols: IP, Internet Control Message Protocol (ICMP), Address Resolution Protocol (ARP), Static routes, Default routes, Routing Information Protocol (RIP), Open Shortest Path First (OSPF), Route Policies (RPS), Virtual Router Redundancy Protocol (VRRP), and the Dynamic Host Configuration Protocol / BootStrap Protocol relay agent.

Using VRF Lite, the switch:

- partitions traffic and data and represents an independent router in the network
- provides virtual routers that are transparent to end-users
- supports overlapping IP address spaces in separate VRFs
- supports addresses that are not restricted to the assigned address space given by host Internet Service Providers (ISP).
- supports SMLT/RSMLT
- supports Border Gateway Protocol

IPv6 is supported on VRF 0 only.

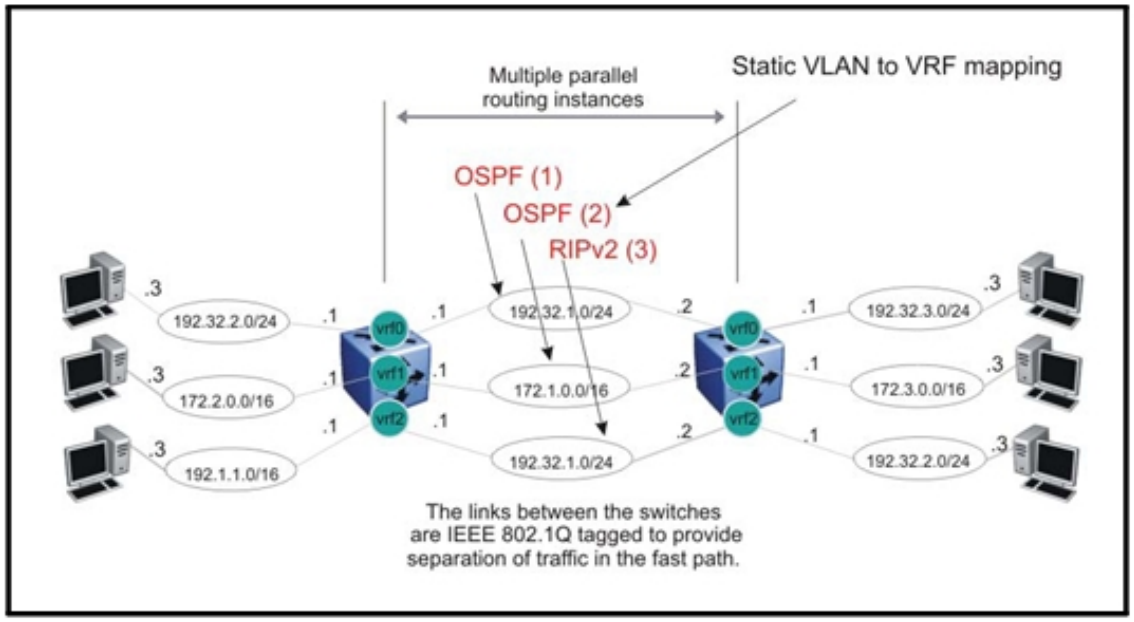
VRF Lite architecture examples

VRF Lite enables a router to act as many routers. This provides virtual traffic separation per user and provides security. For example, you can use VRF Lite to:

- provide different departments within a company with site-to-site connectivity as well as internet access
- extend WAN VPNs into campus LANs without interconnecting VPNs
- provide centralized and shared access to data centers.

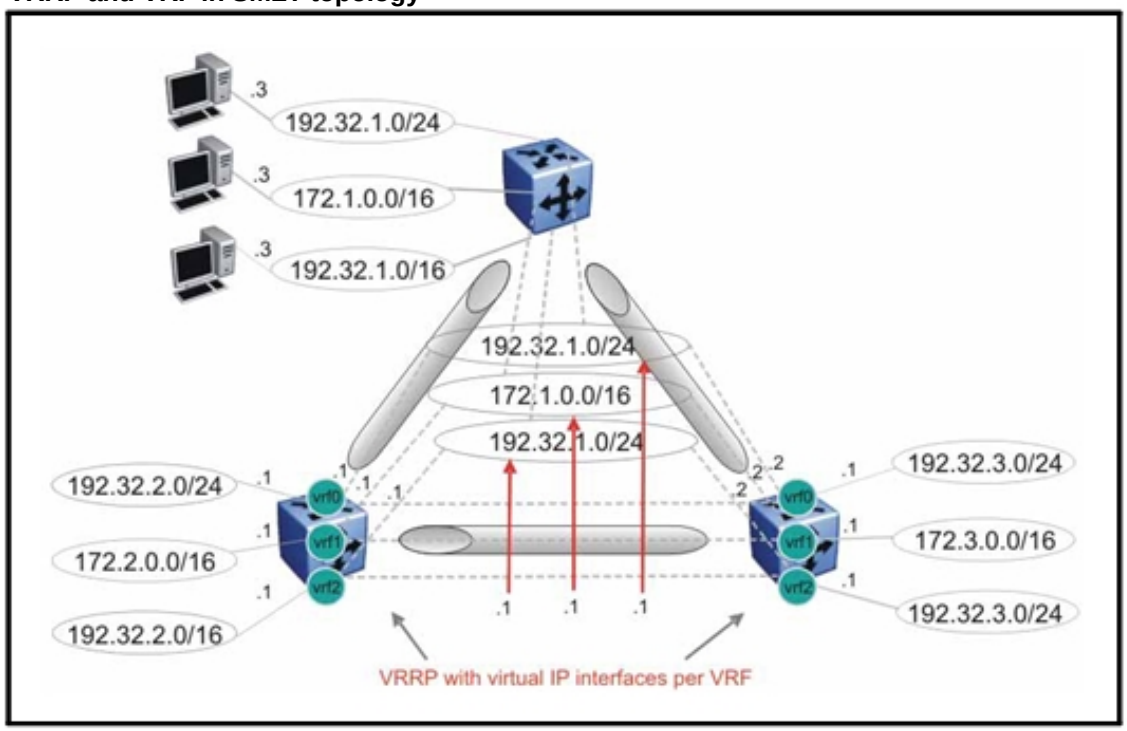
The following figure shows how VRF Lite can be used to emulate IP VPNs.

Figure 43
VRF Lite example



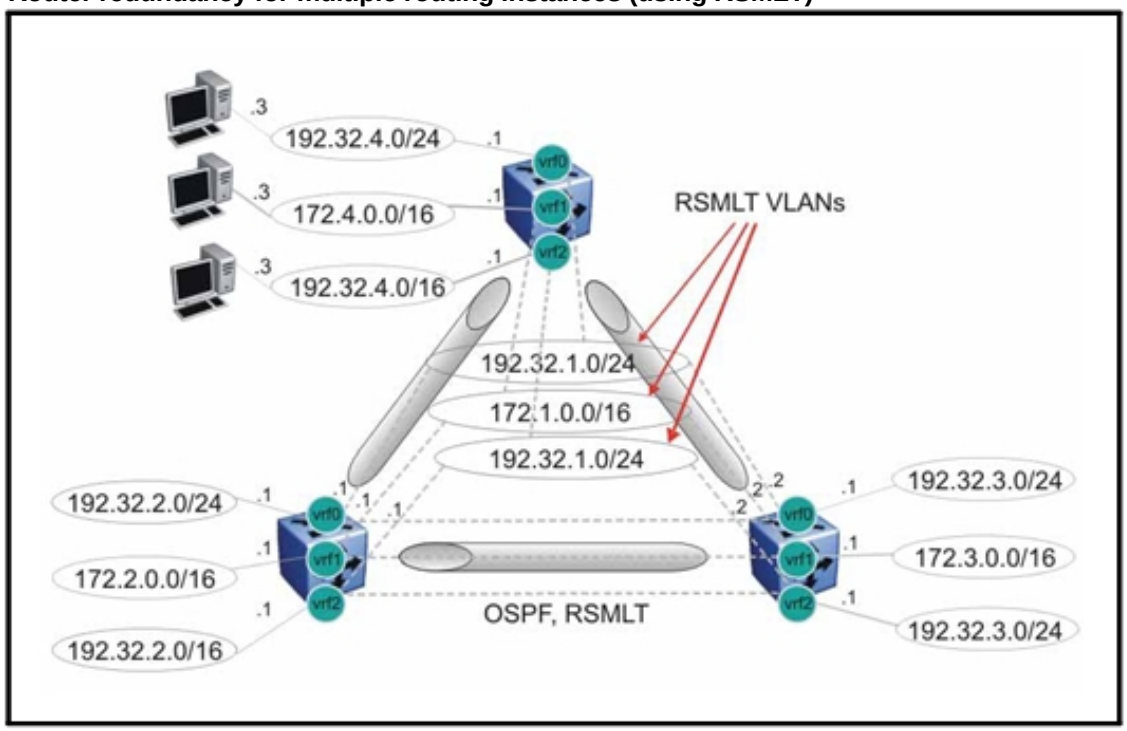
The following figure shows how VRF Lite can be used in an SMLT topology. VRRP is used between the two bottom routers.

Figure 44
VRRP and VRF in SMLT topology



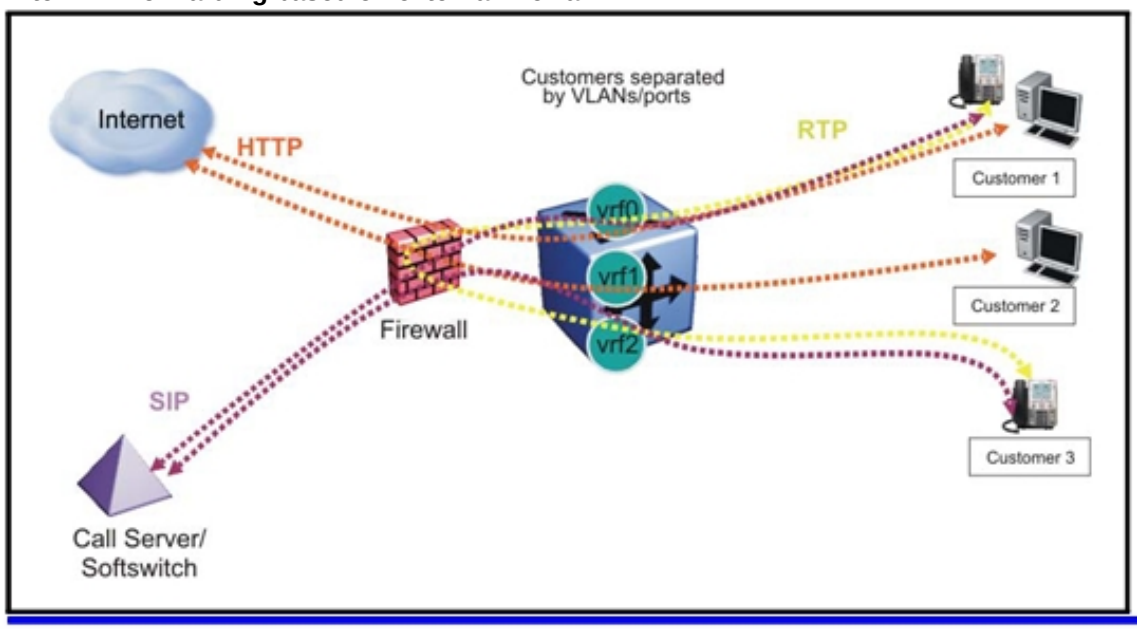
The following figure shows how VRF Lite can be used in an RSMLT topology.

Figure 45
Router redundancy for multiple routing instances (using RSMLT)



The following figure shows how VRFs can interconnect through an external firewall.

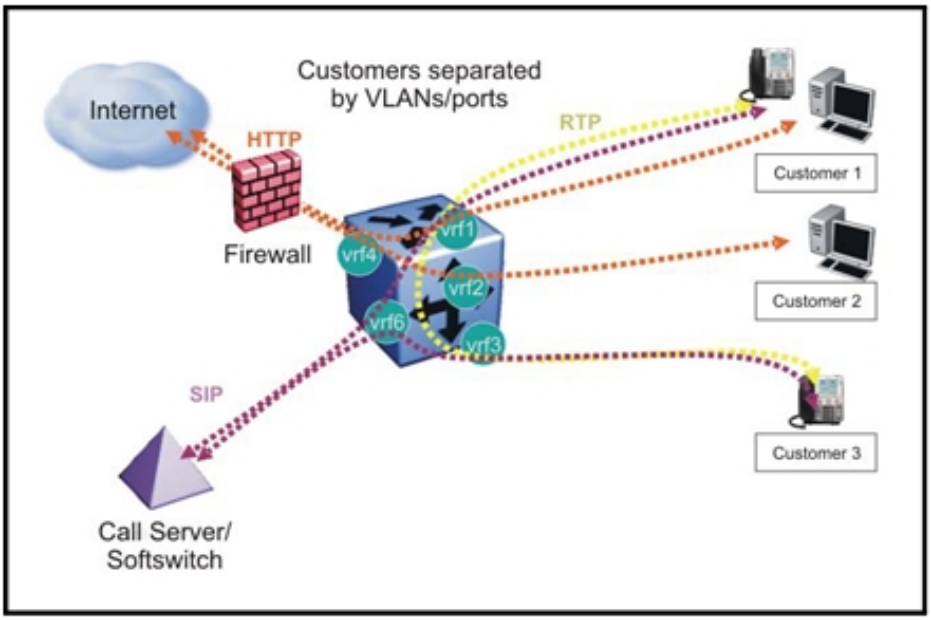
Figure 46
Inter-VRF forwarding based on external firewall



Although customer data separation into Layer 3 virtual routing domains is usually a requirement, sometimes customers must access a common network infrastructure. For example, they want to access the Internet, data storage, VoIP-PSTN, or call signaling services. To interconnect VRF instances, you can use an external firewall that supports virtualization, or use interVRF forwarding for specific services. Using the interVRF solution, routing policies and static routes can be used to inject IP subnets from one VRF instance to another, and filters can be used to restrict access to certain protocols.

The following figure shows inter-VRF forwarding. In this solution, routing policies can be used to leak IP subnets from one VRF to another. Filters can be used to restrict access to certain protocols. This enables hub-and-spoke network designs for, for example, VoIP gateways.

Figure 47
Inter VRF communication, internal inter-VRF forwarding



Virtual Router Redundancy Protocol

The Virtual Router Redundancy Protocol (VRRP) provides a backup router that takes over if a router fails. This is important when you must provide redundancy mechanisms. To configure VRRP so that it works correctly, use the information in the following sections.

VRRP navigation

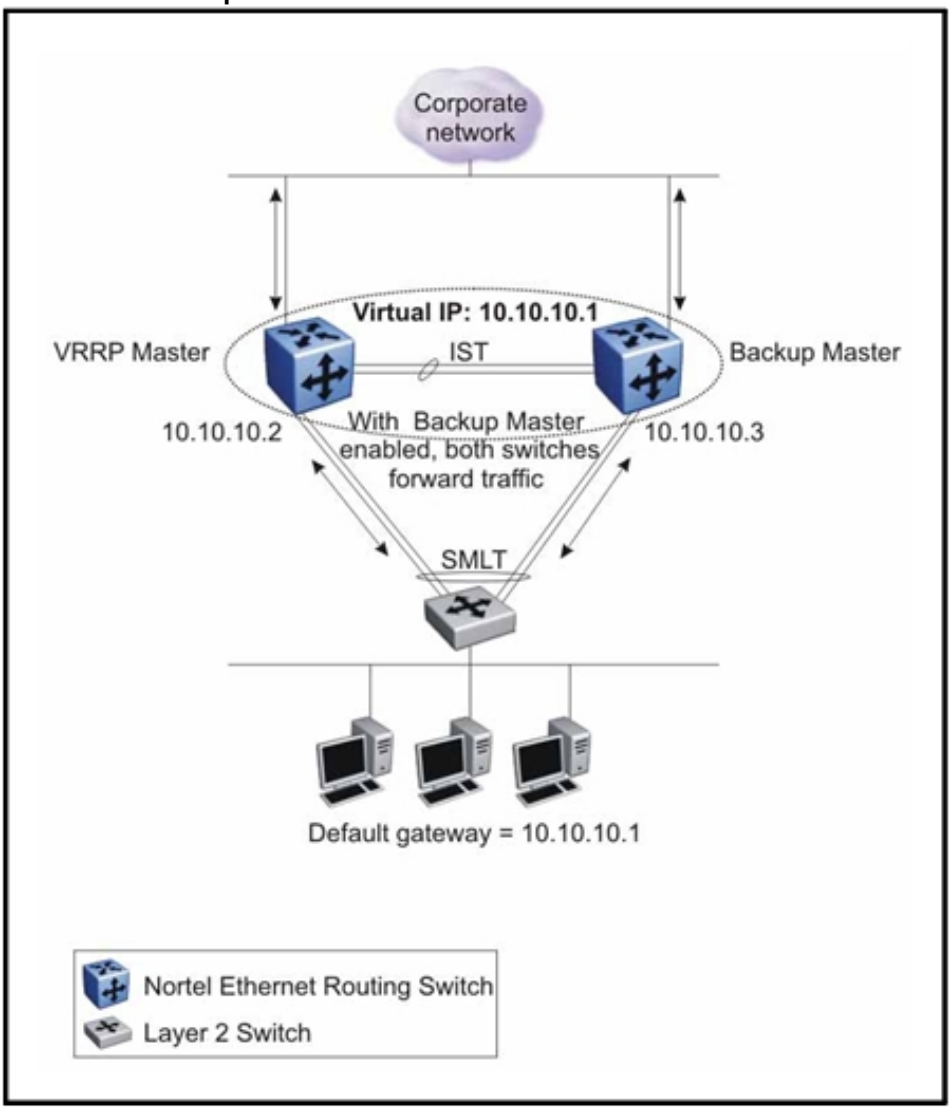
VRRP guidelines

VRRP provides another layer of resiliency to your network design by providing default gateway redundancy for end users. If a VRRP-enabled router connected to the default gateway fails, failover to the VRRP backup router ensures there is no interruption for end users attempting to route from their local subnet.

Typically, only the VRRP Master router forwards traffic for a given subnet. The backup VRRP router does not route traffic destined for the default gateway. Instead, the backup router employs Layer 2 switching on the IST to deliver traffic to the VRRP master for routing.

To allow both VRRP switches to route traffic, Nortel has created an extension to VRRP, BackupMaster, that creates an active-active environment for routing. With BackupMaster enabled on the backup router, the backup router no longer switches traffic to the VRRP Master. Instead the BackupMaster routes all traffic received on the BackupMaster IP interface according to the switch routing table. This prevents the edge switch traffic from unnecessarily utilizing the IST to reach the default gateway.

Figure 48
VRRP with BackupMaster



Nortel recommends that you use a VRRP BackupMaster configuration with any SMLT configuration that has an existing VRRP configuration.

The VRRP BackupMaster uses the VRRP standardized backup switch state machine. Thus, VRRP BackupMaster is compatible with standard VRRP.

When implementing VRRP, follow the Nortel recommended best practices:

- Do not configure the virtual address as a physical interface that is used on any of the routing switches. Instead, use a third address, for example:

— Interface IP address of VLAN a on Switch 1 = x.x.x.2 —

Interface IP address of VLAN a on Switch 2 = x.x.x.3 —

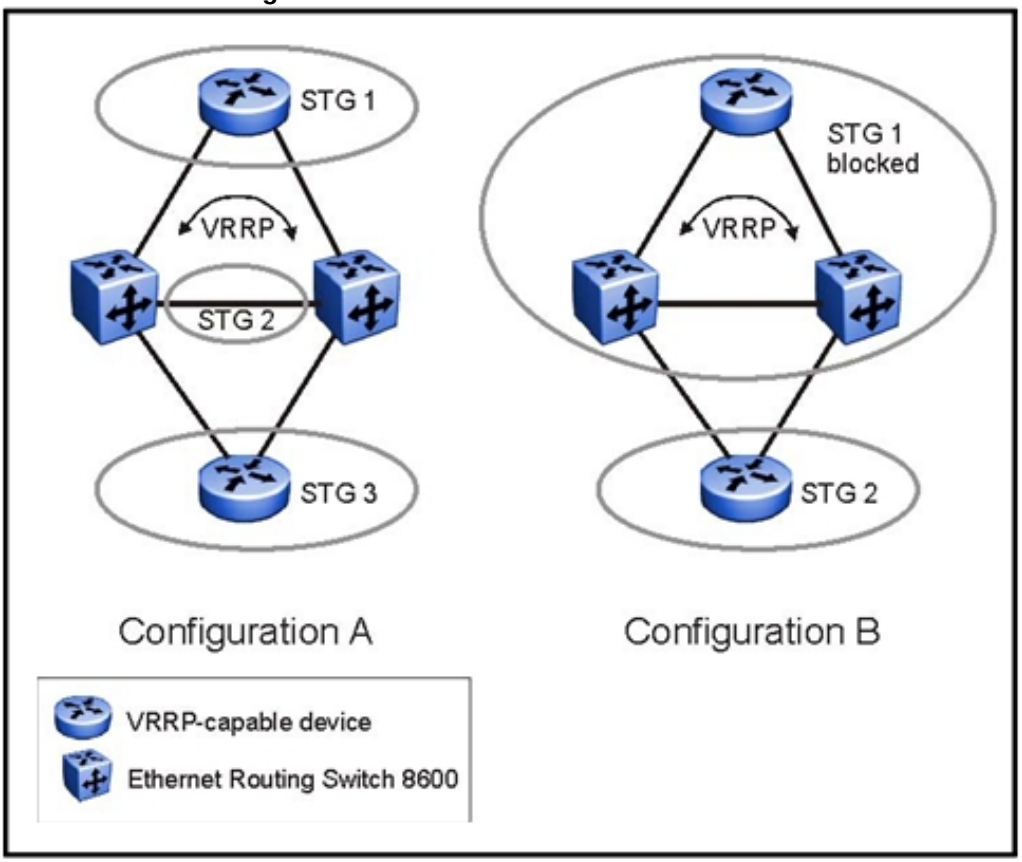
Virtual IP address of VLAN a = x.x.x.1

- Set the VRRP hold down timer long enough such that the IGP routing protocol has time to converge and update the routing table. In some cases, setting the VRRP hold down timer to a minimum of 1.5 times the IGP convergence time is sufficient. For OSPF, it is suggested to use a value of 90 seconds if using the default OSPF timers.
- Implement VRRP BackupMaster for an active-active configuration (BackupMaster works across multiple switches participating in the same VRRP domain.).
- Configure VRRP priority as 200 to set VRRP Master
- Stagger VRRP Masters between Ethernet Routing Switches in the core
- Take care when implementing VRRP Fast as this creates additional control traffic on the network and also creates a greater load on the CPU. To reduce the convergence time of VRRP, the VRRP Fast feature allows the modification of VRRP timers to achieve sub-second failover of VRRP. Without VRRP Fast, normal convergence time is approximately 3 seconds.
- Ensure that both SMLT aggregation switches can reach the same destinations by using a routing protocol. For routing purposes, configure per-VLAN IP addresses on both SMLT aggregation switches.
- Introduce an additional subnet on the IST that has a shortest-route-pat h to avoid issuing Internet Control Message Protocol (ICMP) redirect messages on the VRRP subnets. (To reach the destination, ICMP redirect messages are issued if the router sends a packet back out through the same subnet on which it is received).
- Do not use VRRP BackupMaster and critical IP at the same time. Use one or the other.
- When implementing VRRP on multiple VLANs between the same switches, Nortel recommends that you configure a unique VRID on each VLAN.

VRRP and STG

VRRP protects clients and servers from link or aggregation switch failures. Your network configuration should limit the amount of time a link is down during VRRP convergence. The following figure shows two possible configurations of VRRP and STG; the first is optimal and the second is not.

Figure 49
VRRP and STG configurations



In this figure, configuration A is optimal because VRRP convergence occurs within 2 to 3 seconds. In configuration A, three STGs are configured and VRRP runs on the link between the two routers (R). STG 2 is configured on the link between the two routers, which separates the link between the two routers from the STGs found on the other devices. All uplinks are active.

In configuration B, VRRP convergence takes between 30 and 45 seconds because it depends on spanning tree convergence. After initial convergence, spanning tree blocks one link (an uplink), so only one uplink is used. If an error occurs on the uplink, spanning tree reconverges, which can take up to 45 seconds. After spanning tree reconvergence, VRRP can take a few more seconds to failover.

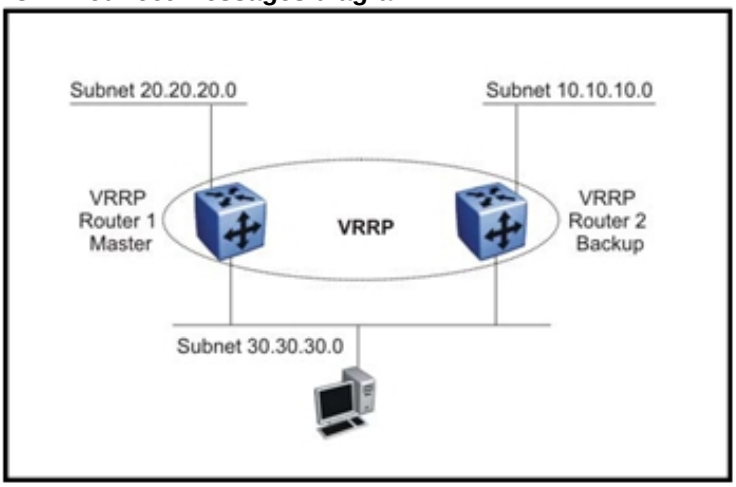
Rather than configuring STG with VRRP, Nortel recommends that you enable SMLT with VRRP to simplify the network configuration and reduce the failover time. For VRRP and SMLT information, see “SMLT and Layer 3 traffic Redundant Default Gateway: VRRP ” (page 107).

VRRP and ICMP redirect messages

You can use VRRP and Internet Control Message Protocol (ICMP) in conjunction. However, doing so may not provide optimal network performance.

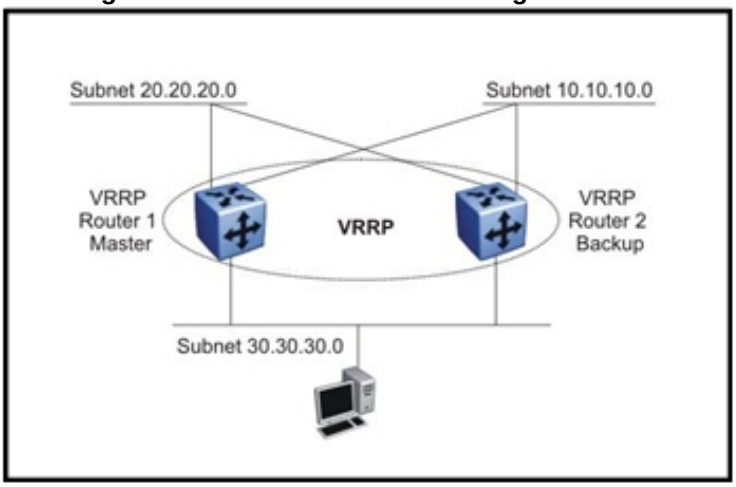
Consider the network shown in the following figure. Traffic from the client on subnet 30.30.30.0, destined for the 10.10.10.0 subnet, is sent to routing switch 1 (VRRP Master). This traffic is then forwarded on the same subnet to routing switch 2 where it is routed to the destination. For each packet received, Routing switch 1 sends an ICMP redirect message to the client to inform it of a shorter path to the destination through routing switch 2.

Figure 50
ICMP redirect messages diagram



To avoid excessive ICMP redirect messages if network clients do not recognize ICMP redirect messages, Nortel recommends the network design shown in the following figure. Ensure that the routing path to the destination through both routing switches has the same metric to the destination. One hop goes from 30.30.30.0 to 10.10.10.0 through routing switch 1 and routing switch 2. Do this by building symmetrical networks based on the network design examples presented in “Redundant network design” (page 75).

Figure 51
Avoiding excessive ICMP redirect messages



VRRP versus RSMLT for default gateway resiliency

A better alternative than VRRP with BackupMaster is to use RSMLT L2 Edge. For Release 5.0 and later, Nortel recommends that you use an RSMLT L2 Edge configuration, rather than VRRP with BackupMaster, for those products that support RSMLT L2 Edge.

RSMLT L2 Edge provides:

- Greater scalability—VRRP scales to 255 instances, while RSMLT scales to the maximum number of VLANs.
- Simpler configuration—Simply enable RSMLT on a VLAN; VRRP requires virtual IP configuration, along with other parameters.

For connections in pure Layer 3 configurations (using a static or dynamic routing protocol), a Layer 3 RSMLT configuration is recommended over VRRP. In these instances, an RSMLT configuration provides faster failover than one with VRRP because the connection is a Layer 3 connection, not just a Layer 2 connection for default gateway redundancy.

Both VRRP and RSMLT can provide resiliency for the end station's default gateway. The configurations of these features are different, but both provide the same end result and are transparent to the end station.

For more information on RSMLT, see "Routed SMLT" (page 115).

Subnet-based VLAN guidelines

You can use subnet-based VLANs to classify end-users in a VLAN based on the end-user source IP addresses. For each packet, the switch performs a lookup, and, based on the source IP address and mask,

determines to which VLAN the traffic belongs. To provide security, subnet-based VLANs can be used to allow only users on the appropriate IP subnet to access to the network.

You cannot classify nonIP traffic using a subnet-based VLAN.

You can enable routing in each subnet-based VLAN by assigning an IP address to the subnet-based VLAN. If no IP address is configured, the subnet-based VLAN is in Layer 2 switch mode only.

You can enable VRRP for subnet-based VLANs. The traffic routed by the VRRP Master interface is forwarded by hardware. Therefore, no throughput impact is expected when you use VRRP on subnet-based VLANs.

You can use subnet-based VLANs to achieve multinetting functionality; however, multiple subnet-based VLANs on a port can only classify traffic based on the sender IP source address. Thus, you cannot multinet by using multiple subnet-based VLANs between routers (Layer 3 devices). Multinetting is supported, however, on all end-user-facing ports.

You cannot classify Dynamic Host Configuration Protocol (DHCP) traffic into subnet-based VLANs because DHCP requests do not carry a specific source IP address; instead, they use an an all broadcast address. To support DHCP to classify subnet-based VLAN members, create an overlay port-based VLAN to collect the bootp/DHCP traffic and forward it to the appropriate DHCP server. After the DHCP response is forwarded to the DHCP client and it learns its source IP address, the end-user traffic is appropriately classified into the subnet-based VLAN.

The switch supports a maximum number of 200 subnet-based VLANs.

Subnet-based VLANs are incompatible with some wireless terminals. This is especially true in those configurations where you use the Ethernet Routing Switch 8600 as a classification device (that is, as an IP subnet-based VLAN and a port-based VLAN configured on the same port). During the roaming phase, wireless terminals may lose the session with their application servers. Terminals lose the session because of the absence of the IP header in the frames that these terminals send. Thus, the frames are sent through the port-based VLAN, not through the IP subnet-based VLAN. Ensure that your wireless access devices operate correctly.

PPPoE-based VLAN design example

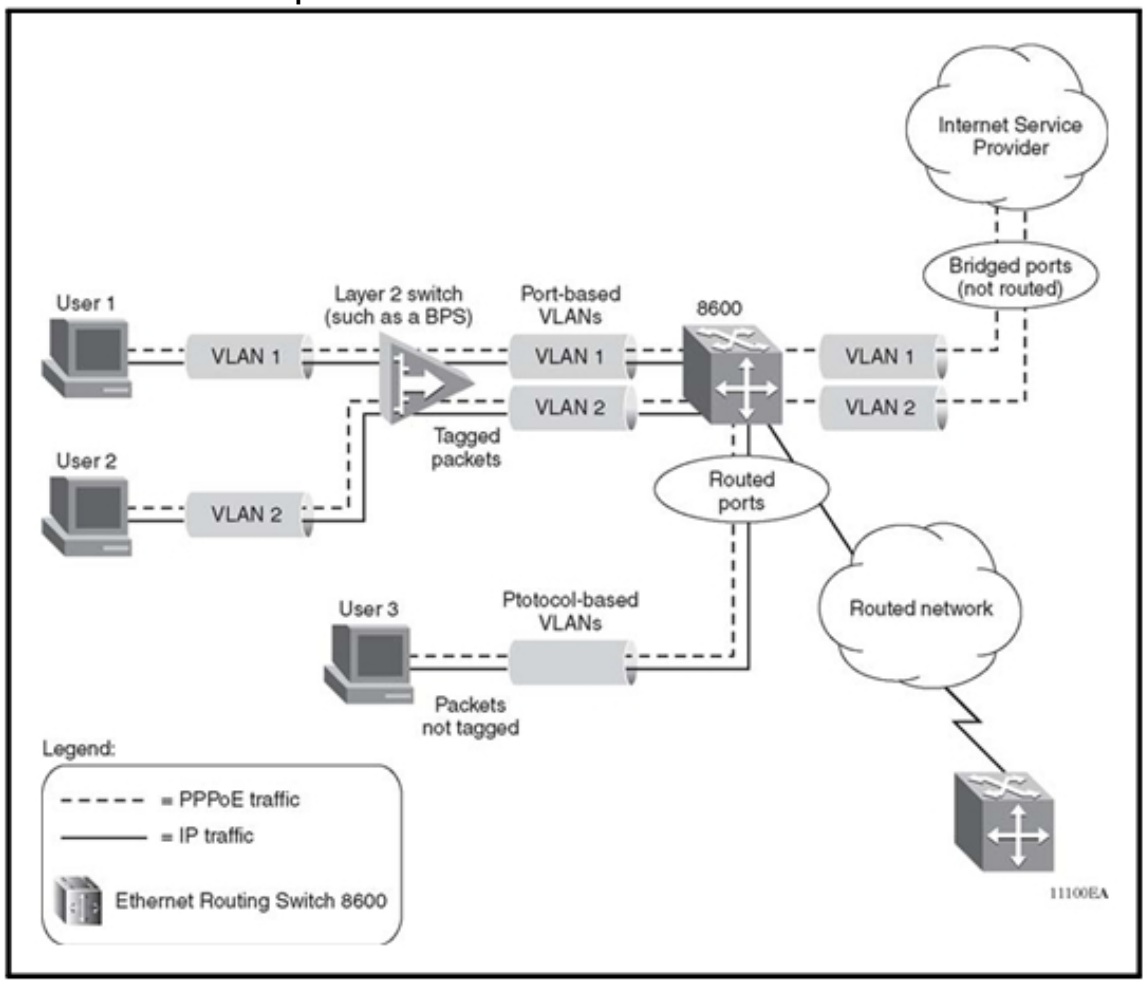
You can connect multiple Ethernet devices to a remote site through a device (such as a modem) using Point-to-Point Protocol over Ethernet (PPPoE). PPPoE allows multiple users to share a common Internet connection (see RFC 2516: *Point-to-Point Protocol over Ethernet* ()).

This example shows how to use PPPoE protocol-based VLANs to redirect PPPoE Internet traffic to a Internet service provider (ISP) network while IP traffic is sent to a routed network. Use this design in a service provider application to redirect subscriber Internet traffic to a separate network from the IP routed network. The design also applies to enterprise networks that need to isolate PPPoE traffic from the routed IP traffic, even when this traffic is received on the same VLAN.

The following figure shows the network design that achieves the following goals:

- users can generate IP and PPPoE traffic. IP traffic is routed, and PPPoE traffic is bridged to the ISP network. If any other type of traffic is generated, it is dropped by the Layer 2 switch or the Ethernet Routing Switch 8600 (when users are attached directly to the 8600).
- Each user is assigned their own VLAN.
- Each user has two VLANs when directly connected to the Ethernet Routing Switch 8600: one for IP traffic and the other for PPPoE traffic.
- PPPoE bridged traffic preserves user VLANs.

Figure 52
PPPoE and IP traffic separation



This configuration uses indirect connections (users are attached to a Layer 2 switch) and direct connections (users are attached directly to the Ethernet Routing Switch 8600). These connections are described in following sections.

Both PPPoE and IP traffic flows through the network. Assumptions and configuration requirements include the following:

- PPPoE packets between users and the ISP are bridged.
- Packets received from the Layer 2 switch are tagged, whereas packets received from the directly connected user (User 3) are not tagged.
- IP packets between the user and the 8600 are bridged, whereas packets between the Ethernet Routing Switch 8600 and the routed network are routed.
- VLANs between the Layer 2 switch and the 8600 are port-based.

- VLANS from the directly connected user (User 3) are protocol-based.
- The connection between the Ethernet Routing Switch 8600 and the ISP is a single port connection.
- The connection between the Layer 2 switch and the Ethernet Routing Switch 8600 can be a single port connection or a MultiLink Trunk (MLT) connection.
- Ethernet Routing Switch 8600 ports connected to the user side (Users 1,2, and 3) and the routed network are routed ports.
- Ethernet Routing Switch 8600 ports connected to the ISP side are bridged (not routed) ports.

Indirect connections

The following figure shows a switch using routable port-based VLANs for indirect connections. When configured in this way:

- Port P1 provides a connection to the Layer 2 switch.

  Port P1 is configured for tagging. All P1 ingress and egress packets are tagged (the packet type can be either PPPoE or IP).

- Port P2 provides a connection to the ISP network.

  Port P2 is configured for tagging. All P2 ingress and egress packets are tagged (the packet type is PPPoE).

- Port P3 provides a connection to the routed network.

  Port P3 can be configured for either tagging or nontagging (if untagged, the header does not carry any VLAN tagging information). All P3 ingress and egress packets are untagged (the packet type is IP).

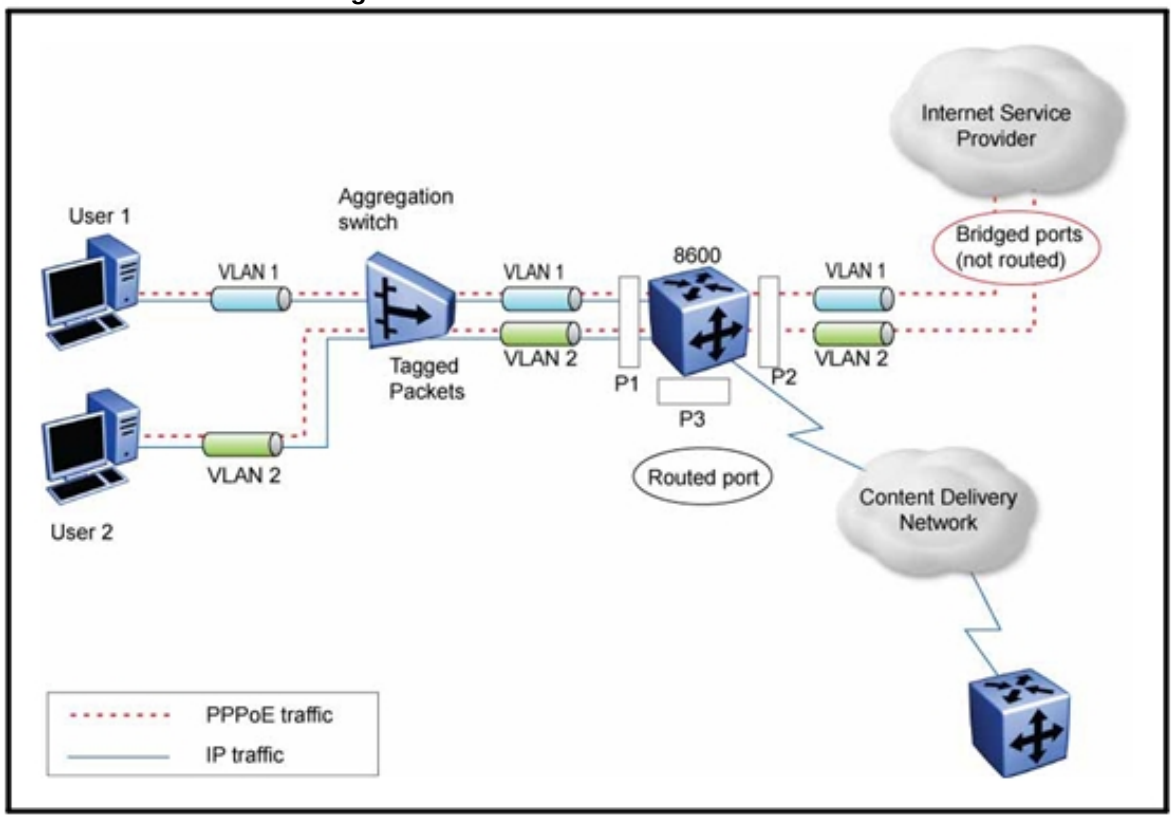
- Ports P1 and P2 must be members of the same VLAN.

  The VLAN must be configured as a routable VLAN. Routing must be disabled on Port P2. VLAN tagging is preserved on P1 and P2 ingress and egress packets.

- Port P3 must be a member of a routable VLAN but cannot be a member of the same VLAN as Ports P1 and P2. VLAN tagging is not preserved on P3 ingress and egress packets.

For indirect user connections, you must disable routing on port P2. This allows the bridging of traffic other than IP and routing of IP traffic outside of port number 2. In the latter case, port 1 has routing enabled and allows routing of IP traffic to port 3. By disabling IP routing on port P2, no IP traffic flows to this port.

Figure 53
Indirect PPPoE and IP configuration



Direct connections

To directly connect to the Ethernet Routing Switch 8600, a user must create two protocol-based VLANs on the port: one for PPPoE traffic and one for IP traffic (see the following figure). When configured in this way:

- Port P1 is an access port.

  Port P1 must belong to both the IP protocol-based VLAN and the PPPoE protocol-based VLAN.

- Port P2 provides a connection to the ISP network.

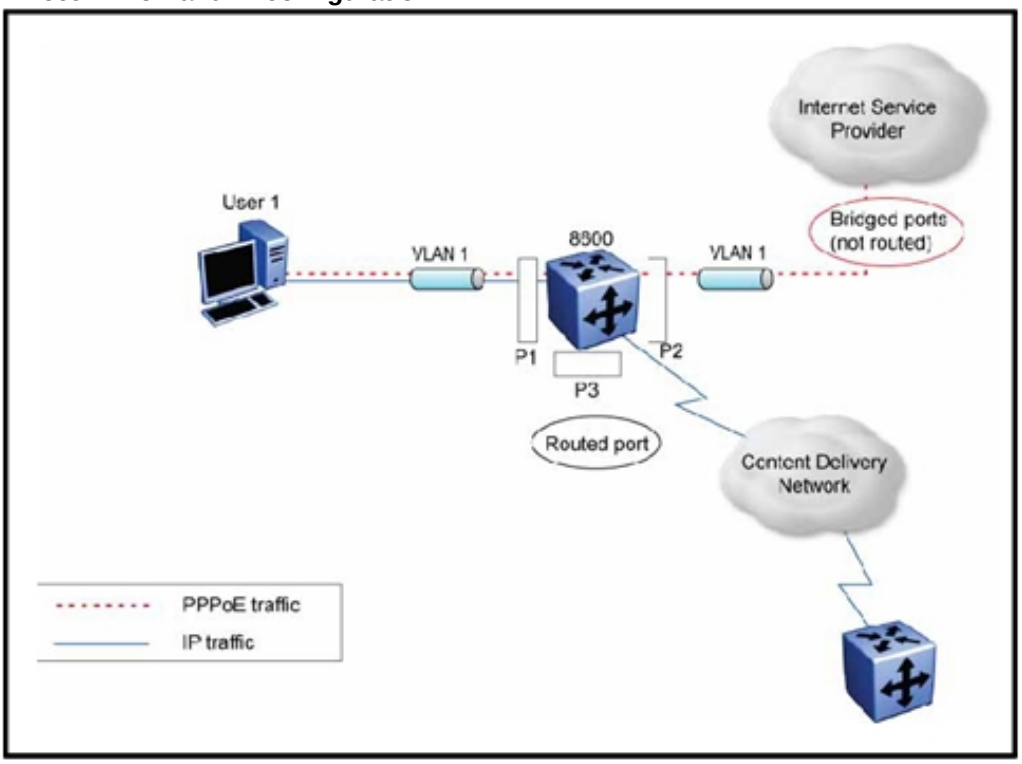
  P2 is configured for tagging to support PPPoE traffic to the ISP for multiple users. P2 ingress and egress packets are tagged (the packet type is PPPoE).

- Port P3 provides a connection to the Content Delivery Network.

  P3 can be configured for either tagging or nontagging (if untagged, the header does not carry any VLAN tagging information). P3 ingress and egress packets are untagged (the packet type is IP). Port P3 must be a member of a routable VLAN, but cannot be a member of the same VLAN as ports P1 and P2.

For the direct connections, protocol-based VLANs (IP and PPPoE) are required to achieve traffic separation. The disabling of routing on each port is not required because routed IP VLANs are not configured on port 2 (they are for indirect connections).

Figure 54
Direct PPPoE and IP configuration



Border Gateway Protocol

Use Border Gateway Protocol (BGP) to ensure that the switch can communicate with other BGP-speaking routers on the Internet backbone. BGP is an exterior gateway protocol designed to exchange network reachability information with other BGP systems in the same or other autonomous systems (AS). This network reachability information includes information about the AS list that the reachability information traverses. By using this information, you can prune routing loops and enforce policy decisions at the AS level.

BGP performs routing between two sets of routers operating in different autonomous systems (AS). An AS can use two kinds of BGP: Interior BGP (IBGP), which refers to the protocol that BGP routers use within an autonomous system, and Exterior BGP (EBGP), which refers to the protocol that BGP routers use across two different autonomous systems. BGP information is redistributed to Interior Gateway Protocols (IGP) running in the autonomous system.

BGPv4 supports classless inter-domain routing. BGPv4 advertises the IP prefix and eliminates the concept of network class within BGP. BGP4 can aggregate routes and AS paths. BGP aggregation does not occur when routes have different multiexit discs or next-hops.

To use BGP, you must have Ethernet Routing Switch 8600 software version 3.3 or later installed. BGP is supported on all interface modules. For large BGP environments, Nortel recommends that you use the 8692 SF/CPU.

BGP Equal-Cost Multipath (ECMP) allows a BGP speaker to perform route balancing within an AS by using multiple equal-cost routes submitted to the routing table by OSPF or RIP. Load balancing is performed on a per-packet basis.

To control route propagation and filtering, RFCs 1772 and 2270 recommends that multihomed, nontransit Autonomous Systems not run BGPv4. To address the load sharing and reliability requirements of a multihomed user, use BGP between them.

For more information about BGP and a list of CLI BGP commands, see *Nortel Ethernet Routing Switch 8600 Configuration — BGP Services* (NN46205-510) .

BGP navigation

- "BGP scaling" (page 175)
- "BGP considerations" (page 175)
- "BGP and other vendor interoperability" (page 176)
- "BGP design examples" (page 177)

BGP scaling

For information about BGP scaling numbers, see Table 5 "Supported scaling capabilities" (page 38) and *Nortel Ethernet Routing Switch 8600 Release Notes* (NN46205-402) . The Release Notes take precedence over this document.

BGP considerations

Be aware of the following BGP design considerations.

Use the max-prefix parameter to limit the number of routes imported from a peer. This parameter prevents nonM mode configurations from accepting more routes than they can handle. Use a setting of 0 to accept an unlimited number of prefixes.

BGP does not operate with an IP router in nonforwarding (host-only) mode. Thus, ensure that the routers which you want BGP to operate with are in forwarding mode.

If you are using BGP for a multi-homed AS (one that contains more than a single exit point), Nortel recommends that you use OSPF for your IGP, and BGP for your sole exterior gateway protocol. Otherwise, use intra-AS IBGP routing.

If OSPF is the IGP, use the default OSPF tag construction. The use of EGP or the modification of the OSPF tags makes network administration and proper configuration of BGP path attributes difficult.

For routers that support both BGP and OSPF, you must set the OSPF router ID and the BGP identifier to the same IP address. The BGP router ID automatically uses the OSPF router ID.

In configurations where BGP speakers reside on routers that have multiple network connections over multiple IP interfaces (the typical case for IBGP speakers), consider using the address of the circuitless (virtual) IP interface as the local peer address. In this way, you ensure that BGP is reachable as long as an active circuit exists on the router.

By default, BGP speakers do not advertise or inject routes into their IGP. You must configure route policies to enable route advertisement.

Coordinate routing policies among all BGP speakers within an AS so that every BGP border router within an AS constructs the same path attributes for an external path.

Configure accept and announce policies on all IBGP connections to accept and propagate all routes. Make consistent routing policy decisions on external BGP connections.

You cannot enable or disable the Multi-Exit Discriminator selection process. You cannot disable the aggregation when routes have different MEDs (MULTI_EXIT_DISC) or NEXT_HOP.

BGP and other vendor interoperability

BGP interoperability has been successfully demonstrated between the Ethernet Routing Switch 8600 Software Release 3.3, Cisco 6500 Software Release IOS 11.3, and Juniper M20 Software Release 5.3R2.4.

For more information about BGP and BGP commands, see *Nortel Ethernet Routing Switch 8600 Configuration — BGP Services* (NN46205-510) . For configuration examples, go to Nortel Technical Support and download the *Border Gateway Protocol (BGP-4) Technical Configuration Guide*.

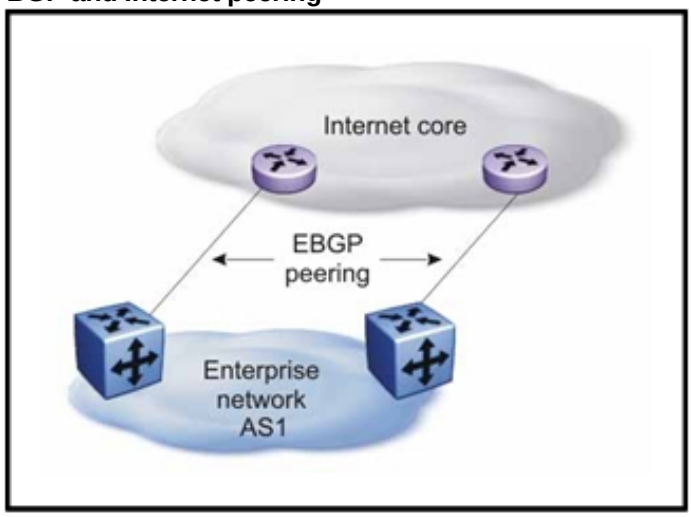
BGP design examples

The following design examples describe typical Ethernet Routing Switch 8600 BGP applications.

BGP and Internet peering

By using BGP, you can perform Internet peering directly between the Ethernet Routing Switch 8600 and another edge router. In such a scenario, you can use each Ethernet Routing Switch 8600 for aggregation and peer it with a Layer 3 edge router, as shown in the following figure.

Figure 55
BGP and Internet peering

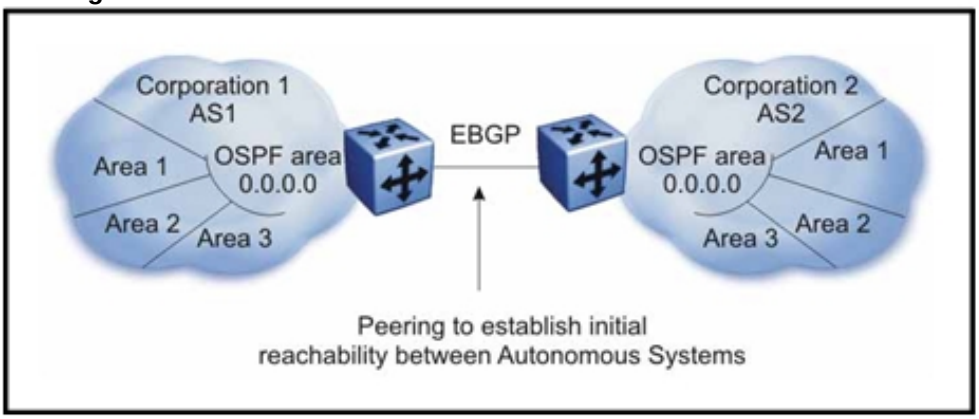


In cases where the Internet connection is single-homed, to reduce the size of the routing table, Nortel recommends that you advertise Internet routes as the default route to the IGP.

Routing domain interconnection with BGP

You can implement BGP so that autonomous routing domains, such as OSPF routing domains, are connected. This allows the two different networks to begin communicating quickly over a common infrastructure, thus giving network designers additional time to plan the IGP merger. Such a scenario is particularly effective when network administrators wish to merge two OSPF area 0.0.0.0s (see the following figure).

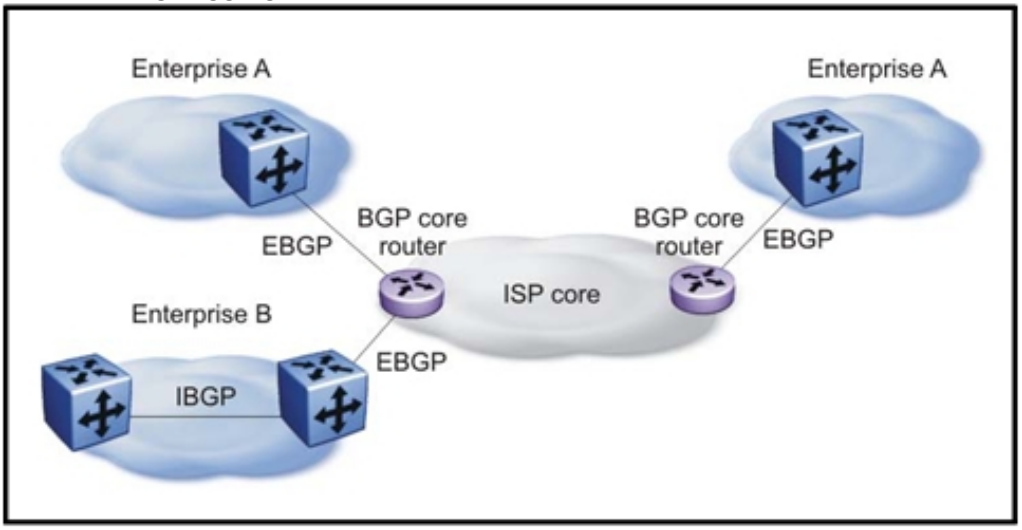
Figure 56
Routing domain interconnection with BGP



BGP and edge aggregation

You can perform edge aggregation with multiple point of presence/edge concentrations. The Ethernet Routing Switch 8600 provides 1000 or 10/100 Mbit/s EBGP peering services. To interoperate with Multiprotocol Label Switching (MPLS) or Virtual Private Network (VPN) (RFC 2547) services at the edge, this particular scenario is ideal. You can use BGP to inject dynamic routes rather than using static routes or RIP (see the following figure).

Figure 57
BGP and edge aggregation



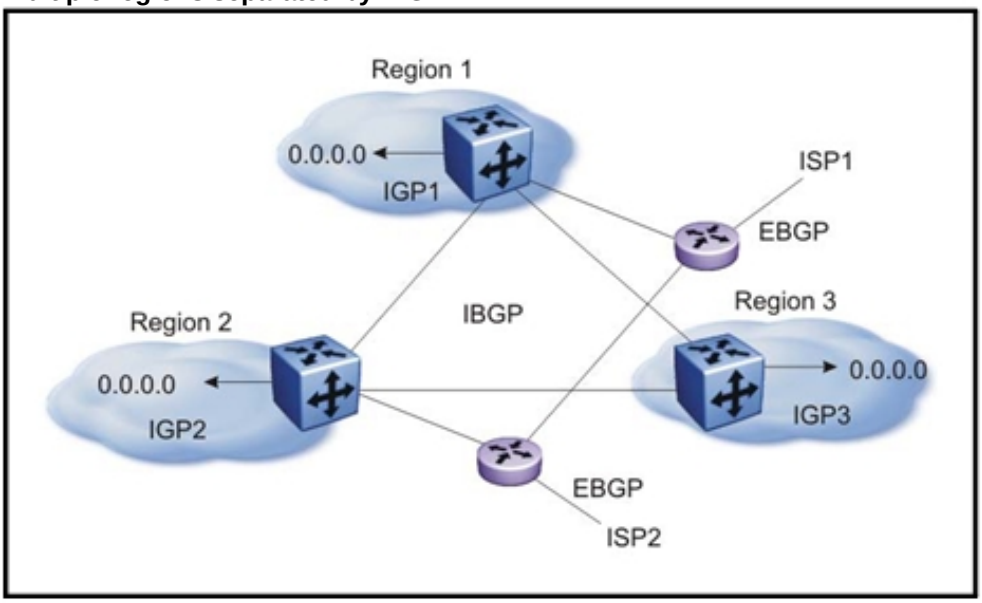
BGP and ISP segmentation

You can use the switch as a peering point between different regions or ASs that belong to the same ISP. In such cases, you can define a region as an OSPF area, an AS, or a part of an AS.

You can divide the AS into multiple regions that each run different Interior Gateway Protocols (IGP). Interconnect regions logically via a full IBGP mesh. Each region then injects its IGP routes into IBGP and also injects a default route inside the region. Thus, for destinations that do not belong to the region, each region defaults to the BGP border router.

Use the community parameter to differentiate between regions. You can use this parameter in conjunction with a route reflector hierarchy to create large VPNs. To provide Internet connectivity, this scenario requires you to make your Internet connections part of the central IBGP mesh (see the following figure).

Figure 58
Multiple regions separated by IBGP



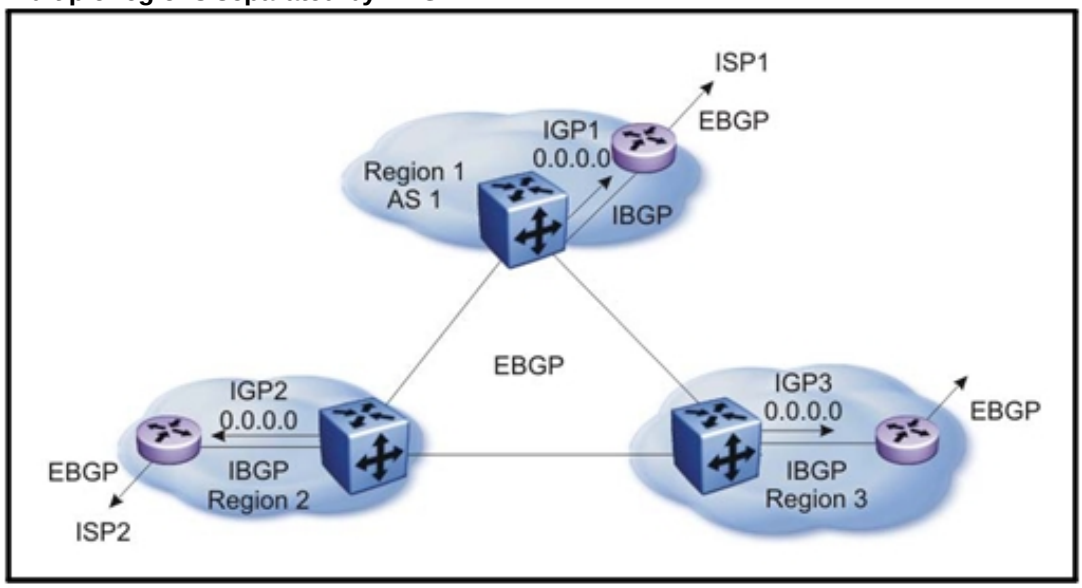
In this figure, consider the following:

- The AS is divided into three regions that each run different and independent IGPs.
- Regions are logically interconnected via a full-mesh IBGP, which also provides Internet connectivity.
- Internal nonBGP routers in each region default to the BGP border router, which contains all routes.
- If the destination belongs to any other region, the traffic is directed to that region; otherwise, the traffic is sent to the Internet connections according to BGP policies.

To set multiple policies between regions, represent each region as a separate AS. Then, implement EBGP between ASs, and implement IBGP within each AS. In such instances, each AS injects its IGP routes into BGP where they are propagated to all other regions and the Internet.

The following figure shows the use of EBGP to join several ASs.

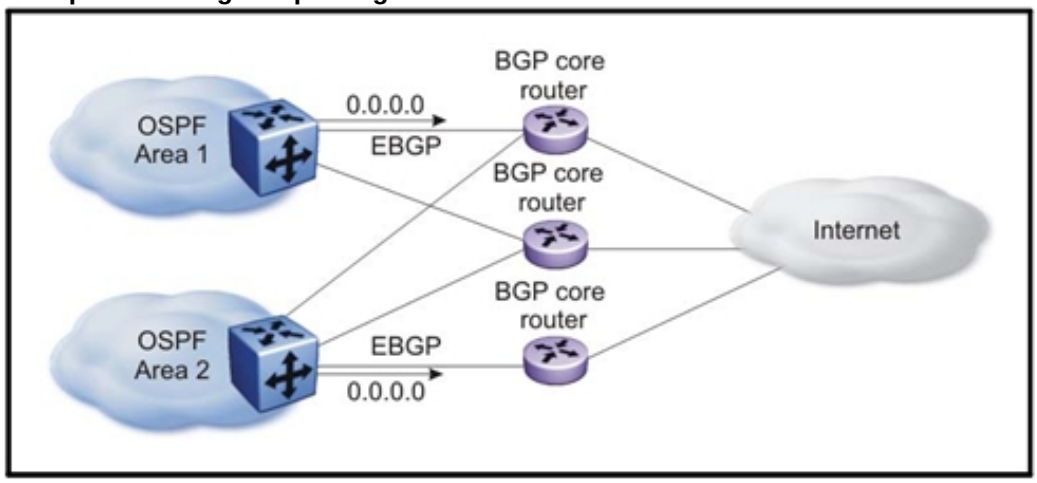
Figure 59
Multiple regions separated by EBGP



You can obtain AS numbers from the Inter-Network Information Center (NIC) or use private AS numbers. If you use private AS numbers, be sure to design your Internet connectivity very carefully. For example, you can introduce a central, well-known AS to provide interconnections between all private ASs and/or the Internet. Before propagating the BGP updates, this central AS strips the private AS numbers to prevent them from leaking to providers.

The following figure illustrates a design scenario in which you use multiple OSPF regions to peer with the Internet.

Figure 60
Multiple OSPF regions peering with the Internet



Open Shortest Path First

Use Open Shortest Path First to ensure that the switch can communicate with other OSPF-speaking routers. This section describes some general design considerations and presents a number of design scenarios for OSPF.

For more information about OSPF and a list of OSPF commands see *Nortel Ethernet Routing Switch 8600 Configuration — OSPF and RIP* (NN46205-522) .

OSPF navigation

- "OSPF scaling guidelines" (page 181)
- "OSPF design guidelines" (page 182)
- "OSPF and CPU utilization" (page 183)
- "OSPF network design examples" (page 183)

OSPF scaling guidelines

For information about OSPF scaling numbers, see Table 5 "Supported scaling capabilities" (page 38) and *Nortel Ethernet Routing Switch 8600 Release Notes* (NN46205-402) . The Release Notes take precedence over this document.

OSPF LSA limits

To determine OSPF link state advertisement (LSA) limits:

1. Use the command **`show ip ospf area`** to determine the LSA_CNT and to obtain the number of LSAs for a given area.
2. Use the following formula to determine the number of areas. Ensure the total is less than 40K:

$$\sum \mathrm{Adj}_N * \mathrm{LSA\_CNT}_N < 40k$$

N=1tothenumberofareasperswitch

Adj_N = number of adjacencies per Area N

$\mathrm{LSA\_CNT}_N$ = number of LSAs per Area N

For example, assume that a switch has a configuration of three areas with a total of 18 adjacencies and 1000 routes. This includes:

- 3 adjacencies with an LSA_CNT of 500 (Area 1)
- 10 adjacencies with an LSA_CNT of 1000 (Area 2)
- 5 adjacencies with an LSA_CNT of 200 (Area 3)

Calculate the number as follows:

3*500+10*1000+5*200=12.5K < 40K

This configuration ensures that the switch operates within accepted scalability limits.

OSPF design guidelines

Follow these additional OSPF guidelines:

- Use OSPF area summarization to reduce routing table sizes.
- Use OSPF passive interfaces to reduce the number of active neighbor adjacencies.
- Use OSPF active interfaces only on intended route paths.

  Configure wiring closet subnets as OSPF passive interfaces unless they form a legitimate routing path for other routes.
- Minimize the number of OSPF areas per switch to avoid excessive shortest path calculations.

  The switch executes the Djikstra algorithm for each area separately.
- Ensure that the OSPF dead interval is at least four times the OSPF hello interval

OSPF and CPU utilization

When you create an OSPF area route summary on an area boundary router (ABR), the summary route can attract traffic to the ABR for which the router does not have a specific destination route. The enabling of ICMP unreachable message generation on the switch may result in a high CPU utilization rate.

To avoid high CPU utilization, Nortel recommends that you use a black hole static route configuration. The black hole static route is a route (equal to the OSPF summary route) with a next-hop of 255.255.255.255. This ensures that all traffic that does not have a specific next-hop destination route is dropped.

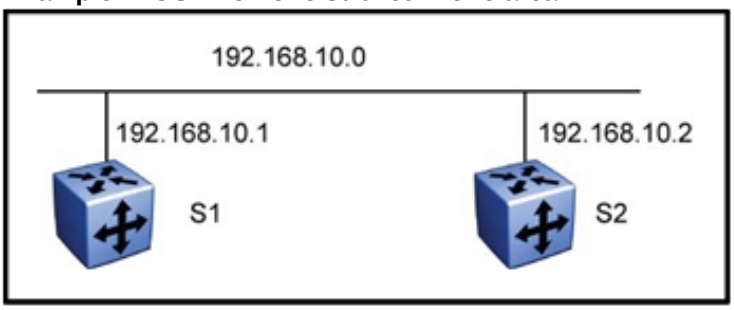
OSPF network design examples

Three OSPF network design examples are presented in the sections that follow.

Example 1: OSPF on one subnet in one area

Example 1 describes a simple implementation of an OSPF network: enabling OSPF on two switches (S1 and S2) that are in the same subnet in one OSPF area. See the following figure.

Figure 61
Example 1: OSPF on one subnet in one area



The routers in example 1 have the following settings:

- S1 has an OSPF router ID of 1.1.1.1, and the OSPF port is configured with an IP address of 192.168.10.1.
- S2 has an OSPF router ID of 1.1.1.2, and the OSPF port is configured with an IP address of 192.168.10.2.

The general method used to configure OSPF on each routing switch is:

1. Enable OSPF globally.
2. Verify that IP forwarding is enabled on the switch.
3. Configure the IP address, subnet mask, and VLAN ID for the port.

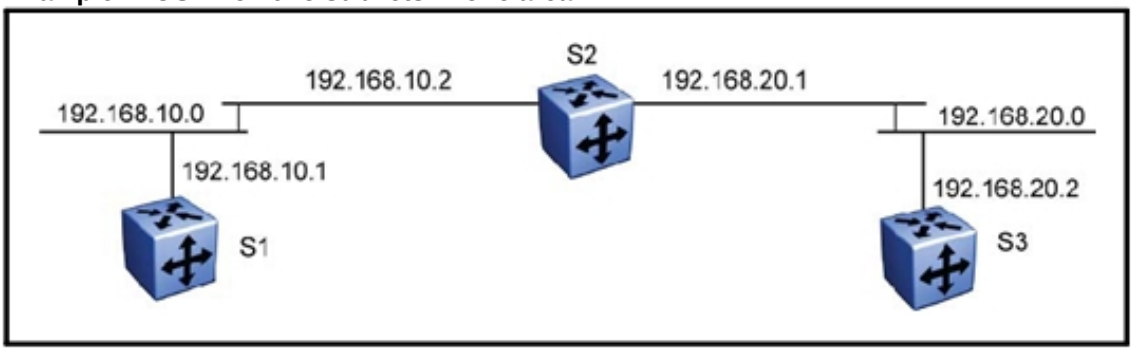
4. If RIP is not required on the port, disable it.
5. Enable OSPF for the port.

After you configure S2, the two switches elect a designated router (DR) and a backup designated router (BDR). They exchange Hello packets to synchronize their link state databases.

Example 2: OSPF on two subnets in one area

The following figure shows a configuration in which OSPF operates on three switches. OSPF performs routing on two subnets in one OSPF area. In this example, S1 directly connects to S2, and S3 directly connects to S2, but any traffic between S1 and S3 is indirect, and passes through S2.

Figure 62
Example 2: OSPF on two subnets in one area



The routers in example 2 have the following settings:

- S1 has an OSPF router ID of 1.1.1.1, and the OSPF port is configured with an IP address of 192.168.10.1.
- S2 has an OSPF router ID of 1.1.1.2, and two OSPF ports are configured with IP addresses of 192.168.10.2 and 192.168.20.1.
- S3 has an OSPF router ID of 1.1.1.3, and the OSPF port is configured with an IP address of 192.168.20.2.

The general method used to configure OSPF on each routing switch is:

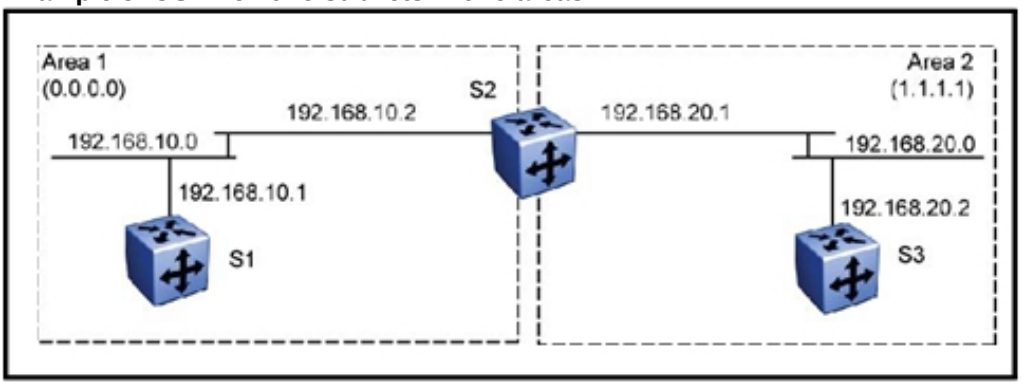
1. Enable OSPF globally.
2. Insert IP addresses, subnet masks, and VLAN IDs for the OSPF ports on S1 and S3 and for the two OSPF ports on S2. The two ports on S2 enable routing and establish the IP addresses related to the two networks.
3. Enable OSPF for each OSPF port allocated with an IP address.

When all three switches are configured for OSPF, they elect a DR and BDR for each subnet and exchange hello packets to synchronize their link state databases.

Example 3: OSPF on two subnets in two areas

The following figure shows an example where OSPF operates on two subnets in two OSPF areas. S2 becomes the ABR for both networks.

Figure 63
Example 3: OSPF on two subnets in two areas



The routers in scenario 3 have the following settings:

- S1 has an OSPF router ID of 1.1.1.1. The OSPF port is configured with an IP address of 192.168.10.1 which is in OSPF area 1.
- S2 has an OSPF router ID of 1.1.1.2. One port has an IP address of 192.168.10.2, which is in OSPF area 1. The second OSPF port on S2 has an IP address of 192.168.20.1 which is in OSPF area 2.
- S3 has an OSPF router ID of 1.1.1.3. The OSPF port is configured with an IP address of 192.168.20.2 which is in OSPF area 2.

The general method used to configure OSPF for this three-switch network is:

1. On all three switches, enable OSPF globally.
2. Configure OSPF on one network.

   On S1, insert the IP address, subnet mask, and VLAN ID for the OSPF port. Enable OSPF on the port. On S2, insert the IP address, subnet mask, and VLAN ID for the OSPF port in area 1, and enable OSPF on the port. Both routable ports belong to the same network. Therefore, by default, both ports are in the same area.
3. Configure three OSPF areas for the network.
4. Configure OSPF on two additional ports in a second subnet.

   Configure additional ports and verify that IP forwarding is enabled for each switch to ensure that routing can occur. On S2, insert the IP address, subnet mask, and VLAN ID for the OSPF port in area 2, and enable OSPF on the port. On S3, insert the IP address, subnet mask, and VLAN ID for the OSPF port, and enable OSPF on the port.

The three switches exchange Hello packets.

In an environment with a mix of Cisco and Nortel switches/routers, you may need to manually modify the OSPF parameter RtrDeadInterval to 40 seconds.

Internetwork Packet Exchange

Internetwork Packet Exchange (IPX) is a datagram networking protocol used by Novell NetWare operating systems. If you must support IPX traffic on your network, use the following guidelines.

IPX is not supported on R series modules.

IPX navigation

- "IPX and R series modules" (page 186)
- "IPX and Get Nearest Server" (page 189)
- "IPX and LLC encapsulation and translation" (page 189)
- "IPX RIP and SAP policies" (page 189)

IPX and R series modules

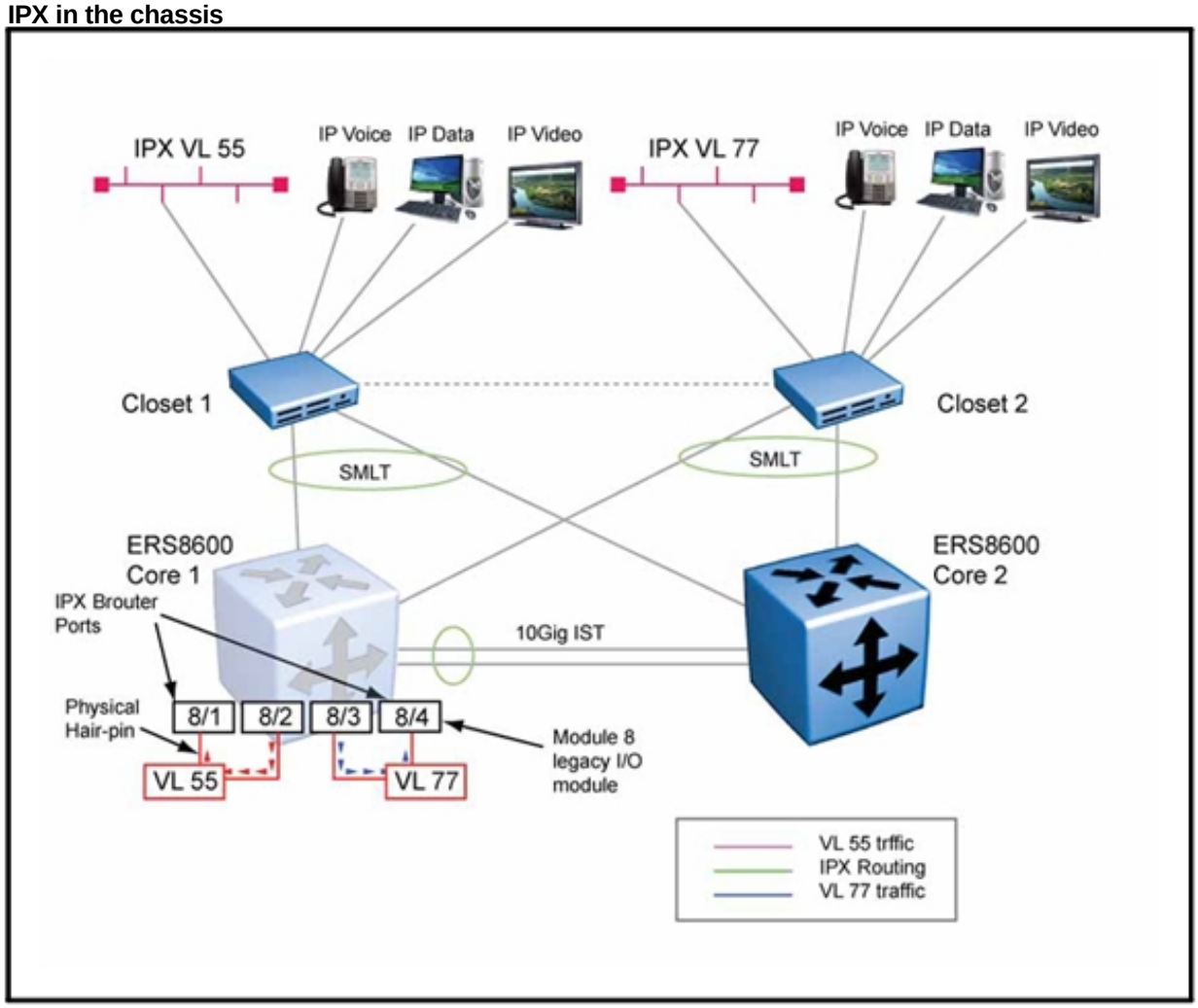
IPX is not supported on R series modules. To route IPX traffic and still use R series modules, you have two primary choices:

- Bridge the IPX traffic to an external IPX router.
- Install a Classic module and use a physical hairpin.

The bridge option is straightforward to design and implement. In this case, tag the IPX VLANs across an MLT/SMLT trunk to another switch with Classic modules, or to another Nortel router that supports IPX routing. Then, enable IPX on the VLANs.

However, if you do not want to add another switch or router in the network for IPX routing only, and R mode is not required, you can run the chassis in mixed mode and install a Classic module in the chassis for the IPX routing (see the following figure).

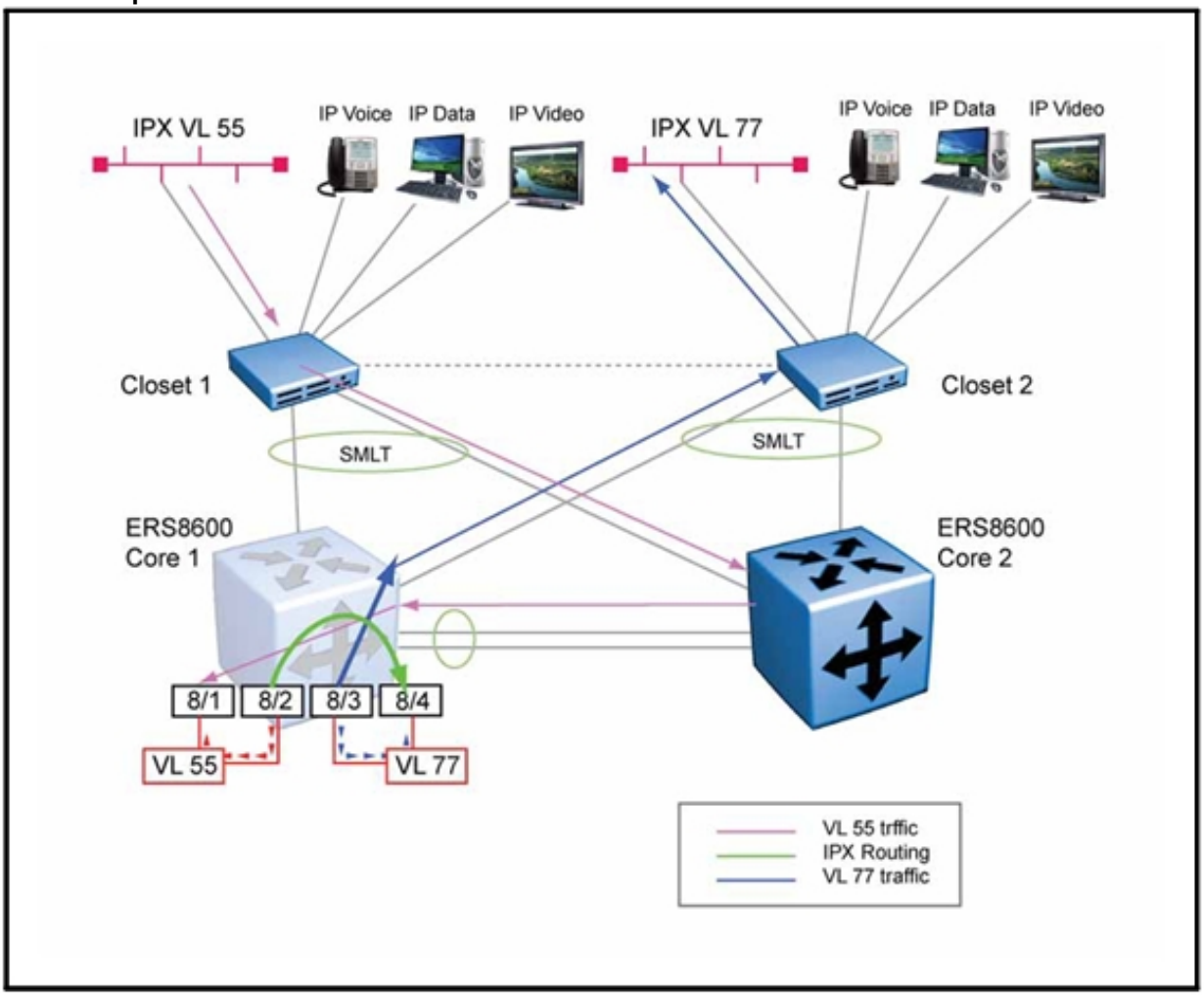
Figure 64
IPX in the chassis



As shown in this figure, it is assumed that the interswitch trunk, which connects the closets to the core, is connected to Ethernet Routing Switch 8600 modules, thereby disabling IPX routing on VLANs 55 and 77. To terminate the IPX traffic on the Classic module (module 8), each IPX VLAN requires a separate physical hairpin, as shown. Port 8/1 is associated with VL55 which is hairpinned to port 8/2 and then is configured as an IPX brouter port. Likewise for VL77, port 8/3 is associated with VL77 and then hairpinned to IPX brouter port 8/4.

If a device on VL55 wants to communicate with a device on VL77, the traffic passes through the SMLT links to Ethernet Routing Switch 8600 Core1 or Core2, is broadcast to port 8/1, flows through the hairpin to the IXP brouter port 8/2, is routed to IPX brouter port 8/4, is broadcast via the physical hairpin to port 8/3, and finally is sent to the SMLT link and on to the IPX VLAN on Closet N. The following figure provides an illustration of the traffic path.

Figure 65
IPX traffic path



Each IPX VLAN needs a physical hairpin; you cannot tag multiple VLANs across a single physical hairpin.

The following key points must be considered:

- The IPX brouter ports must be removed from all VLANs and from STG1, or a Layer 2 loop forms when you connect the hairpin. You cannot simply disable STG on the brouter port; the port must be removed from the STG group.
- For resiliency, you can design an identical hairpin setup on the second Ethernet Routing Switch 8600 Core node, and IPX traffic for a given flow is routed by the switch that responds first.

IPX and Get Nearest Server

IPX clients use the Get Nearest Server (GNS) request to find a server for logon. If a server is available on the same network segment, this server answers the GNS request with a GNS response. If no server is present, the routing device provides the GNS response.

The switch chooses the closest NetWare server by using the following algorithm:

1. The switch checks the route cost.
2. If multiple servers exist with the same RIP route cost, the switch uses the server with the lowest Service Advertisement Protocol (SAP) hop count.
3. If multiple servers with the same SAP cost are available, the switch responds with the servers in alphabetical order. This provides a means to load balance user network logons over multiple servers.

If you encounter connection problems because the switch responds with a nonoptimal NetWare server, increase the hop count to that NetWare server.

IPX and LLC encapsulation and translation

Logical Link Control (LLC) translation to and from Gigabit Ethernet (GbE) ports is not supported by other module types. To avoid network connectivity problems, avoid designs that require LLC translation: use one encapsulation type throughout your network.

If you have client switches with LLC encapsulation and another encapsulation type, do not use LLC encapsulation over the Gigabit Ethernet connection.

The Ethernet Routing Switch 8616SXE module and all other enhanced Gigabit Ethernet modules (E modules) support LLC translation to and from Gigabit Ethernet (GE) ports.

IPX RIP and SAP policies

You can use IPX RIP policies (filters) to shield networks from users on different network segments. Route filters give you control over the routing of IPX packets from one area of an IPX internetwork to another.

Use route filters to help maximize the use of the available bandwidth throughout the IPX internetwork. Filters also help improve network security by restricting a user's view of other networks. You can configure inbound and outbound route filters on a per-interface basis, instructing the interface to advertise/accept or drop filtered RIP packets. The action parameter that

you define for the filter determines whether the router advertises, accepts, or drops RIP packets from routers that match the filter criteria. The same concept applies to SAP packets.

For information about configuring IPX RIP/SAP policies see *Nortel Ethernet Routing Switch 8600 Configuration — IPX Routing Operations* (NN46205-505) .

IP routed interface scaling

The Ethernet Routing Switch 8600 supports up to 1972 IP routed interfaces using SF/CPUs that have 256 MB of memory. You can upgrade SF/CPUs that do not have 256 MB by using the memory upgrade kit (Part # DS1404015). For instructions, see *Nortel Ethernet Routing Switch 8600 Upgrades* (NN46205-400) .

When you configure a large number of IP routed interfaces, use the following guidelines:

- Use passive interfaces on most of the configured interfaces. You can only make very few interfaces active.
- For Distance Vector Multicast Routing Protocol (DVMRP), you can use up to 80 active interfaces and up to 1200 passive interfaces. This assumes that no other routing protocols are running. If you need to run other routing protocols to perform IP routing, you can enable IP forwarding and use routing policies and default route policies . If you use a dynamic routing protocol, enable only a few interfaces with OSPF or RIP. One or two OSPF or RIP interfaces allow the switch to exchange dynamic routes.
- When using Protocol Independent Multicast (PIM), configure a maximum of 10 PIM active interfaces. The remainder can be passive interfaces. Nortel recommends that you use IP routing policies with one or two unicast IP active interfaces.

Internet Protocol version 6

Internet Protocol version 6 (IPv6) enables high-performance, scalable internet communications. This section provides information that you can use to help deploy IPv6 in your network.

For more information about IPv6, see *Nortel Ethernet Routing Switch 8600 Configuration — IPv6 Routing Operations* (NN46205-504) .

IPv6 navigation

- "IPv6 requirements" (page 191)
- "IPv6 design recommendations" (page 191)

IPv6 requirements

To use IPv6, the switch requires:

- Rseriesmodulesforhardwareforwarding
- Enterprise Enhanced CPU daughter card (SuperMezz)
- at least one 8692 SF/CPU module
- Ethernet Routing Switch 8600 Software Release 4.1 or later for IPv6 hardware-based forwarding

IPv6 mode does not route a VLAN that spans Classic and R series modules.

IPv6 design recommendations

Nortel Layer 2 and Layer 3 Ethernet switches support protocol-based IPv6 VLANs. To simplify network configuration with IPv6, Nortel recommends that you use protocol-based IPv6 VLANs from Edge Layer 2 switches. The core switch performs hardware-based IPv6 line-rate routing.

For IPv6 scaling information, see Table 5 "Supported scaling capabilities" (page 38).

Transition mechanisms for IPv6

The Ethernet Routing Switch 8600 helps networks transition from IPv4 to IPv6 by using three primary mechanisms:

- Dual Stack mechanism, where the IPv4 and IPv6 stacks can communicate with both IPv6 and IPv4 devices
- Tunneling, which involves the encapsulation of IPv6 packets to traverse IPv4 networks and the encapsulation of IPv4 packets to traverse IPv6 networks
- Translation mechanisms, which translate one protocol to the other

Dual-stack tunnels

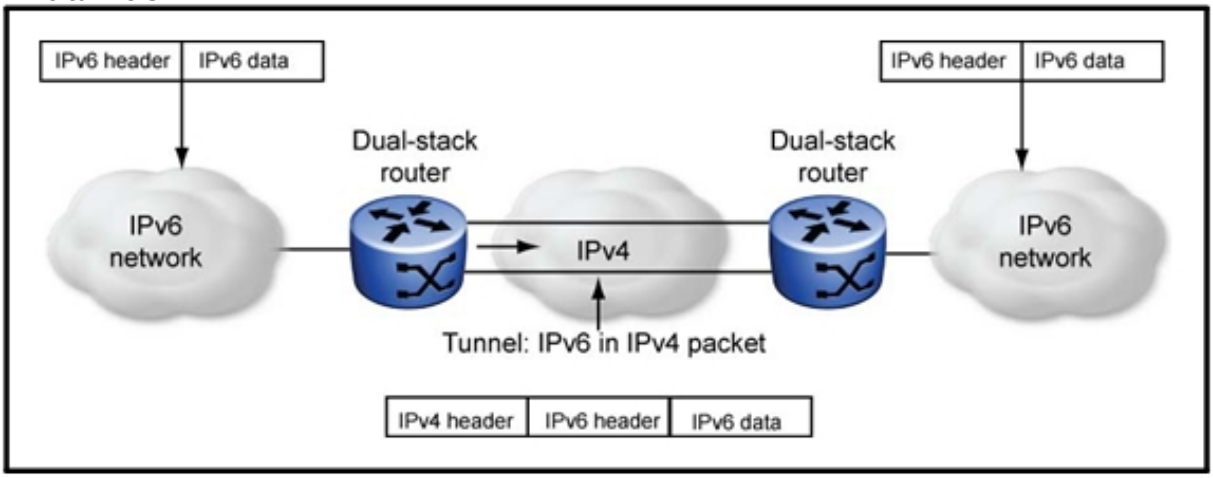
A manually configured tunnel (as per RFC 2893) is equivalent to a permanent link between two IPv6 domains over an IPv4 backbone. Use tunnels to provide stable, secure communication between two edge routers or between an end system and an edge router, or to provide a connection to remote IPv6 networks.

Edge routers and end systems (at the end of the tunnel) must be dual-stack implementations. At each end of the tunnel, configure the IPv4 and IPv6 addresses of the dual-stack routing switch on the tunnel interface and identify the entry and exit (or source and destination) points using IPv4 addresses. For Enterprise networks, your ISP provides you with the appropriate IPv6 address prefix for your site. Your ISP also provides you with the required destination IPv4 address for the exit point of the tunnel.

The following figure shows a manually-configured tunnel.

For more examples, see *Nortel Ethernet Routing Switch 8600 Configuration — IPv6 Routing Operations* (NN46205-504) .

Figure 66
IPv6 tunnels



Because each tunnel exists between only two routing switches and is independently managed, additional tunnels are required whenever you add new routing switches. Each additional tunnel and switch increases management overhead. Network Address Translation (NAT), when applied to the outer IPv4 header, is allowed along the path of the tunnel only if the translation map is stable and preestablished.

Multicast network design

Use multicast routing protocols to efficiently distribute a single data source among multiple users in the network. This section provides information about designing networks that support IP multicast routing.

For more information about multicast routing, see *Nortel Ethernet Routing Switch 8600 Configuration — IP Multicast Routing Protocols* (NN46205-501) .

Navigation

General multicast considerations

Use the following general rules and considerations when planning and configuring IP multicast.

General multicast considerations navigation

Multicast and VRF-lite

PIM-SM, PIM-SSM, and IGMP are supported in VRF-Lite configurations. No other multicast protocols are supported with VRF-lite.

Multicast virtualization provides support for:

- Virtualization of control and data plane
- Multicast routing tables managers (MRTM)
- Virtualized PIM-SM/SSM, IGMPv1/v2/v3
- Support for overlapping multicast address spaces
- Support for Global Routing Table (VRF0) and 255 VRFs
- SMLT/RSMLT support for Multicast VRFs •

64 instances of PIM-SM/SSM

- Total of 4000 multicast routes

Requirements

To support multicast virtualization, the Ethernet Routing Switch 8600 must be equipped with the following:

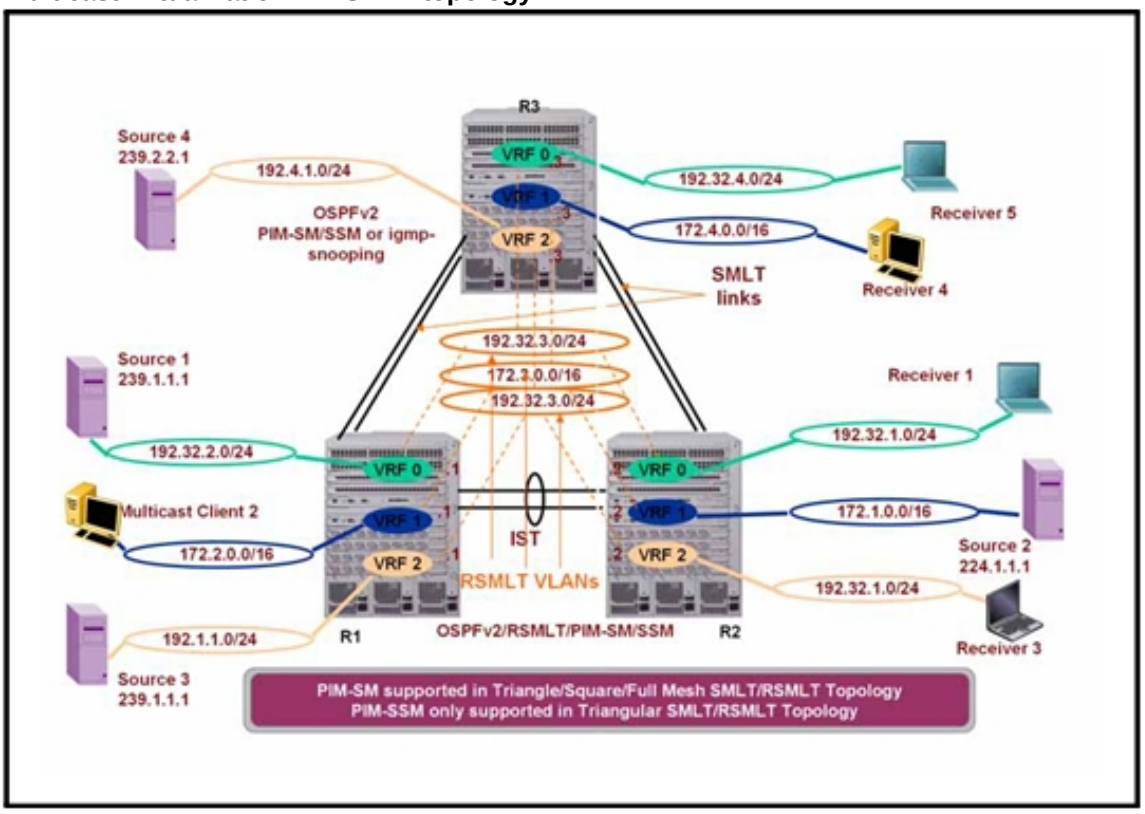
- Release 5.1 (or later) software
- Premier Software License

- R/RS modules
- SuperMezz CPU-Daughter card

Multicast virtualization network scenarios

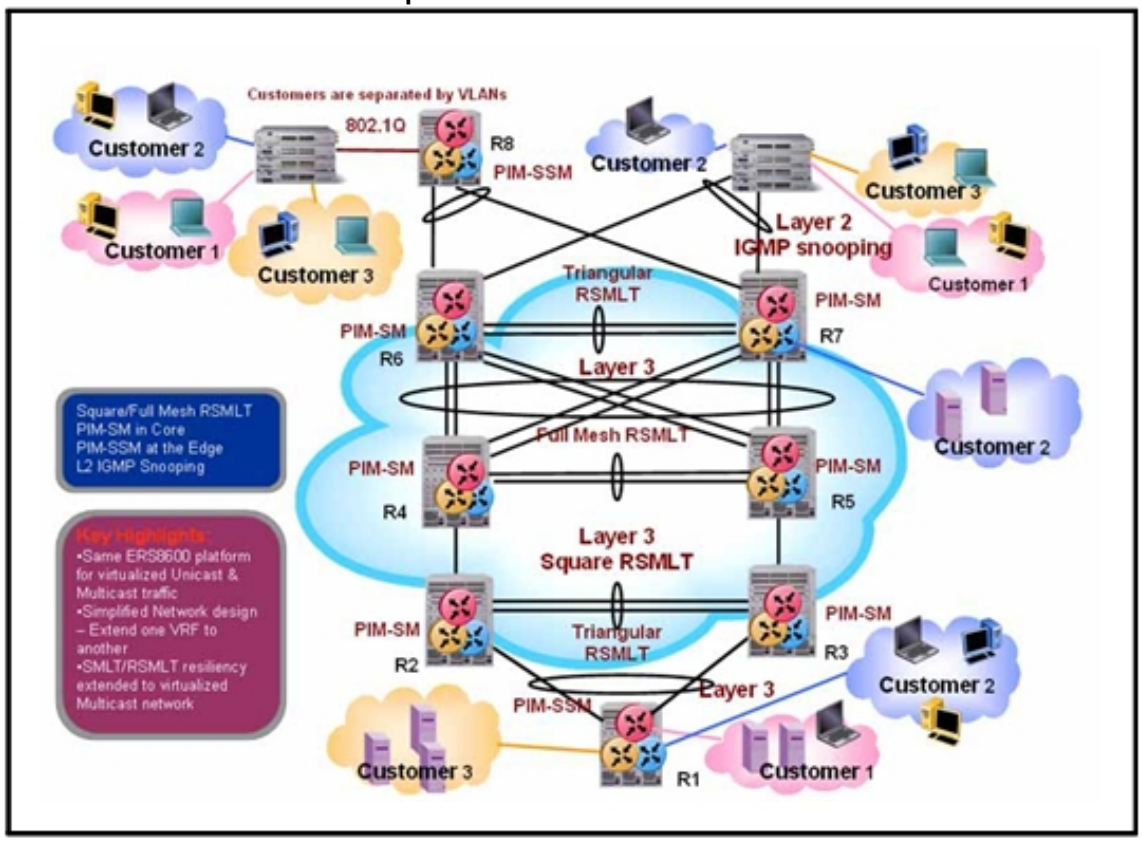
The following figure shows an example of multicast virtualization in an RSMLT topology.

Figure 67
Multicast virtualization in RSMLT topology



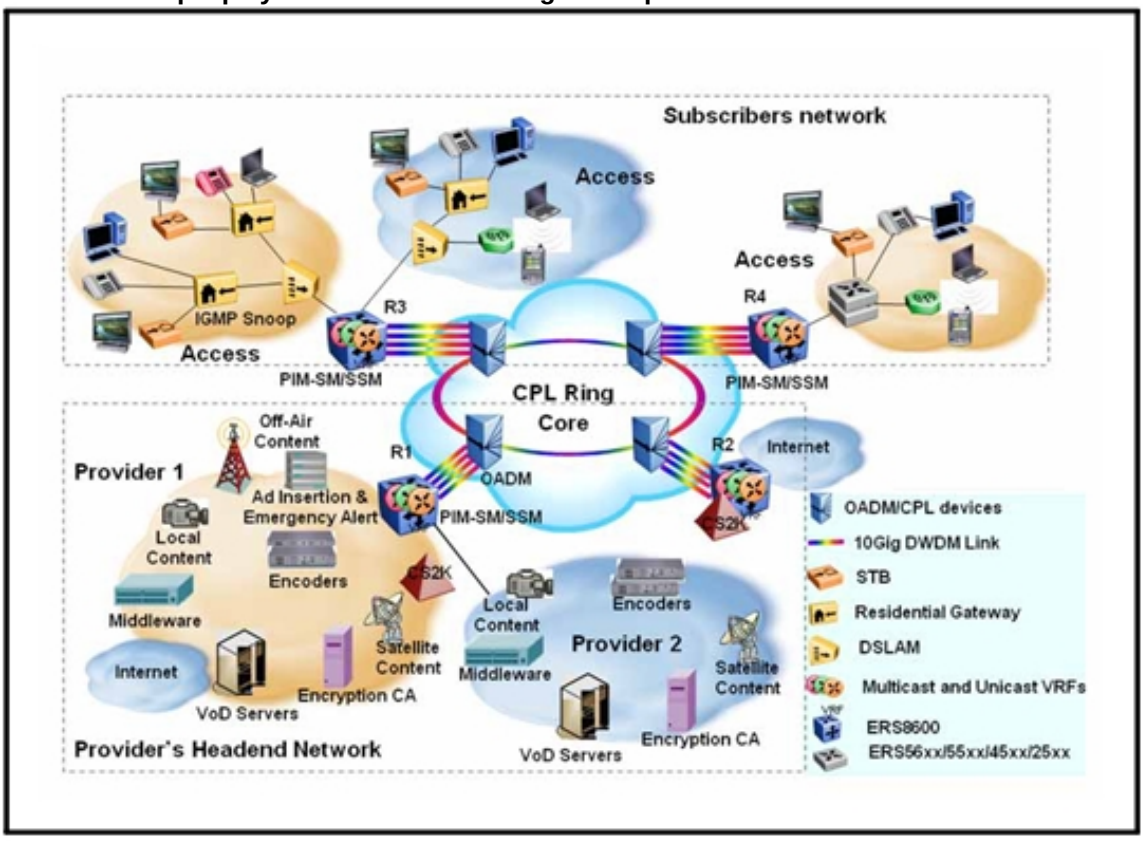
The following figure shows an example of multicast virtualization in an Enterprise/Metro network.

Figure 68
Multicast virtualization for Enterprise/Metro network



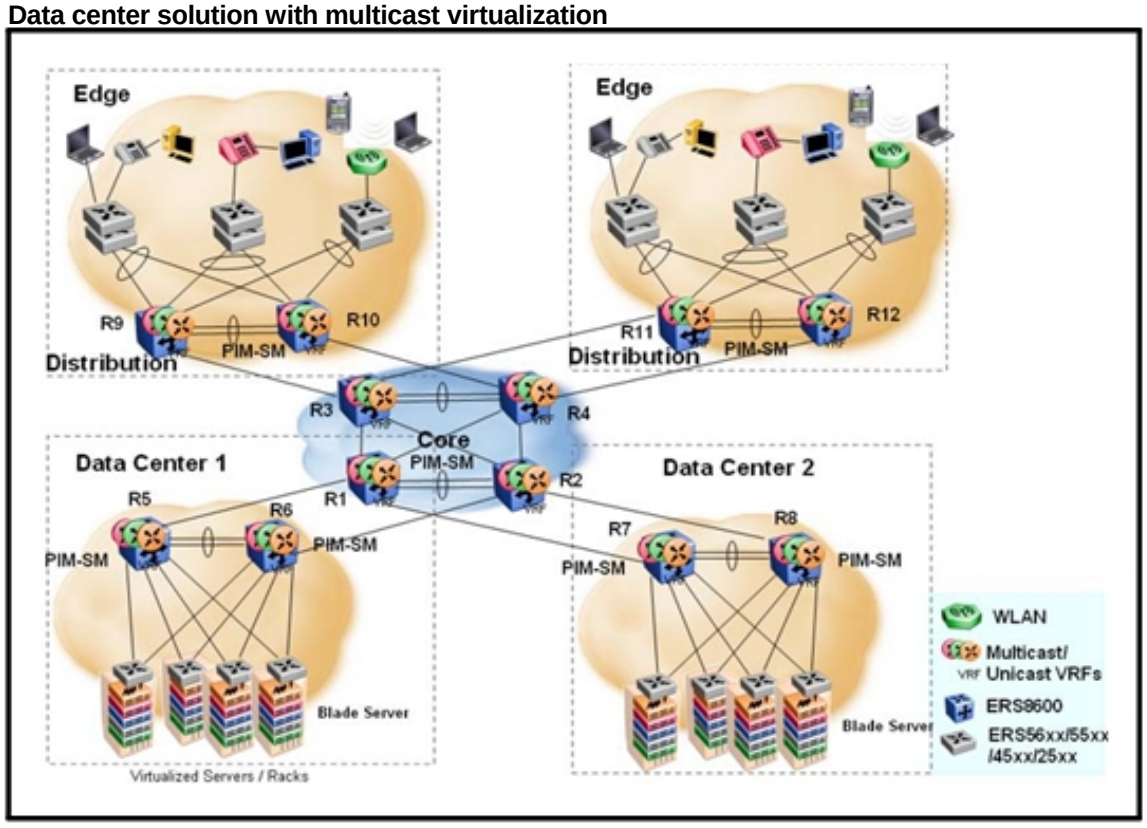
The following figure shows an example of multicast virtualization supporting an end-to-end triple play solution for an MSO/Large Enterprise.

Figure 69
End-to-end triple play solution for MSO/Large Enterprise



The following figure shows an example of multicast virtualization in a data center.

Figure 70
Data center solution with multicast virtualization



Multicast and Multi-Link Trunking considerations

Multicast traffic distribution is important because the bandwidth requirements can be substantial when a large number of streams are employed. The Ethernet Routing Switch 8600 can distribute IP multicast streams over links of a multilink trunk. If you need to use several links to share the load of several multicast streams between two switches, use one of the following:

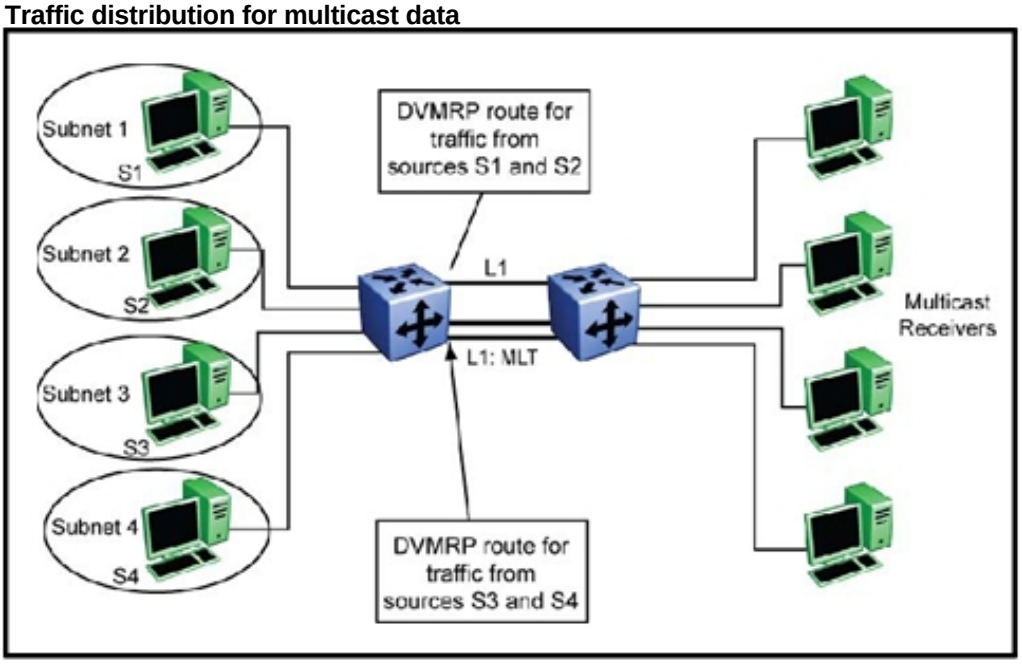
- “DVMRP or PIM route tuning to load share streams” (page 198)
- “Multicast flow distribution over MLT” (page 199)

DVMRP or PIM route tuning to load share streams

You can use Distance Vector Multicast Routing Protocol (DVMRP) or Protocol Independent Multicast (PIM) routing to distribute multicast traffic. With this method, you must distribute sources of multicast traffic on different IP subnets and configure routing metrics so that traffic from different sources flows on different paths to the destination groups.

The following figure illustrates one way to distribute multicast traffic sourced on different subnets and forwarded on different paths.

Figure 71
Traffic distribution for multicast data



The multicast sources S1 to S4 are on different subnets; use different links for every set of sources to send their multicast data. In this case, S1 and S2 send their traffic on a common link (L1) and S3 and S4 use another common link (L2). These links can be MLT links. Unicast traffic is shared on the MLT links, whereas multicast traffic only uses one of the MLT links. Receivers can be located anywhere on the network. This design can be worked in parallel with unicast designs and, in the case of DVMRP, does not impact unicast routing.

In this example, sources must be on the VLAN that interconnects the two switches. In more generic scenarios, you can design the network by changing the interface cost values to force some paths to be taken by multicast traffic.

When multicast routing is used in MLT configurations, Nortel recommends using E, M, or R series modules if the MLT on the Ethernet Routing Switch 8600 is connected to a nonEthernet Routing Switch 8600 device.

Multicast flow distribution over MLT

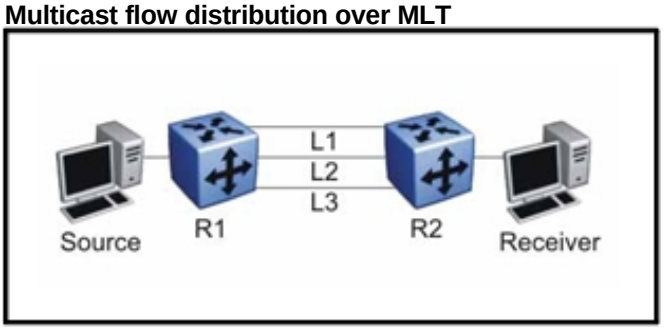
The multicast flow distribution over MLT feature is supported only on E, M, or R series modules. As a result, all the modules that have ports in an MLT must be E, M , or R series modules to enable multicast flow distribution over MLT.

MultiLink Trunking distributes multicast streams over a multilink trunk based on the source-subnet and group addresses of the packets. You can choose the address parameters that the distribution algorithm uses. As a result, you can distribute the load on different ports of the MLT and achieve an even stream distribution.

To determine the egress port for a particular Source, Group (S,G) pair, the number of active ports of the MLT is used to MOD the number generated by the XOR of each byte of the masked group address with the masked source address. (The MOD function returns the remainder after a number is divided by divisor; the XOR [or exclusive-or function] operates such that *a* XOR *b* is true if *a* is true, or if *b* is true, but not if both are false, or both are true.)

Flow distribution and stream failover considerations This section describes a traffic interruption issue that can occur in a PIM domain that has the multicast MLT flow redistribution feature enabled. The following figure illustrates a normal scenario where multicast streams flow from R1 to R2 through an MLT. The streams are distributed on links L1, L2 and L3.

Figure 72
Multicast flow distribution over MLT



If link L1 goes down, the affected streams are distributed on links L2 and L3. However, with redistribution enabled, the unaffected streams (flowing on L2 and L3) also start distributing. Because the switch does not update the corresponding RPF (Reverse Path Forwarding) ports on switch R2 for these unaffected streams, this causes the activity check for these streams to fail (because of an incorrect RPF port). Then, the switch improperly prunes these streams.

To avoid this issue, make sure that the **`activity-chk-interval`** parameter is set to its default of 210 seconds. If the activity check fails when the (S,G) entry timer expires (210 seconds), the switch deletes the (S,G) entry. The (S,G) entry is recreated when packets corresponding to the (S,G) pair reach the switch again. There can be a short window of traffic interruption during this deletion-creation period.

Multicast scalability design rules

To increase multicast route scaling, follow these eight design rules:

1. Whenever possible, use simple network designs that do not use VLANs that span several switches. Instead, use routed links to connect switches.

2. Whenever possible, group sources should send to the same group in the same subnet. The Ethernet Routing Switch 8600 uses a single egress forwarding pointer for all sources in the same subnet sending to the same group. Be aware that these streams have separate hardware forwarding records on the ingress side.

   To obtain information about the ingress and egress port information for IP multicast streams flowing through your switch, use the CLI command **`show ip mroute-hw group trace`**.

   In the NNCLI, the command is **`show ip mroute hw-group-trace`**.

3. Do not configure multicast routing on edge switch interfaces that do not contain multicast senders or receivers. By following this rule, you:

   — Provide secured control over multicast traffic that enters or exits the interface.

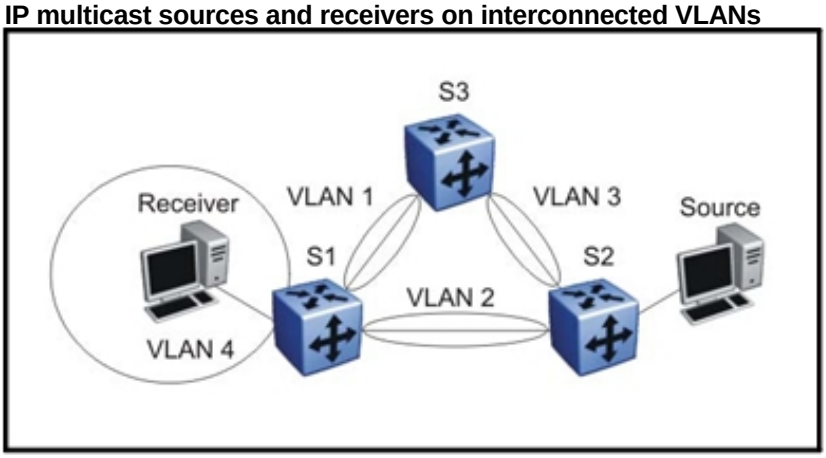
   — Reduce the load on the switch, as well as the number of routes. This improves overall performance and scalability.

4. Avoid initializing many (several hundred) multicast streams simultaneously. Initial stream setup is a resource-intensive task, and initializing a large number may slow down the setup time. In some cases, this can result in some stream loss.

5. Whenever possible, do not connect IP multicast sources and receivers by using VLANs that interconnect switches (see the following figure). In some cases, this can result in excessive hardware record use. By placing the source on the interconnected VLAN, traffic takes two paths to the destination, depending on the RPF checks and the shortest path to the source.

   For example, if a receiver is placed on VLAN 1 on switch S1 and another receiver is placed on VLAN 2 on this switch, traffic may be received from two different paths to the two receivers. This results in the use of two forwarding records. When the source on switch S2 is placed on a different VLAN than VLAN 3, traffic takes a single path to switch S1 where the receivers are located.

Figure 73
IP multicast sources and receivers on interconnected VLANs



6. Use default timer values for PIM and DVMRP. When timers are decreased for faster convergence, they usually adversely affect scalability because control messages are sent more frequently. If faster network convergence is required, configure the timers with the same values on all switches in the network. Also, in most cases, you must perform baseline testing to achieve optimal values for timers versus required convergence times and scalability. For more information, see "DVMRP timer tuning" (page 213).
7. For faster convergence, configure the Bootstrap and Rendezvous Point routers on a circuitless IP. See "Circuitless IP for PIM-SM" (page 228).
8. For faster convergence, Nortel recommends using a static Rendezvous Point (RP) router.

IP multicast address range restrictions

IP multicast routers use D class addresses, which range from 224.0.0.0 to 239.255.255.255. Although subnet masks are commonly used to configure IP multicast address ranges, the concept of subnets does not exist for multicast group addresses. Consequently, the usual unicast conventions—where you reserve the *all 0s* subnets, *all 1s* subnets, *all 0s* host addresses, and *all 1s* host addresses—do not apply.

Addresses from 224.0.0.0 through 224.0.0.255 are reserved by the Internet Assigned Numbers Authority for link-local network applications. Packets with an address in this range are not forwarded by multicast-capable routers. For example, OSPF uses 224.0.0.5 and 224.0.0.6, and VRRP uses 224.0.0.18 to communicate across local broadcast network segments.

IANA has also reserved the range of 224.0.1.0 through 224.0.1.255 for well-known applications. These addresses are also assigned by IANA to specific network applications. For example, the Network Time Protocol (NTP) uses 224.0.1.1, and Mtrace uses 224.0.1.32. RFC 1700 contains a complete list of these reserved addresses.

Multicast addresses in the 232.0.0.0/8 (232.0.0.0 to 232.255.255.255) range are reserved only for source-specific multicast (SSM) applications, such as one-to-many applications. (See draft-holbrook-ssm-00.txt). While this is the publicly reserved range for SSM applications, private networks can use other address ranges for SSM.

Finally, addresses in the range 239.0.0.0/8 (239.0.0.0 to 239.255.255.255) are administratively scoped addresses; they are reserved for use in private domains and should not be advertised outside that domain. This multicast range is analogous to the 10.0.0.0/8, 172.16.0.0/20, and 192.168.0.0/16 private address ranges in the unicast IP space.

A private network should only assign multicast addresses from 224.0.2.0 through 238.255.255.255 to applications that are publicly accessible on the Internet. Multicast applications that are not publicly accessible should be assigned addresses in the 239.0.0.0/8 range.

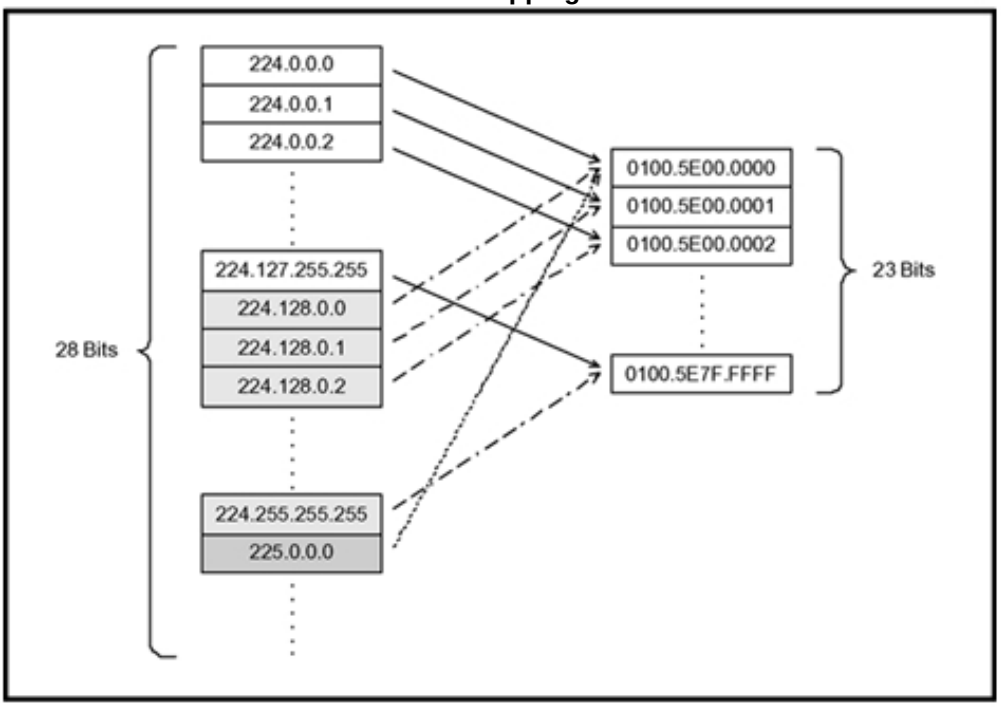
Although you can use any multicast address you choose on your own private network, it is generally not good design practice to allocate public addresses to private network entities. Do not use public addresses for unicast host or multicast group addresses on private networks. To prevent private network addresses from escaping to a public network, you may wish to use announce and accept policies as described in “Announce and accept policy examples” (page 213).

Multicast MAC address mapping considerations

Like IP, Ethernet has a range of multicast MAC addresses that natively support Layer 2 multicast capabilities. While IP has a total of 28 addressing bits available for multicast addresses, Ethernet has only 23 addressing bits assigned to IP multicast. The Ethernet multicast MAC address space is much larger than 23 bits, but only a subrange of that larger space is allocated to IP multicast. Because of this difference, 32 IP multicast addresses map to one Ethernet multicast MAC address.

IP multicast addresses map to Ethernet multicast MAC addresses by placing the low-order 23 bits of the IP address into the low-order 23 bits of the Ethernet multicast address 01:00:5E:00:00:00. Thus, more than one multicast address maps to the same Ethernet address (see the following figure). For example, all 32 addresses 224.1.1.1, 224.129.1.1, 225.1.1.1, 225.129.1.1, 239.1.1.1, 239.129.1.1 map to the same 01:00:5E:01:01:01 multicast MAC address.

Figure 74
Multicast IP address to MAC address mapping



Most Ethernet switches handle Ethernet multicast by mapping a multicast MAC address to multiple switch ports in the MAC address table. Therefore, when you design the group addresses for multicast applications, take care to efficiently distribute streams only to hosts that are receivers. The Ethernet Routing Switch 8600 switches IP multicast data based on the IP multicast address, not the MAC address, and thus, does not have this issue.

As an example, consider two active multicast streams using addresses 239.1.1.1 and 239.129.1.1. Suppose that two Ethernet hosts, receiver A and receiver B, are connected to ports on the same switch and only want the stream addressed to 239.1.1.1. Suppose also that two other Ethernet hosts, receiver C and receiver D, are also connected to the ports on the same switch as receiver A and B and wish to receive the stream addressed to 239.129.1.1. If the switch utilizes the Ethernet multicast MAC address to make forwarding decisions, then all four receivers receive both streams—even though each host only wants one stream. This increases the load on both the hosts and the switch. To avoid this extra load, Nortel recommends that you manage the IP multicast group addresses used on the network.

The switch does not forward IP multicast packets based on multicast MAC addresses—even when bridging VLANs at Layer 2. Thus, the switch does not encounter this problem. Instead, it internally maps IP multicast group addresses to the ports that contain group members.

When an IP multicast packet is received, the lookup is based on the IP group address, regardless of whether the VLAN is bridged or routed. Be aware that while the Ethernet Routing Switch 8600 does not suffer from the problem described in the previous example, other switches in the network can. This is particularly true of pure Layer 2 switches.

In a network that includes nonEthernet Routing Switch 8600 equipment, the easiest way to ensure that this issue does not arise is to use only a consecutive range of IP multicast addresses corresponding to the lower order 23 bits of that range. For example, use an address range from 239.0.0.0 through 239.127.255.255. A group address range of this size can still easily accommodate the needs of even the largest private enterprise.

Dynamic multicast configuration changes

Nortel recommends that you do not perform dynamic multicast configuration changes when multicast streams are flowing in a network. For example, do not change the routing protocol running on an interface, or the IP address, or the subnet mask for an interface until multicast traffic ceases.

For such changes, Nortel recommends that you temporarily stop all multicast traffic. If the changes are necessary and you have no control over the applications that send multicast data, it may be necessary for you to disable the multicast routing protocols before performing the change. For example, consider disabling multicast routing before making interface address changes. In all cases, these changes result in traffic interruptions because they impact neighbor state machines and stream state machines.

IGMPv2 back-down to IGMPv1

The DVMRP standard states that when a router operates in Internet Group Management Protocol version 2 mode (IGMPv2) and another router is discovered on the same subnet in IGMPv1 mode, the router must back down to IGMPv1 mode. When the Ethernet Routing Switch 8600 detects an IGMPv1-only router, it automatically downgrades from IGMPv2 to IGMPv1 mode.

Automatic back-down saves network down time and configuration effort. However, the switch cannot dynamically change back to IGMPv2 mode because multiple routers now advertise their capabilities as limited to IGMPv1 only. To return to IGMPv2 mode, the switch must first lose its neighbor relationship. Subsequently, when the switch reestablishes contact with its neighboring routers, it operates in IGMPv2 mode.

IGMPv3 backward compatibility

Beginning with Release 5.1, IGMPv3 for PIM-SSM is backward compatible with IGMPv1/v2. According to RFC 3376, the multicast router with IGMPv3 can use one of two methods to handle older query messages:

- If an older version of IGMP is present on the router, the querier must use the lowest version of IGMP present on the network.
- If a router that is not explicitly configured to use IGMPv1 or IGMPv2, hears an IGMPv1 query or IGMPv2 general query, it logs a rate-limited warning.

You can configure whether the switch downgrades the version of IGMP to handle older query messages. If the switch downgrades, the host with IGMPv3 only capability does not work. If you do not configure the switch to downgrade the version of IGMP, the switch logs a warning.

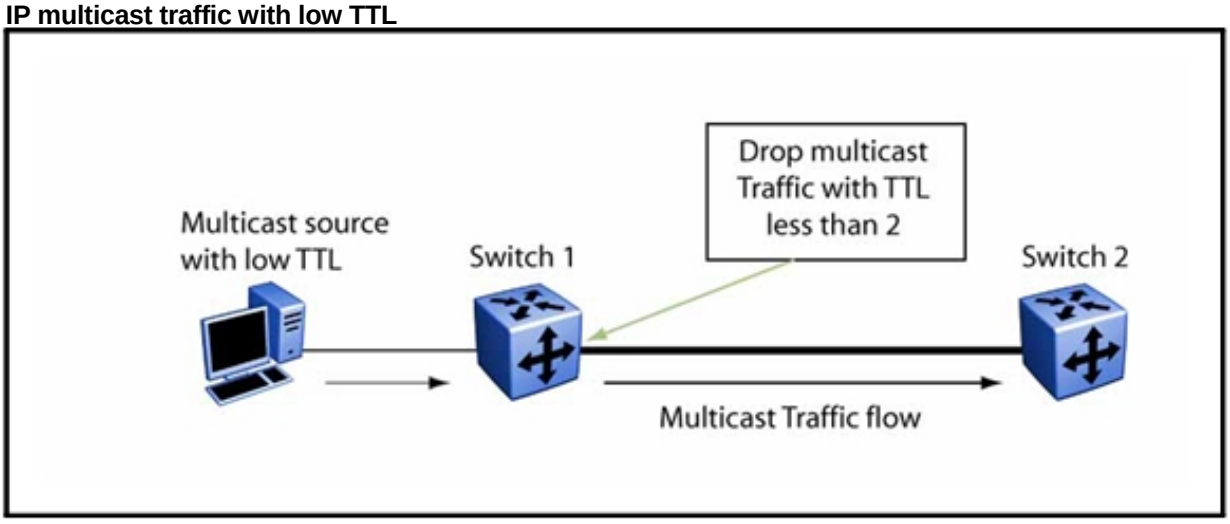
TTL in IP multicast packets

The Ethernet Routing Switch 8600 treats multicast data packets with a Time To Live (TTL) of 1 as expired packets and sends them to the CPU before dropping them. To avoid this, ensure that the originating application uses a hop count large enough to enable the multicast stream to traverse the network and reach all destinations without reaching a TTL of 1. Nortel recommends using a TTL value of 33 or 34 to minimize the effect of looping in an unstable network.

To avoid sending packets with a TTL of 1 to the CPU, the switch prunes multicast streams with a TTL of 1 if they generate a high load on the CPU. In addition, the switch prunes all multicast streams with a TTL of 1 to the same group for sources on the same originating subnet as the stream.

To ensure that a switch does not receive multicast streams with a TTL of 1, thus pruning other streams that originate from the same subnet for the same group, you can configure the upstream Ethernet Routing Switch 8600 (Switch 1) to drop multicast traffic with a TTL of less than 2 (see Figure 75 "IP multicast traffic with low TTL" (page 207)). In this configuration, all streams that egress the switch (Switch 1) with a TTL of 1 are dropped.

Figure 75
IP multicast traffic with low TTL



A change in the accepted egress TTL value does not take effect dynamically on active streams. To change the TTL, disable DVMRP and then enable it again on the interface with a TTL of greater than 2. Use this workaround for an Ethernet Routing Switch 8600 network that has a high number of multicast applications with no control on the hop count used by these applications.

In all cases, an application should not send multicast data with a TTL lower than 2. Otherwise, all of that application traffic is dropped, and the load on the switch is increased. Enhanced modules (E, M, or R series modules), which provide egress mirroring, do not experience this behavior.

Multicast MAC filtering

Certain network applications, such as the Microsoft Network Load Balancing Solution, require multiple hosts to share a multicast MAC address. Instead of flooding all ports in the VLAN with this multicast traffic, you can use the Multicast MAC Filtering feature to forward traffic to a configured subset of the ports in the VLAN. This multicast MAC address is not an IP multicast MAC address.

At a minimum, map the multicast MAC address to a set of ports within the VLAN. In addition, if traffic is routed on the local Ethernet Routing Switch 8600, you must configure an Address Resolution Protocol (ARP) entry to map the shared unicast IP address to the shared multicast MAC address. You must configure an ARP entry because the hosts can also share a virtual IP address, and packets addressed to the virtual IP address need to reach each host.

Nortel recommends that you limit the number of such configured multicast MAC addresses to a maximum of 100. This number is related to the maximum number of possible VLANs you can configure because for

every multicast MAC filter that you configure the maximum number of configurable VLANs reduces by one. Similarly, configuring large numbers of VLANs reduces the maximum number of configurable multicast MAC filters downwards from 100.

Although you can configure addresses starting with 01.00.5E, which are reserved for IP multicast address mapping, do not enable IP multicast with streams that match the configured addresses. This may result in incorrect IP multicast forwarding and incorrect multicast MAC filtering.

Guidelines for multicast access policies

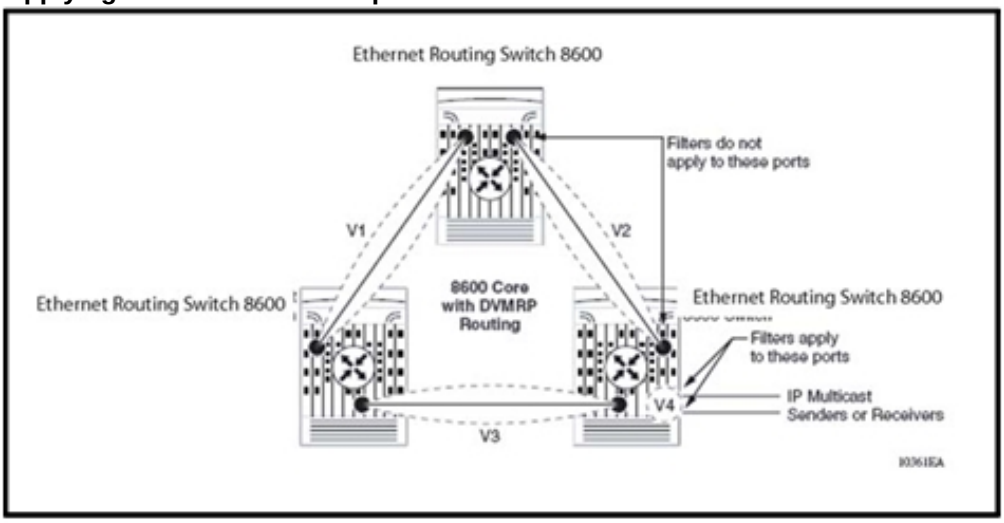
Use the following guidelines when you configure multicast access policies:

- Use masks to specify a range of hosts. For example, 10.177.10.8 with a mask of 255.255.255.248 matches hosts addresses 10.177.10.8 through 10.177.10.15. The host subnet address and the host mask must be equal to the host subnet address. An easy way to determine this is to ensure that the mask has an equal or fewer number of trailing zeros than the host subnet address. For example, 3.3.0.0/255.255.0.0 and 3.3.0.0/255.255.255.0 are valid. However, 3.3.0.0/255.0.0.0 is not.
- Receive access policies should apply to all eligible receivers on a segment. Otherwise, one host joining a group makes that multicast stream available to all.
- Receive access policies are initiated when reports are received with addresses that match the filter criteria.
- Transmit access policies are applied when the first packet of a multicast stream is received by the switch.

Multicast access policies can be applied to a DVMRP or PIM routed interface if IGMP reports the reception of multicast traffic. In the case of DVMRP routed interfaces where no IGMP reports are received, some access policies cannot be applied. The static receivers work properly on DVMRP or PIM switch-to-switch links.

With the exception of the static receivers that work in these scenarios, and the other exceptions noted at the end of this section, Figure 76 "Applying IP multicast access policies for DVMRP" (page 209) illustrates where access policies can and cannot be applied. On VLAN 4, access policies can be applied and take effect because IGMP control traffic can be monitored for these access policies. The access policies do not apply on the ports connecting switches together on V1, V2, or V3 because multicast data forwarding on these ports depends on DVMRP or PIM and does not use IGMP.

Figure 76
Applying IP multicast access policies for DVMRP



The following rules and limitations apply to IGMP access policy parameters when used with IGMP versus DVMRP and PIM:

- The static member parameter applies to IGMP snooping, DVMRP, and PIM on both interconnected links and edge ports.
- The Static Not Allowed to Join parameter applies to IGMP snooping, DVMRP, and PIM on both interconnected links and edge ports.
- For multicast access control, the denyRx parameter applies to IGMP snooping, DVMRP, and PIM. The DenyTx and DenyBoth parameters apply only to IGMP snooping.

Split-subnet and multicast

The split-subnet issue arises when a subnet is divided into two unconnected sections in a network. This results in the production of erroneous routing information about how to reach the hosts on that subnet. The split-subnet problem applies to any type of traffic. However, it has a larger impact on a PIM-SM network.

To avoid the split-subnet problem in PIM networks, ensure that the Rendezvous Point (RP) router is not located in a subnet that can become a split subnet. Also, avoid having receivers on this subnet. Because the RP is an entity that must be reached by all PIM-enabled switches with receivers in a network, placing the RP on a split-subnet can impact the whole multicast traffic flow. Traffic can be affected even for receivers and senders that are not part of the split-subnet.

Pragmatic General Multicast guidelines

Pragmatic General Multicast (PGM) is a reliable multicast transport protocol for applications that require ordered, duplicate free, multicast data delivery from multiple sources to multiple receivers. PGM guarantees that a receiver in a multicast group can receive all data from transmissions and retransmissions or can detect unrecoverable packet loss.

The Ethernet Routing Switch 8600 implements the Network Element part of PGM. Hosts running PGM implement the other PGM features. PGM operates on a session basis, so every session requires state information. Therefore, control both the number of sessions that the switch allows and the window size of these sessions. The window size controls the number of possible retransmissions for a given session and also influences the memory size in the network element that handles these sessions.

The following examples can help you design PGM-based parameters for better scalability. These examples are based on memory consumption calculations for sessions with a given window size. They assume that a maximum of 32 MBytes is used by PGM. The examples are based on session creation observations with a window_size of 5000 and a given amount of system memory. The number of bytes allocated in the system for each session is (4 bytes x [win_size*2] + overhead) where overhead is 236 bytes. The total number of sessions possible is the available memory divided by the number of bytes required for each session.

These guidelines can help you develop an estimate of the needed memory requirements. For a network with high retransmissions, be aware that memory requirements can be greater than these values indicate.

Example 1

If 32 MBytes of system memory is available for PGM, the number of sessions the switch can create is (32 MB/ 40 236) = 795 sessions. To avoid impacting other protocols running on the switch, do not allow more than 795 sessions.

Example 2

If 1.6 MB of system memory is available for PGM, the number of sessions the switch can create is (1.6 MB/ 40 236) = 40 sessions. In this case, ensure that the window size of the application is low (usually below 100). The window size is related to client and server memory and affects the switch only when retransmission errors occur.

In addition to window size, also limit the total number of PGM sessions to control the amount of memory that PGM uses. Specifically, ensure that PGM does not consume the memory required by the other protocols. The default value for the maximum number of sessions is 100.

Distance Vector Multicast Routing Protocol guidelines

Distance Vector Multicast Routing Protocol (DVMRP) is an Interior Gateway Protocol (IGP) that routes multicast packets through a network. DVMRP is based on RIP, but unlike RIP, it keeps track of return paths to the source of multicast packets. DVMRP uses the Internet Group Management Protocol (IGMP) to exchange routing packets.

For more information about DVMRP, see *Nortel Ethernet Routing Switch 8600 Configuration — IP Multicast Routing Protocols* (NN46205-501) .

DVMRP navigation

DVMRP scalability

IP multicast scaling depends on several factors. Some limitations are related to the system itself (for example, CPU and memory resources); other limitations are related to your network design.

Scaling information for DVMRP is based on test results for a large network under different failure conditions. Unit testing of such scaling numbers provides higher numbers, particularly for the number of IP multicast streams. The numbers specified in this section are recommended for general network design.

No VLAN IDs restrictions exist as to what can be configured with DVMRP. You can configure up to 500 VLANs for DVMRP. If you configure more than 300 DVMRP interfaces, you require a CPU with suitable RAM memory. You can use the 8691 SF/CPU, which has 128 MB of RAM, or the 8692 SF/CPU, which can have up to 256 MB. You can also use the CPU Memory Upgrade Kit to upgrade to 256 MB.

Software Release 4.1 and later supports up to 1200 DVMRP interfaces. Configure most interfaces as passive DVMRP interfaces and keep the number of active interfaces to under 80. If the number of DVMRP

interfaces approaches the 1200 interface limit, Nortel recommends that you configure only a few interfaces as active DVMRP interfaces (configure the rest as passive).

The number of DVMRP multicast routes can scale up to 2500 when deployed with other protocols, such as OSPF or RIP. With the proper use of DVMRP routing policies, your network can support a large number of routes. For information about using policies, see "DVMRP policies" (page 213).

The recommended maximum number of active multicast source/group pairs (S,G) is 2000.

Nortel recommends that the number of source subnets times the number of receiver groups not exceed 500. If you need more than 500 active streams, group senders into the same subnets to achieve higher scalability. Give careful consideration to traffic distribution to ensure that the load is shared efficiently between interconnected switches (for more information, see "Multicast and Multi-Link Trunking considerations" (page 198)).

ATTENTION
For DVMRP scaled configurations with more than thousand streams, to avoid multicast traffic loss, you may have to increase routing protocol timeouts (for example, dead interval for OSPF, and so on).

The scaling limits given in this section are not hard limits; they are a result of scalability testing with switches under load with other protocols running in the network. Depending on your network design, these numbers can vary.

DVMRP design guidelines

As a general rule, design your network with routed VLANs that do not span several switches. Such a design is simpler and easier to troubleshoot and, in some cases, eliminates the need for protocols such as the Spanning Tree Protocol (STP). In the case of DVMRP enabled networks, such a configuration is particularly important. When DVMRP VLANs span more than two switches, temporary multicast delayed record aging on the nondesignated forwarder may occur after receivers leave.

DVMRP uses not only the hop count metric but also the IP address to choose the reverse path forwarding (RPF) path. Thus, to ensure the utilization of the best path, assign IP addresses carefully.

As with any other distance vector routing protocol, DVMRP suffers from count-to-infinity problems when loops occur in the network. This makes the settling time for the routing table higher.

Avoid connecting senders and receivers to the subnets/VLANs that connect core switches. To connect servers that generate multicast traffic or act as multicast receivers to the core, connect them to VLANs different from the ones that connect the switches. As shown in Figure 76 "Applying IP multicast access policies for DVMRP" (page 209), V1, V2, and V3 connect the core switches, and the IP multicast senders or receivers are placed on VLAN V4, which is routed to other VLANs using DVMRP.

The Nortel Ethernet Routing Switch 8600 does not support DVMRP in SMLT full-mesh designs.

DVMRP timer tuning

You can configure several DVMRP timers. These timers control the neighbor state updates (nbr-timeout and nbr-probe-interval timer), route updates (triggered-update-interval and update-interval), route maintenance (route-expiration-timeout, route-discard-timeout, route-switch-timeout) and stream forwarding states (leaf-timeout and fwd-cache-timeout).

For faster network convergence in the case of failures or route changes, you may need to change the default values of these timers. If so, Nortel recommends that you follow these rules:

- Ensure that all timer values match on all switches in the same DVMRP network. Failure to do so may result in unpredictable network behavior and troubleshooting difficulties.
- Do not use low timer values, especially low route update timers because this can result in a high CPU load: the CPU must process frequent messages. Also, setting lower timer values, such as those for the route-switch timeout, can result in a flapping condition in cases where routes time out very quickly.
- Follow the DVMRP standard (RFC 1075) as per the relationship between correlated timers. For example, the Route Hold-down equals twice the Route Report Interval.

DVMRP policies

DVMRP policies include announce and accept, do not advertise self, and default route policies. By filtering routes that are not necessary to advertise, you can use policies to scale to very large DVMRP networks.

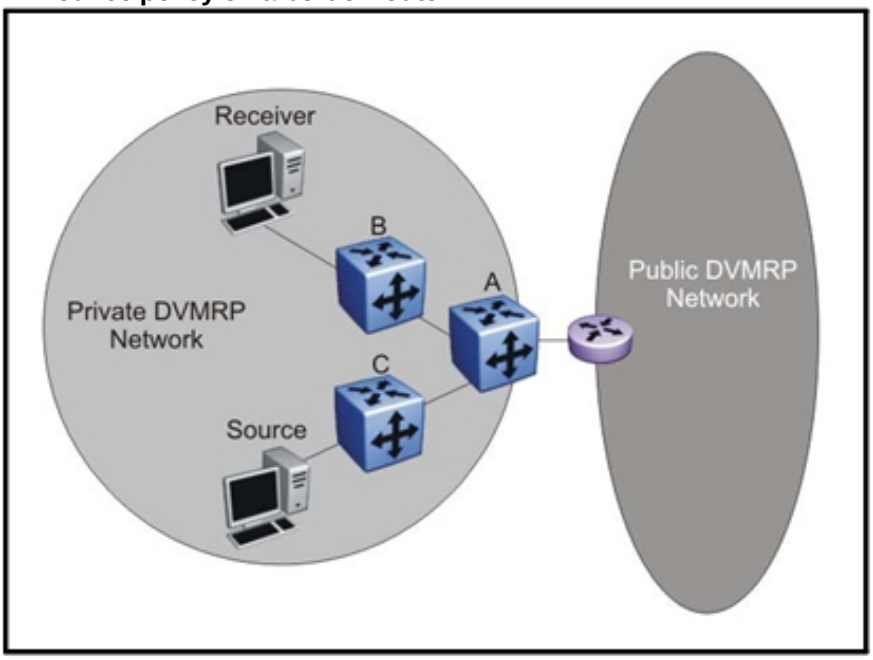
Announce and accept policy examples

By using accept or announce policies, you can filter out subnets that only have multicast receivers without impacting the ability to deliver streams to those subnets.

The following figure shows an example of a network boundary router that connects a public multicast network to a private multicast network. Both networks contain multicast sources and use DVMRP for routing. The goal is to receive and distribute public multicast streams on the private network, while not forwarding private multicast streams to the public network.

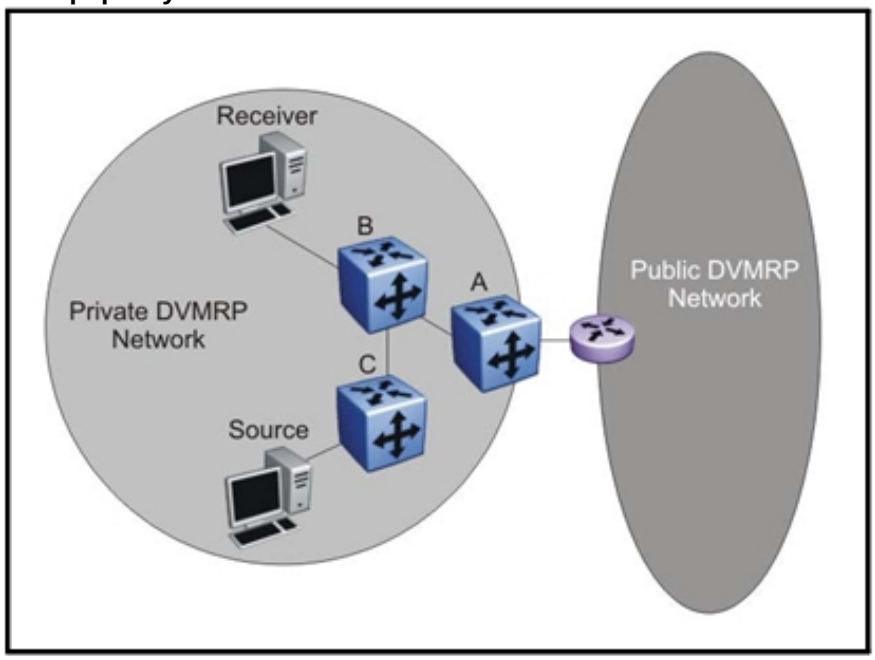
Given the topology, an appropriate solution is to use an announce policy on the public network interface of Router A. This prevents the public network from receiving the private multicast streams, while allowing Router A to still act as a transit router within the private network. Public multicast streams are forwarded to the private network as desired.

Figure 77
Announce policy on a border router



The following figure illustrates a similar scenario. As before, the goal is to receive and distribute public multicast streams on the private network, while not forwarding private multicast streams to the public network. This time, Router A has only one multicast-capable interface connected to the private network. Because one interface precludes the possibility of intradomain multicast transit traffic, private multicast streams do not need to be forwarded to Router A. In this case, it is inefficient to use an announce policy on the public interface because private streams are forwarded to Router A and then are dropped (and pruned) by Router A. In such circumstances, it is appropriate to use an accept policy on the private interface of Router A. Public multicast streams are forwarded to the private network as desired.

Figure 78
Accept policy on a border router

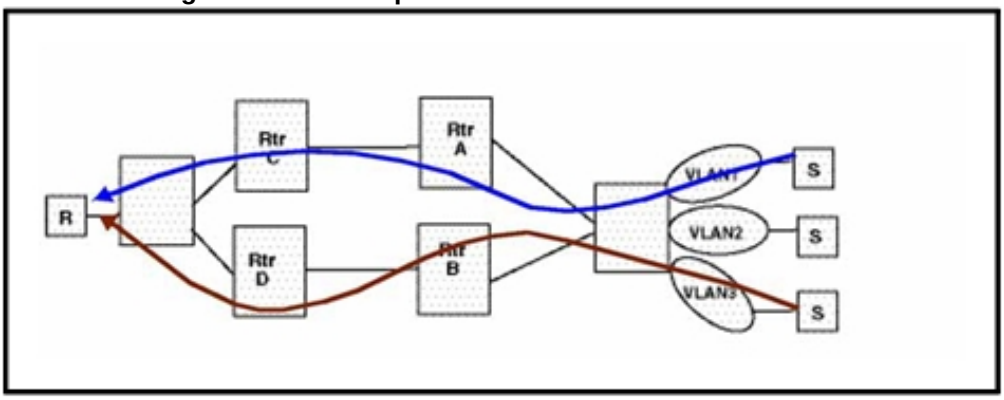


Accept policies are useful when you cannot control routing updates on the neighboring router. For example, a service provider cannot directly control the routes advertised by its neighboring router, so the provider can configure an accept policy to only accept certain agreed-on routes.

You can use an accept policy to receive a default route over an interface. If a neighbor supplies a default route, you can accept only that route and discard all others, which reduces the size of the routing table. In this situation, the default route is accepted and poison-reversed, whereas the more specific routes are filtered and not poison-reversed.

You can also use announce or accept policies (or both) to implement a form of traffic engineering for multicast streams based on the source subnet. The following figure shows a network where multiple potential paths exist through the network. According to the default settings, all multicast traffic in this network follows the same path to the receivers. Load balancing can distribute the traffic to the other available links. To make the path between Routers B and D more preferable, use announce policies on Router A to increase the advertised metric of certain routes. Thus, traffic that originates from those subnets takes the alternate route between B and D.

Figure 79
Load balancing with announce policies

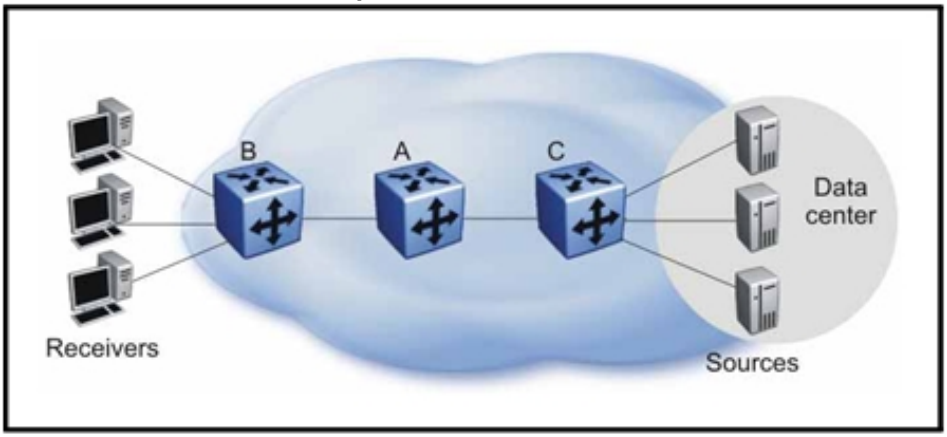


Do not advertise self policy examples

Do not advertise self policies are easier to configure than regular announce policies, while providing a commonly-used policy set. When you enable this feature, DVMRP does not advertise any local interface routes to its neighbors. However, it still advertises routes that it receives from neighbors. Because this disables the ability of networks to act as a source of multicast streams, do not enable it on any routers that are directly connected to senders.

The following figure shows a common use of this policy. Router A is a core router that has no senders on any of its connected networks. Therefore, it is unnecessary for its local routes to be visible to remote routers, so Router A is configured not to advertise any local routes. This makes it purely a transit router. Similarly, Router B is an edge router that is connected only to potential receivers. None of these hosts are allowed to be a source. Thus, configure Router B in a similar fashion to ensure it also does not advertise any local routes.

Figure 80
Do not advertise local route policies



Because all multicast streams originate from the data center, Router C must advertise at least some of its local routes. Therefore, you cannot enable the do not advertise self feature on all interfaces. If certain local routes (that do not contain sources) should not be advertised, you can selectively enable do not advertise self policies on a per-interface basis or you can configure announce policies.

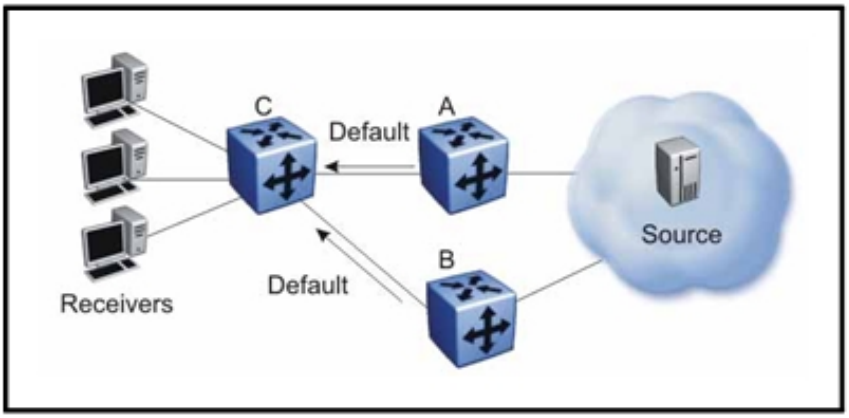
Default route policy examples

Use a default route policy to reduce the size of the multicast routing table for parts of the network that contain only receivers. You can configure an interface to supply (inject) a default route to a neighbor.

The default route does not appear in the routing table of the supplier. You can configure an interface to not listen for the default route. Once a default route is learned from a neighbor, it is placed in the routing table and potentially advertised to its other neighbors, depending on whether or not you configure the outgoing interfaces to advertise the default route. Advertising a default on an interface is different from supplying a default on an interface. The former only advertises a default if it has learned a default on another interface, whereas the latter always advertises a default. The default setting for interfaces is to listen and advertise, but not supply a default route.

The metric assigned to an injected default route is 1 by default. However, you can alter it. Changing metrics is useful in situations where two or more routers are advertising the default route to the same neighbor, but one link or path is preferable over the other. For example, in the following figure, Router A and B both advertise the default route to Router C. Because Router A is the preferred path for multicast traffic, configure it with a lower metric (a value of 1 in this case) than that of Router B, which is configured with a value of 2. Router C then chooses the lower metric and poison-reverses the route to Router A.

Figure 81
Default route



Nortel recommends that you configure announce policies on Routers A and B to suppress the advertisement of all other routes to Router C. Alternatively, you can configure accept policies on Router C to prevent all routes from Router A and Router B, other than the default, from installation in the routing table.

DVMRP passive interfaces

A DVMRP passive interface acts like an IGMP interface: no DVMRP neighbors, and hence no DVMRP routes, are learned on that interface. However, multicast sources and receivers exist on the interface.

The passive interface feature is useful if you wish to use IGMP Snoop and DVMRP on the same switch. IGMP Snoop and Layer 3 IGMP (with DVMRP and PIM) operate independently of each other. If you configure DVMRP on interface 1 and IGMP Snoop on interface 2 on Switch A, multicast data with sources from interface 1 is not forwarded to the receivers learned on interface 2 (and vice versa). To overcome this communication problem, use a DVMRP passive interface.

Configure passive interfaces only on interfaces that contain potential sources of multicast traffic. If the interfaces are connected to networks that only have receivers, Nortel recommends that you use a do not advertise self policy on those interfaces.

Do not attempt to disable a DVMRP interface if multicast receivers exist on that interface.

If you must support more than 512 potential sources on separate local interfaces, configure the vast majority as passive interfaces. Ensure that only 1 to 5 total interfaces are active DVMRP interfaces.

You can also use passive interfaces to implement a measure of security on the network. For example, if an unauthorized DVMRP router is attached to the network, a neighbor relationship is not formed, and thus, no routing information from the unauthorized router is propagated across the network. This feature also has the convenient effect of forcing multicast sources to be directly attached hosts.

Protocol Independent Multicast-Sparse Mode guidelines

Protocol Independent Multicast-Sparse Mode (PIM-SM) uses an underlying unicast routing information base to perform multicast routing. PIM-SM builds unidirectional shared trees rooted at a Rendezvous Point (RP) router per group and can also create shortest-path trees per source.

PIM-SM navigation

PIM-SM and PIM-SSM scalability

PIM-SM and PIM-SSM support VRF-lite. You can configure up to 64 instances of PIM-SM or PIM-SSM.

You can configure up to 1500 VLANs for PIM. If you configure more than 300 PIM interfaces, you require a CPU with suitable RAM memory space. You can use the 8691 SF/CPU, which has 128 MB of RAM, or the 8692 SF/CPU, which can have up to 256 MB. You can also use the CPU Memory Upgrade Kit to upgrade to 256 MB.

Interfaces that run PIM must also use a unicast routing protocol (PIM uses the unicast routing table), which puts stringent requirements on the system. As a result, 1500 interfaces may not be supported in some scenarios, especially if the number of routes and neighbors is high. With a high number of interfaces, take special care to reduce the load on the system.

Use few active IP routed interfaces. You can use IP forwarding without a routing protocol enabled on the interfaces, and enable only one or two with a routing protocol. You can configure proper routing by using IP routing policies to announce and accept routes on the switch. Use PIM passive interfaces on the majority of interfaces. Nortel recommends a maximum of ten active PIM interfaces on a switch when the number of interfaces exceeds 300. The PIM passive interface has the same uses and advantages as the DVMRP passive interface. For more details, see "DVMRP passive interfaces" (page 218).

ATTENTION
Nortel does not support more than 80 interfaces and recommends the use of not more than 10 PIM active interfaces in a large-scale configuration of more than 500 VLANs. If you configure more interfaces, they must be passive.

When using PIM-SM, the number of routes can scale up to the unicast route limit because PIM uses the unicast routing table to make forwarding decisions. For higher route scaling, Nortel recommends that you use OSPF rather than PIM.

As a general rule, a well-designed network should not have many routes in the routing table. For PIM to work properly, ensure that all subnets configured with PIM are reachable and that PIM uses the information in the unicast routing table. For the RPF check, to correctly reach the source of any multicast traffic, PIM requires the unicast routing table. For more information, see “PIM network with nonPIM interfaces” (page 235).

Nortel recommends that you limit the maximum number of active multicast (S,G) pairs to 2000. Ensure that the number of source subnets times the number of receiver groups does not exceed 500.

ATTENTION
With R/RS modules, use the show sys mgid-usage command to verify (S,G) scaling. In SMLT environments, each (S,G) entry will use two egress records, hence "show sys record-reservation" shows two records per (S,G) entry. "show sys mgid-usage" command displays one MGID per (S,G) entry.

PIM general requirements

Nortel recommends that you design simple PIM networks where VLANs do not span several switches.

PIM relies on unicast routing protocols to perform its multicast forwarding. As a result, your PIM network design should include a unicast design where the unicast routing table has a route to every source and receiver of multicast traffic, as well as a route to the Rendezvous Point (RP) router and Bootstrap router (BSR) in the network. Ensure that the path between a sender and receiver contains PIM-enabled interfaces. Receiver subnets may not always be required in the routing table.

Nortel recommends that you follow these guidelines:

- Ensure that every PIM-SM domain is configured with a RP and a BSR.
- Ensure that every group address used in multicast applications has an RP in the network.
- As a redundancy option, you can configure several RPs for the same group in a PIM domain.
- As a load sharing option, you can have several RPs in a PIM-SM domain map to different groups.
- Configure an RP to map to all IP multicast groups. Use the IP address of 224.0.0.0 and the mask of 240.0.0.0.
- Configure an RP to handle a range of multicast groups by using the mask parameter. For example, an entry for group value of 224.1.1.0 with a mask of 255.255.255.192 covers groups 224.1.1.0 to 224.1.1.63.
- In a PIM domain with both static and dynamic RP switches, you cannot configure one of the (local) interfaces for the static RP switches as the RP. For example, in the following scenario:

  (static rp switch) Sw1 ------ Sw2 (BSR/Cand-RP1) -----Sw3

  you cannot configure one of the interfaces on switch Sw1 as static RP because the BSR cannot learn this information and propagate it to Sw2 and Sw3. PIM requires that you consistently set RP on all the routers of the PIM domain, so you can only add the remote interface Candidate-RP1 (Cand-RP) to the static RP table on Sw1.
- If a switch needs to learn an RP-set, and has a unicast route to reach the BSR through this switch, Static RP cannot be enabled or configured on a switch in a mixed mode of candidate RP and static RP switches. For examples, see the following two figures.

Figure 82
Example 1

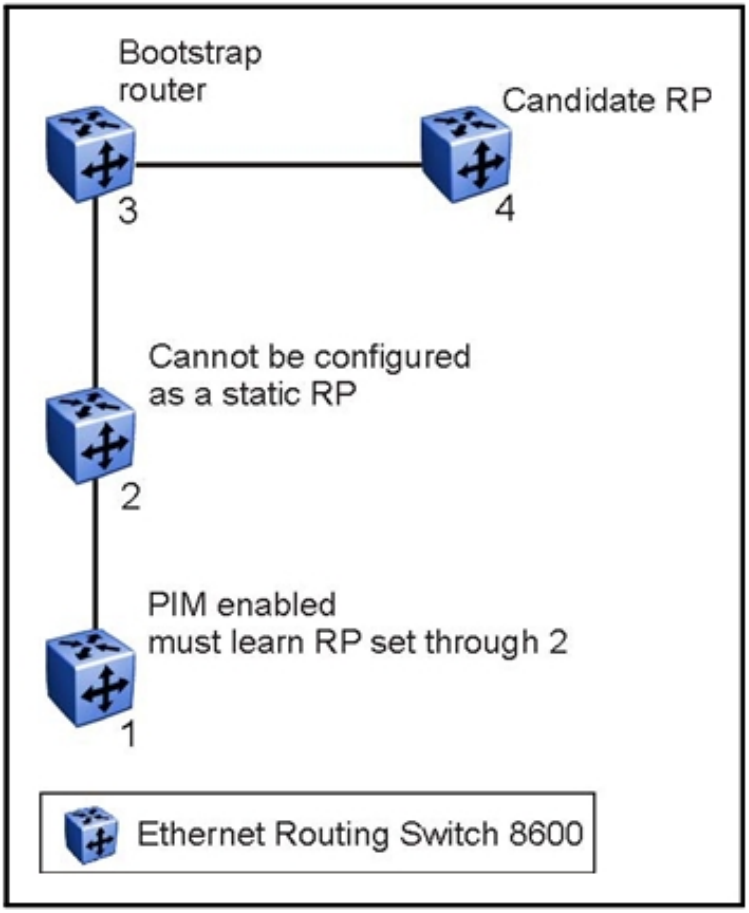
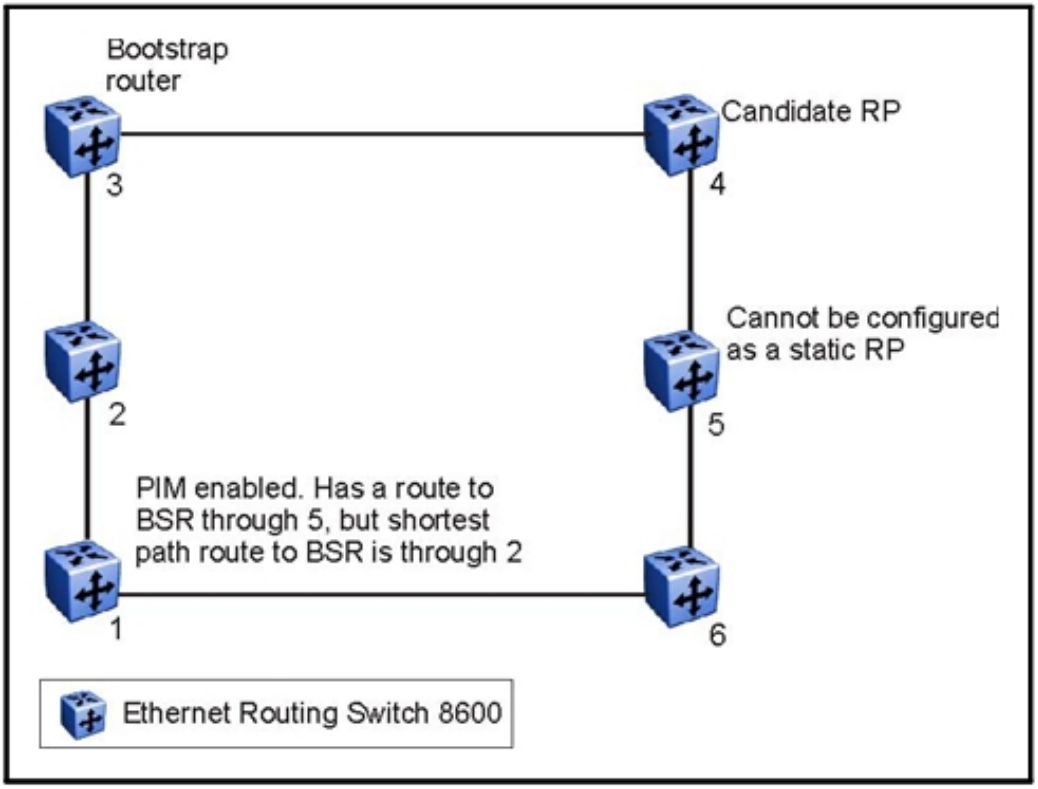


Figure 83
Example 2



PIM and Shortest Path Tree switchover

When an IGMP receiver joins a multicast group, it first joins the shared tree. Once the first packet is received on the shared tree, the router uses the source address information in the packet to immediately switch over to the shortest path tree (SPT).

To guarantee a simple, yet high-performance implementation of PIM-SM, the switch does not support a threshold bit rate in relation to SPT switchover. Intermediate routers (that is, not directly connected IGMP hosts) do not switch over to the SPT until directed to do so by the leaf routers.

Other vendors may offer a configurable threshold, such as a certain bit rate at which the SPT switch-over occurs. Regardless of their implementation, no interoperability issues with the Ethernet Routing Switch 8600 result. Switching to and from the shared and shortest path trees is independently controlled by each downstream router. Upstream routers relay Joins and Prunes upstream hop-by-hop, building the desired tree

as they go. Because any PIM-SM compatible router already supports shared and shortest path trees, no compatibility issues should arise from the implementation of configurable switchover thresholds.

PIM traffic delay and SMLT peer reboot

PIM uses a Designated Router (DR) to forward data to receivers on the DR VLAN. The DR is the router with the highest IP address on a LAN. If this router is down, the router with the next highest IP address becomes the DR.

The reboot of the DR in a Split Multilink Trunking (SMLT) VLAN may result in data loss because of the following actions:

- When the DR is down, the nonDR switch assumes the role and starts forwarding data.
- When the DR comes back up, it has priority (higher IP address) to forward data so the nonDR switch stops forwarding data.
- The DR is not ready to forward traffic due to protocol convergence and because it takes time to learn the RP set and create the forwarding path. This can result in a traffic delay of 2 to 3 minutes (because the DR learns the RP set after OSPF converges).

To avoid this traffic delay, a workaround is to configure static RP on the peer SMLT switches. This avoids the process of selecting an active RP router from the list of candidate RPs, and also of dynamically learning about RPs through the BSR mechanism. Then, when the Designated Router comes back, traffic resumes as soon as OSPF converges. This workaround reduces the traffic delay.

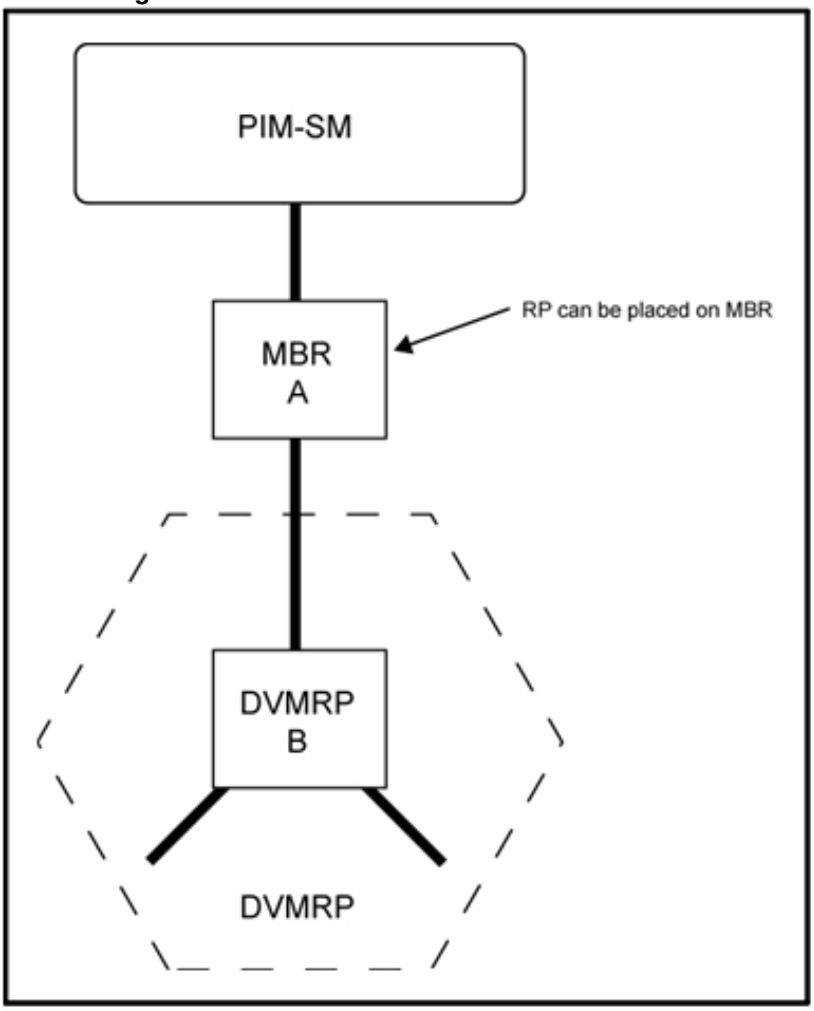
PIM-SM to DVMRP connection: MBR

ATTENTION
Software Release 5.0 does not support PIM multicast border router (MBR) functionality over SMLT.

Use the Multicast Border Router (MBR) functionality to connect a PIM-SM domain to a DVMRP domain. A switch configured as an MBR has both PIM-SM and DVMRP interfaces.

The easiest way to configure an MBR is to use one switch to connect a PIM-SM domain to a DVMRP domain, although you can use redundant switches for this purpose. You can use more than one interface on the switch to link the domains together. The following figure illustrates this basic scenario.

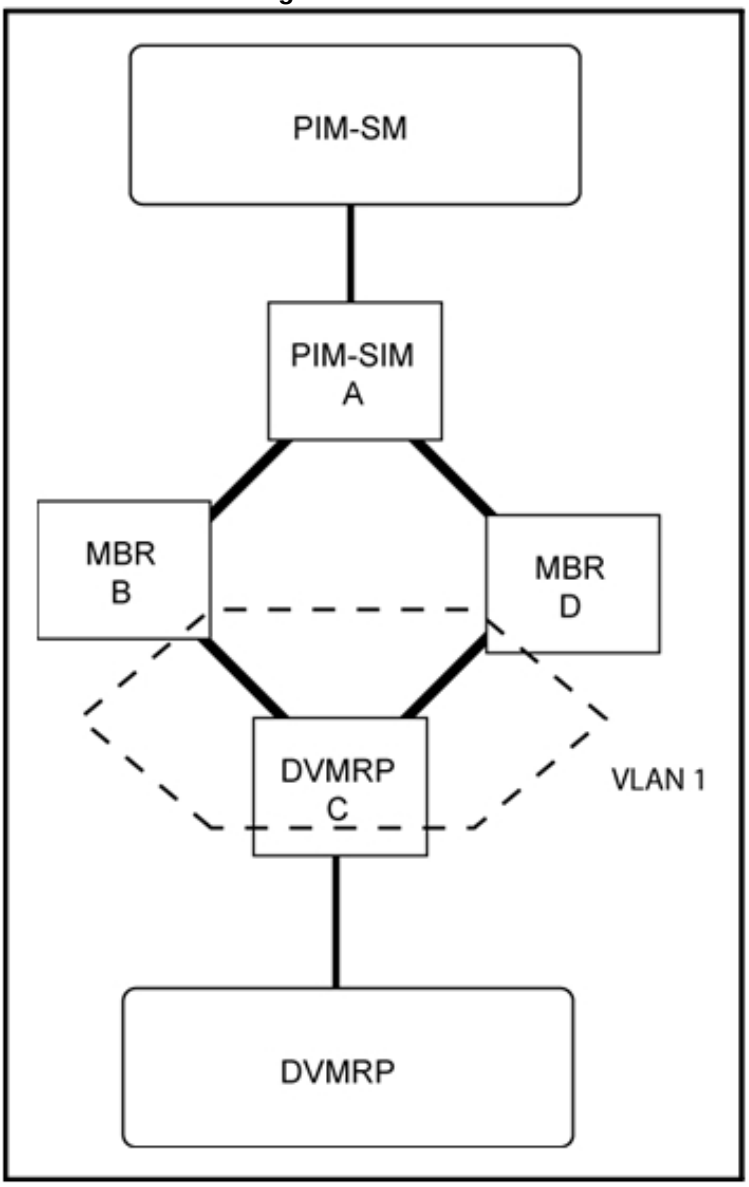
Figure 84
MBR configuration



With the Ethernet Routing Switch 8600 implementation you can place the RP anywhere in the network.

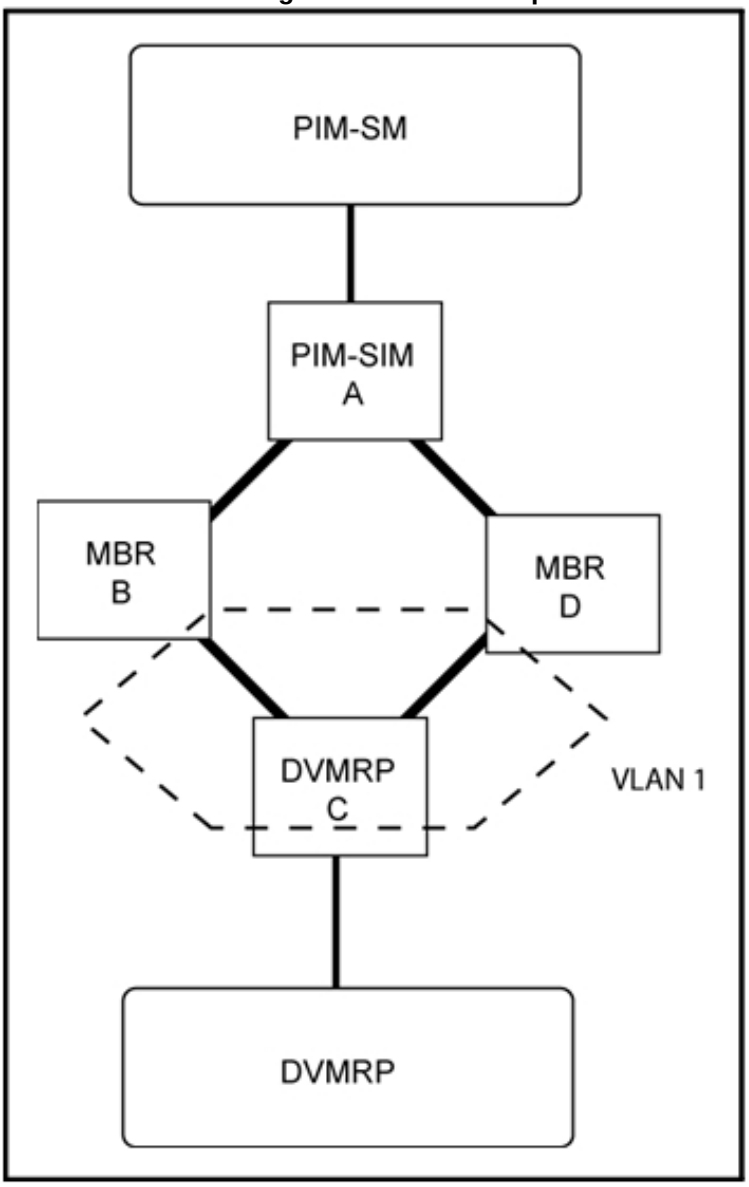
The following figure shows a redundant MBR configuration, where two MBR switches connect a PIM to a DVMRP domain. This configuration is not a supported configuration; MBRs that connect two domains should not span the same VLAN on the links connected to the same domain.

Figure 85
Redundant MBR configuration



For a proper redundant configuration, ensure that the links use two separate VLANs (see the following figure). Ensure that the unicast routes and DVMRP routes always point to the same path.

Figure 86
Redundant MBR configuration with two separate VLANs



The following paragraphs describe a failure scenario possible with this configuration.

Assume that switch A has a multicast sender, and switch C has a receiver. The RP is at D. Then, suppose that the unicast route on C allows data to reach source A through B, and that DVMRP tells upstream switch B to reach the source on A. If so, data flows from A to B to C and traffic that comes from D is discarded.

If the link between C and B fails, the unicast route on switch C indicates that the path to reach the source is through D. If DVMRP has not yet learned the new route to the source, then it cannot create an mroute for the stream when traffic is received and the stream is discarded.

Even after learning the route, DVMRP does not create an mroute for the stream. Thus, data is discarded. To resolve this issue, stop the affected streams until DVMRP ages out the entries. Another alternative is to reinitialize DVMRP (disable and reenable) and then restart the multicast streams.

If you cannot disable DVMRP or the streams, lower the DVMRP timers for faster convergence. Then DVMRP learns its routes before PIM learns the new unicast routes and reroutes the stream.

If DVMRP and unicast routes diverge while traffic flows, the same problem may occur. As a result, for safe MBR network operation, Nortel recommends that you use the simple design proposed in "PIM-SM to DVMRP connection: MBR" (page 224).

MBR and path cost considerations

When using the MBR to connect PIM-SM domains to DVMRP domains, ensure that the unicast path cost metric is not greater than 32, or issues may occur in the network. The DVMRP maximum metric value is 32. On the MBR, DVMRP obtains metric information for the PIM domain routes from unicast protocols. If DVMRP finds a route with a metric higher than 32 on the MBR, this route is considered to be unreachable. The reverse path check (RPF) check fails and data is not forwarded.

To avoid this issue, make sure that your unicast routes do not have a metric higher than 32, especially when using OSPF for routing. OSPF can have reachable routes with metrics exceeding 32.

Circuitless IP for PIM-SM

Use circuitless IP (CLIP) to configure a resilient RP and BSR for a PIM network. When you configure an RP or BSR on a regular interface, if it becomes nonoperational, the RP and BSR also become nonoperational. This results in the election of other redundant RPs and BSRs, if any, and may disrupt IP multicast traffic flow in the network. As a sound practice for multicast networks design, always configure the RP and BSR on a circuitless IP interface to prevent a single interface failure from causing these entities to fail.

Nortel also recommends that you configure redundant RPs and BSRs on different switches and that these entities be on CLIP interfaces. For the successful setup of multicast streams, ensure that a unicast route to all CLIP interfaces from all locations in the network exists. A unicast route

is mandatory because, for proper RP learning and stream setup on the shared RP tree, every switch in the network needs to reach the RP and BSR. PIM-SM circuitless IP interfaces can only be utilized for RP and BSR configurations, and are not intended for other purposes.

PIM-SM and static RP

Use static RP to provide security, interoperability, and/or redundancy for PIM-SM multicast networks. In some networks, the administrative ease derived from using dynamic RP assignment may not be worth the security risks involved. For example, if an unauthorized user connects a PIM-SM router that advertises itself as a candidate RP (CRP or cand-RP), it may possibly take over new multicast streams that would otherwise be distributed through an authorized RP. If security is important, static RP assignment may be preferable.

You can use the static RP feature in a PIM environment with devices that run legacy PIM-SMv1 and auto-RP (a proprietary protocol that the Ethernet Routing Switch 8600 does not support). For faster convergence, you can also use static RP in a PIM-SMv2 environment. If static RP is configured with PIM-SMv2, the BSR is not active.

Static RP and auto-RP

Some legacy PIM-SMv1 networks may use the auto-RP protocol. Auto-RP is a Cisco proprietary protocol that provides equivalent functionality to the standard Ethernet Routing Switch 8600 PIM-SM RP and BSR. You can use the static RP feature to interoperate in this environment. For example, in a mixed-vendor network, you can use auto-RP among routers that support the protocol, while other routers use static RP. In such a network, ensure that the static RP configuration mimics the information that is dynamically distributed to guarantee that multicast traffic is delivered to all parts of the network.

In a mixed auto-RP and static RP network, ensure that the Ethernet Routing Switch 8600 does not serve as an RP because it does not support the auto-RP protocol. In this type of network, the RP must support the auto-RP protocol.

Static RP and RP redundancy

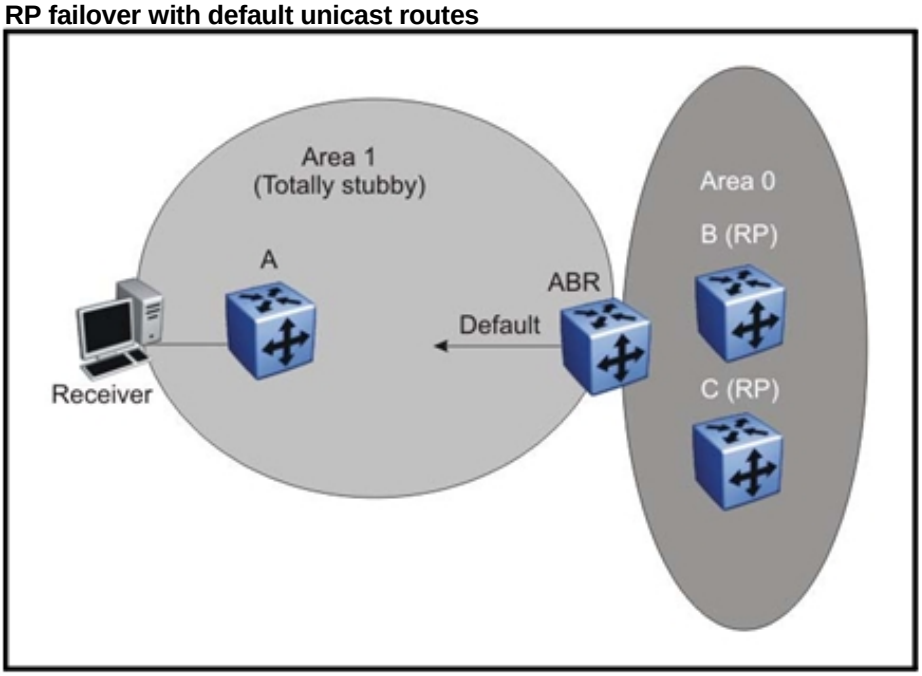
You can provide RP redundancy through static RPs. To ensure consistency of RP selection, implement the same static RP configuration on all PIM-SM routers in the network. In a mixed vendor network, ensure that the same RP selection criteria is used among all routers. For example, to select the active RP for each group address, the switch uses a hash algorithm defined in the PIM-SMv2 standard. If a router from another vendor selects the active RP based on the lowest IP address, then the inconsistency preventss the stream from being delivered to certain routers in the network.

When a group address-to-RP discrepancy occurs among PIM-SM routers, network outages occur. Routers that are unaware of the true RP cannot join the shared tree and cannot receive the multicast stream.

Failure detection of the active RP is determined by the unicast routing table. As long as the RP is considered reachable from a unicast routing perspective, the local router assumes that the RP is fully functional and attempts to join the shared tree of that RP.

The following figure shows a hierarchical OSPF network where a receiver is located in a totally stubby area. If RP B fails, PIM-SM router A does not switch over to RP C because the injected default route in the unicast routing table indicates that RP B is still reachable.

Figure 87
RP failover with default unicast routes



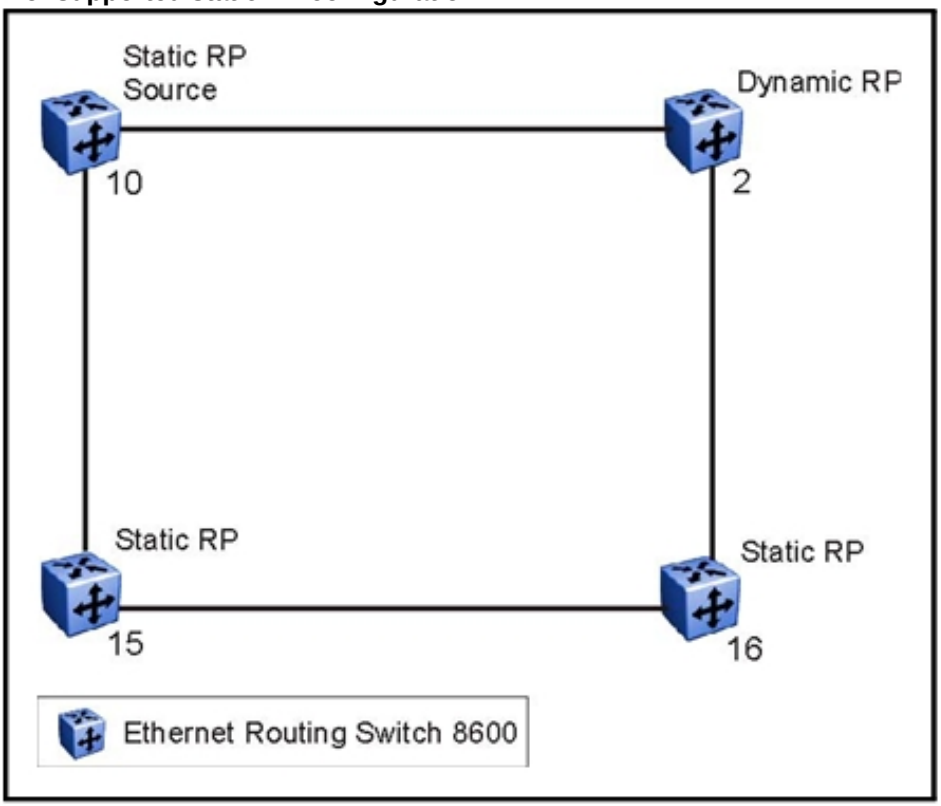
Because failover is determined by unicast routing behavior, carefully consider the unicast routing design, as well as the IP address you select for the RP. Static RP failover performance depends on the convergence time of the unicast routing protocol. For quick convergence, Nortel recommends that you use a link state protocol, such as OSPF. For example, if you are using RIP as the routing protocol, an RP failure may take minutes to detect. Depending on the application, this situation can be unacceptable.

Static RP failover time does not affect routers that have already switched over to the SPT; failover time only affects newly-joining routers.

Nonsupported static RP configurations

If you use static RP, dynamic RP learning is disabled. The following figure shows a nonsupported configuration for static RP. In this example because of interoperation between static RP and dynamic RP, no RP exists at switch 2. However, (S,G) creation and deletion occurs every 210 seconds at switch 16.

Figure 88
Nonsupported static RP configuration



Switches 10, 15, and 16 use Static RP, whereas Switch 2 uses dynamic RP. The source is at Switch 10, and the receivers are Switch 15 and 16. The RP is at Switch 15 locally. The Receiver on Switch 16 cannot receive packets because its SPT goes through Switch 2.

Switch 2 is in a dynamic RP domain, so it cannot learn about the RP on Switch 15. However, (S, G) records are created and deleted on Switch 16 every 210 seconds.

Rendezvous Point router considerations

You can place an RP on any switch when VLANs extend over several switches. Indeed, you can place your RP on any switch in the network. However, when using PIM-SM,, Nortel recommends that you not span VLANs on more than two switches.

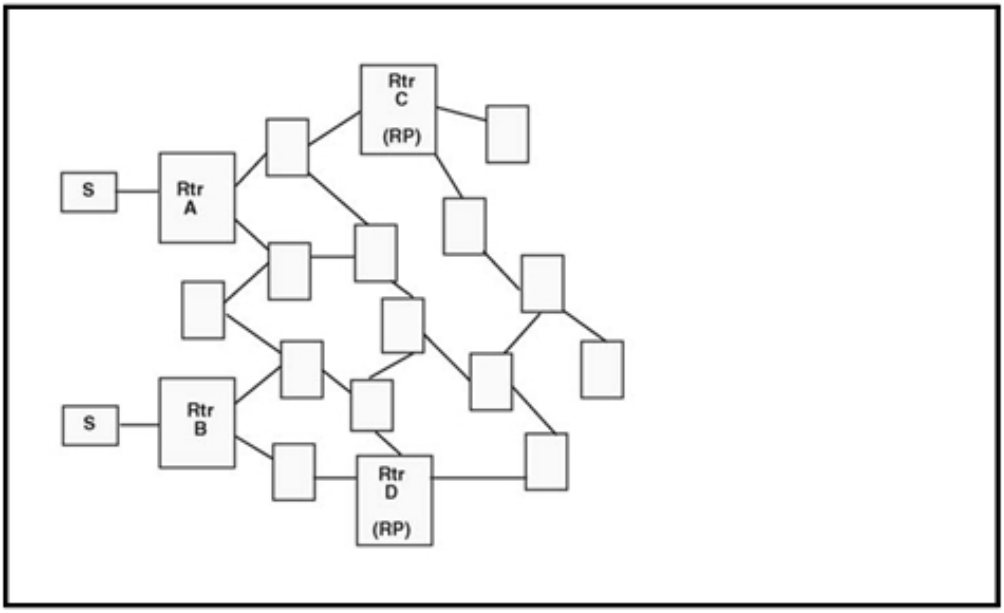
PIM-SM design and the BSR hash algorithm

To optimize the flow of traffic down the shared trees in a network that uses bootstrap router (BSR) to dynamically advertise candidate RPs, consider the hash function. The hash function used by the BSR to assign multicast group addresses to each candidate RP (CRP).

The BSR distributes the hash mask used to compute the RP assignment. For example, if two RPs are candidates for the range 239.0.0.0 through 239.0.0.127, and the hash mask is 255.255.255.252, that range of addresses is divided into groups of four consecutive addresses and assigned to one or the other candidate RP.

The following figure illustrates a suboptimal design where Router A sends traffic to a group address assigned to RP D. Router B sends traffic assigned to RP C. RP C and RP D serve as backups for each other for those group addresses. To distribute traffic, it is desirable that traffic from Router A use RP C and that traffic from Router B use RP D.

Figure 89
Example multicast network



While still providing redundancy in the case of an RP failure, you can ensure that the optimal shared tree is used by using the following methods.

1. Use the hash algorithm to proactively plan the group-address-to-RP assignment.

   Use this information to select the multicast group address for each multicast sender on the network and to ensure optimal traffic flows.

This method is helpful for modeling more complex redundancy and failure scenarios, where each group address has three or more CRPs.

2. Allow the hash algorithm to assign the blocks of addresses on the network and then view the results using the command **`show ip pim active-rp`**.

   Use the command output to assign multicast group addresses to senders that are located near the indicated RP. The limitation to this approach is that while you can easily determine the current RP for a group address, the backup RP is not shown. If more than one backup for a group address exists, the secondary RP is not obvious. In this case, use the hash algorithm to reveal which of the remaining CRPs take over for a particular group address in the event of primary RP failure.

The hash algorithm works as follows:

1. For each CRP router with matching group address ranges, a hash value is calculated according to the formula:

   Hash value [G, M, C(i)] = {1 103 515 245 * [(1 103 515245 * (G&M) +12 345) XOR C(i)] + 12 345} mod 2^31

   The hash value is a function of the group address (G), the hash mask (M), and the IP address of the CRP C(i). The expression (G&M) guarantees that blocks of group addresses hash to the same value for each CRP, and that the size of the block is determined by the hash mask.

   For example, if the hash mask is 255.255.255.248, the group addresses 239.0.0.0 through 239.0.0.7 yield the same hash value for a given CRP. Thus, the block of eight addresses are assigned to the same RP.

2. The CRP with the highest resulting hash value is chosen as the RP for the group. In the event of a tie, the CRP with the highest IP address is chosen.

   This algorithm is run independently on all PIM-SM routers so that every router has a consistent view of the group-to-RP mappings.

Candidate RP considerations

The CRP priority parameter helps to determine an active RP for a group. The hash values for different RPs are only compared for RPs with the highest priority. Among the RPs with the highest priority value and the same hash value, the CRP with the highest RP IP address is chosen as the active RP.

You cannot configure the CRP priority. Each RP has a default CRP priority value of 0, and the algorithm uses the RP if the group address maps to the grp-prefix that you configure for that RP. If a different router in the network has a CRP priority value greater than 0, the switch uses this part of the algorithm in the RP election process.

Currently, you cannot configure the hash mask used in the hash algorithm. Unless you configure a different PIM BSR in the network with a nondefault hash mask value, the default hash mask of 255.255.255.252 is used. Static RP configurations do not use the BSR hash mask; they use the default hash mask.

For example:

RP1 = 128.10.0.54 and RP2 = 128.10.0.56.
The group prefix for both RPs is 238.0.0.0/255.0.0.0.
Hash mask = 255.255.255.252.

The hash function assigns the groups to RPs in the following manner:

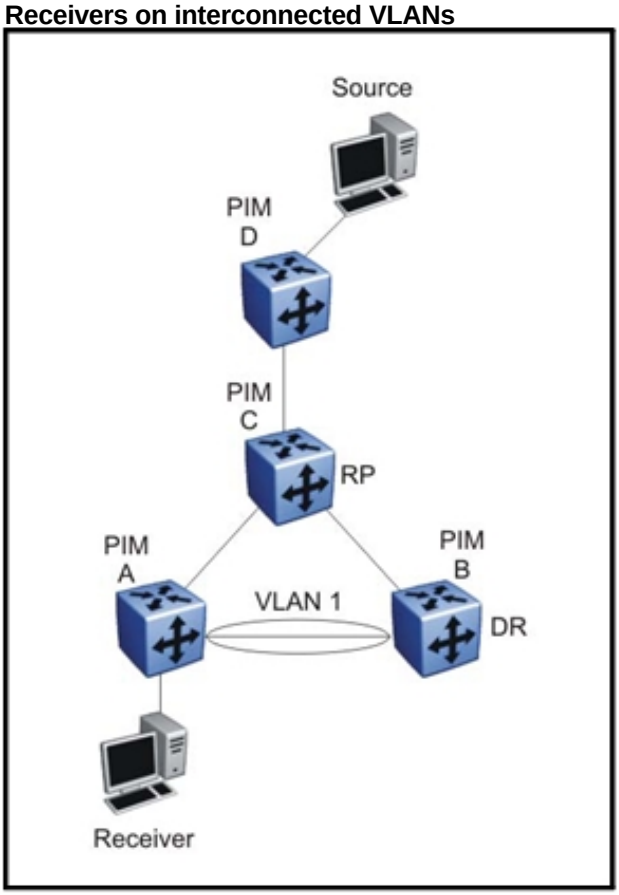
The group range 238.1.1.40 to 238.1.1.51 (12 consecutive groups) maps to 128.10.0.56.
The group range 238.1.1.52 to 238.1.1.55 (4 consecutive groups) maps to 128.10.0.54.
The group range 238.1.1.56 to 238.1.1.63 (8 consecutive groups) maps to 128.10.0.56.

PIM-SM receivers and VLANs

Some designs cause unnecessarily traffic flow on links in a PIM-SM domain. In these cases, traffic is not duplicated to the receivers, but waste bandwidth.

The following figure shows such a situation. Switch B is the Designated Router (DR) between switches A and B. Switch C is the RP. A receiver R is placed on the VLAN (V1) that interconnects switches A and B. A source sends multicast data to receiver R.

Figure 90
Receivers on interconnected VLANs



IGMP reports sent by R are forwarded to the DR, and both A and B create (*,G) records. Switch A receives duplicate data through the path from C to A, and through the second path from C to B to A. Switch A discards the data on the second path (assuming the upstream source is A to C).

To avoid this waste of resources, Nortel recommends that you do not place receivers on V1. This guarantees that no traffic flows between B and A for receivers attached to A. In this case, the existence of the receivers is only learned through PIM Join messages to the RP [for (*,G)] and of the source through SPT Joins.

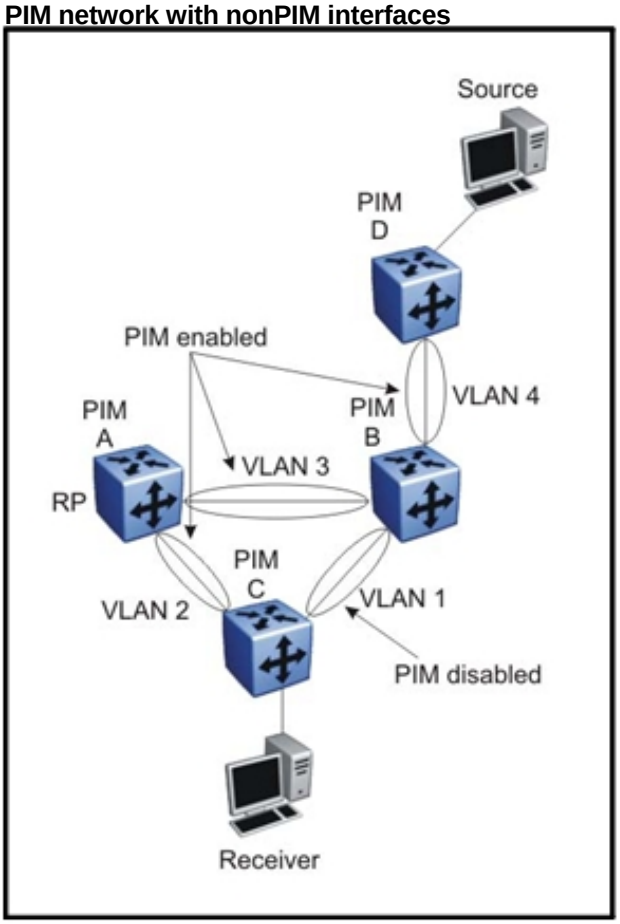
PIM network with nonPIM interfaces

For proper multicast traffic flow in a PIM-SM domain, as a general rule, enable PIM-SM on all interfaces in the network (even if paths exist between all PIM interfaces). Enable PIM on all interfaces because PIM-SM relies on the unicast routing table to determine the path to the RP, BSR, and multicast sources. Ensure that all routers on these paths have PIM-SM enabled interfaces.

Figure 91 "PIM network with nonPIM interfaces" (page 236) provides an example of this situation. If A is the RP, then initially receiver R receives data from the shared tree path (that is, through switch A).

If the shortest path from C to the source is through switch B, and the interface between C and B does not have PIM-SM enabled, then C cannot switch to the SPT. C discards data that comes through the shared path tree (that is, through A). The simple workaround is to enable PIM on VLAN1 between C and B.

Figure 91
PIM network with nonPIM interfaces



Protocol Independent Multicast-Source Specific Multicast guidelines

PIM Source Specific Multicast (SSM) is a one-to-many model that uses a subset of the PIM-SM features. In this model, members of an SSM group can only receive multicast traffic from a single source, which is more efficient and puts less load on multicast routing devices.

IGMPv3 supports PIM-SSM by enabling a host to selectively request or filter traffic from individual sources within a multicast group.

IGMPv3 and PIM-SSM operation

Release 3.5 introduces an SSM-only implementation of IGMPv3. This is not a full implementation, and it processes messages according to the following rules:

- When an IGMPv2 report is received on an IGMPv3 interface, the switch drops the IGMPv2 report. IGMPv3 is not backward compatible with IGMPv2.
- In dynamic mode, when an IGMPv3 report is received with several nonSSM sources, but matches a configured SSM range, the switch does not process the report.
- When an IGMPv2 router sends queries on an IGMPv3 interface, the switch downgrades this interface to IGMPv2 (backward compatibility).

  This may cause traffic interruption, but the switch recovers quickly.
- When an IGMPv3 report is received for a group with a different source than the one in the SSM channels table, the switch drops the report.

PIM-SSM design considerations

Considerations the following information when designing an SSM network:

- When SSM is configured, it affect SSM groups only. The switch handles other groups in sparse mode (SM).
- You can configure PIM-SSM only on switches at the edge of the network. Core switches use PIM-SM if they do not have receivers for SSM groups.
- For networks where group addresses are already in use, you can change the SSM range to match the groups.
- One switch has a single SSM range.
- You can have different SSM ranges on different switches.

  Configure the core switches that relay multicast traffic so that they cover all of these groups in their SSM range, or use PIM-SM.
- One group in the SSM range can have a single source for a given SSM group.
- You can have different sources for the same group in the SSM range (different channels) if they are on different switches.

  Two different devices in a network may want to receive data from a physically closer server for the same group. Hence, receivers listen to different channels (still same group).

For PIM-SSM scaling information, see “PIM-SM and PIM-SSM scalability” (page 219).

MSDP

Multicast Source Discovery Protocol (MSDP) allows rendezvous point (RP) routers to share source information across Protocol Independent Multicast Sparse-Mode (PIM-SM) domains. RP routers in different domains use MSDP to discover and distribute multicast sources for a group.

MSDP-enabled RP routers establish MSDP peering relationships with MSDP peers in other domains. The peering relationship occurs over a TCP connection. When a source registers with the local RP, the RP sends out Source Active (SA) messages to all of its MSDP peers. The Source Active message identifies the address of the source, the multicast group address, and the address of the RP that originates the message.

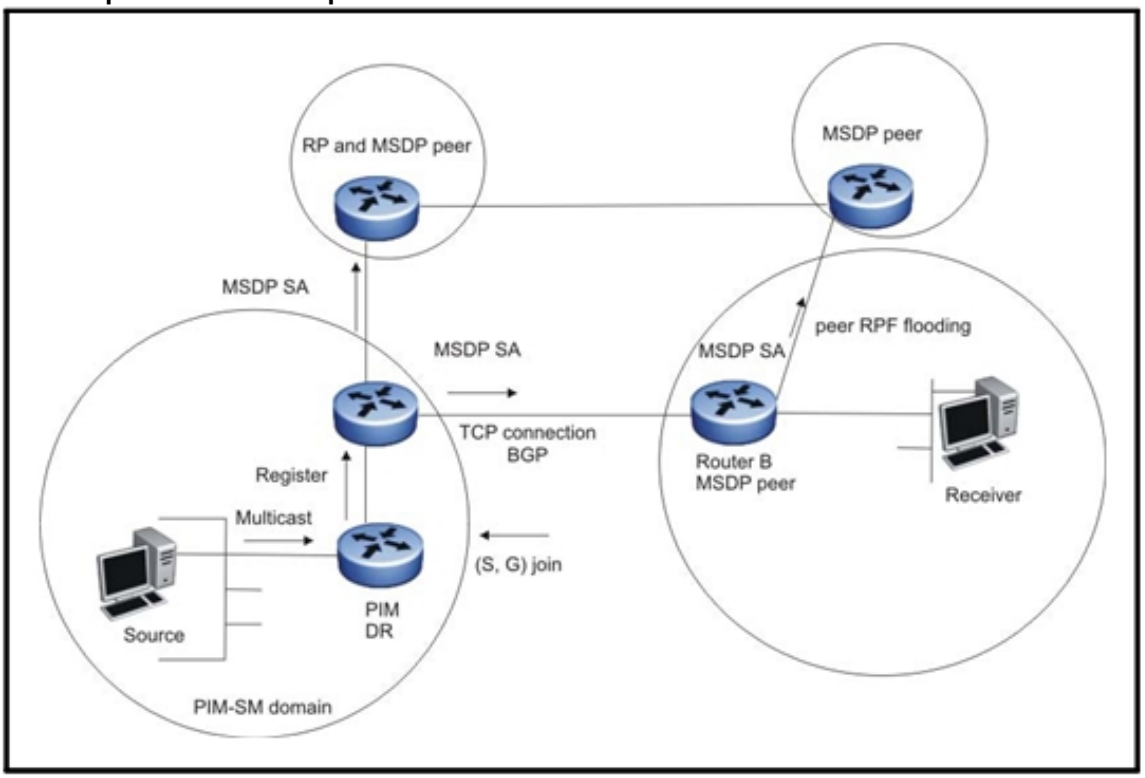
Each MSDP peer that receives the SA floods it to all MSDP peers that are downstream from the originating RP. To prevent loops, each receiving MSDP peer examines the BGP routing table to determine which peer is the next hop towards the RP that originated the SA. This peer is the Reverse Path Forwarding (RPF) peer. Each MSDP peer drops any SAs that are received on interfaces other than the one connecting to the RPF peer.

MSDP is similar to BGP and in deployments it usually follows BGP peering.

When receivers in a domain belong to a multicast group whose source is in a remote domain, the normal PIM-SM source-tree building mechanism delivers multicast data over an interdomain distribution tree. However, with MSDP, group members continue to obtain source information from their local RP. They are not directly dependent on the RPs in other domains.

The following figure shows an example MSDP network.

Figure 92
MSDP operation between peers



MSDP routers cache SA messages by default. The cache reduces join latency for new receivers and reduces storms by advertising from the cache at a period of no more than twice for the SA advertisement timer interval and not less than once for the SA advertisement period. The SA advertisement period is 60 seconds.

Peers

Configure neighboring routers as the MSDP peers of the local router to explicitly define the peer relationships. You must configure at least one peer. MSDP typically runs on the same router as the PIM-SM RP. In a peering relationship, the MSDP peer with the highest IP address listens for new TCP connections on port 639. The other side of the peer relationship makes an active connection to this port.

Default peers

Configure a default MSDP peer when the switch is not in a BGP-peering relationship with an MSDP peer. If you configure a default peer, the switch accepts all SA messages from that peer.

MSDP configuration considerations

Nortel recommends that you configure MSDP on RPs for sources that send to global groups to announce to the Internet.

You cannot configure the MSDP feature for use with the Virtual Router Forwarding (VRF) feature. You can configure MSDP for the base router only.

You can configure the RP to filter which sources it describes in SA messages. You can use Message Digest (MD) 5 authentication to secure control messages.

Nortel Ethernet Routing Switch 8600 supports the MSDP management information base (MIB) as described in RFC 4624.

Static mroute

The Ethernet Routing Switch 8600 supports a static IP route table to separate the paths for unicast and multicast streams. Only multicast protocols use this table. Adding a route to this table does not affect the switching or routing of unicast packets.

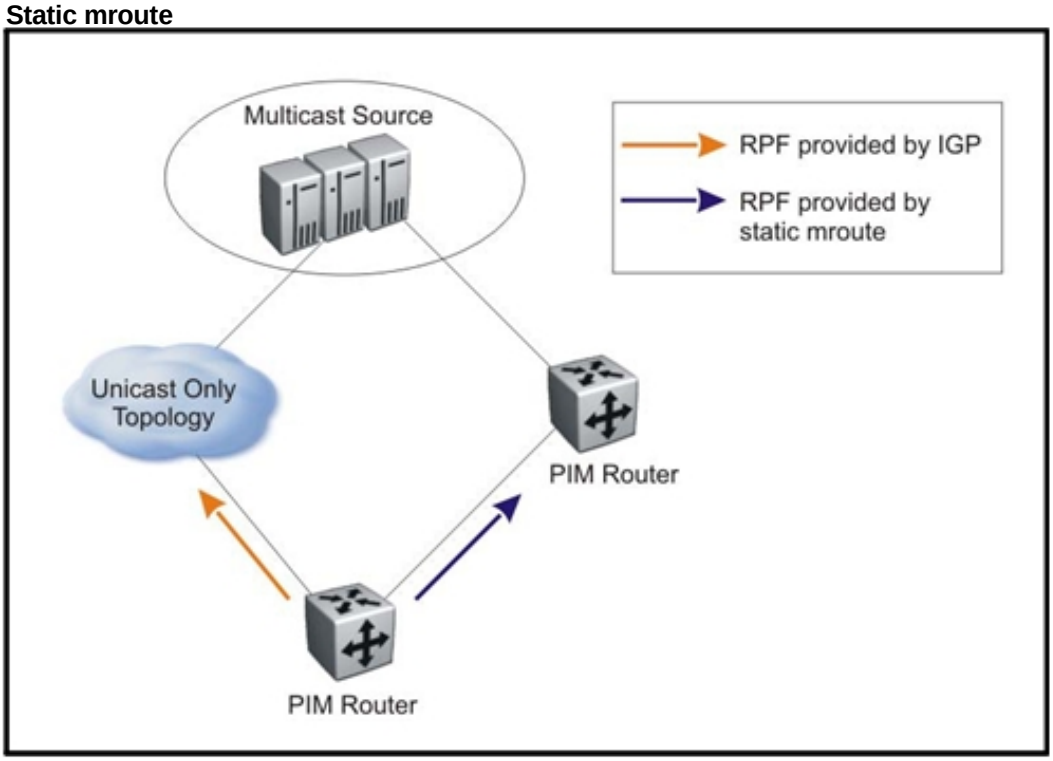
The entries in this table use the following attributes:

- IP prefix or IP mask—the destination network for the added route
- Reverse Path Forwarding (RPF) address—the IP address of the RPF neighbor towards the rendezvous point (RP) or source
- route preference—the administrative distance for the route

  If the unicast routing table and the multicast-static IP route table use different routes for the same destination network, the system compares the administrative distance with that of the protocol that contributed the route in the unicast routing table.
- route status—the status, either enabled or disabled, of the route in the table

The following figure shows an example of static mroute configured in a network.

Figure 93
Static mroute



The system does not advertise or redistribute routes from the multicast-static IP route table. The system uses these routes only for RPF calculation. The system uses the following rules to determine RPF:

- Direct or local routes for a destination take precedence over a route for the same destination in the static route table.
- If a route exists in the static route table, and no route exists in the unicast routing table for the destination, the system uses the route in the static route table.
- If a route is available in both the unicast routing table and the static route table, the system uses the route from the static route table only if the administrative distance is less than or equal to that of the unicast route entry.
- If no route exists in the static route table for the destination, the system uses the route from the unicast routing table, if available.
- The system performs a longest prefix match during a lookup in the static route table. The lookup ignores routes that are administratively disabled.
- After the system performs a lookup within the static mroute table, if multiple routes exist for a matching prefix, the system chooses the route with the least preference. If multiple routes exist with a matching prefix and the same preference, the system chooses the route with the

highest RPF address. This selection method only occurs within the static mroute table; the system still compares the selected route with a route from RTM, if one exists.

DVMRP and PIM comparison

DVMRP and PIM have some major differences in the way they operate and forward IP multicast traffic. Choose the protocol that is better adapted to your environment. If necessary, you can use a mix of the two protocols in different sections of the network and link them together with the MBR feature.

DVMRP and PIM comparison navigation

- "Flood and prune versus shared and shortest path trees" (page 242)
- "Unicast routes for PIM versus DMVRP own routes" (page 243)
- "Convergence and timers" (page 243)
- "PIM versus DVMRP shutdown" (page 243)

Flood and prune versus shared and shortest path trees

DVMRP uses flood and prune operations whereas PIM-SM uses shared and shortest-path trees. DVMRP is suitable for use in a dense environment where receivers are present in most parts of the network. PIM-SM is better suited for a sparse environment where few receivers are spread over a large area, and flooding is not efficient.

If DVMRP is used In a network where few receivers exist, much unnecessary network traffic results, especially for those branches where no receivers exist. DVMRP also adds additional state information about switches with no receivers.

In PIM-SM, all initial traffic must flow to the RP before reaching the destination switches. This makes PIM-SM vulnerable to RP failure, which is why redundant RPs are used with PIM-SM. Even with redundant RPs, the DVMRP convergence time can be faster than that of PIM, depending on where the failure occurs.

In PIM-SM, initially, traffic must flow to the RP before dat acan flow to the receivers. This action means that the RP can become a bottleneck, resulting in long stream initialization times. To reduce the probability of an RP bottleneck, the switch allows immediate switching to the SPT after the first packet is received.

Unicast routes for PIM versus DMVRP own routes

DVMRP uses its own RIPv2-based routing protocol and its own routing table. Therefore, DVMRP can build different paths for multicast traffic than for unicast traffic. PIM-SM relies on unicast routing protocols to build its routing table, so its paths are always linked to unicast paths.

In DVMRP, multicast route policies can be applied regardless of any existing unicast route policies. PIM must follow unicast routing policies, which limits flexibility in tuning PIM routes.

PIM-SM can scale to the unicast routing protocol limits (several thousand), whereas DVMRP has limited route scaling (two to three thousand) because of the nature of its RIPv2-based route exchange. This makes PIM-SM more scalable than DVMRP in large networks where the number of routes exceed the number supported by DVMRP (assuming DVMRP policies cannot be applied to reduce the number of routes).

Convergence and timers

DVMRP includes configurable timers that provide control of network convergence time in the event of failures. PIM requires unicast routing protocol convergence before it can converge, thus, it can take longer for PIM to converge.

PIM versus DVMRP shutdown

If you disable PIM on an interface, ensure that all paths to the RP, BSR, and sources for any receiver on the network have PIM enabled. PIM must be enabled because the BSR router sends an RP-set message to all PIM-enabled interfaces. In turn, this can cause a PIM-enabled switch to receive RP-set from multiple PIM neighbors towards the BSR. A PIM-enabled switch only accepts the BSR message from the RPF neighbor towards the BSR.

DVMRP does not operate with the same constraint because the existence of one path between a source and a receiver is enough to obtain the traffic for that receiver. In Figure 91 "PIM network with nonPIM interfaces" (page 236), if DVMRP replaces PIM, the path through A to the receiver is used to obtain the traffic. DVMRP uses its own routing table, and thus, is not impacted by the unicast routing table.

IGMP and routing protocol interactions

The following cases provide design tips for those situations where Layer 2 multicast is used with Layer 3 multicast protocols. The interoperation of Layer 2 and 3 multicast typically occurs when a Layer 2 edge device connects to one or several Layer 3 devices.

To prevent the switch from dropping some multicast traffic, configure the IGMP Query Interval to a value higher than five.

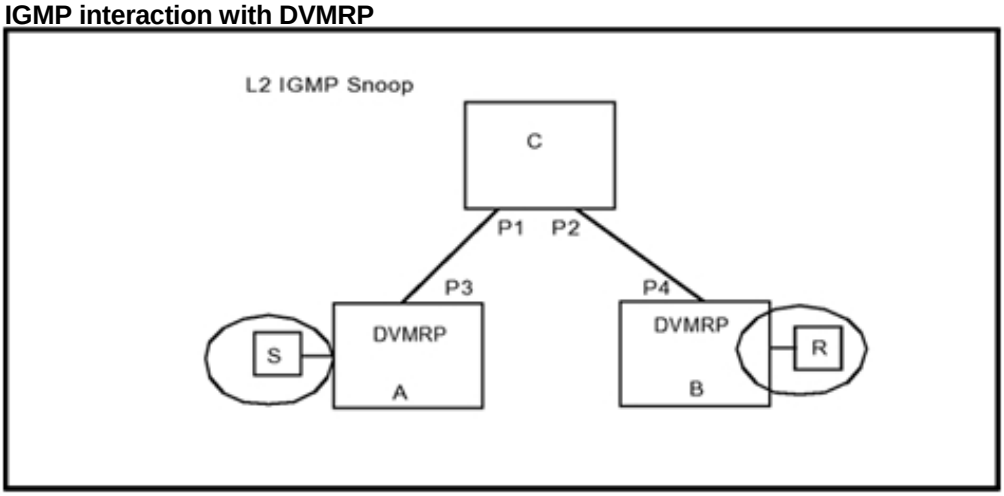
IGMP and routing protocol interactions navigation

- "IGMP and DVMRP interaction" (page 244)
- "IGMP and PIM-SM interaction" (page 245)

IGMP and DVMRP interaction

This section describes a possible problem that can arise when IGMP Snoop and DVMRP interact. In the following figure, switches A and B run DVMRP, and switch C runs IGMP Snoop. Switch C connects to A and B through ports P1 and P2 respectively. Ports P1, P2, P3, and P4 are in the same VLAN. Source S is attached to switch A on a VLAN different than the one that connects A to C. A receiver (R) is attached to switch B on another VLAN.

Figure 94
IGMP interaction with DVMRP



Switch C is not configured with any multicast ports (that is, is a nonmulticast router, or mrouter). If switch A is the querier, it becomes the mrouter (multicast router port) port for C. The receiver cannot receive data from source S because C does not forward data on the link between C and B.

You can surmount this problem by using one of the following methods:

- Configure ports P1 and P2 as mrouter ports on the IGMP Snoop VLAN.
- Configure switches A, B, and C to run Multicast Router Discovery (MRDISC) on their common VLANs.

MRDISC allows the Layer 2 switch to dynamically learn the location of switches A and B and thus, add them as mrouter ports. If you connect switches A and B together, no specific configuration is required because the issue does not arise.

IGMP and PIM-SM interaction

This section describes a possible problem that can arise when IGMP Snoop and PIM-SM interact. In this example, switches A and B run PIM-SM, and switch C runs IGMP Snoop. A and B interconnect with VLAN 1, and C connects A and B with VLAN 2.

If a receiver (R) is placed in VLAN 2 on switch C, it does not receive data. PIM chooses the router with the higher IP address as the Designated Router (DR), whereas IGMP chooses the router with the lower IP address as the querier. Thus, if B becomes the DR, A becomes the querier on VLAN 2. IGMP reports are forwarded only to A on the mrouter port P1. A does not create a leaf because reports are received on the interface towards the DR.

As in the previous IGMP interaction with DVMRP, you can surmount this problem in two different ways:

- Configure ports P1 and P2 as mrouter ports on the IGMP Snoop VLAN.
- Configure switches A, B, and C to run Multicast Router Discovery on their common VLANs.

MRDISC allows the Layer 2 switch to dynamically learn the location of switches A and B and thus, add them as mrouter ports. This issue does not occur when DVMRP uses the same switch as the querier and forwarder, for example, when IGMPv2 is used.

Multicast and SMLT guidelines

Software Release 4.1 and 5.0 does not support the MLT multicast distribution feature in SMLT configurations.

For more information about SMLT topologies, see “SMLT topologies” (page 113) or *Nortel Ethernet Routing Switch 8600 Configuration — Link Aggregation, MLT, and SMLT* (NN46205-518) .

Multicast and SMLT guidelines navigation

Triangle topology multicast guidelines

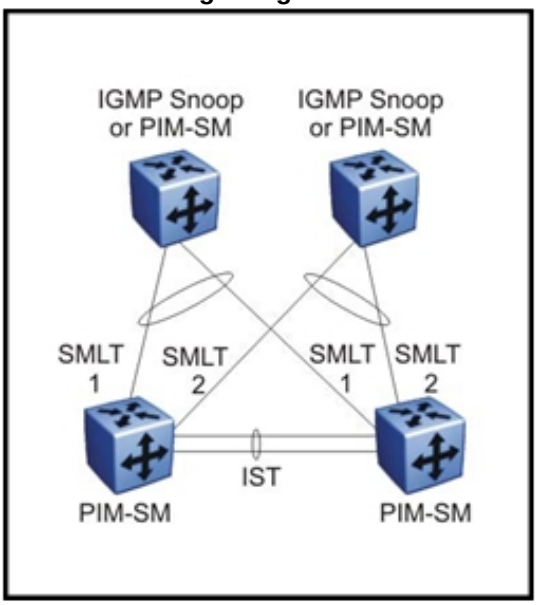
A triangle design is an SMLT configuration in which you connect edge switches or SMLT clients to two aggregation switches. Connect the aggregation switches together with an interswitch trunk that carries all the split multilink trunks configured on the switches.

The following triangle configurations are supported:

- aconfigurationwithLayer3PIM-SMroutingonboththeedgeand aggregation switches
- aconfigurationwithLayer2snoopingontheclientswitchesandLayer 3 routing with PIM-SM on the aggregation switches

To avoid using an external querier to provide correct handling and routing of multicast traffic to the rest of the network, Nortel recommends that you use the triangle design with IGMP Snoop at the client switches. Then use multicast routing (DVMRP or PIM)at the aggregation switches as shown in the following figure.

Figure 95
Multicast routing using PIM-SM



Client switches run IGMP Snoop or PIM-SM, and the aggregation switches run PIM-SM. This design is simple and, for the rest of the network, IP multicast routing is performed by means of PIM-SM. The aggregation switches are the queriers for IGMP, thus, an external querier is not required to activate IGMP membership. These switches also act as redundant switches for IP multicast.

Multicast data flows through the IST link when receivers are learned on the client switch and senders are located on the aggregation switches, or when sourced data comes through the aggregation switches. This data is destined for potential receivers attached to the other side of the IST. The data does not reach the client switches through the two aggregation switches because only the originating switch forwards the data to the client switch receivers.

Always place any multicast receivers and senders on the core switches on VLANs different from those that span the IST.

Square and full-mesh topology multicast guidelines

In a square design, you connect a pair of aggregation switches to another pair of aggregation switches. If you connect the aggregation switches in a full-mesh, it is a full-mesh design. Prior to release 4.1.1, the full-mesh design does not support SMLT and IP multicast. Releases 4.1.1 and later support Layer 3 IP multicast (PIM-SM only) over a full-mesh SMLT or Routed SMLT (RSMLT) configuration. The Nortel Ethernet Routing Switch 8600 does not support DVMRP in SMLT full-mesh designs.

In a square design, you must configure all switches with PIM-SM. Nortel recommends that you place the BSR and RP in one of the four core switches. For both full-mesh and square topologies that use multicast, you must set the multicast square-smlt flag.

SMLT and multicast traffic issues

This section describes potential traffic issues that can occur in multicast/SMLT networks.

When PIM-SM or other multicast protocols are used in an SMLT environment, the protocol should be enabled on the IST. Although, in general, routing protocols should not run over an IST, multicast routing protocols are an exception.

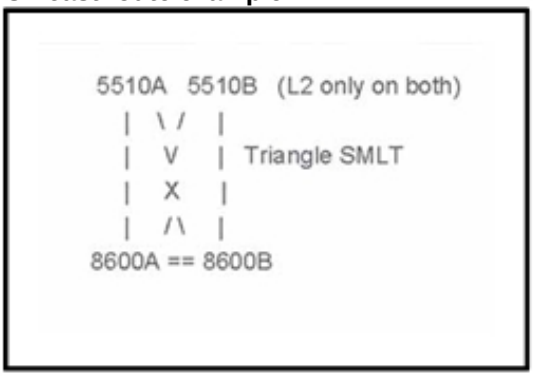
In a single PIM domain with an MBR (Multicast Border Router), Nortel does not support a configuration of DVMRP in a triangle SMLT and PIM-SM in a square SMLT.

When you use PIM and a unicast routing protocol, ensure that the unicast route to the BSR and RP has PIM active and enabled. If multiple OSPF paths exist, and PIM is not active on each path, the BSR is learned on a path that does not have PIM active. The unicast route issue can be described as follows. In the network shown in the following figure, the switches are configured with the following:

- 5510A VLAN is VLAN 101.
- 5510B VLAN is VLAN 102.

- BSR is configured on 8600B.
- Both Ethernet Routing Switch 8600s have OSPF enabled, and PIM is enabled and active on VLAN 101.
- Both Ethernet Routing Switch 8600s have OSPF enabled, and PIM is either disabled or passive on VLAN 102.

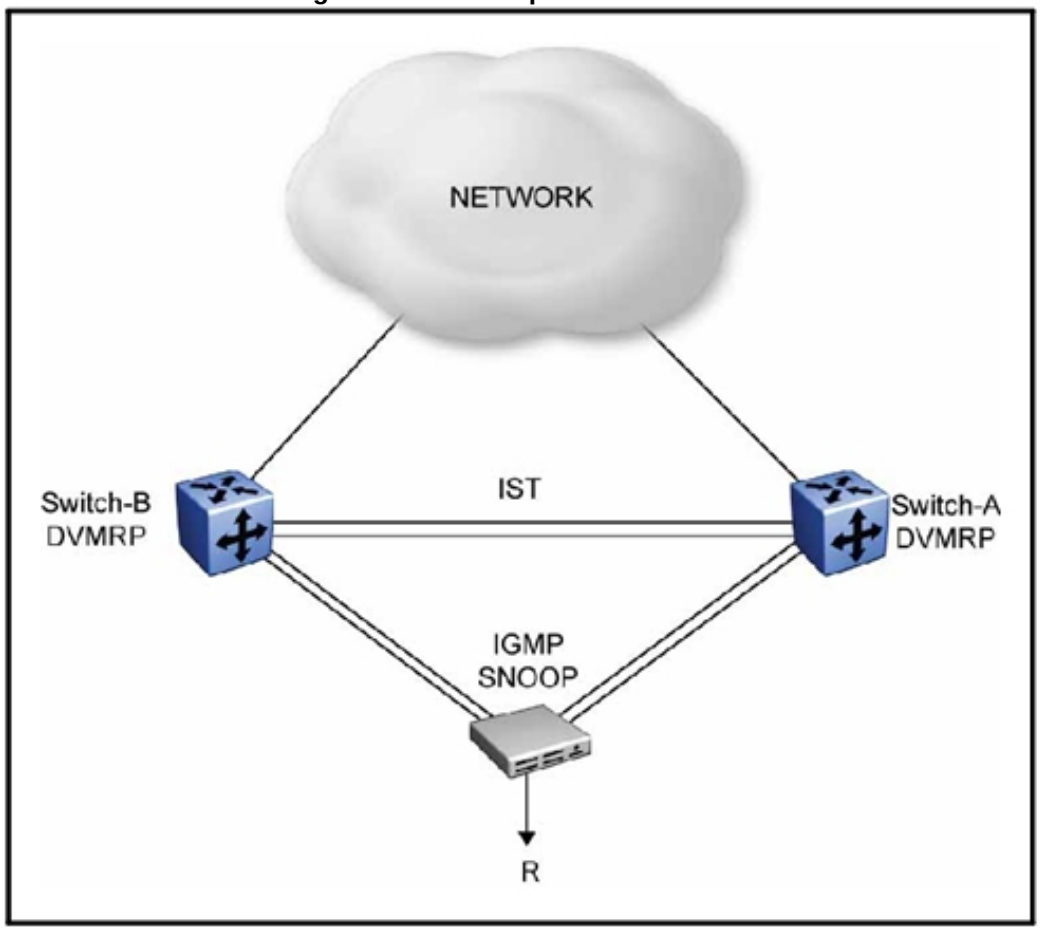
Figure 96
Unicast route example



In this example, the unicast route table on 8600A learns the BSR on 8600B through VLAN 102 via OSPF. The BSR is either not learned or does not provide the RP to 8600A.

Another traffic issue can occur when the path to a source network on the aggregation switches is the same for both switches. When the path is the same, duplicate traffic can result. The following figure illustrates the issue and the solution.

Figure 97
Multicast and SMLT design that avoids duplicate traffic



Assume that the source network is 10.10.10.0/24, switches A and B know the DVMRP metric for the IST interface, the interfaces towards NETWORK are all configured as 10, and the total cost to the source is the same.

- AhasaDVMRProute10.10.10.0withametricof10andanupstream neighbor through the interface connecting to NETWORK.
- BhasaDVMRProute10.10.10.0withametricof10andanupstream neighbor through the interface connecting to NETWORK.

A and B learn the DVMRP route for the sender (S) network with the same metric:

- Assume that A is the querier for the interface connected to the IGMP Snoop-enabled switch.
- When receiver R sends an IGMP report, A learns the receiver on the SMLT port and forwards the IST-IGMP message to B.

- After B receives the message from A, B learns the receiver on its SMLT port connected to the IGMP switch. So, both A and B have local receivers on their SMLT port.
- SsendsdatathatisreceivedbybothAandBthroughtheinterface connected to NETWORK. Because both A and B have a local receiver on the SMLT port, the IGMP switch receives data from both the routers, causing R to receive duplicate traffic.

In this configuration, both A and B forward traffic to the IGMP SNOOP switch, and the receiver receives duplicate traffic.

The solution to this issue is to configure the metrics on the DVMRP interfaces so that either A or B learns the source network route through the IST. In this way, the router that receives traffic from the IST blocks traffic from the SMLT (receiver) port so that the IGMP switch receives traffic from only one router.

Configure the metric of the DVMRP interface towards NETWORK on either A or B. For example, configure Switch B so that the route metric through the DVMRP interface is greater than the metric through the IST interface. Therefore, the NETWORK interface metric on B should be greater than 2.

If the metric of the NETWORK interface on B is configured to 3, B can learn route 10.10.10.0 through the NETWORK interface with a metric of 12 (because the metric is incremented by 2), and through the IST interface with a metric of 11. So B learns route 10.10.10.0 with a cost of 11 to the upstream neighbor through the IST link.

With these metrics, traffic from S goes from A to B only on the IST link. Because traffic received on the IST cannot go to the SMLT link, the IGMP switch does not receive traffic from B. Therefore, R no longer receives duplicate traffic; it receives traffic from switch A only.

Multicast for multimedia

The Ethernet Routing Switch 8600 provides a flexible and scalable multicast implementation for multimedia applications. Several features are dedicated to multimedia applications and in particular, to television distribution.

Multicast for multimedia navigation

- “Static routes” (page 251)
- “Join and leave performance” (page 251)
- “Fast Leave” (page 251)
- “Last Member Query Interval tuning” (page 252)

Static routes

You can configure DVMRP static mroutes. This feature is useful in cases where streams must flow continuously and not become aged. Be careful in using this feature—ensure that the programmed entries do not remain on a switch when they are no longer necessary.

You can also use IGMP static receivers for PIM static (S,G)s. The main difference between static mroutes and static (S,G) pairs is that static mroute entries only require the group address. You can use static receivers in edge configurations or on interconnected links between switches.

Join and leave performance

For TV applications, you can attach several TV sets directly, or through Business Policy Switch 2000, to the Ethernet Routing Switch 8600. Base this implementation on IGMP; the set-top boxes use IGMP reports to join a TV channel and IGMP Leaves to exit the channel. When a viewer changes channels, an IGMPv2 Leave for the old channel (multicast group) is issued, and a membership report for the new channel is sent. If viewers change channels continuously, the number of joins and leaves can become large, particularly when many viewers are attached to the switch.

The Ethernet Routing Switch 8600 supports more than a thousand Joins/Leaves per second, which is well adapted to TV applications.

ATTENTION
For IGMPv3, Nortel recommends that you ensure a Join rate of 250 per second or less. If the Ethernet Routing Switch 8600 must process more than 250 Joins per second, users may have to resend Joins.

When you use the IGMP proxy functionality in the Business Policy Switch 2000, you reduce the number of IGMP reports received by the Ethernet Routing Switch 8600. This provides better overall performance and scalability.

Fast Leave

IGMP Fast Leave supports two modes of operation: Single User Mode and Multiple User Mode.

In Single User Mode, if more than one member of a group is on the port and one of the group members leaves the group, everyone stops receiving traffic for this group. A Group-Specific-Query is not sent before the effective leave takes place.

Multiple User Mode allows several users on the same port/VLAN. If one user leaves the group and other receivers exist for the same stream, the stream continues. The switch achieves this by tracking the number

of receivers that join a given group. For Multiple User Mode to operate properly, do not suppress reports. This ensures that the switch properly tracks the correct number of receivers on an interface.

The Fast Leave feature is particularly useful in IGMP-based TV distribution where only one receiver of a TV channel is connected to a port. In the event that a viewer changes channels quickly, considerable bandwidth savings are obtained if Fast Leave is used.

You can implement Fast Leave on a VLAN and port combination; a port that belongs to two different VLANs can have Fast Leave enabled on one VLAN (but not on the other). Thus, with the Fast Leave feature enabled, you can connect several devices on different VLANs to the same port.
This strategy does not impact the traffic when one device leaves a group to which another device is subscribed. For example, you can use this feature when two TVs are connected to a port through two set-top boxes, even if you use the Single User Mode.

Last Member Query Interval tuning

When an IGMPv2 host leaves a group, it notifies the router by using a Leave message. Because of the IGMPv2 report suppression mechanism, the router is unaware of other hosts that require the stream. Thus, the router broadcasts a group-specific query message with a maximum response time equal to the Last Member Query Interval (LMQI).

Because this timer affects the latency between the time that the last member leaves and when the stream actually stops, you must properly tune this parameter. This timer can especially affect TV delivery or other large-scale, high-bandwidth multimedia applications. For instance, if you assign a value that is too low, this can lead to a storm of membership reports if a large number of hosts are subscribed. Similarly, assigning a value that is too high can cause unwanted high-bandwidth stream propagation across the network if users change channels rapidly. Leave latency is also dependent on the robustness value, so a value of two equates to a leave latency of twice the LMQI.

Determine the proper LMQI setting for your particular network through testing. If a very large number of users are connected to a port, assigning a value of three may lead to a storm of report messages when a group-specific query is sent. Conversely, if streams frequently start and stop in short intervals, as in a TV delivery network, assigning a value of ten may lead to frequent congestion in the core network.

Another performance-affecting factor that you need to be aware of is the error rate of the physical medium. It also affects the proper choice of LMQI values. For links that have high packet loss, you may find it necessary to adjust the robustness variable to a higher value to compensate for the possible loss of IGMP queries and reports.

In such cases, leave latency is adversely impacted as numerous group-specific queries are unanswered before the stream is pruned. The number of unanswered queries is equal to the robustness variable (default two). The assignment of a lower LMQI may counterbalance this effect. However, if you set it too low it may actually exacerbate the problem by inducing storms of reports on the network. Keep in mind that LMQI values of three and ten, with a robustness value of two, translate to leave latencies of six tenths of a second and two seconds, respectively.

When you choose a LMQI, consider all of these factors to determine the best setting for the given application and network. Test that value to ensure that it provides the best performance.

ATTENTION
In networks that have only one user connected to each port, Nortel recommends that you use the Fast Leave feature instead of LMQI, since no wait is required before the stream stops. Similarly, the robustness variable does not impact the Fast Leave feature, which is an additional benefit for links with high loss.

Internet Group Membership Authentication Protocol

Internet Group Membership Authentication Protocol (IGAP) is a multicast authentication and accounting protocol. With IGAP authentication and accounting features, service providers and enterprises can manage and control multicast groups on their networks.

IGAP is an IETF Internet draft that extends the functionality of the Internet Group Management Protocol (IGMPv2) and uses a standard authentication server with IGAP extensions.

The Ethernet Routing Switch 8600 processes messages according to the following rules:

- On IGAP-enabled interfaces, the switch processes IGAP messages and ignores all other IGMP messages.
- On IGMP-enabled interfaces, the switch processes IGMP messages and ignores IGAP messages.
- IGAP operates with Fast Leave only and does not generate Group-Specific-Queries as IGMPv2 does. The Ethernet Routing Switch 8600 supports the Single User and Multiple User Fast Leave modes for IGAP.

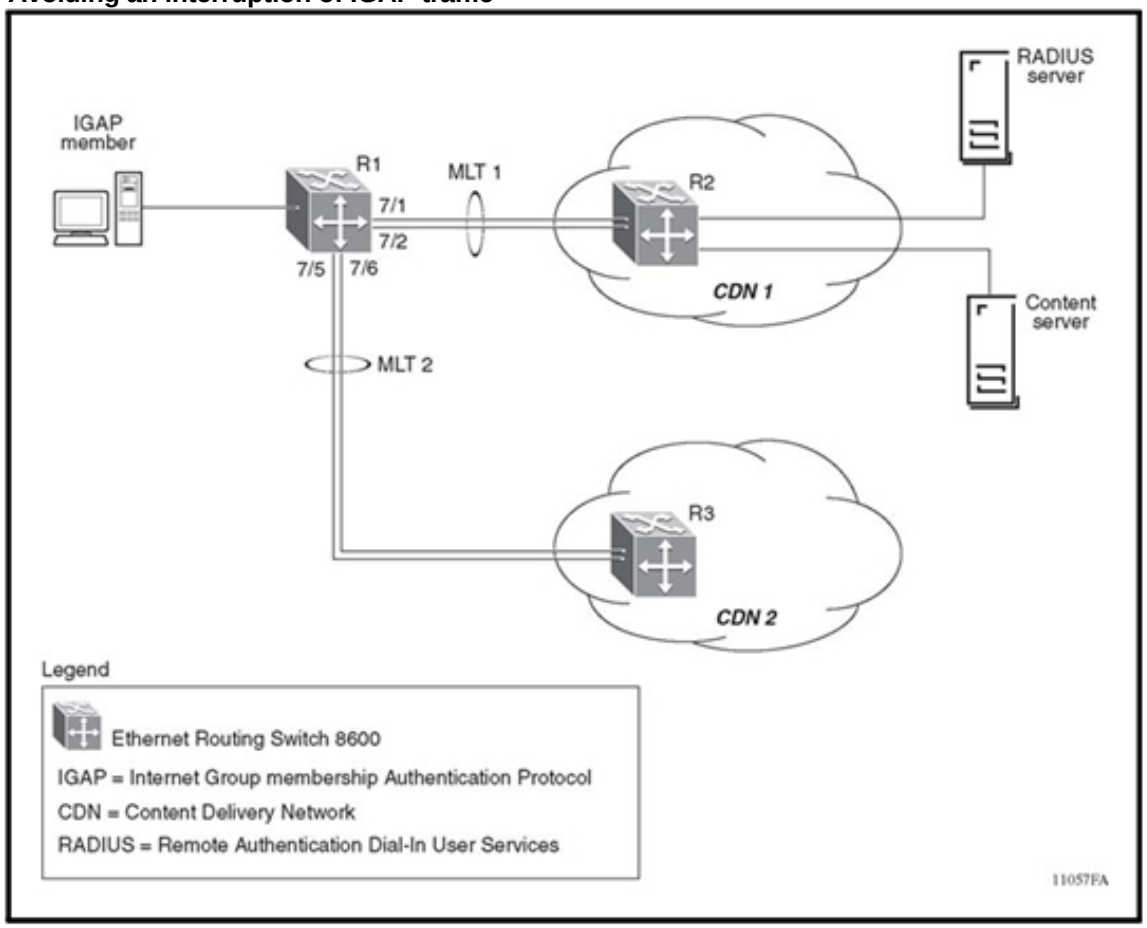
For more information about IGAP, see *Nortel Ethernet Routing Switch Configuration — IGAP* (NN46205-512) .

IGAP and MLT

In an IGAP/MLT environment, if an MLT link goes down, it can potentially interrupt IGAP traffic.

The following figure shows an IGAP member connected to an Ethernet Routing Switch 8600 edge switch (R1) that has two MLT links. The MLT links provide alternative routes to the RADIUS authentication server and the Content Delivery Network (CDN) server.

Figure 98
Avoiding an interruption of IGAP traffic



The following scenario shows how a potential traffic interruption can occur:

1. An authenticated IGAP member receives multicast traffic. Accounting starts.
2. R1 uses MLT1 to transfer data and accounting messages.
3. MLT1 goes down.

Because the (S,G) entry is deleted, an Accounting Stop message is triggered.

4. MLT2 redistributes the traffic that exists on MLT1.

   Because a new (S,G) entry is created with a different session ID, an Accounting Start message is triggered.

MLT1 is down, so both the Accounting Stop and Accounting Start messages are sent to the RADIUS server on MLT2. If the Accounting Stop message is sent before OSPF can recalculate the route change and send an Accounting Start message, the switch drops the User Datagram Protocol (UDP) packets.

This scenario does not cause an accounting error because RADIUS uses the session ID to calculate accounting time. Even though the route loss and OSPF recalculation caused the packets to be sent out-of-sequence, IGAP and RADIUS process the events in the correct order.

To avoid traffic loss if an MLT link must be disabled, use the following workaround:

- Enable Equal Cost Multicast Protocol (ECMP) on the edge switch (R1) and on both of the CDN switches (R2 and R3).
- Set the route preference (path cost) of the alternative link (MLT2) to equal or higher than MLT1.

With this workaround, the switchover is immediate. Traffic is not interrupted and accounting does not have to be stopped and restarted.

MPLS IP VPN and IP VPN Lite

The Ethernet Routing Switch 8600 supports Multiprotocol Label Switching (MPLS) and IP Virtual Private Networks (VPN) to provide fast and efficient data communications. In addition, to support IP VPN capabilities without the complexities associated with MPLS deployments, the Ethernet Routing Switch 8600 supports IP VPN Lite.

Use the design considerations provided in this section to help you design optimum MPLS IP VPN, and IP VPN Lite networks.

Navigation

MPLS IP VPN

Beginning with Release 5.0, the Ethernet Routing Switch supports MPLS networking based on RFC 4364 (RFC 4364 obsoletes RFC 2547). RFC 4364 describes a method by which a Service Provider can use an IP backbone to provide IP Virtual Private Networks (VPNs) for its customers. This method uses a peer model, in which the customer's edge routers (CE routers) send their routes to the service provider's edge routers (PE routers). Data packets are tunneled through the backbone, so that the core routers (P routers) do not need to know the VPN routes. This means that the P routers can scale to an unlimited number of IP VPNs and also that no configuration change is required on the P nodes when IP VPN services are added or removed. VPN routes are exchanged between PE routers using Border Gateway Protocol (BGP) with Multiprotocol extensions (BGP-MP).

There is no requirement for the CE routers at different sites to peer with each other or to have knowledge of IP Virtual Private Networks (VPNs) across the service provider's backbone. The CE device can also be a Layer 2 switch connected to the PE router.

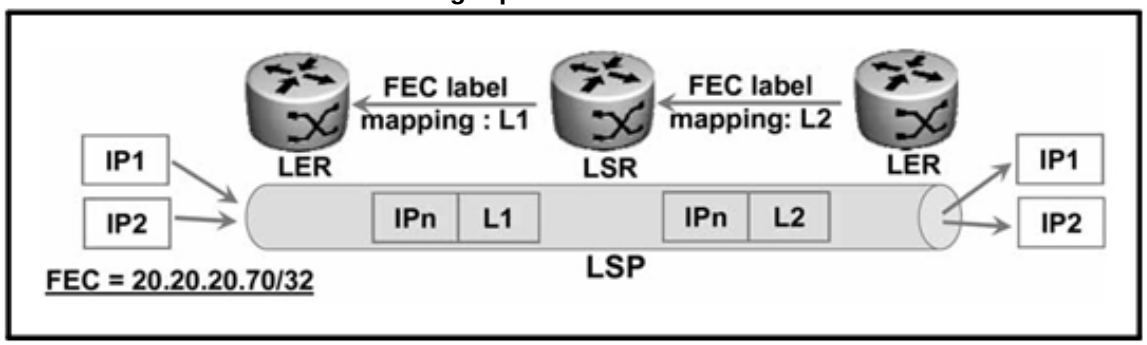
RFC 4364 defines a framework for layer 3 VPNs over an IP backbone with BGP. It is commonly deployed over MPLS but can use IPSec or GRE tunnels.

Nortel IP-VPN uses MPLS for transport.

MPLS overview

Multi-Protocol Label Switching (MPLS) [RFC3031] is primarily a service provider technology where IP traffic can be encapsulated with a label stack and then label switched across a network via Label Switched Routers (LSR) using Label Switched Paths (LSP). An LSP is an end-to-end unidirectional tunnel set up between MPLS-enabled routers. Data travels through the MPLS network over LSPs from the network ingress to the network egress. The LSP is determined by a sequence of labels, initiated at the ingress node. Packets that require the same treatment for transport through the network are grouped into a forwarding equivalence class (FEC).

Figure 99
Label Switched Path and Forwarding Equivalent Class

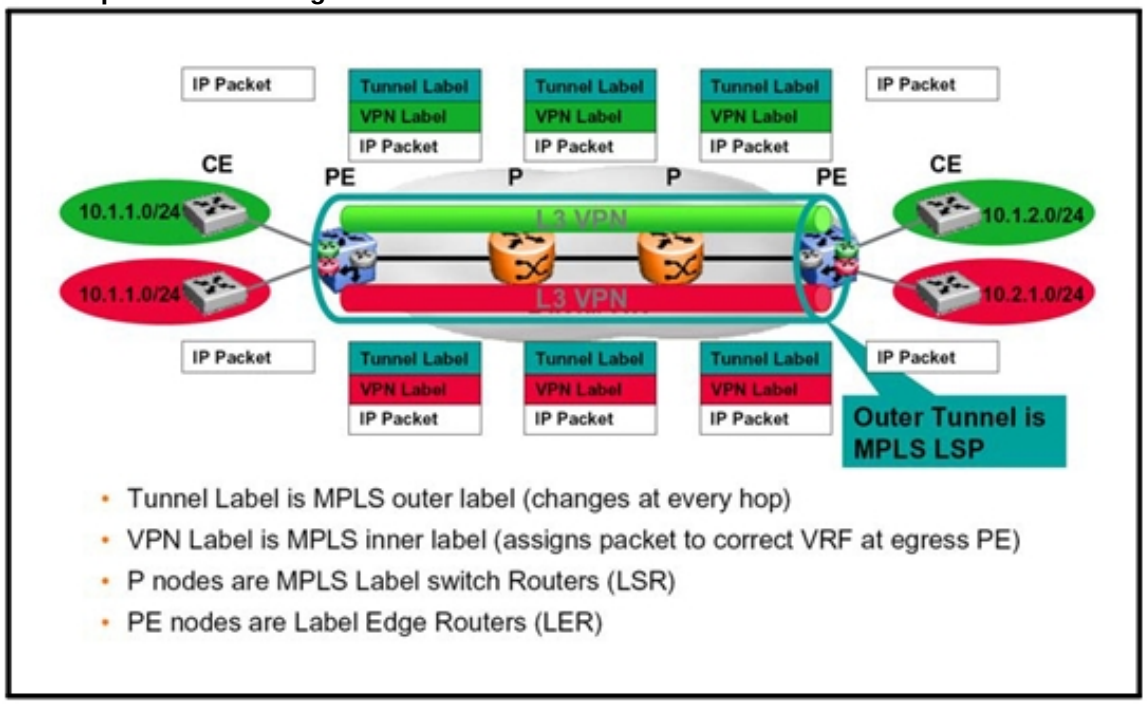


The FECs are identified by the destination subnet of the packets to be forwarded. All packets in the same FEC use the same LSP to travel across the network. Packets are classified once, as they enter the network; all subsequent forwarding decisions are based on the FEC to which each packet belongs (that is, each label corresponds to a FEC).

Operation of MPLS IP VPN

MPLS IP-VPN enabled routers use two labels as shown in the following figure. The Ethernet Routing Switch 8600 uses LDP for IP VPN. LDP generates and distributes an outer label referred as a tunnel label, which is in fact the LSP. BGP-MP generates and distributes the inner label referred to as the VPN label.

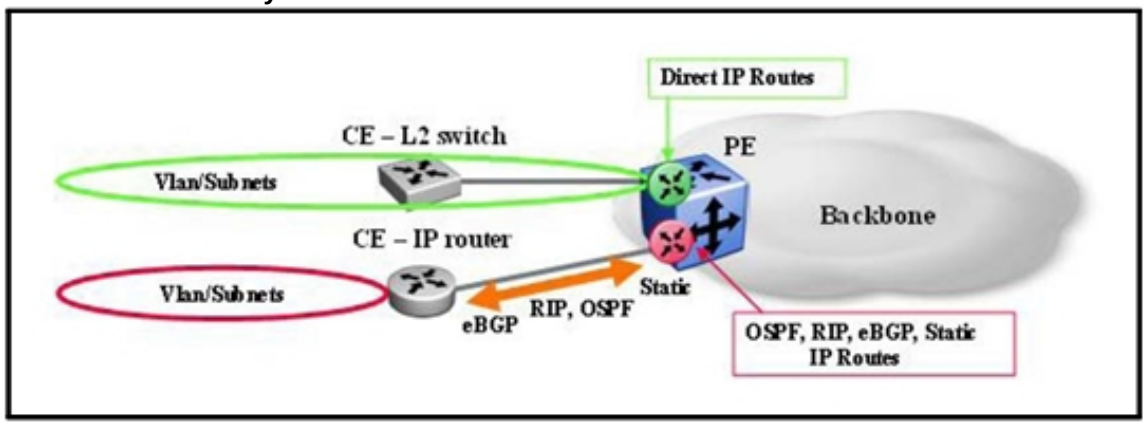
Figure 100
IP VPN packet forwarding



Within a VPN, there can be no overlapping addresses. However, if two VPNs have no common sites, then they may have overlapping address spaces. To support this capability, the PE router must maintain separate forwarding routing tables. To provide multiple independent IPv4 routing and forwarding tables, the Ethernet Routing Switch 8600 supports a default routing instance (VRF0) and up to 255 Virtual Routing and Forwarding (VRF) instances (VRF1 to VRF255).

The PE router maintains separate route tables for each VRF and isolates the traffic into distinct VPNs. Each VRF is associated with one customer, connecting to one or more CE devices but all belonging to the same customer. As shown in the following figure, if the CE is a Layer 3 device, the VRFs exchange routes with the locally connected device using any suitable routing protocol (eBGP, OSPF, RIP, Static Routes). If the CE is a Layer 2 switch, then the customer routes are local (direct) routes configured directly on the relevant VRF of the PE node.

Figure 101
CE to PE connectivity



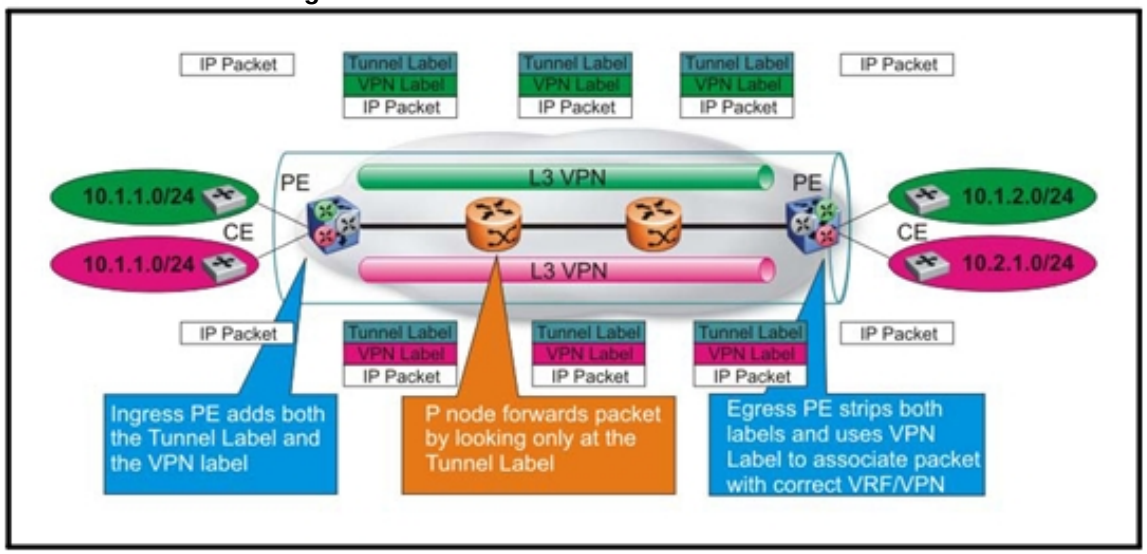
The PE nodes must exchange local VRF customer IPv4 routes with other remote PE nodes that are also configured with a VRF for the same customer (that is, the same IP VPN) while still ensuring that routes from different customers and IP VPNs are kept separate and any identical IPv4 routes originating from two different customers can both be advertised and kept separate. This is achieved via the use of iBGP peering between the PE nodes. These iBGP sessions are terminated on a single circuitless IP (CLIP) interface (belonging to the Backbone Global Routing Table (GRT) on the PE nodes. Because BGP runs over TCP, it can be run directly between the PE nodes across the backbone (there is no BGP requirement on the P nodes).

A full iBGP peering mesh is required between all PEs. In order to scale to a large number of PE devices, BGP Route reflectors are recommended.

Upon receiving traffic from a CE router, the PE router performs a route lookup in the corresponding VRF route table. If there is a match in the VRF route table with a BGP nexthop entry, the PE router adds the IP packet into an MPLS label stack consisting of an inner and outer label. The inner VPN label is associated with the customer VPN. The BGP next-hop is the circuitless IP (CLIP) address of the upstream PE router. The outer LDP tunnel label is used by the P routers to label switch the packet through the network to the appropriate upstream PE router. The P routers are unaware of the inner label.

As shown in the following figure, upon receiving the packet, the upstream PE router removes the top LDP label and performs a lookup based on the VPN label to determine the outgoing interface associated with the corresponding VRF. The VPN label is removed and the packet is forwarded to the CE router.

Figure 102
PE router label switching

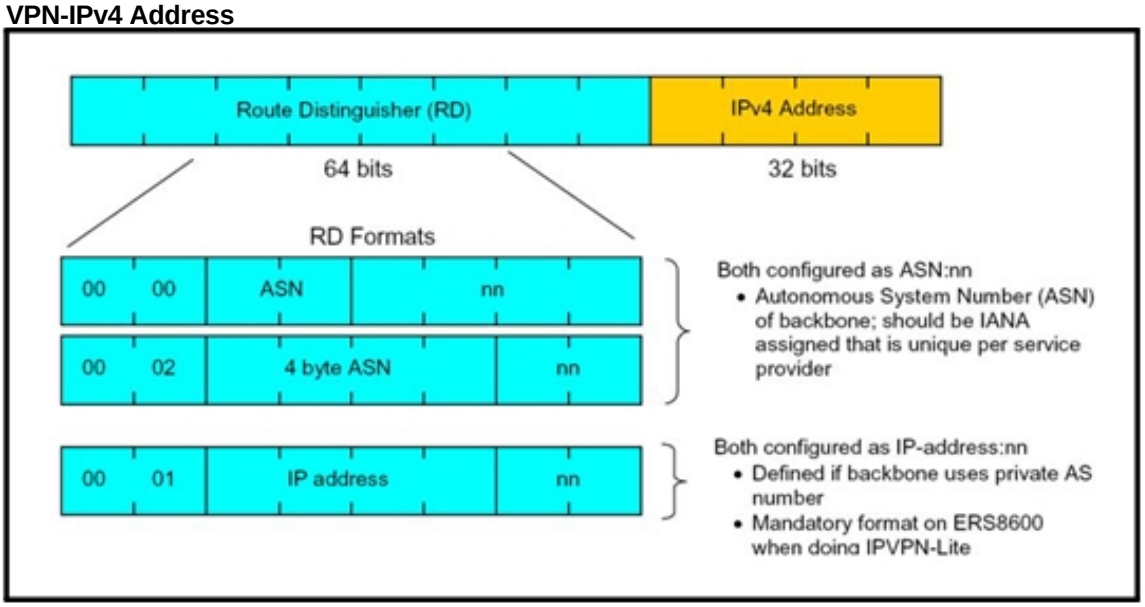


The VPN-IPv4 routes are distributed by MPLS labels. The MPLS label switched paths are used as the tunneling mechanism. Hence, all nodes in the network must support Label Distribution Protocol (LDP) and in particular Downstream Unsolicited mode must be supported for Ethernet interfaces. LDP uses implicit routing, thus it relies on the underlying IGP protocol to determine the path between the various nodes in the network. Hence, LDP uses the same path as that selected by the IGP protocol used.

Route distinguishers

PE routers use BGP to allow distribution of VPN routes to other PE routers. BGP Multiprotocol Extensions (BGP-MP) allows BGP to forward routes from multiple address families, in this case, VPN-IPv4 addresses. The BGP-MP address contains a 12-byte VPN-IPv4 address which in turn contains an 8-byte Route Distinguisher (RD) and a 4-byte IPv4 address. The Route Distinguisher makes the IPv4 address globally unique. As a result, each VPN can be distinguished by its own RD, and the same IPv4 address space can be used over multiple VPNs.

Figure 103
VPN-IPv4 Address



The RD is configured on each and every VRF created on the PE nodes and must be configured such that no other VRF on any other PE in the backbone has the same value. RDs are encoded as part of the Network Layer Reachability Information (NLRI) in the BGP Update messages.

Please note that the RD is simply a number that you configure. It provides a means to identify a PE node which may contain one or more VRFs. It does not identify the origin of the route nor does it specify the set of VPNs or VRFs to which the routes are distributed. Its sole purpose is to provide a mechanism to support distinct routes to a common IPv4 address prefix. By allowing different RDs to support the same IPv4 addresses, overlapping addresses are supported.

Route targets

When an VPN-IPv4 route advertised from a PE router is learned by a given PE router, it is associated with one or more Route Target (RT) attributes. The RT, which is configured on the PE router as either import, export, or both, is the glue which determines whether a customer VPN-IPv4 route being advertised by one PE router can be accepted by another remote PE router resulting in the formation of a logical IP VPN end to end. These routes are accepted by a remote PE providing the remote PE has a matching import RT configured on one of its VRFs.

A Route Target attribute can be thought of as identifying a set of sites, though it would be more precise to think of it as identifying a set of VRFs. Each VRF instance is associated with one or more Route Target (RT) attributes. Associating a particular Route Target attribute with a route allows that route to be placed in the VRFs that are used for routing traffic

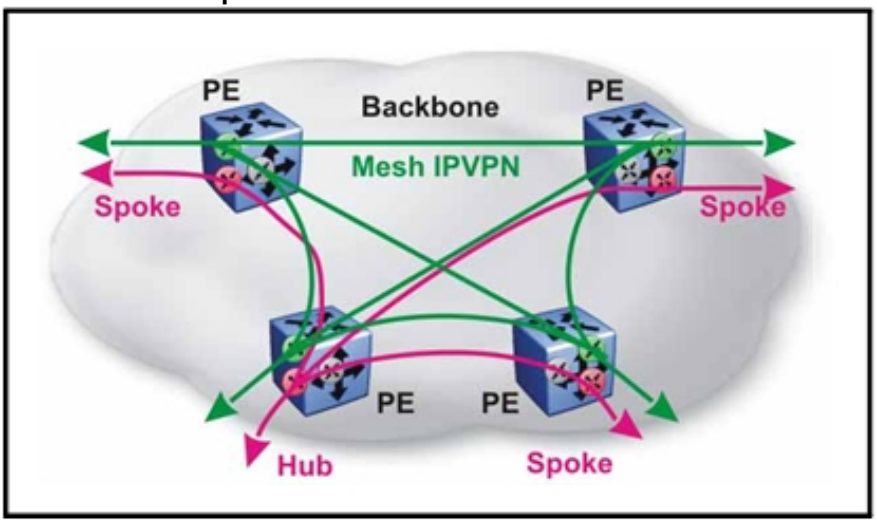
among the sites in that VPN. Note that a route can only have one RD, but it can have multiple Route Targets. RTs also enhance the PE scaling capability since a given PE node only accepts VPN-IPv4 routes for which it has local VRFs belonging to that IP VPN; any other VPN-IPv4 routes are not accepted.

Each VPN-IPv4 route contains a route target extended community that is advertised or exported by the PE router export policy. Any PE router in the network configured with a matching route target in its import policy imports the route for that particular VRF.

RTs must be configured in such a way as to be unique for each IP VPN.

Since each VRF can be configured with any number of RTs (either as import, export or both) this allows each VRF to be part of any number of overlapping IP VPNs. The use of RT can also be exploited to achieve a number of different IP VPN topologies, from any-to-any (meshed) where all VRFs in the same IP VPN have the same import and export RT, to hub and spoke topologies where the hub nodes use one export RT (configured as import RT on spokes) and a different import RT (configured as export RT on the spokes). Topologies with multiple hub sites can also be achieved.

Figure 104
IP VPN hub and spoke



In terms of configuration both the RD and the RT are configured with the same format and are usually configured in the same VRF context on the PE device.

The route target frame format is identical to the route distinguisher as shown in Figure 103 "VPN-IPv4 Address" (page 262).

IP VPN requirements and recommendations

To use IP VPN, you require R or RS modules, as well as the 8692 SF/CPU with SuperMezz. You also require the appropriate license. IP VPN can run in an R mode or mixed mode chassis. Classic modules (E or M modules) can be in the Global Routing Engine, but they cannot connect to provider edge or provider core devices. E and M modules do not support MPLS label switching.

The Ethernet Routing Switch 8600 supports IP VPN over 802.3ad (MLT)

Partial-HA (P-HA) is supported by IP VPN. P-HA means that a module can be enabled and configured when system is running in the HA mode. The configuration database is synchronized between the Master and Slave CPU, so that, on failover, the module starts with the same configuration that the Master CPU executed. After failover, the Standby CPU (new Master) starts the module with the synchronized configuration, and does not carry over the module's runtime state information from the previous Master CPU. If a modules communicates with peers externally, the session is reestablished. P-HA support allows modules to run in the HA mode. Although the module is restarted with most recent configuration, the failover time is improved compared to a single SF/CPU reboot in nonHA mode.

VPN tunnel dampening is not supported.

The Ethernet Routing Switch 8600 requires that a unique VRF be associated with a unique VPN in a single PE device. This means that no two VRFs are attached to the same VPN, thus requiring forwarding between VRFs in single PE. All the CE devices that belong to a single VPN in a single PE device must be part of a single VRF.

The throughput for all standard packet sizes for VPN routed traffic is minimum 90% (depends on egress queue behavior). For IP VPN scalability information, see Table 5 "Supported scaling capabilities" (page 38) or the Release Notes. The Release Notes take precedence over this document.

IP VPN prerequisites

Before you use IP VPN:

- Choose an Interior Gateway Protocol: OSPF and RIP are supported.
- Choose a Route Distinguisher (RD): a unique RD per VRF is supported.
- Select an access topology, an access routing protocol (static routes, RIP, OSPF, or EBGP, or a mix of these), and provide provider edge to customer edge router addressing.

- Define site backup and resiliency options (for example, dual access lines to a single provider edge (PE) router, dual access lines with dual PEs, dual access lines with two CEs and two PEs).
- Set up an Autonomous System Number (ASN). ASNs are usually allocated by service providers for customers that need to connect to the provider edge router using eBGP.

IP VPN deployment scenarios

When the Ethernet Routing Switch 8600 is used as a PE device, the following are the means by which a CE device can connect to PE device:

- One CE connects to a single PE using a single GbE, 10 GbE, or 10/100/1000 Mbit/s port.
- One CE multilink trunks to a single PE using multiple (up to eight) GbE, 10 GbE, or 10/100/1000 Mbit/s ports.
- One CE connects to two PEs (two VRFs but same VPN) using RSMLT.
- Multiple CEs connect to a single PE using VRF, and packets are locally forwarded.

A CE device exchanges routing information with PE devices using static routes and an Interior Gateway Protocol (IGP), for example, OSPF and RIP. The CE device routing engine works with the routing protocol running in the context of a VRF in the PE device. This generally occurs in Enterprise environments.

A CE device exchanges routing information with a PE device using EBGP. The routing engine in the CE device works with EBGP running in the context of a VRF in the PE device. This suits carrier deployments.

When the Ethernet Routing Switch 8600 is used as a PE device, the following are the means by which a PE device can connect to a provider core device:

- One PE connect to a single provider core router using a single GbE, 10 GbE, or 10/100/1000 Mbit/s port
- One PE multilink trunks to a single provider core using multiple (up to eight) GbE, 10 GbE, or 10/100/1000 Mbit/s ports
- One PE connects to two Ps (without SMLT support).
- PE directly connects to PE

A PE device exchanges routing information with a provider core device using an IGP and static routes. The global routing engine in the PE device works with the routing protocol running in the context of a global routing engine in the provider core device.

For detailed IP VPN configuration examples, see *IP-VPN (MPLS) for ERS 86000 Technical Configuration Guide* (NN48500-569) and *IP-VPN and IP-LER Interoperability for Ethernet Routing Switch Technical Configuration Guide* (NN48500-571) . For detailed VRF Lite configuration examples, see *VRF-Lite for Ethernet Routing Switch 8600 Technical Configuration Guide* (NN48500-570) .

MPLS interoperability

The Ethernet Routing Switch 8600 MPLS implementation has been verified with:

- Cisco 7500 (with RSVP, Cisco cannot function as the RSVP egress LER when used with the Ethernet Routing Switch 8600)
- Juniper M10

MTU and Retry Limit

The MPLS maximum transmission unit (MTU) is dynamically provisioned (1522 or 1950 bytes) and it supports jumbo frames (9000 bytes). Packets that exceed the MTU are dropped. The allowed data CE frame size is MTU size minus MPLS encapsulation (header) size. For control frames (for example, LDP) the frame size is 1522 or 1950 bytes.

For the Ethernet Routing Switch 8600, the MPLS RSVP LSP Retry Limit is infinite by design (a setting of zero means infinite). When the limit is infinite, should a Label Switched Path (LSP) go down, it is retried using exponential backoff. The Retry Limit is not configurable.

IP VPN Lite

With Nortel IP VPN-Lite, the Ethernet Routing Switch 8600 can provide a framework for delivering RFC4364 IP VPNs over an IP backbone, rather than over MPLS.

In terms of Data Plane packet forwarding across the same backplane, RFC 4364 defines an implementation based on MPLS where the backbone must be MPLS capable and a full mesh of MPLS Label Switched Paths (LSPs) must already be in place between the PE nodes.

While still leveraging the same identical RFC 4364 framework at the control plane level, Nortel IP VPN-Lite delivers the same IP VPN capabilities over a IP routed backbone using simple IP in IP encapsulation with no requirement for MPLS and the complexities involved with running and maintaining an MPLS backbone.

With IP VPN-Lite a second Circuitless IP (CLIP) address is configured on the PE nodes (in the Backbone GRT and re-advertised across the Backbone by the IGP). This second CLIP address is used to provide

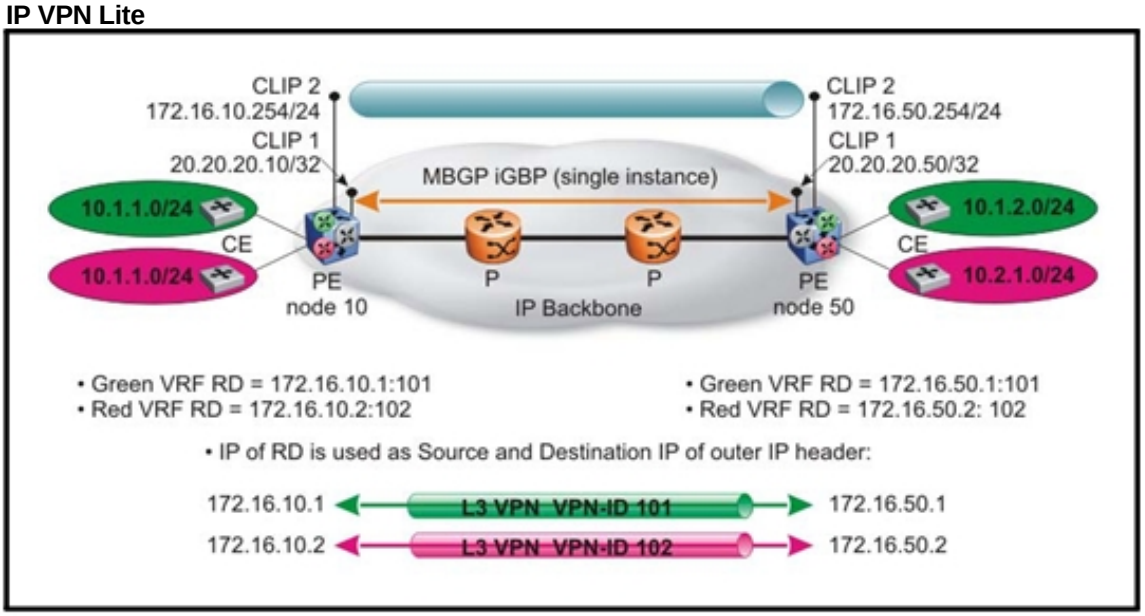
address space for the outer header of IP-in-IP encapsulation for all IP VPNs packets terminating to and originating from the PE. This second Circuitless address is therefore ideally configured as a network route (in other words, not as a 32 bit mask host route) with enough address space to accommodate every VRF configured on the PE. A 24 bit mask provides sufficient address space for 252 VRFs. Furthermore, as these networks only need to be routed within the provider backbone and no further, public address space can be used. When this second CLIP address is configured it must also be enabled for IP VPN services.

With Nortel IP VPN-Lite, the RD is now used to convey one extra piece of information over and above its intended use within the RFC 4364 framework. In the RFC, the only purpose of the RD is to ensure that identical IPv4 routes from different customers are rendered unique so that BGP can treat them as separate VPN-IPv4 routes. With IP VPN-Lite, the RD is now also used to advertise to remote PE devices what IP address needs to be used as the outer IP-in-IP encapsulation when those remote PE devices need to deliver a customer packet over the IP VPN back to the PE node which owns the destination route to which the packet is addressed.

Therefore, when configuring RD for IP VPN-Lite, the RD must always be configured as Type 1 format (IPaddress:number), and the IP address configured in the RD must allocate one host IP address defined by the second CLIP interface for each VRF on the PE. Again, the RD must still be configured to ensure that no other VRF on any other PE has the same RD.

In the following example the second CLIP interface is configured as a private address, with a 24 bit mask, where the third octet identifies the PE node-id and the fourth octet (the host portion) defines the VRF on that PE node. The number following the IP address is then simply allocated to uniquely identify the VPN-id.

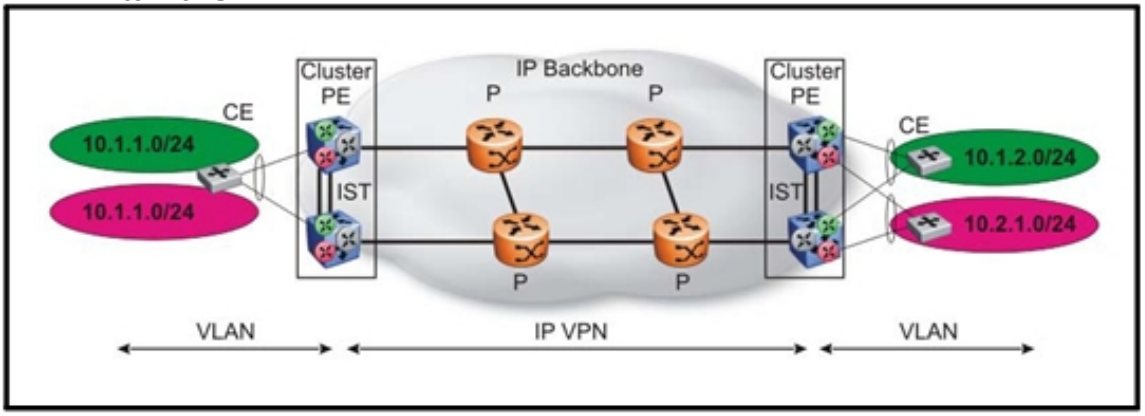
Figure 105
IP VPN Lite



IP VPN-Lite can therefore easily be deployed on any enterprise existing IP routed network and automatically leverage the existing backbone architecture in terms of load balancing and subsecond failover resiliency. While MPLS struggles to achieve these goals and only does so by bringing in exponential complexity, Nortel IP VPN-Lite can simply leverage these capabilities from either a pure IP OSPF routed core where ECMP is enabled or a network core designed with Nortel SMLT/RSMLT clustering.

Furthermore, PEs can be just as easily deployed with SMLT clustering towards the CE edge devices thus delivering a very attractive clustered PE solution. This is easily achievable whether the CE is a L2 device (using SMLT Clustering) or an L3 device (where the SMLT cluster needs to be RSMLT enabled).

Figure 106
IP VPN Lite with SMLT



Overall, IP VPN-Lite provides support for the following:

- 256 VPNs per each system
- Filtering support (UNI side)
- Overlapping addresses
- MP-BGP extensions
- BGP route refresh
- BGP route reflection
- Peering to multiple route reflectors
- Route reflection server (NNI side)
- Full mesh and hub and spoke designs
- Extended community Type 0 and 1
- import and export route targets and route distinguishers
- IP-BGP extensions
- IEEE 802.3ad/MLT
- Split MultiLink Trunking (SMLT) and Routed Split MultiLink Trunking (RSMLT) for CE connectivity
- ECMP
- VRF-based ping and traceroute
- UNI packet classification (port, VLAN, IP, VRF, and VPN)
- VRF UNI routing protocols (RIP, OSPF, eBGP)

An IP VPN-Lite PE device provides four functions:

- An IGP protocol, such as OSPF, across the core network to connect remote PE devices
- VRFs to provide traffic separation
- MP-BGP to exchange VPN routes and service IP addresses with remote PE devices
- The forwarding plane to encapsulate the customer IP packet into the revise IP header

IP VPN Lite deployment scenarios

The following sections describe how you can use the IP VPN Lite capability on the Ethernet Routing Switch 8600 to design a sample network interconnecting five separate sites while meeting the following requirements:

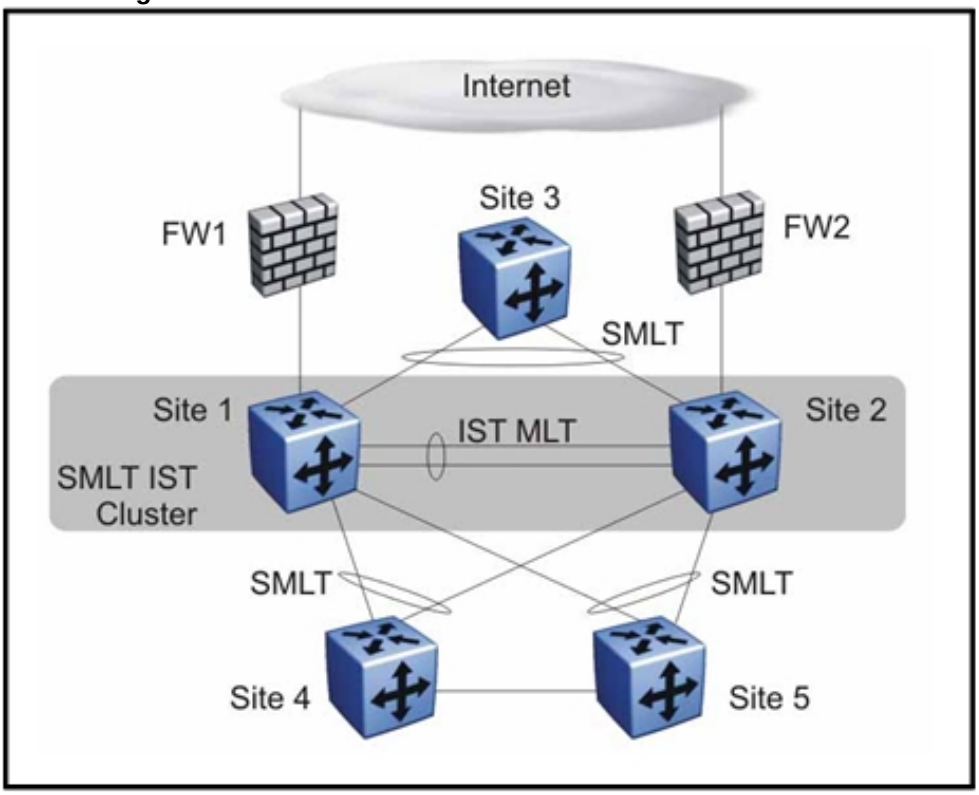
- 10 gigabit connectivity between sites (over dark fiber or DWDM circuits)
- Capability of Layer 2 VPN connectivity between any number of sites
- Capability of Layer 3 VPN connectivity between any number of sites
- VPN scalability to scale up to ~100 Layer 2 VPNs and ~100 Layer 3 VPNs
- Two main sites to provide Internet connectivity to every other site
- Ability to provide Internet connectivity to each Layer 3 VPN while not allowing any connectivity between Layer 3 VPNs (no overlapping address space between different VPNs)
- Resilient design with subsecond failover times (No Spanning Tree)
- Low latency, high bandwidth, non-blocking design where all traffic is hardware switched

For detailed configuration steps for these examples, see *IP VPN-Lite for Ethernet Routing Switch 8600 Technical Configuration Guide* (NN48500-562) .

SMLT design

To meet the design requirements, an Ethernet Routing Switch 8600 is deployed at each site. As shown in the following figure, the five Ethernet Routing Switch 8600s are interconnected using 10 gigabit Ethernet links in an SMLT cluster configuration. The Ethernet Routing Switch 8600s in the two main sites, which provide Internet connectivity to the network, are the SMLT cluster nodes (which logically act as one switch) and are interconnected by a DMLT IST connection. The remaining sites are connected as SMLT edge devices using an SMLT triangle topology. VLACP is enabled on all links using long timers on IST links and short timers on SMLT links. The maximum number of hops for traffic to reach a remote site is at most 2 hops and in some cases 1 hop only.

Figure 107
SMLT design



With Nortel's advanced packet processor architecture, the Ethernet Routing Switch 8600 always hardware switches all traffic flows including IP VPN traffic used in this design. This means that if a non-blocking 10 gigabit hardware configuration is used (for example, using 8683XLR or 8683XZR 3-port 10GBASE-X LAN/WAN XFP modules), then full 10 gigabit bandwidth and extremely low latency is available from site to site.

Furthermore, if 10 gigabit later becomes insufficient between any sites, you can increase the bandwidth in this design by adding additional 10 gigabit links to the existing MLTs.

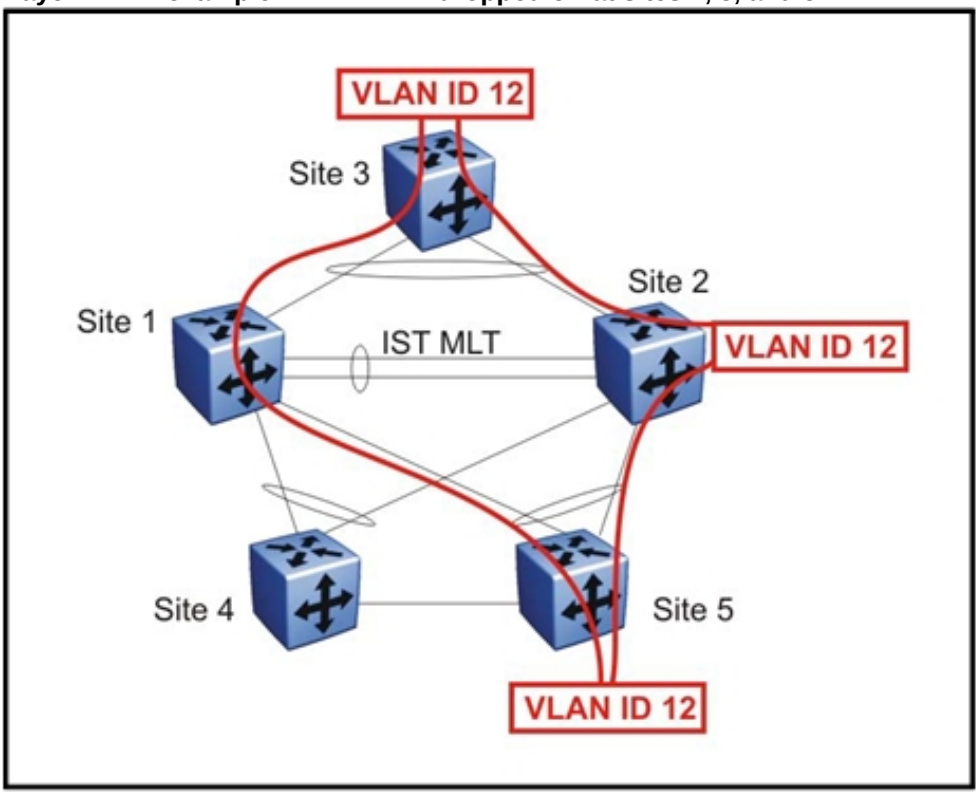
ATTENTION
To support the VRF and IP VPN functionalities used in this design, you must equip the Ethernet Routing Switch 8600 with R or RS I/O Modules, 8692 SF/CPU card with Super-Mezzanine daughter card, and Release 5.0 or higher software with the Premium Software License.

Layer 2 VPN design

To provide Layer 2 VPN services, native VLANs are created on top of the SMLT design. These VLANs do not have an IP assigned and can be added or dropped at any site. A suitable range of VLAN IDs are reserved for these Layer 2 VLANs. In this example, VLAN IDs 2-99 are reserved for

this purpose. As illustrated in the following figure, VLAN ID 12 is spanned across 3 sites. Please note that any Layer 2 VLANs that are added to this design must always be configured on both main sites 1 and 2 (the SMLT IST cluster) but only on the Ethernet Routing Switch 8600 SMLT edge switches that require the VLANs. In this example, VLAN 12 is added to the SMLT IST cluster switches at sites 1 and 2 and then added at Sites 3 and 5. At sites 2, 3 and 5, Layer 2 VLAN 12 is also configured on one or more edge facing interfaces.

Figure 108
Layer 2 VPN example: VLAN ID 12 dropped off at sites 2, 3, and 5

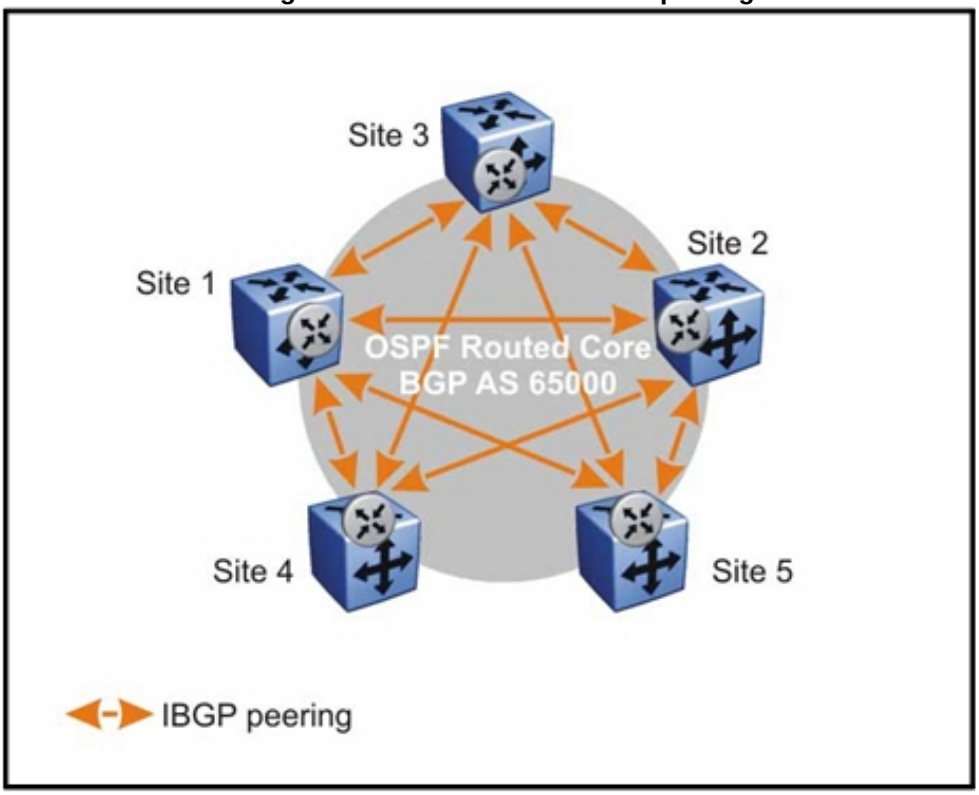


As part of best design guidelines, do not use VLAN ID 1 (the default VLAN).

Inter-site IGP routing design

As shown in the following figure, Layer 3 IGP connectivity between all five sites is provided using two routed VLANs where an OSPF backbone area is enabled on all five Ethernet Routing Switch 8600s. This routing instance constitutes the default routing instance of the Ethernet Routing Switch 8600 platform which is know as the Global Routing Table (GRT) or VRF0. The purpose of this routed GRT routing instance is purely to provide IP connectivity between a number of Circuitless IP (CLIP) interfaces that must created on each Ethernet Routing Switch 8600.

Figure 109
GRT IGP VLAN running OSPF and full mesh of IBGP peerings



Each Ethernet Routing Switch 8600 is configured with a Circuitless IP address (CLIP) host address using a 32-bit mask. From these CLIP interfaces, a full mesh of IBGP peerings is configured between the Ethernet Routing Switch 8600s in each site. The IBGP peerings are enabled for VPNv4 and IP VPN Lite capability and are used to populate the IP routing tables within the VRF instances used to terminate the Layer 3 VPNs.

To support a larger number of sites, Nortel recommends the use of BGP Route-Reflectors. This can be accomplished by making the Ethernet Routing Switch 8600 at site 1 and site 2 redundant Route-Reflectors and every other site a Route-Reflector client.

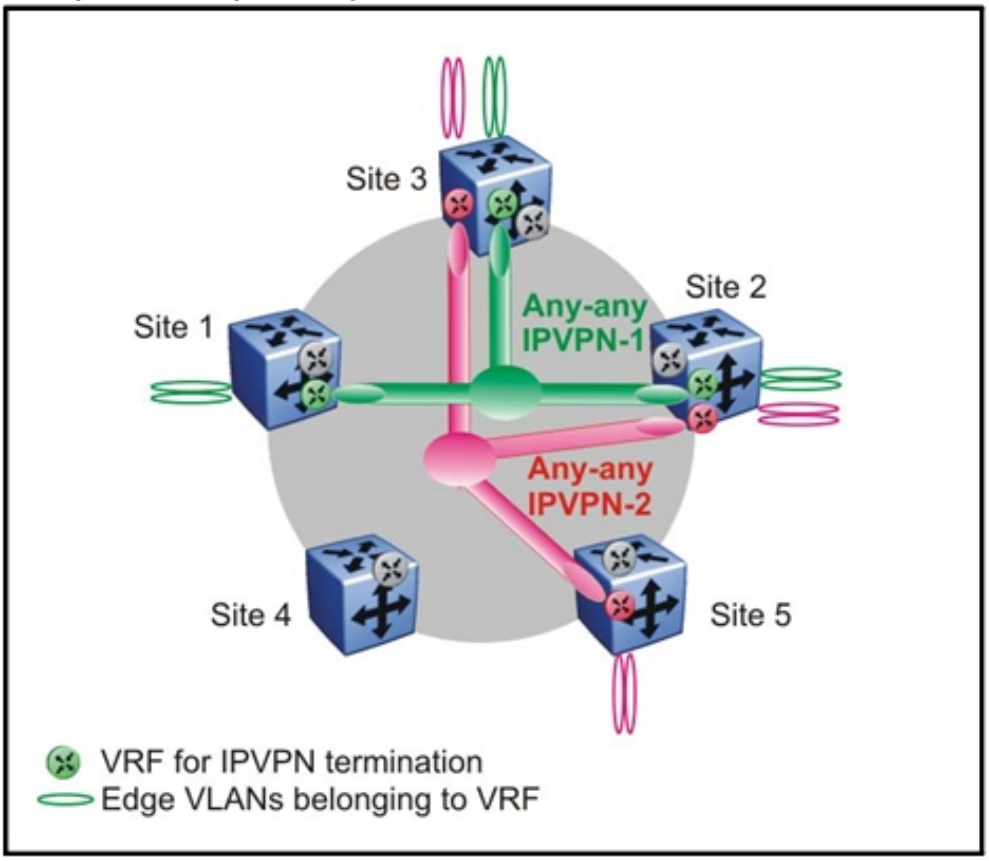
Layer 3 VPN design

The Layer 3 VPNs are implemented using Nortel IP VPN Lite.

To provide address space for the IPinIP encapsulation, each Ethernet Routing Switch 8600 is also configured with a second CLIP network address (the Service IP) which is created using a 24-bit mask rather than a host 32-bit mask.

Layer 3 VPNs are then configured by first creating a VRF instance at all the sites where the VPN must terminate. As shown in the following figure, IP VLANs local to each site can then be assigned to the relevant VRF, thus ensuring IP routing connectivity between VLANs assigned only to the same VRF instance, but no IP routing towards other IP VLANs assigned to other VRF instances. Each VRF then has IP VPN functionality enabled which allows it to belong to one or more Layer 3 VPNs. This configuration is done by assigning an appropriate Route Distinguisher (RD) and import and export Route Targets (RT) to the VRF IP VPN configuration. The end result being that BGP automatically installs remote IP routes from remote VRFs belonging to the same VPN into the local VRF and vice-versa. Furthermore each Layer 3 VPN can be created as any-any, hub-spoke or multihub-spoke by simple manipulation of the import and export RTs as per the RFC 4364 framework.

Figure 110
Example of two separate Layer 3 VPNs

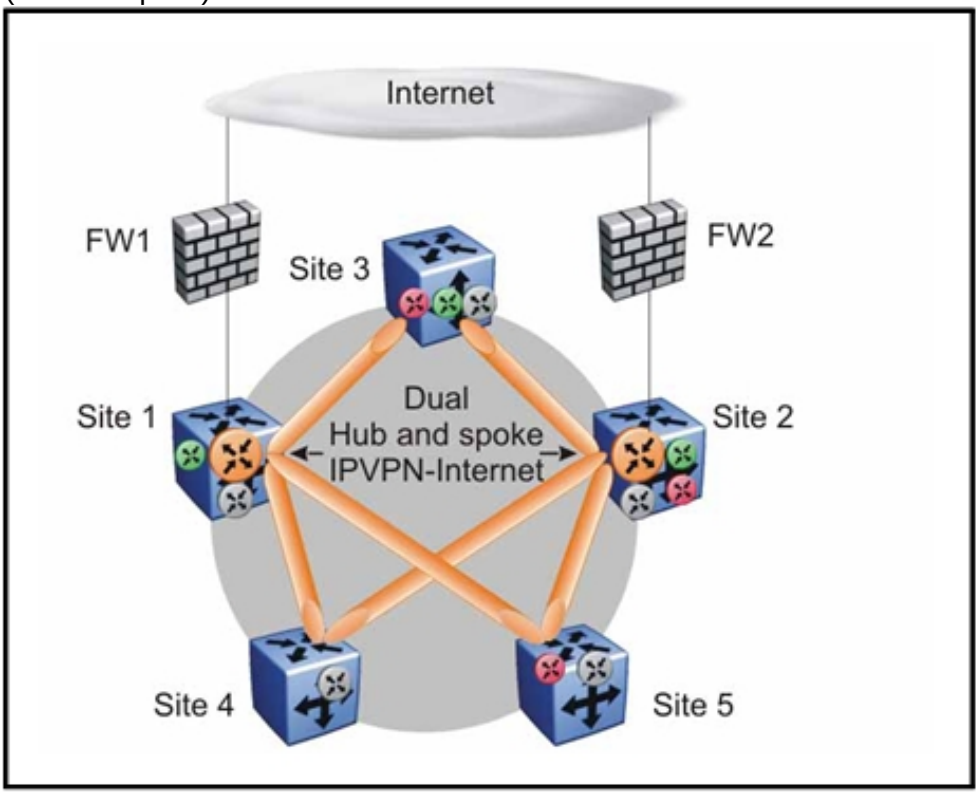


Internet Layer 3 VPN design

The two Ethernet Routing Switch 8600s in the main sites 1 and 2 also have a third CLIP address (also a Service IP) which is made the same at both sites. This CLIP address also uses a 24-bit mask and is only used

for IPinIP encapsulated Layer 3 VPN traffic destined for the Internet. This allows both the site 1 and site 2 Ethernet Routing Switch 8600s to handle Internet bound traffic from Site 3, 4 or 5 regardless of the MLT hash used by these SMLT edge sites (this eliminates the need for site 1 to forward some Internet bound traffic to site 2 over the IST and vice-versa).

To this effect RSMLT functionality is also enabled on Site 1 and 2 on the GRT OSPF VLANs. The Internet VPN is configured as a multihub-spoke (dualhub-spoke).



For more information, see *IP VPN-Lite for Ethernet Routing Switch 8600 Technical Configuration Guide* (NN48500-562) .

Layer 1, 2, and 3 design examples

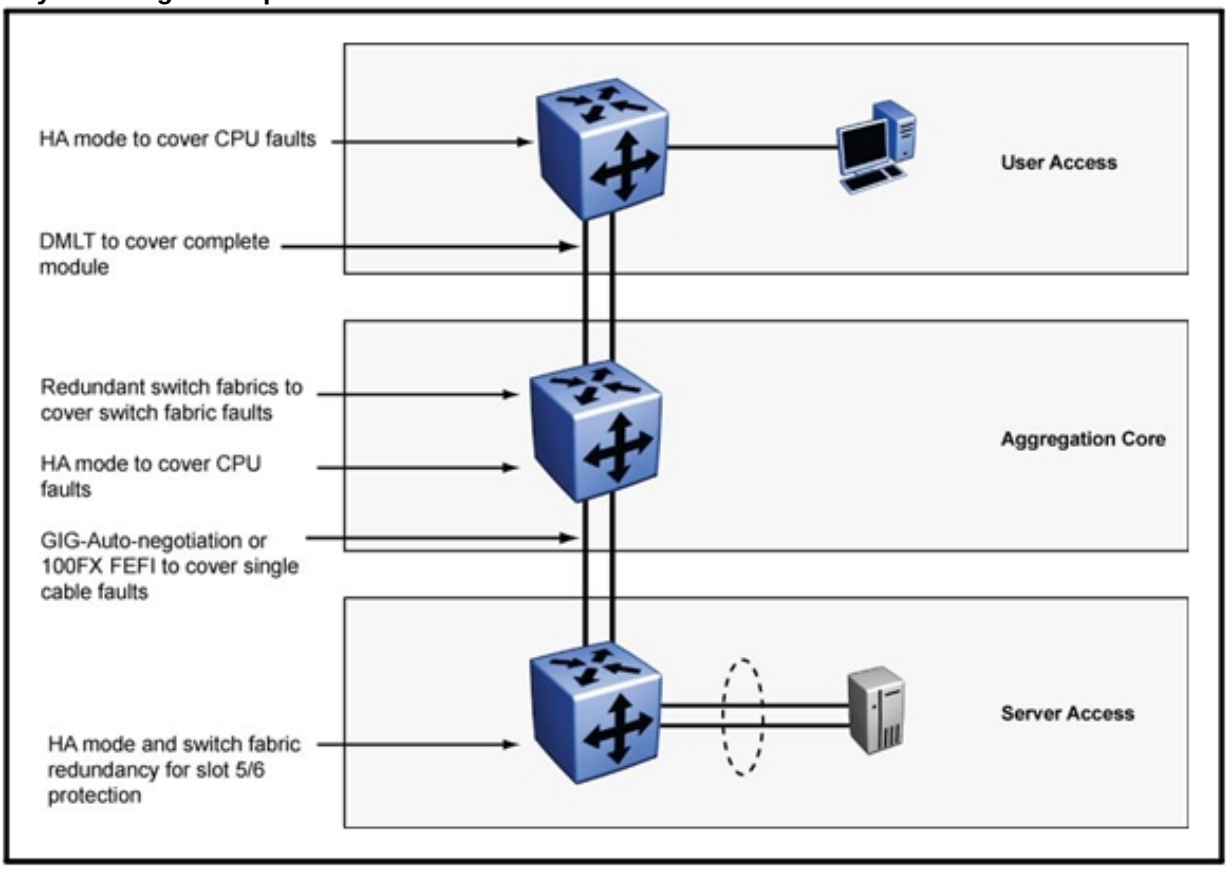
This section provides examples to help you design your network. Layer 1 examples deal with the physical network layouts; Layer 2 examples map Virtual Local Area Networks (VLAN) on top of the physical layouts; and Layer 3 examples show the routing instances that Nortel recommends to optimize IP for network redundancy.

Navigation

Layer 1 examples

The following figures are a series of Layer 1 examples that illustrate the physical network layout.

Figure 111
Layer 1 design example 1



All the Layer 1 redundancy mechanisms are described in example 2.

Figure 112
Layer 1 design example 2

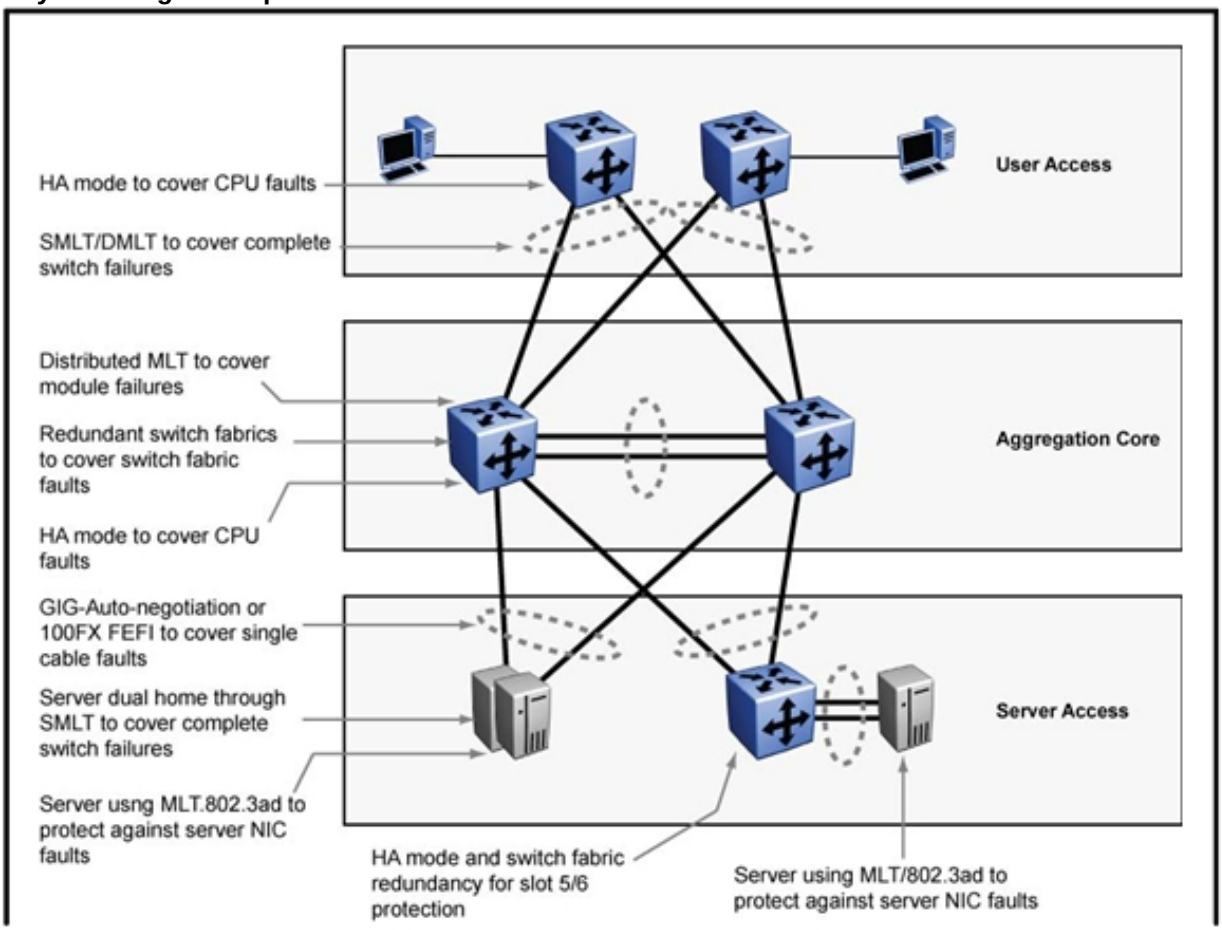


Figure 113
Layer 1 design example 3

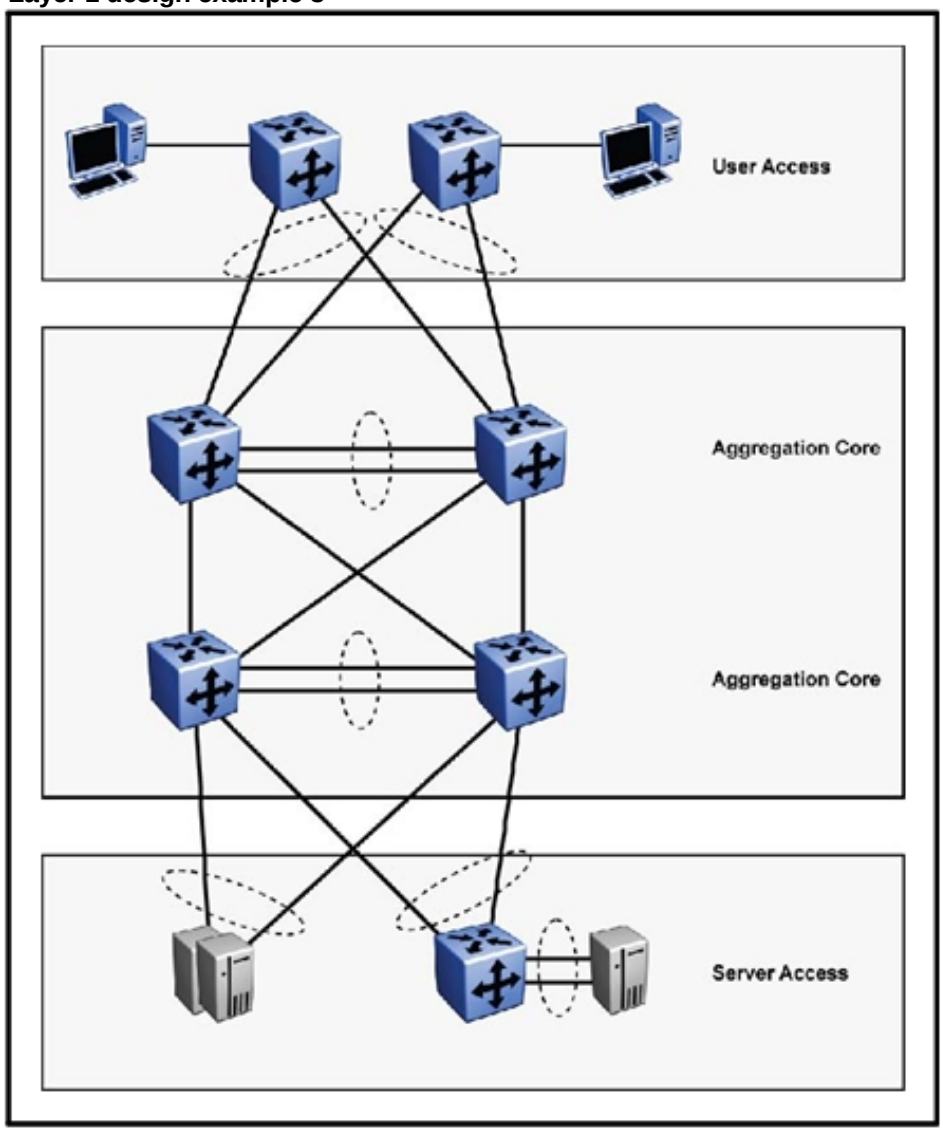
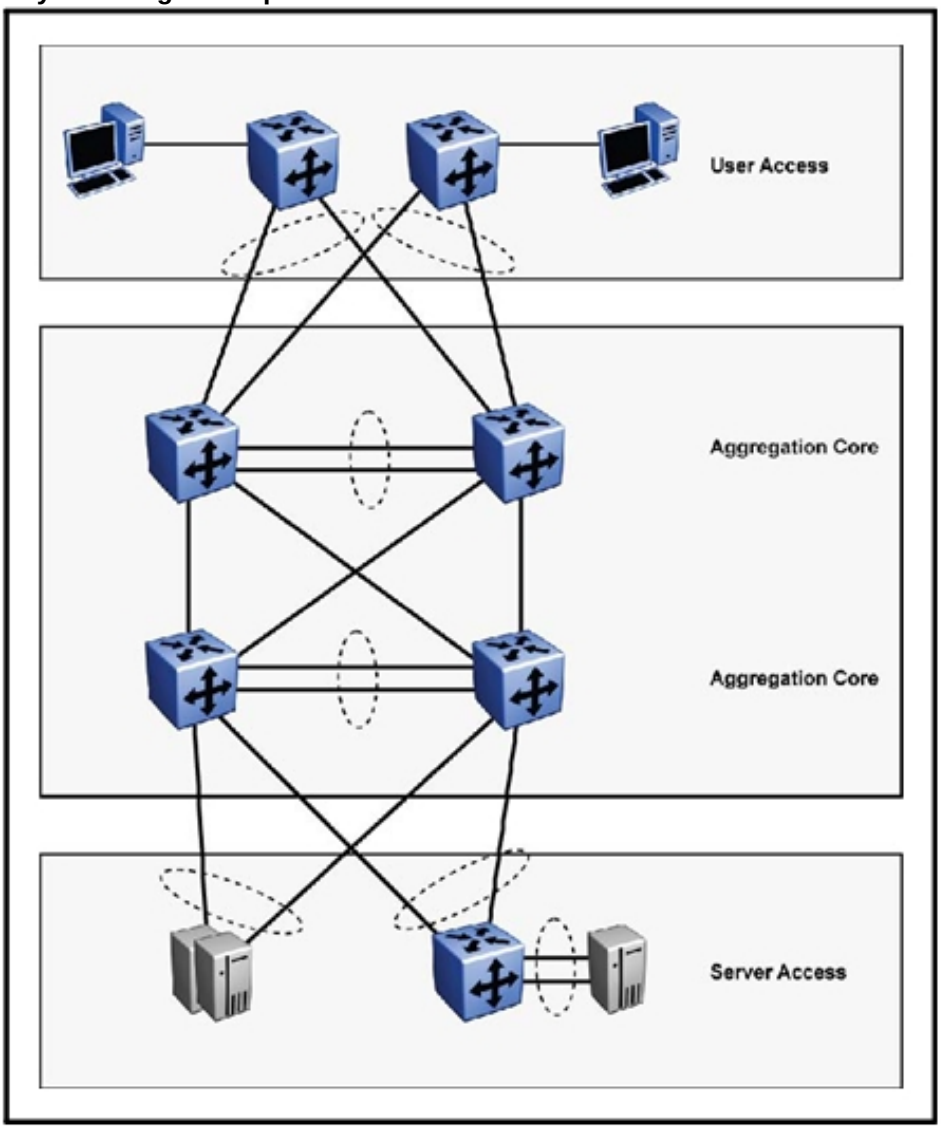


Figure 114
Layer 1 design example 4

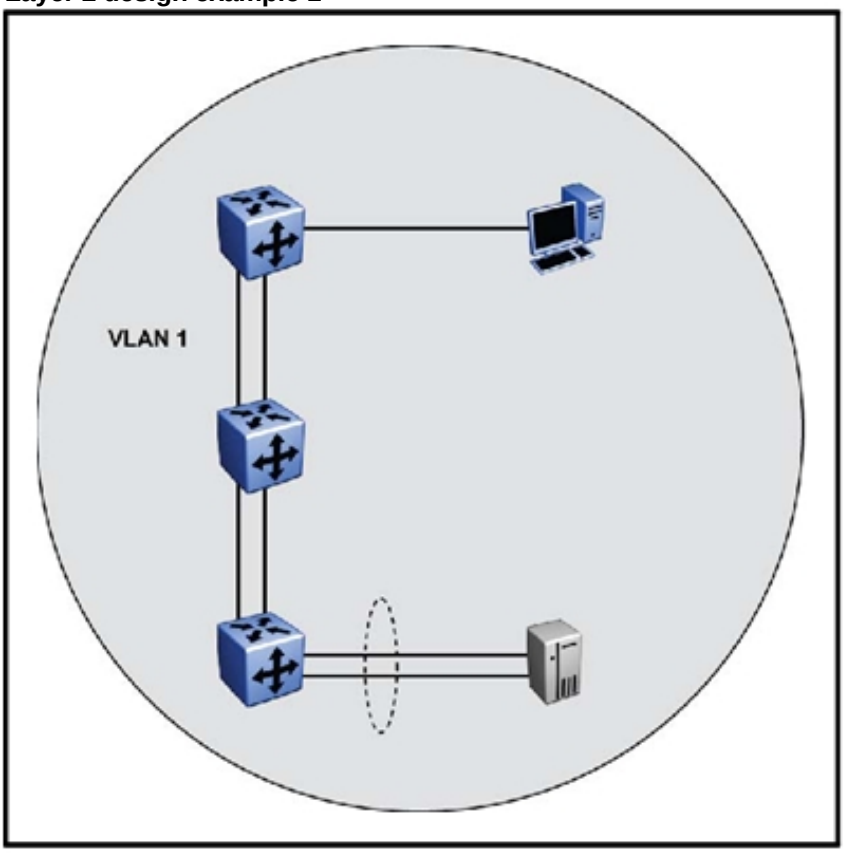


Layer 2 examples

The following figures are a series of Layer 2 network design examples that map VLANs over the physical network layout.

Example 1 shows a redundant device network that uses one VLAN for all switches. To support multiple VLANs, 802.1Q tagging is required on the links with trunks.

Figure 115
Layer 2 design example 1



Example 2 depicts a redundant network using Split Multilink Trunking (SMLT). This layout does not require the use of Spanning Tree Protocol: SMLT prevents loops and ensures that all paths are actively used. Each wiring closet (WC) can have up to 8 Gbit/s access to the core. This SMLT configuration example is based on a three-stage network.

Figure 116
Layer 2 design example 2

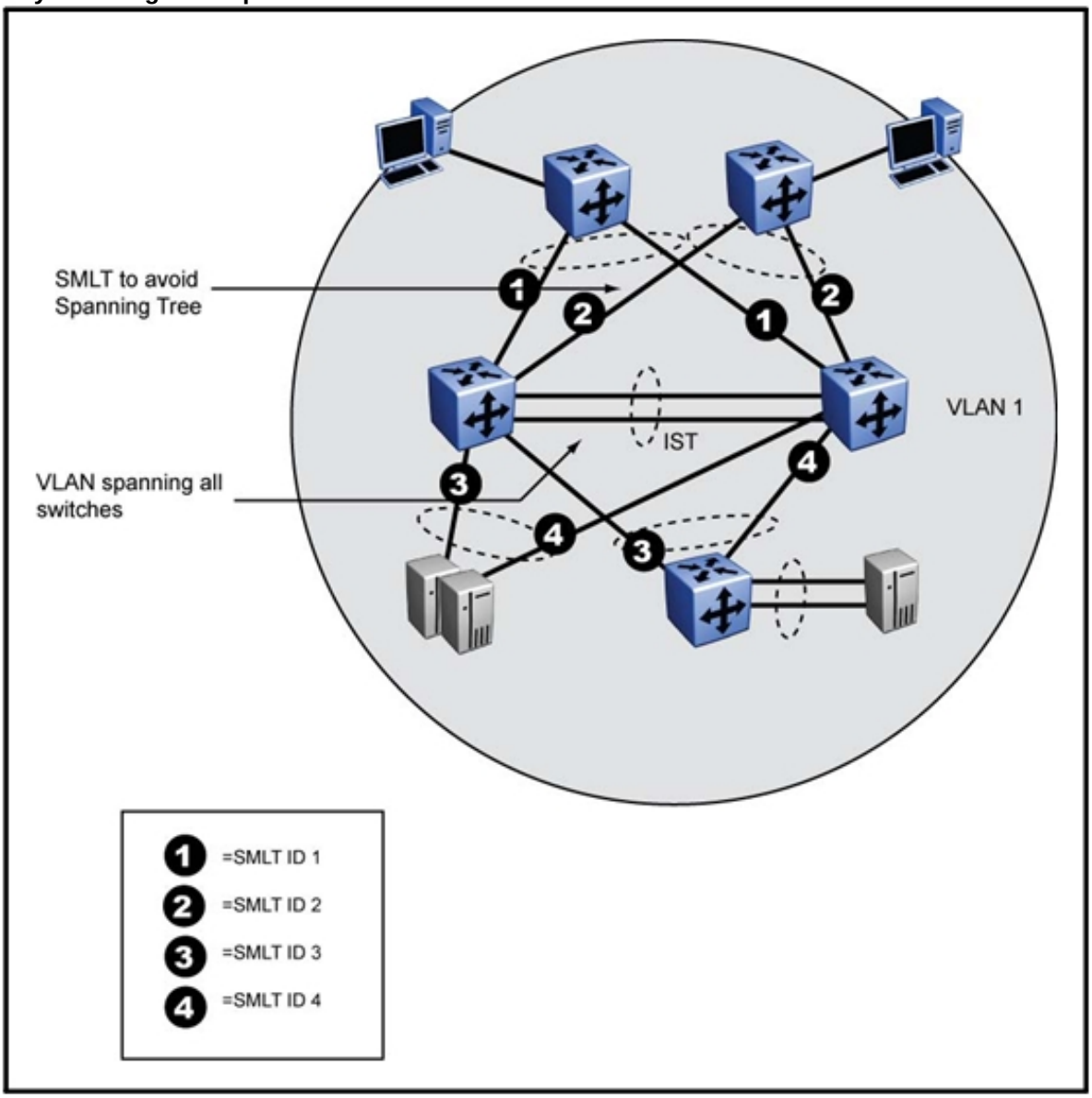
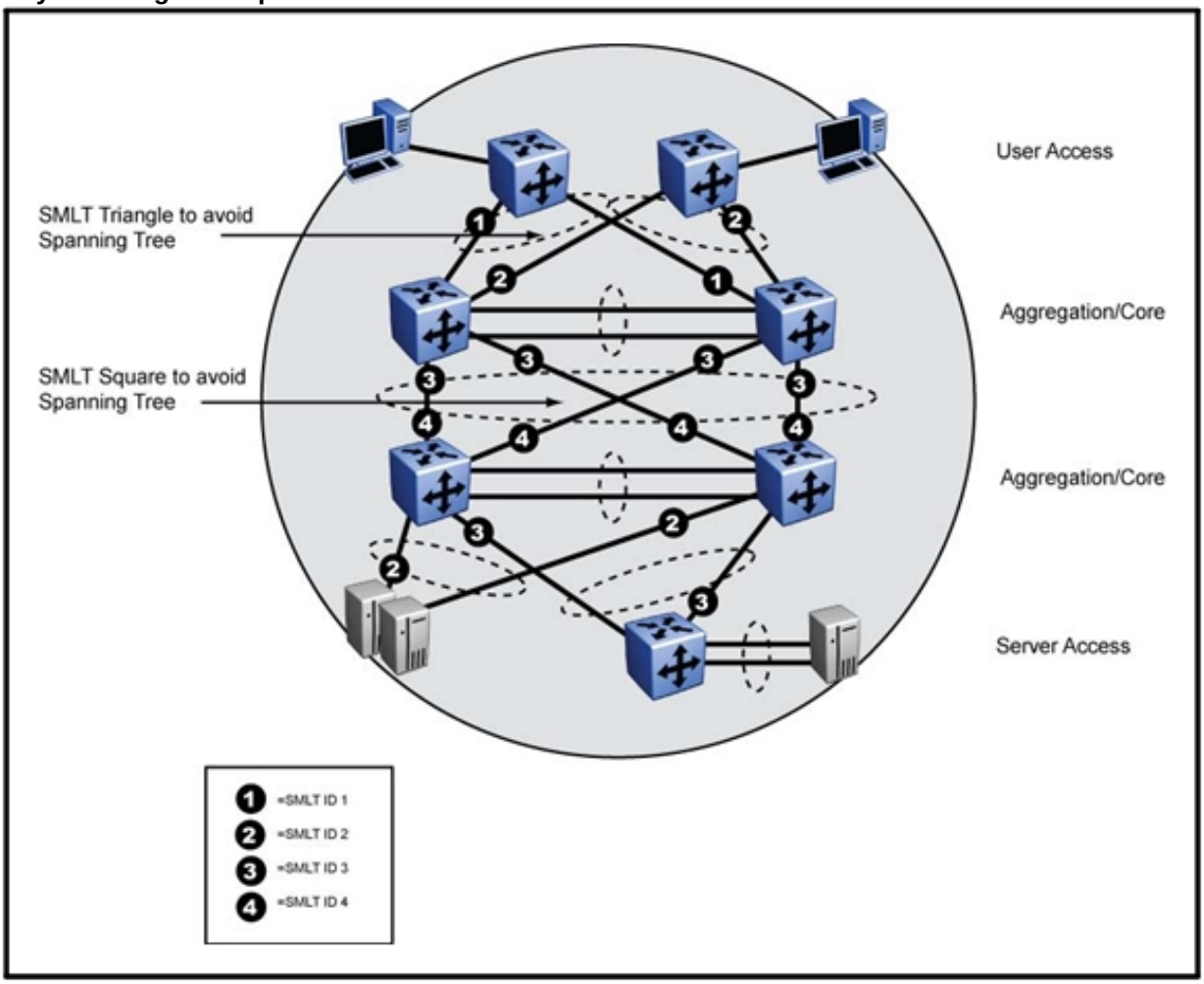


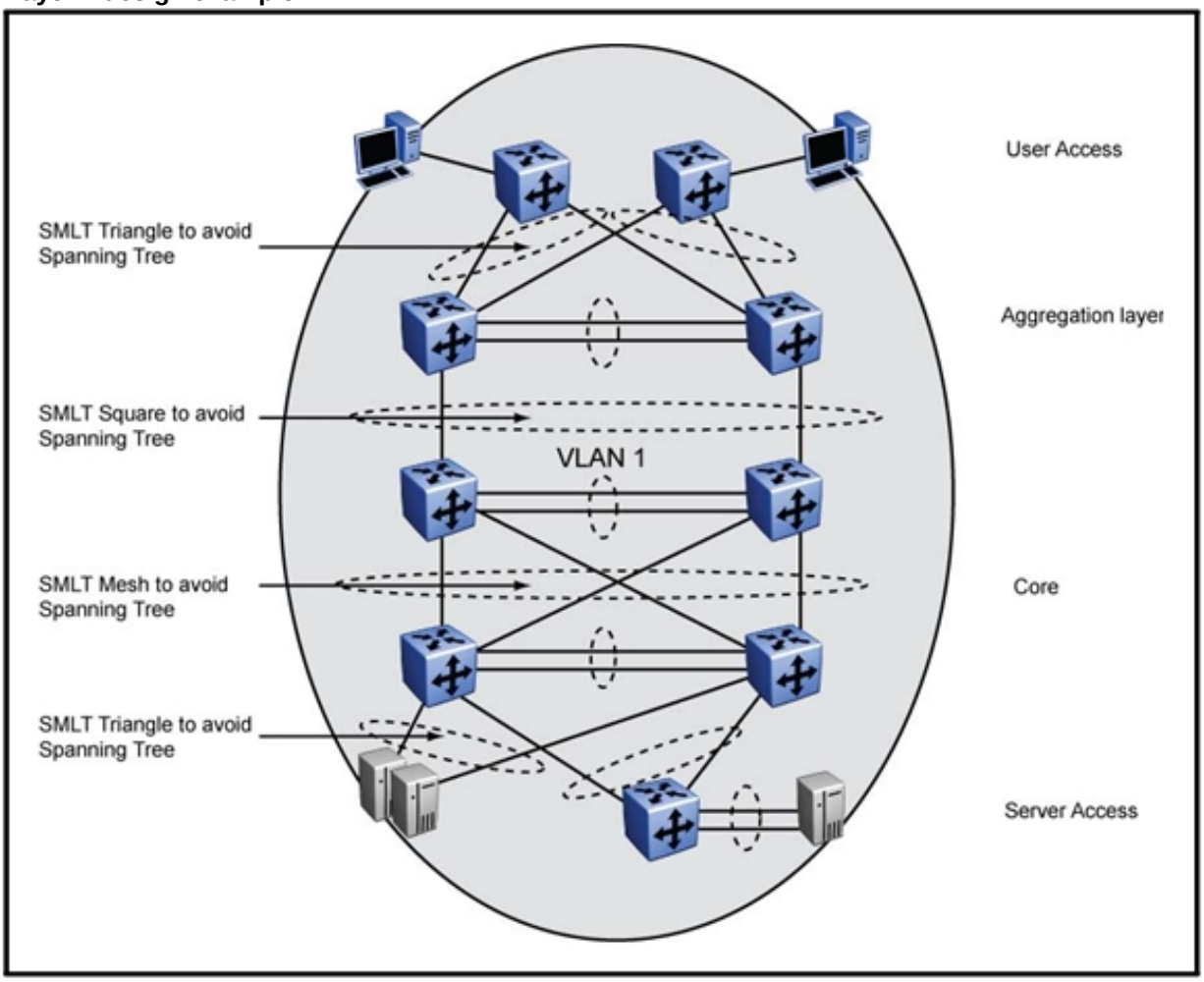
Figure 117
Layer 2 design example 3



In Example 3, a typical SMLT ID setup is shown.

Because SMLT is part of MLT, all SMLT links have an MLT ID. The SMLT and MLT ID can be the same, but this is not necessary.

Figure 118
Layer 2 design example 4



Layer 3 examples

The following figures are a series of Layer 3 network design examples that show the routing instances that Nortel recommends you use to optimize IP for network redundancy.

Figure 119
Layer 3 design example 1

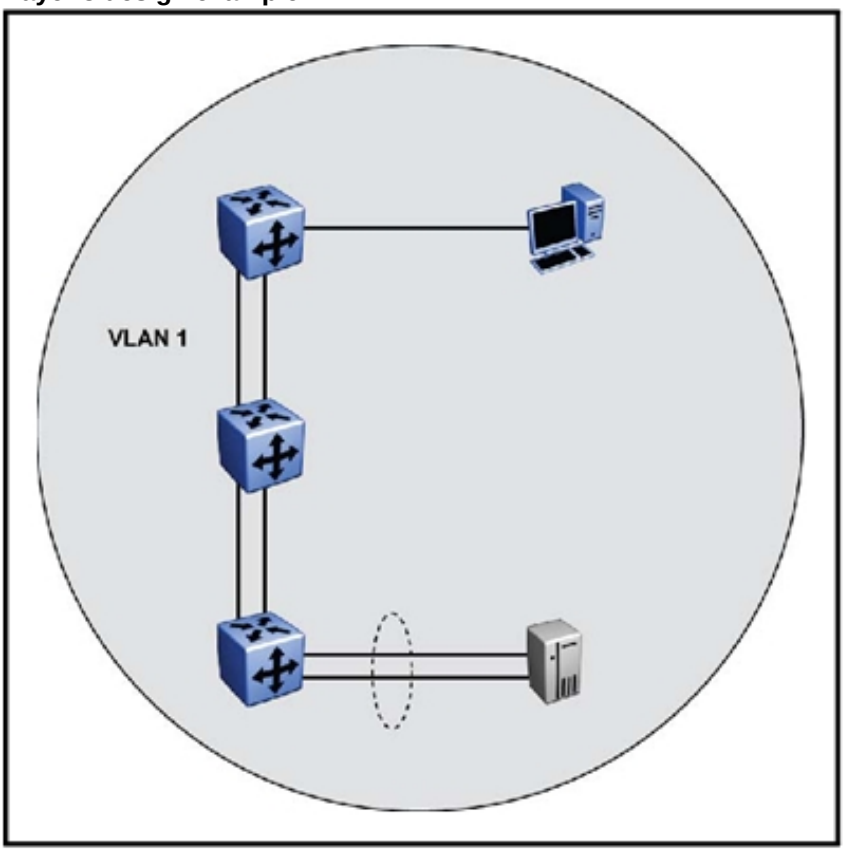
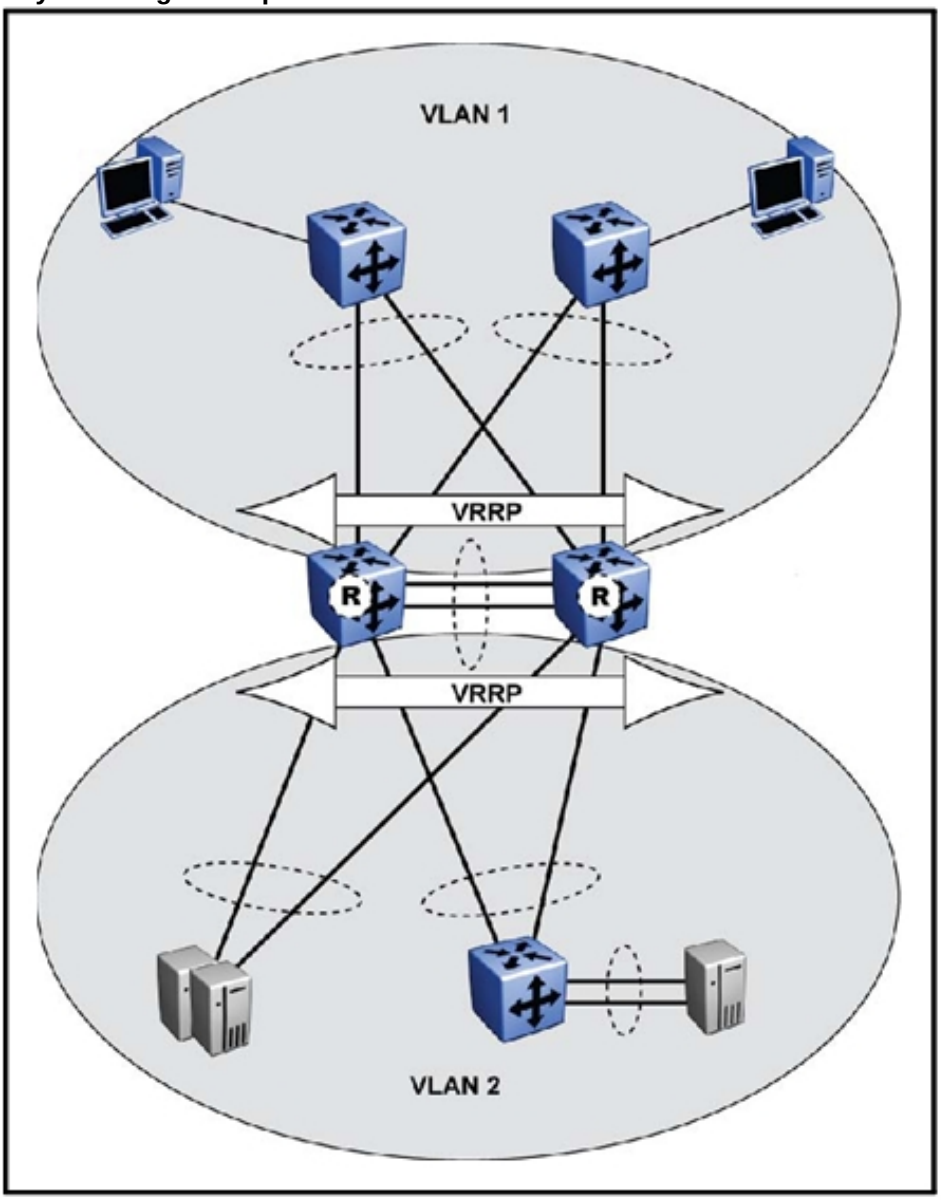


Figure 120
Layer 3 design example 2



In the following figures, DGW denotes Data GateWay.

Figure 121
Layer 3 design example 3

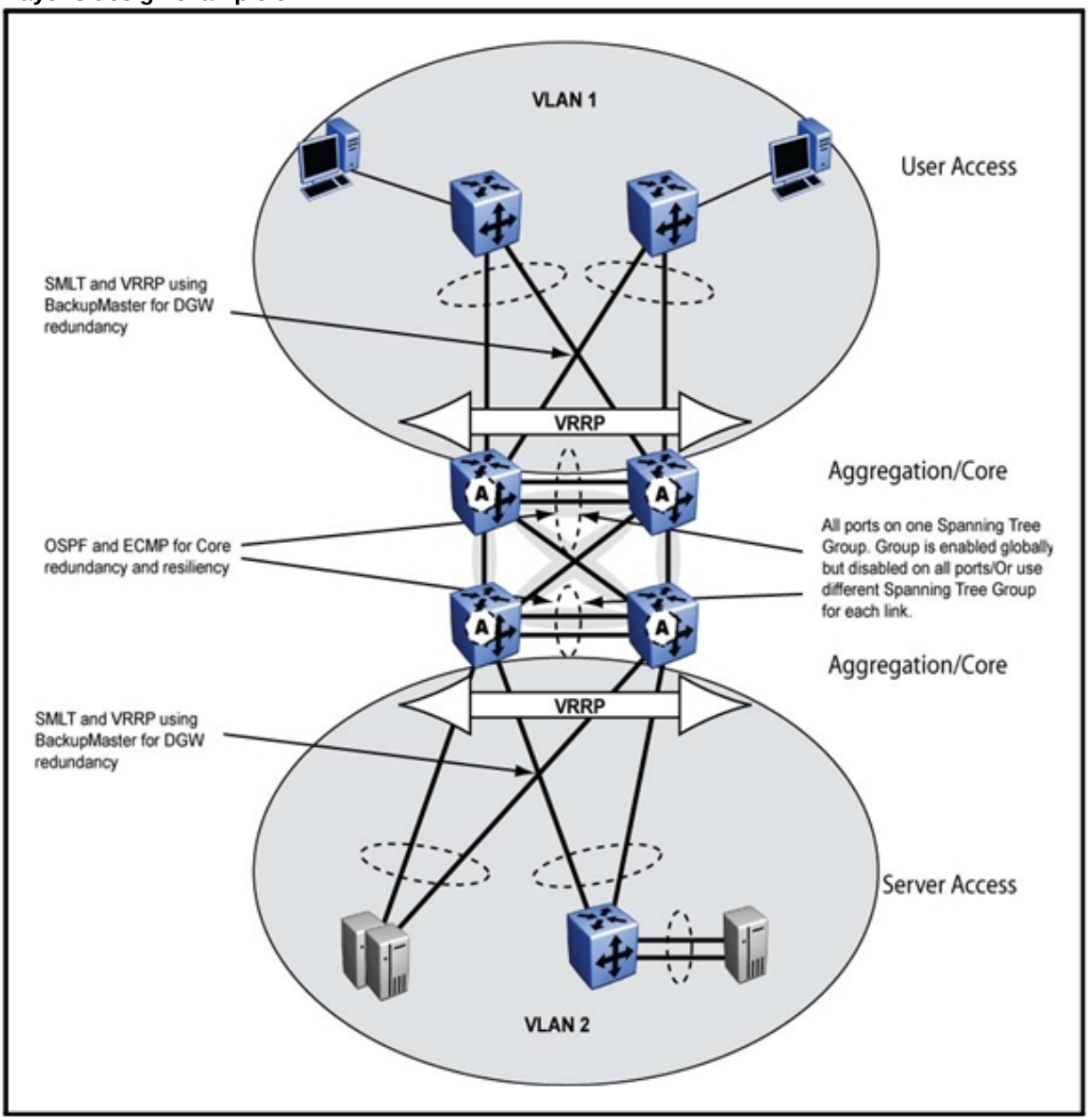
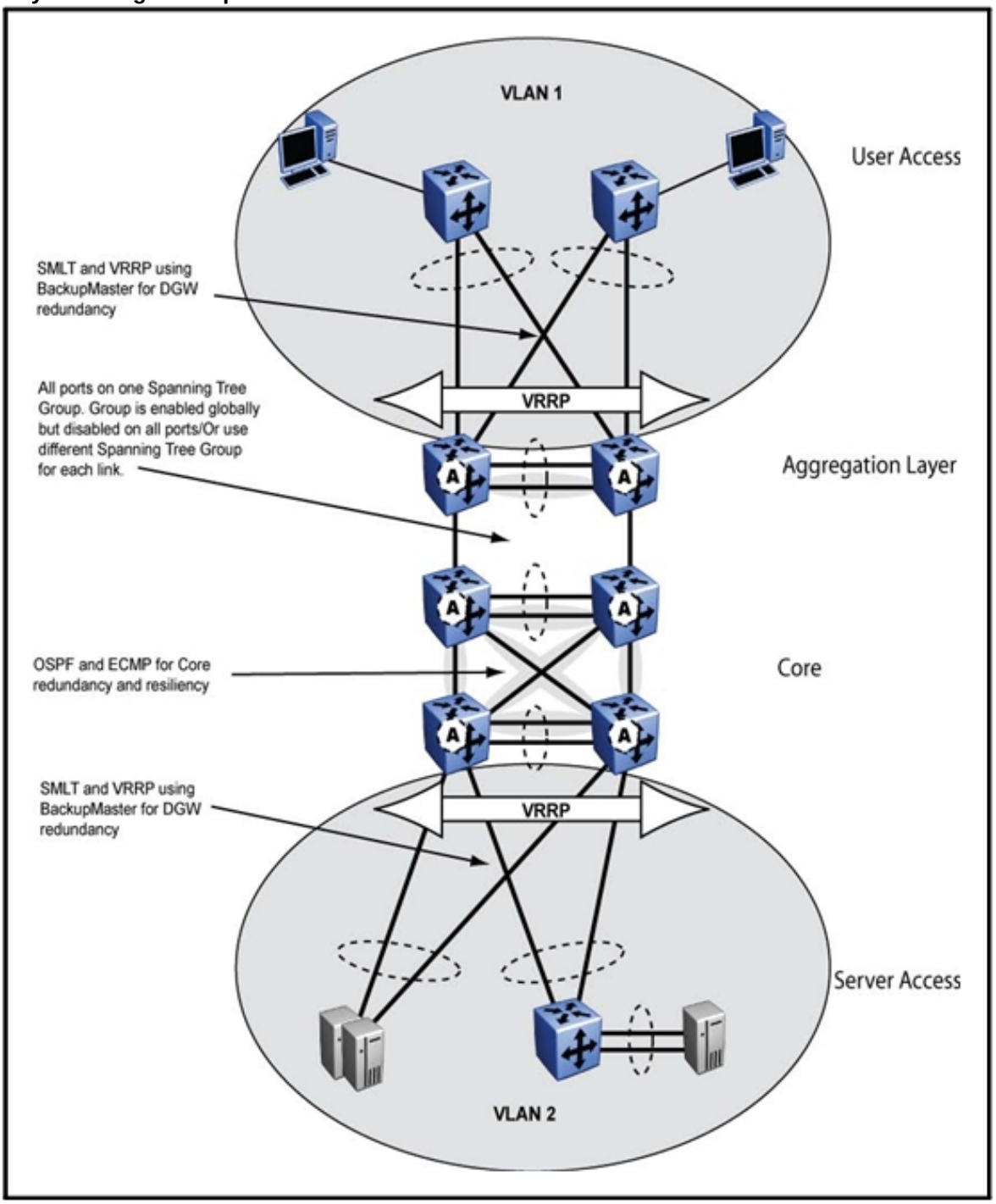


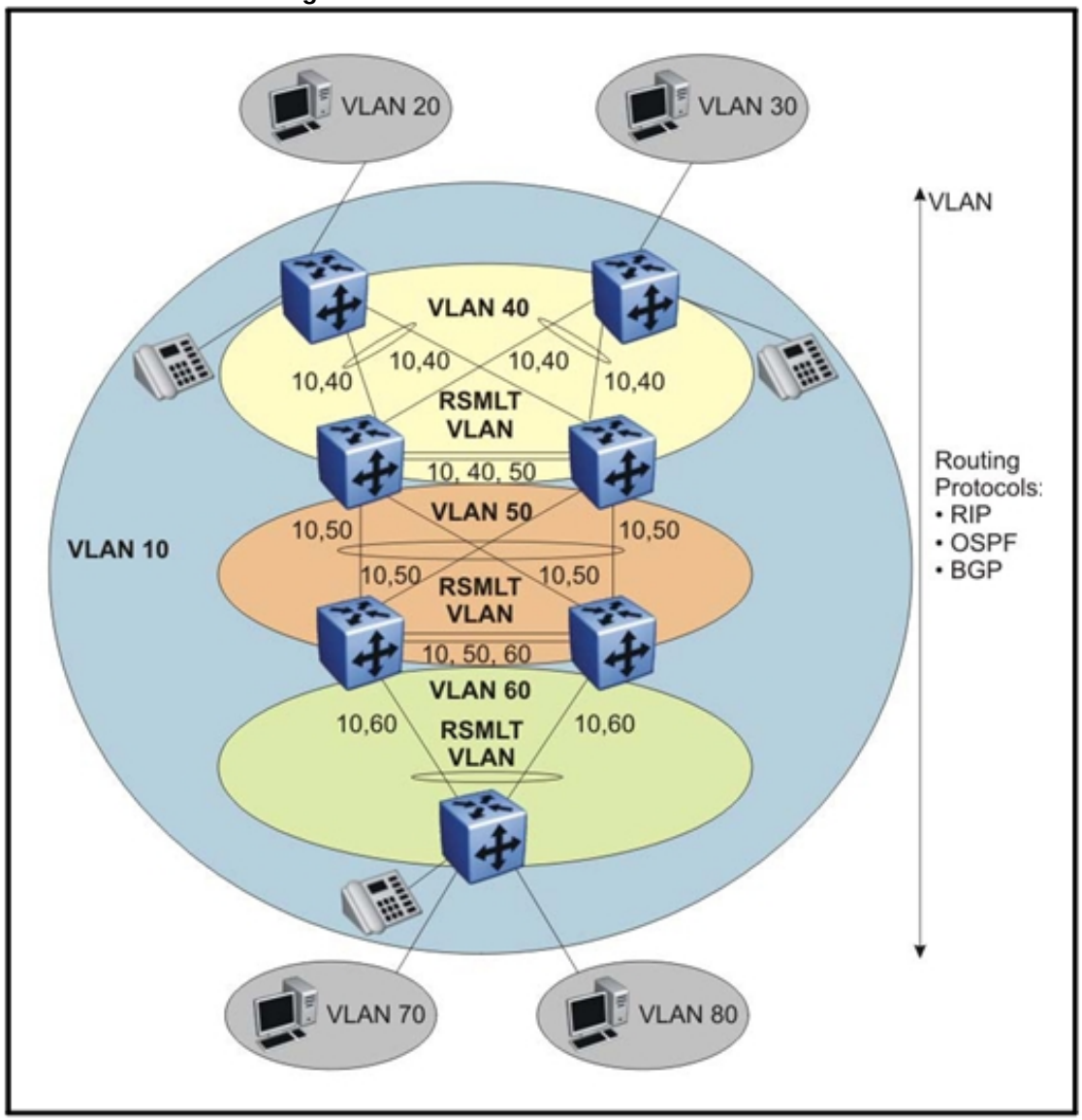
Figure 122
Layer 3 design example 4



RSMLT redundant network with bridged and routed VLANs in the core

In some networks, it is required or desired that a VLAN be spanned through the core of a network (for example, a VoIP VLAN or guest VLAN) while routing other VLANs to reduce the amount of broadcasts or to provide separation. The following figure shows a redundant network design that can perform these functions.

Figure 123
Redundant network design



In this figure, VLAN-10 spans the complete campus network, whereas VLAN 20, 30, 70, and 80 are routed at the wiring closet. VLANs 40, 50, and 60 are core VLANs with RSMLT enabled. These VLANs and their IP

subnets provide subsecond failover for the routed edge VLANs. You can use Routing Information Protocol (RIP), Open Shortest Path First (OSPF) or Border Gateway Protocol (BGP) to exchange routing information. RSMLT and its protection mechanisms prevent the routing protocol convergence time from impacting network convergence time.

All client stations that are members of a VLAN receive every broadcast packet. Each station analyzes each broadcast packet to decide whether the packet are destined for itself or for another node in the VLAN. Typical broadcast packets are Address Resolution Protocol (ARP) requests, RIP updates, NetBios broadcasts, or Dynamic Host Control Protocol (DHCP) requests. Broadcasts increase the CPU load of devices in the VLAN.

To reduce this load, and to lower the impact of a broadcast storm (potentially introduced through a network loop), keep the number of VLAN members below 512 in a VLAN/IP subnet (you can use more clients per VLAN/IP subnet). Then, use Layer 3 routing to connect the VLANs/IP subnets.

You can enable IP routing at the wiring-closet access layer in networks where many users connect to wiring-closets. Most late-model high-end access switches support Layer 3 routing in hardware.

To reduce network convergence time in case of a failure in a network with multiple IP client stations, Nortel recommends that you distribute the ARP request/second load to multiple IP routers/switches. Enabling routing at the access layer distributes the ARP load, which reduces the IP subnet sizes. Figure 123 "Redundant network design" (page 290) shows how to enable routing at the access layer while keeping the routing protocol design robust and simple.

The WSM and Layer 4 to 7 services

Use the Web Switching Module (WSM) to provide layer 4 to 7 services. This section provides information that you need to be aware of when you design networks that use the WSM.

For more information about the WSM, see the following:

- *Nortel Ethernet Routing Switch 8600 Configuration — Web Switching Module using Device Manager* (NN46205-502)
- *Nortel Ethernet Routing Switch 8600 Installation — Modules* (NN46205-304)
- *Nortel Ethernet Routing Switch 8600 Configuration — 8661 SSL Acceleration Module with the Web Switching Module* (NN46205-513)
- *Web OS Switch Software 10.0 Application Guide* () (212777-A)

Navigation

Layer 4 to 7 switching

Layer 4 to 7 switching means that switching is based on higher level protocol header information in the packet. By facilitating deep-packet inspection on Transport Control Protocol (TCP) and User Datagram Protocol (UDP) headers, Layer 4 to 7 switching allows intelligent routing for common applications, including Hypertext Transfer Protocol (HTTP), File Transfer Protocol (FTP), domain name server (DNS), secure socket layer (SSL), Real-Time Streaming Protocol (RTSP), and Lightweight Directory Access Protocol (LDAP).

Layer 4 to 7 switching deals with the intelligent distribution of network traffic and requests across multiple servers or network devices. It permits applications and services to scale, while simultaneously eliminating single points of failure on the network. Layer 4 to 7 switching brings availability, scalability, and fault tolerance to high-performance networks. Intelligent traffic managementcan segregate content across multiple servers and devices, accelerate it, and then prioritize it for delivery across available network resources.

Layer 4 to 7 switching enables at least four major applications for high-performance networks, including:

- Server load balancing
- Global server load balancing
- Firewall and Virtual Private Network (VPN) load balancing
- Transparent cache redirection

The WSM speeds application performance and facilitates the availability and scalability of critical network services by migrating high-level networking functions from software to hardware. By using the WSM, you can perform wire-speed, deep-packet inspection, TCP session analysis, and Intelligent Traffic management.

The WSM provides all the necessary Layer 4 to 7 services including:

- local/global server load balancing
- Web cache redirection
- firewall load balancing
- VPN load balancing
- streaming media load balancing
- Intrusion Detection System (IDS) load balancing
- bandwidth management
- Denial-of-Service (DoS) attack protection
- session persistence
- direct server return
- network failure recovery

The WSM resides inside the Ethernet Routing Switch 8600 as an intelligent module and transforms the Ethernet Routing Switch 8600 into a complete Layer 2-7 intelligent routing solution. Enterprises, service

providers, hosts, content providers, and e-businesses can obtain WebOS traffic management services in a cost-effective, easily customizable input/output (I/O) module.

At the same time, the WSM can aggregate large numbers of 10/100/1000 Ethernet connections to servers, routers, firewalls, caches, and other essential networking devices. The WSM meets the demands of high-performance networks by managing network sessions and real-time load conditions appropriately.

WSM architecture

The WSM takes advantage of the density and robustness of Layer 2 and 3 capabilities on the Ethernet Routing Switch 8600. The WSM provides high-performance intelligent routing based on Layer 4 to 7 information.

The WSM can:

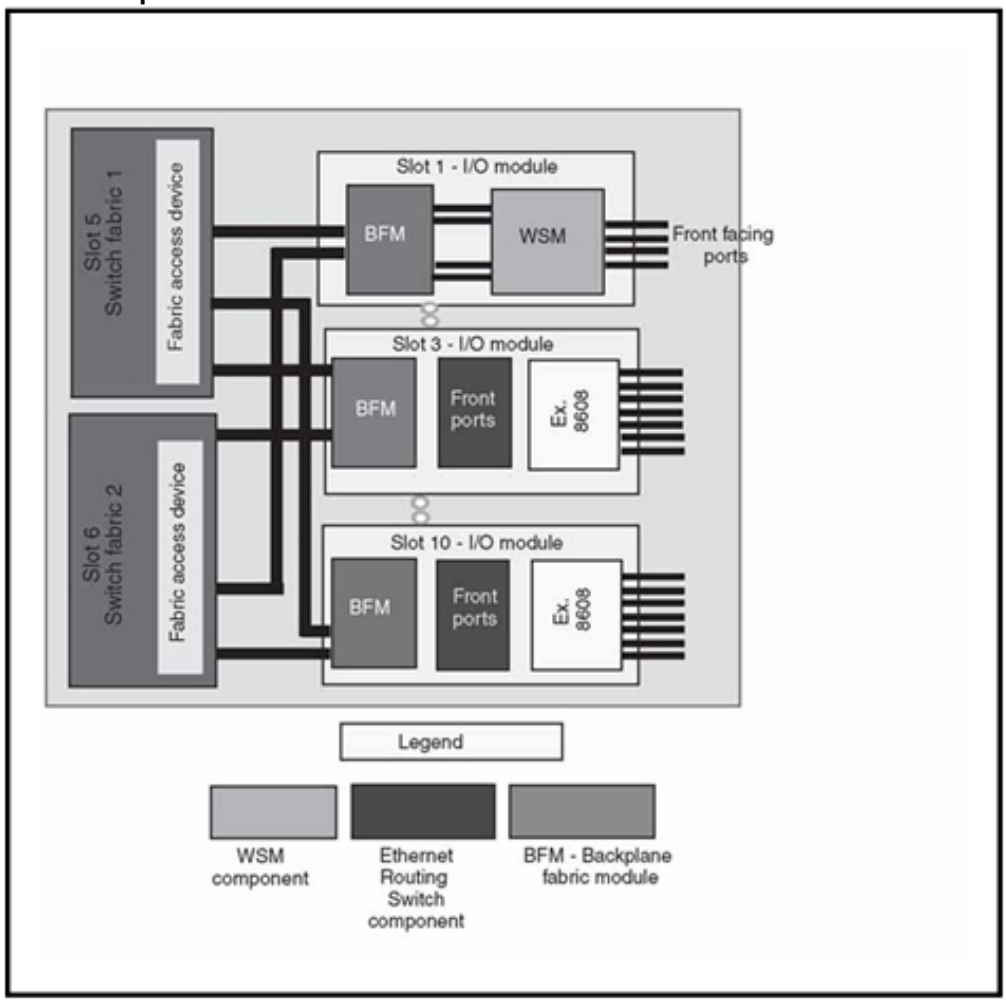
- represent groups of real servers or network devices with a single instance (Virtual IP)
- balance the traffic to a cluster of network devices (server load balancing)
- limit traffic to individual devices or servers (persistent connections) and clusters via specific Layer 4 to 7 policies

Client and server connections through the WSM can use Layer 2 or Layer 3 communication with the Ethernet Routing Switch 8600. Clients connect to the client-side VLAN and servers connect to a unique server-side VLAN. This ensures that traffic does not loop.

Servers and clients can exist on different subnets. Along with the unique two-VLAN approach to processing client and server traffic, the overall configuration process is simplified. By using the WSM default configuration, elements have also been automated to enable easy integration into the Ethernet Routing Switch 8600 environment.

The simplified data path architecture (see the following figure) shows that traffic from a Ethernet Routing Switch 8600 I/O module traverses the Ethernet Routing Switch 8600 to the backplane fabric module (BFM) of the WSM. The Ethernet Routing Switch 8600 connects to the WSM using two dynamically created MLTs tagged as 802.1q. Each MLT consists of two Gigabit links. These MLTs are set up automatically by the Ethernet Routing Switch 8600 when the WSM is initialized.

Figure 124
WSM data path architecture

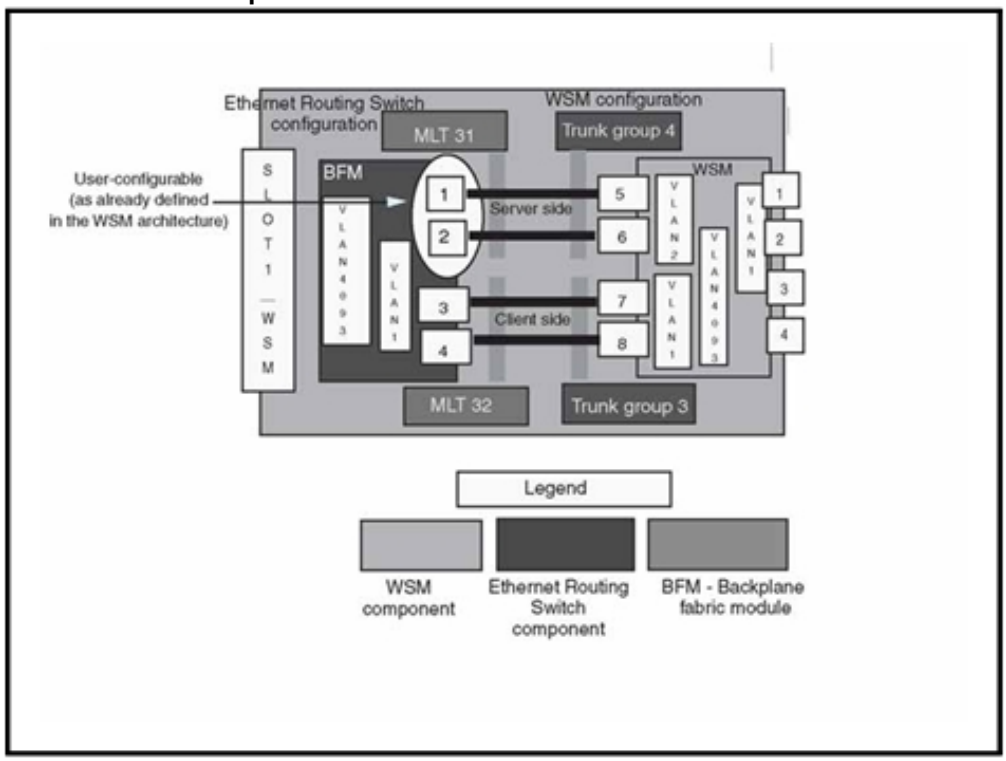


If you connect servers and clients to the Ethernet Routing Switch 8600 I/O module, Nortel recommends that you create two separate VLANs, one for the clients and one for the servers. Then, assign one dynamically-created MLT to each VLAN.

The WSM has four front-facing ports (1, 2, 3, and 4). You can configure each of these at 10/100 Mbit/s via an RJ-45 port or 1000 Mbit/s via an SX port, but not both. The WSM also has four rear-facing Gigabit ports that are used for connectivity to the Ethernet Routing Switch 8600 through the backplane. The WSM has two preconfigured trunks, each of which contains two rear-facing ports.

The following figure shows the detailed WSM data path architecture.

Figure 125
Detailed WSM data path architecture



WSM applications and services

The WSM can improve the performance, scalability, and availability of critical applications and devices in your network.

WSM applications and services navigation

- "WSM and local server load balancing" (page 297)
- "WSM and global server load balancing" (page 299)
- "WSM health metrics" (page 300)
- "WSM and application redirection" (page 301)
- "WSM and VLAN filtering" (page 301)
- "WSM and application abuse protection" (page 302)
- "WSM and Layer 7 deny filters" (page 302)

WSM and local server load balancing

Load balancing offers a cost-effective method to resolve scalability and manageability challenges.

Use server load balancing (SLB) to configure the WSM to balance user-session traffic among a pool of available devices that provide services. SLB benefits your network by providing:

- increased efficiency for server utilization and network bandwidth

  With SLB, your Ethernet Routing Switch 8600 is aware of the shared services provided by your server pool and can balance user session traffic among the available and appropriate resource. Important session traffic is given priority, thus reducing user competition for connections on overutilized devices. For greater control, traffic is distributed according to a variety of user-selectable rules.

- increased reliability and availability of services to users

  If any device in a server pool fails, the remaining servers continue to provide access to vital applications and data. You can bring the failed device back without interrupting access to services.

- increased scalability of services

  As users are added and server capabilities become saturated, you can seamlessly add new servers to the existing network

The WSM acts as the front end to servers and network devices by interpreting user sessions requests and distributing them among the available and appropriate resources. Load balancing via the WSM is performed in the following three ways.

Virtual server-based load balancing

Virtual server-based load balancing is the traditional load balancing method. You can configure the WSM to act as a virtual server. Iit is given a virtual server IP address (or range of addresses) for each collection of services it distributes. You can have as many as 255 virtual servers on the switch, each distributing up to eight different services (up to a total of 2048 services).

Each virtual server is assigned a list of IP addresses of the real servers in the pool where its services reside. When you request a connection to a service, you communicate with a virtual server on the WSM.

When the WSM receives your request, it binds the session to the IP address of the best available resource and remaps the fields in each frame from virtual addresses to real addresses. IP, FTP, RTSP, and static session WAP are examples of some of the services that use virtual servers for load balancing.

Filter-based load balancing

use a filter to control the types of traffic permitted through the WSM. Configure filters to allow, deny, or redirect traffic according to IP address, protocol, or Layer 4 port criteria. In filtered-based load balancing, use a filter to redirect traffic to a real server group.

If you configure the group with more than one real server entry, redirected traffic is load balanced among the available real servers in the group. Firewall load balancing, WAP with RADIUS snooping, and IDS and WAN links use redirection filters to load balance traffic

Content-based load balancing

Content-based load balancing uses Layer 7 application data such as URLs, cookies, and host headers to make intelligent load balancing and routing decisions. URL-based load balancing, browser-smart load balancing, and cookie-based preferential load balancing are a few examples of content load balancing.

Another key element of SLB is the determination of the health and availability of each real server or device. By default, the WSM checks each service on each real server every two seconds. If a service does not respond to four consecutive health checks, the WSM declares the service unavailable.

WSM and global server load balancing

You can enable global server load balancing (GSLB) via a license on the WSM. GSLB overcomes many scalability, availability, and performance issues that are inherent in distributing content across multiple geographic locations. use GSLB to balance server traffic load across multiple physical sites. The WSM GSLB implementation takes into account individual site health, response time, and geographic location. It then integrates the resources of the dispersed server sites for complete global performance.

GSLB also enables enterprises to meet the demand for higher content availability by distributing content and decision making. In this way, it ensures that the best-performing site receives the majority of traffic, thus enabling network administrators to build and control content by user, location, target application, and so on.

On the WSM, GSLB is based on the domain name server (DNS) and proximity by source IP address. Each WSM is capable of responding to client resolution requests with a list of addresses of distributed sites, prioritized by performance, geography, and other criteria.

WSM health metrics

The WSM uses health metrics to determine the most appropriate real server to receive and service client connections. The following table provides information about several of the available metrics. For more information about these and the other available metrics, see *Web OS Switch Software 10.0 Application Guide* () (212777-A).

Table 29
Health checking metrics

Metric	Description
Minmisses	Optimized for application redirection. This metric uses the IP address information in the client request to select a server. Based on its calculated score, the server that is most available is assigned the connection. This metric attempts to minimize the disruption of persistency when servers are removed from service. Use this metric only when persistence is a must.
Hash	Uses the destination IP address for application redirection, the source IP for SLB, and both for firewall load balancing. It ensures that requests are sent to the same server to: • maximize successful cache hits • ensure that client information is retained between sessions • ensure that unidirectional flows of a given session are redirected to the same firewall
Least connections	Uses the number of connections currently open on each real server in real time to determine which one receives the request. The server with the fewest connections is considered the best choice.
Round robin	Issues new connections to each server in turn. When all the real servers in a group receive at least one connection, the issuing process starts over.
Response time	Uses real server response time to assign sessions to servers. The WSM monitors and records the amount of time it takes for each server to reply to the health check and adjusts the real server weights. In such a scenario, a server with half the response time as another server receives a weight twice as high. Thus, the server receives more requests.
Bandwidth	Uses the octet counts to assign sessions. The servers that process more octets are considered to have less available bandwidth. The higher the bandwidth used, the smaller the weight assigned to the server. The next request goes to the real server with the highest amount of free bandwidth. This bandwidth metric requires identical servers with identical connections.

WSM and application redirection

Application redirection improves network bandwidth utilization and provides unique network solutions. You can create filters to redirect traffic to cache and application servers, improving speed of access to common Web or application content, which in turn frees up valuable network bandwidth.

Application redirection helps to reduce traffic congestion by intercepting outbound client requests and redirecting them to a group of application or cache servers on a local networks. If the WSM recognizes the request as one that a local network device can handle, it routes it locally instead of sending it across the Internet.

In addition to increasing the efficiency of a network, the WSM allows clients to access information quickly and lowers WAN access costs.

The WSM also supports content-intelligent application redirection. A network administrator can redirect requests based on HTTP header information. The following table lists the available types of application redirection.

Table 30
Application redirection types

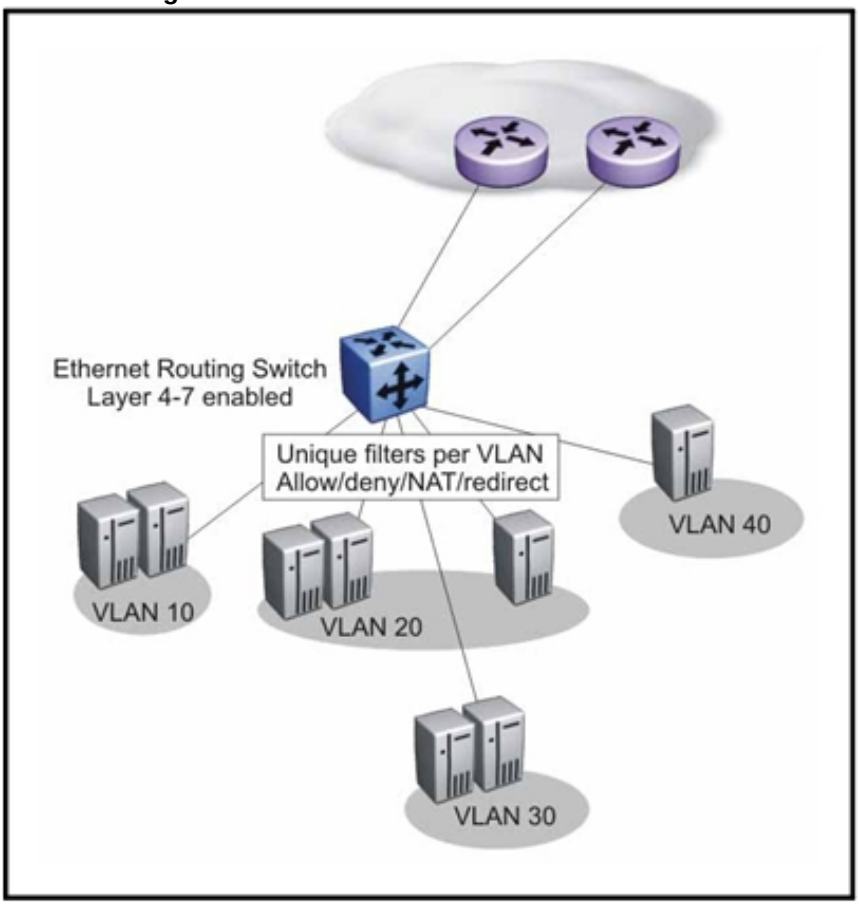
Application redirection type	Description
URL-based	Separates static and dynamic content requests and provides you with the ability to send requests for specific URLs or URL strings to designated cache devices. The WSM off loads the overhead processing from the cache server and only sends appropriate requests to the cache server farm.
HTTP header-based	Defines host names and string IDs that are redirected to cache server farms.
Browser-based	Configure the user-agent to determine if a client request is redirected to a cache or server farm. Thus, you can send different browser types to the appropriate sites locally and on the internet (see the following figure).

WSM and VLAN filtering

On the WSM, you can apply filters per-switch, per-port, or per-VLAN. With VLAN-based filtering, a single WSM can provide differentiated services for multiple groups, customers, users, or departments.

For example, you can define separate filters for the finance department and the marketing department by using the same WSM for two different VLANs. The following figure shows how you can assign different filters to unique VLANs that allow, deny, or redirect client requests, thus enabling differentiated service per group.

Figure 126
VLAN filtering



WSM and application abuse protection

You can use the WSM to prevent a client or group of clients from claiming all the TCP or application resources on servers. You do so by monitoring the rate of incoming requests for connections to a virtual IP address and limiting the client request with a known set of IP addresses.

You ensure application abuse protection by defining the maximum number of TCP connection requests that are allowed within a configured time window. The WSM monitors the number of new TCP connections. If the number exceeds the configured limit, any new TCP connections are blocked or held down.

WSM and Layer 7 deny filters

The WSM can secure your network from virus attacks by monitoring for potential offending string patterns (for example, HTTP URL requests). The WSM examines the HTTP content of the incoming client request for a matching pattern.

If the matching virus pattern is found, the packet is dropped, and a reset frame is sent to the offending client. Syslog messages and an SNMP trap are generated to warn of a possible attack, while back-end devices and servers are automatically protected because the request is denied at the WSM ingress port.

WSM network architectures

This section describes various network architectures that are available for you to use when configuring the Ethernet Routing Switch 8600 and WSM for Layer 2 to Layer 7 processing.

These architectures are not exhaustive. However, they do reflect the most common configurations. In most cases, you can mix and match the methods to accommodate specific requirements. The purpose of this section is to provide you with a framework of the various methods available.

The following architectures are based on a server load balancing example. They use VLAN 1 for client processing and VLAN 2 for server processing.

WSM network architectures navigation

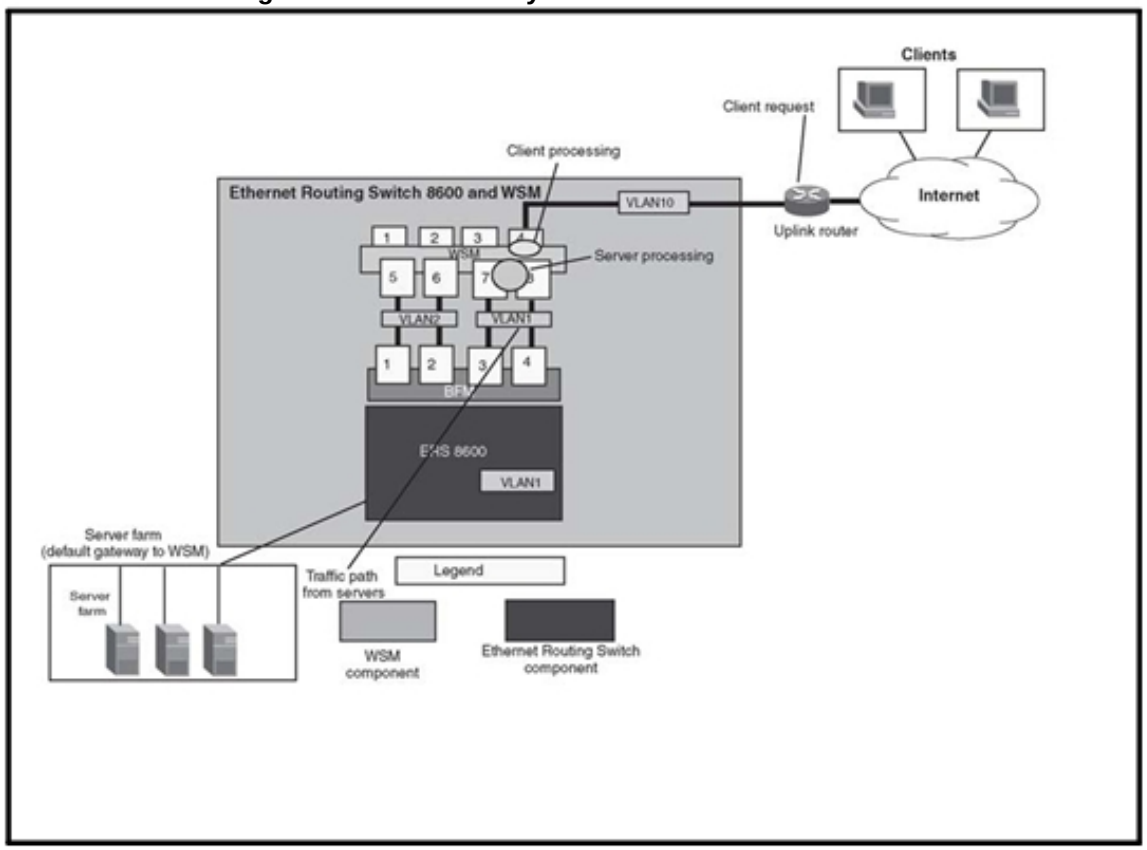
- "Ethernet Routing Switch 8600 as a Layer 2 switch" (page 303)
- "Layer 3 routing" (page 304)
- "Layer 4 to 7 service implementation with a single Ethernet Routing Switch 8600" (page 305)
- "Layer 4 to 7 service implementation with dual Ethernet Routing Switch 8600s" (page 306)

Ethernet Routing Switch 8600 as a Layer 2 switch

Most architectures use the Ethernet Routing Switch 8600 as a Layer 3 device to route traffic from the client and server to the WSM. Occasionally, you may need to implement Layer 4 to 7 services and applications using the Ethernet Routing Switch 8600 as a Layer 2 switch. Such occasions arise if you aggregate optical Ethernet connections.

The sample architecture shown in the following figure shows traffic entering an Ethernet Routing Switch 8600 I/O module and traversing the backplane at Layer 2 to the WSM. In this example, client requests come from the Internet using an uplink router connected to the WSM front-facing port server farm. In turn, this server farm is connected to the Ethernet Routing Switch 8600 I/O module.

Figure 127
The Ethernet Routing Switch 8600 as a Layer 2 switch



VLAN 1 is created in the Ethernet Routing Switch 8600, and the backplane forwarding module ports 3 and 4 (WSM dynamic MLT) are assigned to VLAN 1. An IP address is assigned to VLAN 1 in the WSM, consisting of Ports 7 and 8. The servers point to the IP interface in the WSM as their default gateway. The Ethernet Routing Switch 8600 provides a Layer 2 switching path for the servers that are connected to the I/O module.

Layer 3 routing

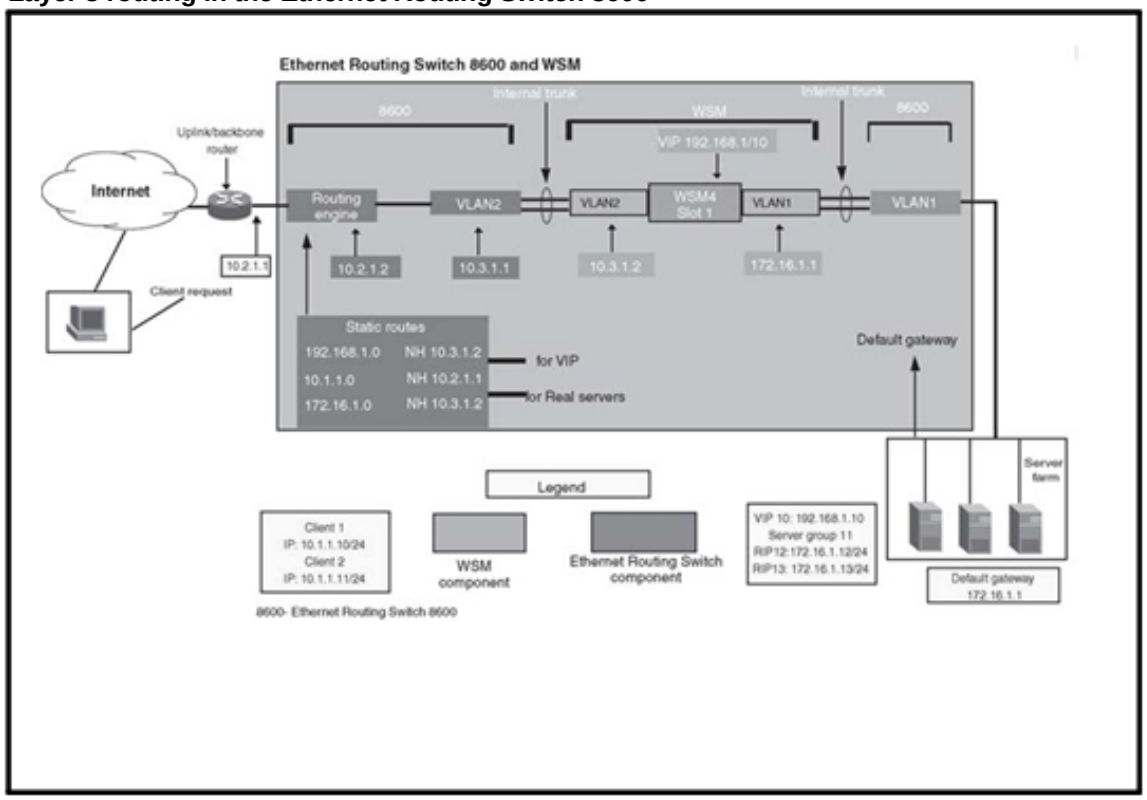
This architecture uses the Ethernet Routing Switch 8600 Layer 3 routing engine to direct traffic to the WSM. In this configuration, client traffic is aggregated elsewhere and is routed or switched to the Ethernet Routing Switch 8600 and WSM. The routing engine of the Ethernet Routing Switch 8600 appropriately routes traffic to the WSM.

In the following figure, the client initiates a request to access a VIP link that first traverses the uplink router. This request is forwarded to the Ethernet Routing Switch 8600 and enters the switch on one of the I/O modules. The routing engine makes a decision on the next-hop based on static-route entries. A static route is created so that all traffic destined for the VIP is forwarded to the WSM.

In this example, the routing engine forwards the packet to a WSM interface in VLAN 2 where Layer 4 to 7 processing occurs. The WSM selects a real server and routes the request out of the VLAN that houses the server. On egress, the traffic is sent out of VLAN 1 and across the backplane to the appropriate server connected to the Ethernet Routing Switch 8600 I/O module in VLAN 1.

This design utilizes the Layer 2 switching and the Layer 3 routing engine of the Ethernet Routing Switch 8600, as well as the Layer 4 to 7 switching and server load balancing capabilities of the WSM.

Figure 128
Layer 3 routing in the Ethernet Routing Switch 8600



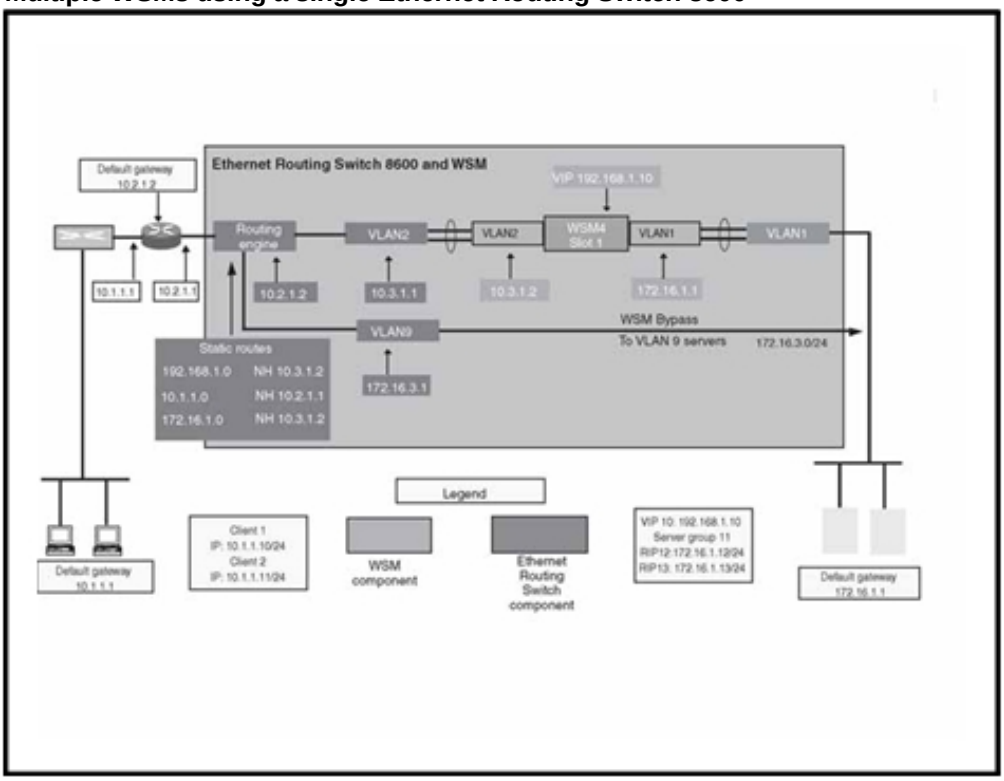
Layer 4 to 7 service implementation with a single Ethernet Routing Switch 8600

The following architecture provides you with a high availability scenario by using a single Ethernet Routing Switch 8600 with multiple WSMs operating in active/standby redundancy mode. From a price standpoint, it is common to use redundant modules and fabrics instead of an entire switch. Sometimes module failover is preferable to an entire network path switch failover.

To offer high-availability, this architecture runs over two instances of Virtual Router Redundancy Protocol (VRRP—one for client access and one for server access) on the WSMs (see the following figure). In this configuration, VRRP on the WSMs communicates over the Ethernet Routing Switch 8600 backplane; thus it reconfigures dynamic MLT connections to every WSM installed in the chassis.

Ensure that VRRP communications occur over an available data path in the event a WSM VRRP Master fails. If it does fail, the Standby WSM can become the Master.

Figure 129
Multiple WSMs using a single Ethernet Routing Switch 8600



Layer 4 to 7 service implementation with dual Ethernet Routing Switch 8600s

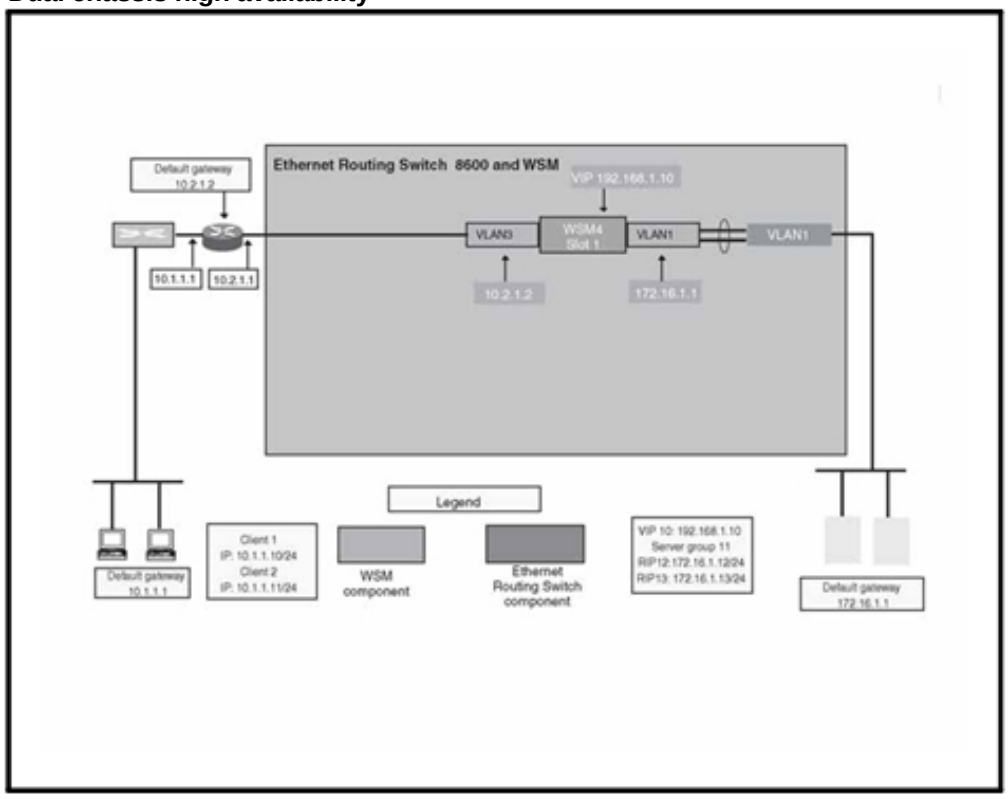
The following architecture utilizes a pair of Ethernet Routing Switch 8600s with multiple WSMs installed to offer a full nodal redundancy, high-availability solution.

This architecture creates a single network route that provides hot-standby access to the Ethernet Routing Switch 8600 for Layer 4 to 7 services. Each of the client-side and server-side routers communicates with a VRRP

instance that runs between the WSMs. These instances determine which Ethernet Routing Switch 8600 and WSM is the Master and which is the Backup.

In the following figure, VRRP is implemented along the data path on the front-end, and out of the data path on the back-end. This ensures that a failure of any component along the data path triggers a failover. This implementation avoids the situation where the inter-switch link on VLAN 1 fails, causing a failover when a failover is not required.

Figure 130
Dual chassis high availability



The simplest method to use to configure servers in a high-availability mode is to employ Network Interface Card (NIC) teaming. In NIC teaming, two NICs share the same IP address, which permits switchover to a live element if the interfacing switch, line, or NIC fails. In this implementation, configure a single IP address that corresponds to a single virtual MAC address. Because the IP address and MAC address do not change, upstream and downstream network devices do not need to perform updates. Even if the Master WSM fails, traffic traverses the VLAN 1 link as long as the Ethernet Routing Switch 8600 runs.

WSM considerations

Only Device Manager version 5.5.x and later supports the Ethernet Routing Switch 8600 and WSM. If you use any version earlier than 5.5, you can adversely affect the automatic configuration of the WSM.

The unknown MAC discard feature discards and prevents any unknown MAC addresses from accessing specific ports. However, If you enable the unknown MAC discard feature on BFM ports 3 and/or 4, connectivity to the WSM is lost.

This results in warning messages similar to the following:

```
[09/13/02 16:56:49] WARNING Task=tRcIpTask An intrusion MAC
address:00:60:cf:50:52:60 at port 2/3
[09/13/02 16:57:37] WARNING Task=tCppRxTask An intrusion MAC
address:00:50:8b:d3:4e:fd at port 2/4
```

If you enable unknown MAC discard on BFM ports, this prevents you from connecting to the WSM. To restore the connection, you must disable the feature on both BFM ports x/3 and x/4, or you must configure the switch to allow specific MAC addresses. You can manually enable WSMs with the unknown MAC discard feature by configuring known MAC addresses (add-allow-mac parameter). This prevents unwanted network devices (such as sniffers) from accessing the network.

WAN link load balancing is only supported though the front-facing ports of the WSM: WAN link load balancing requires a proxy IP address (PIP). You cannot apply a PIP to a trunk group or MLT 31 or 32 of the BFM ports.

Hot-standby mode is not supported on MLT 4 or on rear-facing ports 7 and 8 of the WSM because it causes the switch to lose connectivity.

The WSM allows you to define a port as *hotstan* (hot-standby). When hot-standby is enabled on a port, the hot-standby algorithm controls the forwarding state. Essentially, this algorithm puts the Master VRRP switch in forwarding mode and blocks the Backup switch.

If you configure hot-standby mode on backplane ports 7 or 8, the backup switch loses connectivity because the hot-standby algorithm disables the backup switch management ports.

Network security

The information in this section helps you to design and implement a secure network.

You must provide security mechanisms to prevent your network from attack. If links become congested due to attacks, you can immediately halt end-user services. During the design phase, study availability issues for each layer. See “Redundant network design” (page 75) for more information. Without redundancy, all services can be brought down.

To provide additional network security, you can use the Nortel Contivity VPN product suite, the Shasta 5000 BSN, or the Ethernet Routing Switch Firewall and Intrusion Sensor. They offer differing levels of protection against Denial of Service (DoS) attacks through either third party IDS partners, or through their own high-performance stateful firewalls.

Navigation

DoS protection mechanisms

The Ethernet Routing Switch is protected against Denial-of-Service (DoS) attacks by several internal mechanisms and features.

DoS protection mechanisms navigation

Broadcast and multicast rate limiting

To protect the switch and other devices from excessive broadcast traffic, you can use broadcast and multicast rate limiting on a per-port basis. Use broadcast and multicast rate limiting on Classic modules.

For more information about setting the rate limits for broadcast or multicast packets on a port, see *Nortel Ethernet Routing Switch 8600 Configuration — QoS and IP Filtering for Classic Modules* (NN46205-508) .

Directed broadcast suppression

You can enable or disable forwarding for directed broadcast traffic on an IP-interface basis. A directed broadcast is a frame sent to the subnet broadcast address on a remote IP subnet. By disabling or suppressing directed broadcasts on an interface, you cause all frames sent to the subnet broadcast address for a local router interface to be dropped. Directed broadcast suppression protects hosts from possible DoS attacks.

To prevent the flooding of other networks with DoS attacks, such as the Smurf attack, the Ethernet Routing Switch 8600 is protected by directed broadcast suppression. This feature is enabled by default. Nortel recommends that you not disable it.

For more information about directed broadcast suppression, see *Nortel Ethernet Routing Switch 8600 Security* (NN46205-601) .

Prioritization of control traffic

The Ethernet Routing Switch 8600 uses a sophisticated prioritization scheme to schedule control packets on physical ports. This scheme involves two levels with both hardware and software queues to guarantee proper handling of control packets regardless of the switch load. In turn, this guarantees the stability of the network. Prioritization also guarantees that applications that use many broadcasts are handled with lower priority.

You cannot view, configure, or modify control traffic queues.

CP-Limit recommendations

CP-Limit prevents the CPU from overload by excessive multicast or broadcast control or exception traffic. This ensures that broadcast storms do not impact the stability of the system. By default, CP-Limit protects the CPU from receiving more than 14 000 broadcast/multicast control or exception packets per second within a duration that exceeds 2 seconds.

You can disable CP-Limit and instead, configure the amount of broadcast and/or multicast control or exception frames per second that are allowed to reach the CPU before the responsible interface is blocked and disabled. Based on your environment (severe corresponds to a high-risk environment), the recommended values are shown in the following figure.

Table 31
Recommended CP-Limit values

	Broadcast	Multicast
Severe:		
Workstation (PC)	1000	1000
Server	2500	2500
NonIST Interconnection	7500	6000
Moderate:		
Workstation (PC)	2500	2500
Server	5000	5000
NonIST Interconnection	9000	9000
Relaxed:		
Workstation (PC)	4000	4000
Server	7000	7000
NonIST Interconnection	10000	10000

ARP request threshold recommendations

The Address Resolution Protocol (ARP) request threshold limits the ability of the Ethernet Routing Switch 8600 to source ARP requests for workstation IP addresses it has not learned within its ARP table. The default setting for this function is 500 ARP requests per second. To avoid excessive amounts of subnet scanning caused by a virus (like Welchia), Nortel recommends that you change the ARP request threshold to a value between 100 to 50. This helps to protect the CPU from causing excessive

ARP requests, helps to protect the network, and lessens the spread of the virus to other PCs. The following list gives further ARP threshold recommendations:

- Default: 500
- Severe conditions: 50
- Continuous scanning conditions: 100
- Moderate: 200
- Relaxed: 500

For Ethernet Routing Switch 8600 Release 3.2.2.2 to 3.5.0.0, you can access ARP request threshold feature only through VxWorks shell. Within the shell, the designation is arq_threshold. For more information, contact a Nortel support engineer.

From Release 3.5.0 and later, you can access the feature through the CLI. For more information about the **`config ip arp arpreqthreshold`** command, see *Nortel Ethernet Routing Switch 8600 Configuration — IP Routing Operations* (NN46205-523) .

Multicast Learning Limitation

The Multicast Learning Limitation feature protects the CPU from multicast data packet bursts generated by malicious applications. If more than a certain number of multicast streams enter the CPU through a port during a sampling interval, the port is shut down until the user or administrator takes the appropriate action.

For more information and configuration instructions, see *Nortel Ethernet Routing Switch 8600 Configuration — IP Multicast Routing Protocols* (NN46205-501) .

Damage prevention

To further reduce the chance that your network can be used to damage other existing networks, take the following actions:

1. Prevent IP spoofing.

   You can use the spoof-detect feature.

2. Prevent your network from being used as a broadcast amplification site.

3. To block illegal IP addresses, enable the **`hsecure`** flag (High Secure mode).

   For more information, see "High Secure mode" (page 314) or *Nortel Ethernet Routing Switch 8600 Security* (NN46205-601) .

Packet spoofing

You can stop spoofed IP packets by configuring the switch to only forward IP packets that contain the correct source IP address of your network. By denying all invalid source IP addresses, you minimize the chance that your network is the source of a spoofed DoS attack.

A spoofed packet is one that comes from the Internet into your network with a source address equal to one of the subnet addresses used on your network. Its source address belongs to one of the address blocks or subnets used on your network. To provide spoofing protection, you can use a filter that examines the source address of all outside packets. If that address belongs to an internal network or a firewall, the packet is dropped.

To prevent DoS attack packets that come from your network with valid source addresses, you need to know the IP network blocks that are in use. You can create a generic filter that:

- permits valid source addresses
- denies all other source addresses

To do so, configure an ingress filter that drops all traffic based on the source address that belongs to your network.

If you do not know the address space completely, it is important that you at least deny Private (see RFC1918) and Reserved Source IP addresses. The following table lists the source addresses that you should filter.

Table 32
Source addresses that need to be filtered

Address	Description
0.0.0.0/8	Historical Broadcast. High-Secure mode blocks addresses 0.0.0.0/8 and 255.255.255.255/16. If you enable this mode, you do not have to filter these addresses.
10.0.0.0/8	RFC1918 Private Network
127.0.0.0/8	Loopback
169.254.0.0/16	Link Local Networks
172.16.0.0/12	RFC1918 Private Network
192.0.2.0/24	TEST-NET
192.168.0.0/16	RFC1918 Private Network
224.0.0.0/4	Class D Multicast
240.0.0.0/5	Class E Reserved

Table 32
Source addresses that need to be filtered (cont'd.)

Address	Description
248.0.0.0/5	Unallocated
255.255.255.255/32	Broadcast1

You can also enable the spoof-detect feature on a port.

For more information about the spoof-detect feature, see *Nortel Ethernet Routing Switch 8600 Configuration — VLANs and Spanning Tree* (NN46205-517) .

You can also use the R series module predefined Access Control Template (ACT) for ARP spoof detection. For more information about this ACT, see *Nortel Ethernet Routing Switch 8600 Configuration — QoS and IP Filtering for R and RS Modules* (NN46205-507) .

High Secure mode

To ensure that the Ethernet Routing Switch 8600 does not route packets with an illegal source address of 255.255.255.255 (per RFC 1812 Section 4.2.2.11 and RFC 971 Section 3.2), you can enable High Secure mode.

By default, this feature is disabled. When you enable this flag, the feature is applied to all ports belonging to the same OctaPid (group of 8 10/100 Mbit/s ports [8648 modules], 1 Gbit/s port [8608 modules] or 2 1-Gbit/s ports [8616 modules]).

For more information about hsecure, see *Nortel Ethernet Routing Switch 8600 Security* (NN46205-601) .

For more information about Classic modules and OctaPID assignments, see *Nortel Ethernet Routing Switch 8600 Configuration — VLANs and Spanning Tree* (NN46205-517) .

Security and redundancy

Redundancy in hardware and software is one of the key security features of the Ethernet Routing Switch 8600. High availability is achieved by eliminating single points of failure in the network and by using the unique features of the Ethernet Routing Switch 8600 including:

- acomplete,redundanthardwarearchitecture(switchingfabricsin load sharing, CPU in redundant mode or High Availability [HA] mode, redundant power supplies)
- hot swapping of all elements (I/O blades, switching fabrics/CPUs, power supplies)

- flash cards (PCMCIA) to save multiple config/image files
- alistofsoftwarefeaturesthatallowhighavailabilityincluding:
  — link aggregation (MLT, distributed MLT, and 802.3ad)
  — dual-homing of edge switches to two core switches (SMLT and RSMLT)
  — unicast dynamic routing protocols (RIPv1, RIPv2, OSPF, BGP-4) — multicast dynamic routing protocols (DVMRP, PIM-SM, PIM-SSM) — distribution of routing traffic along multiple paths (ECMP)
  — router redundancy (VRRP)

For a review of various security attacks that could occur in a Layer 2 network, and solutions, see *Layer Security Solutions for ES and ERS Switches Technical Configuration Guide*. This document is available on the Nortel Technical Support Web site in the Ethernet Routing Switch 8600 documentation.

Data plane security

Data plane security mechanisms include the Extended Authentication Protocol (EAP) 802.1x, VLANs, filters, routing policies, and routing protocol protection. Each of these is described in the sections that follow.

Data plane security navigation

- "EAP" (page 315)
- "VLANs and traffic isolation" (page 317)
- "Security at layer 2" (page 317)
- "Security at Layer 3: announce and accept policies " (page 318)
- "Routing protocol security" (page 319)

EAP

To protect the network from inside threats, the switch supports the 802.1x standard. EAP separates user authentication from device authentication. If EAP is enabled, end-users must securely logon to the network before obtaining access to any resource.

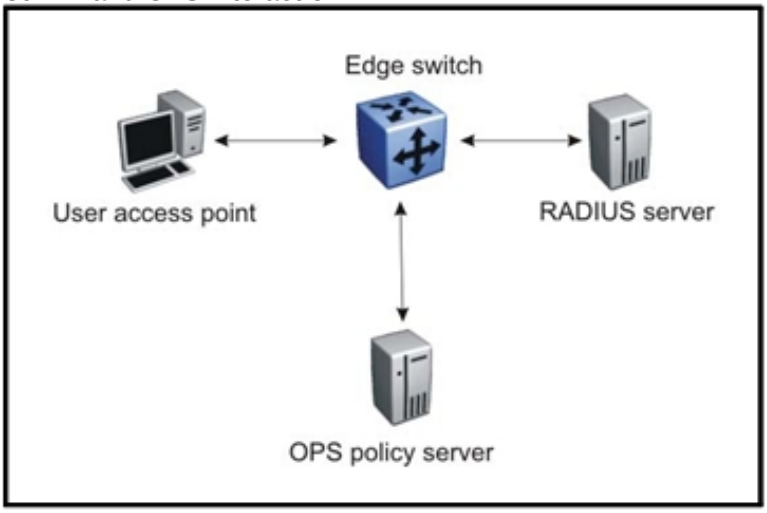
Interaction between 802.1x and Optivity Policy Server v4.0

User-based networking links EAP authorization to individual user-based security policies based on individual policies. As a result, network managers can define corporate policies and configure them on a per-port basis. This adds additional security based on a logon and password.

The Nortel Optivity Policy Server supports 802.1x EAP authentication against RADIUS and other authentication, authorization, and accounting (AAA) repositories. This support helps authenticate the user, grants access to specific applications, and provides real time policy provisioning capabilities to mitigate the penetration of unsecured devices.

The following figure shows the interaction between 802.1x and Optivity Policy Server. First, the user initiates a logon from a user access point and receives a request/identify request from the switch (EAP access point). The user is presented with a network logon. Prior to DHCP, the user does not have network access because the EAP access point port is in EAP blocking mode. The user provides User/Password credentials to the EAP access point via Extensible Authentication Protocol Over LAN (EAPoL). The client PC is considered both a RADIUS peer user and an EAP supplicant.

Figure 131
802.1x and OPS interaction



Software support is included for the Preside (Funk) and Microsoft IAS RADIUS servers. Additional RADIUS servers that support the EAP standard should also be compatible with the Ethernet Routing Switch 8600. For more information, contact your Nortel representative.

802.1x and the LAN Enforcer or VPN TunnelGuard

The Sygate LAN Enforcer or the Nortel VPN TunnelGuard enables the Ethernet Routing Switch 8600 to use the 802.1x standard to ensure that a user connecting from inside a corporate network is legitimate. The LAN Enforcer/TunnelGuard also checks the endpoint security posture, including anti-virus, firewall definitions, Windows registry content, and specific file content (plus date and size). Noncompliant systems that attempt to obtain switch authentication can be placed in a remediation VLAN, where updates can be pushed to the internal user's station, and users can subsequently attempt to join the network again.

VLANs and traffic isolation

You can use the Ethernet Routing Switch 8600 to build secure VLANs. When you configure port-based VLANs, each VLAN is completely separated from the others.

The Ethernet Routing Switch 8600 analyzes each packet independently of preceding packets. This mode, as opposed to the cache mode that some competitors use, allows complete traffic isolation.

For more information about VLANs, see *Nortel Ethernet Routing Switch 8600 Configuration — VLANs and Spanning Tree* (NN46205-517) .

Security at layer 2

At Layer 2, the Ethernet Routing Switch 8600 provides the following security mechanisms:

- **Filters**

  The Ethernet Routing Switch 8600 provides Layer 2 filtering based on the MAC destination and source addresses. This is available per-VLAN.

  For more information about these filters, see *Nortel Ethernet Routing Switch 8600 Configuration — QoS and IP Filtering for Classic Modules* (NN46205-508) .

- **Global MAC filters**

  This feature eliminates the need for you to configure multiple per-VLAN filter records for the same MAC address. By using a Global MAC filter, you can discard ingress MAC addresses that match a global list stored in the switch. You can also apply global MAC filtering to any multicast MAC address. However, you cannot apply it to Local, Broadcast, BPDU MAC, TDP MAC, or All-Zeroes MAC addresses. Once a MAC address is added to this Global list, it cannot be configured statically or learned on any VLAN. In addition, no bridging or routing is performed on packets to or from this MAC address on any VLAN.

  For more information and configuration examples, see *Release Notes for the Ethernet Routing Switch 8600 Release 3.5.2* () .

  For more information about the Layer 2 MAC filter, see *Nortel Ethernet Routing Switch 8600 Configuration — IP Multicast Routing Protocols* (NN46205-501) .

- **Unknown MAC Discard**

  Unknown MAC Discard secures the network by learning allowed MAC addresses during a certain time interval. The switch locks these learned MAC addresses in the forwarding database (FDB) and does not accept any new MAC addresses on the port.

- **Limited MAC learning**

This feature limits the number of FDB-entries learned on a particular port to a user-specified value. After the number of learned FDB-entries reaches the maximum limit, packets with unknown source MAC addresses are dropped by the switch. If the count drops below a configured minimum value due to FDB aging, learning is reenabled on the port.

You can configure various actions like logging, sending traps, and disabling the port when the number of FDB entries reaches the configured maximum limit.

For more information and configuration examples, see the *Release Notes for the Ethernet Routing Switch 8600 Release 3.5.2* () .

Security at Layer 3: filtering

At Layer 3 and above, the Ethernet Routing Switch 8600 provides enhanced filtering capabilities as part of its security strategy to protect the network from different attacks.

You can configure two types of Classic filters on the Ethernet Routing Switch 8600: global filters and source/destination address filters.

R and RS modules support advanced filters based on Access Control Templates (ACT). You can use predefined ACTs designed to prevent, for example, ARP Spoofing, or you can design custom ACTs.

Customer Support Bulletins (CSBs) are available on the Nortel Technical Support Web site to provide information and configuration examples about how to block some attacks.

Security at Layer 3: announce and accept policies

You can use route policies to selectively accept/announce some networks and to block the propagation of some routes. Route policies enhance the security in a network by hiding the visibility of some networks (subnets) to other parts of the network.

You can apply one policy for one purpose. For example, you can apply a RIP announce policy on a given RIP interface. In such cases, all sequence numbers under the given policy are applied to that filter. A sequence number also acts as an implicit preference (that is, a lower sequence number is preferred).

For more information about routing policies, see “DVMRP policies” (page 213).

Routing protocol security

You can protect OSPF and BGP updates with an MD5 key on each interface. At most, you can configure two MD5 keys per interface. You can also use multiple MD5 key configurations for MD5 transitions without bringing down an interface.

For more information, see *Nortel Ethernet Routing Switch 8600 Configuration — OSPF and RIP* (NN46205-522) and *Nortel Ethernet Routing Switch 8600 Configuration — BGP Services* (NN46205-510) .

Control plane security

The control plane physically separates management traffic using the out of band (OOB) interface. The control plane facilitates High Secure mode, management access control, access policies, authentication, Secure Shell and Secure Copy, and SNMP, each of which is described in the sections that follow.

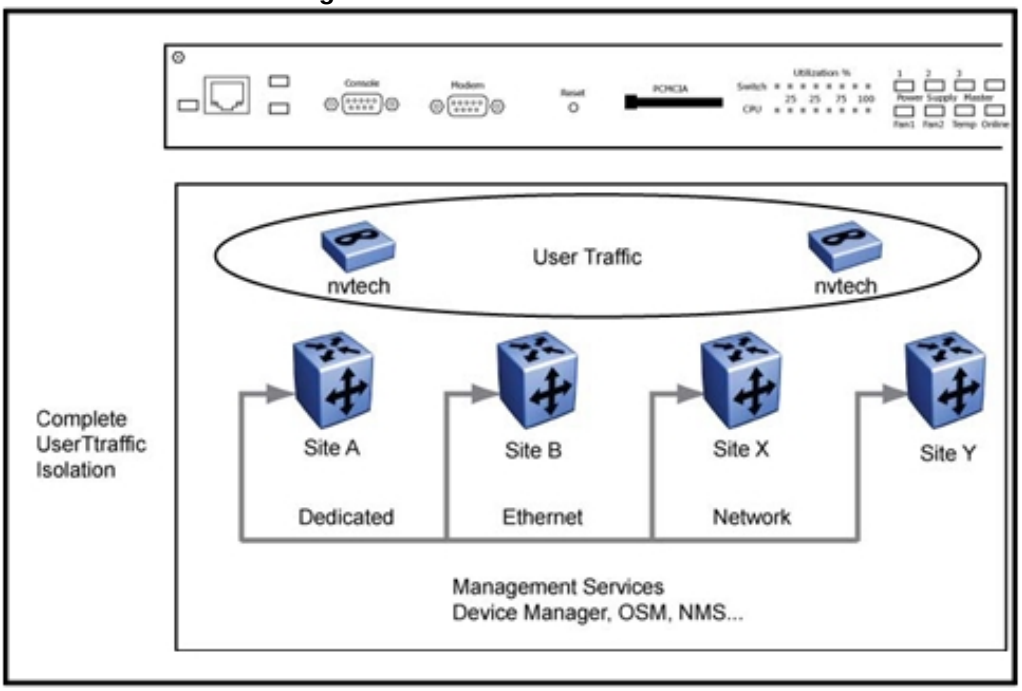
Control plane security navigation

- "Management port" (page 319)
- "Management access control" (page 321)
- "High Secure mode" (page 314)
- "Security and access policies" (page 322)
- "RADIUS authentication" (page 323)
- "TACACS+" (page 325)
- "Encryption of control plane traffic" (page 326)
- "SNMP header network address" (page 327)
- "SNMPv3 support" (page 328)
- "Other security equipment" (page 328)

Management port

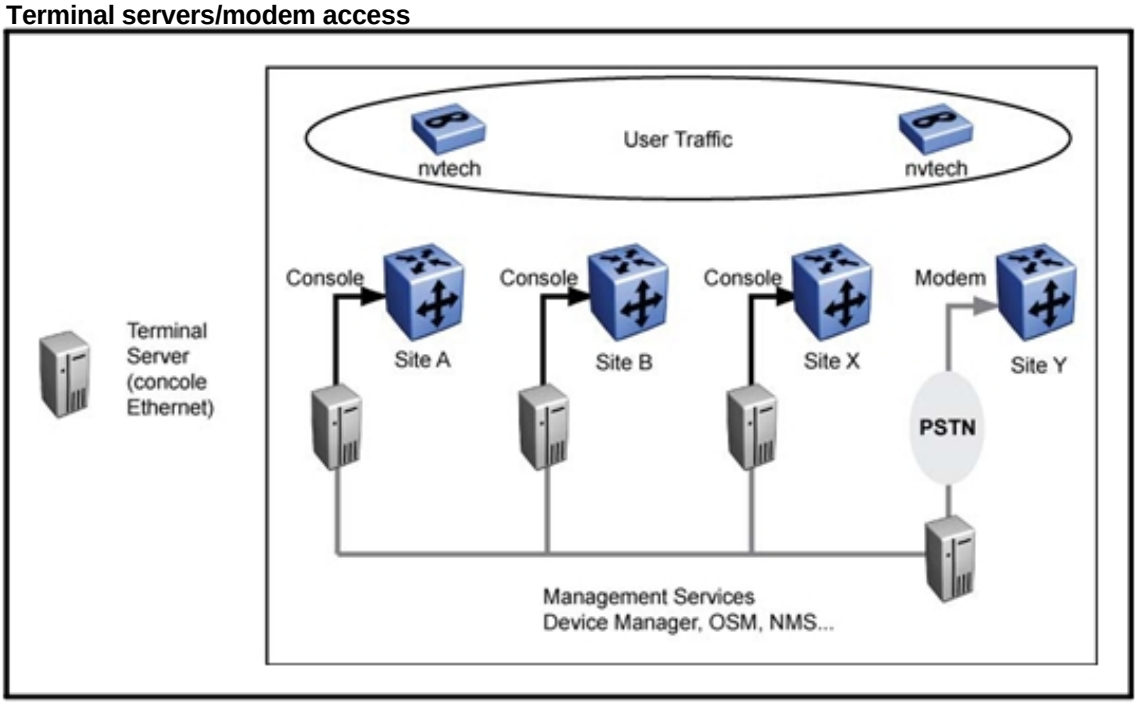
The Ethernet Routing Switch 8600 provides an isolated management port on the switch fabric/CPU. This separates user traffic from management traffic in highly sensitive environments, such as brokerages and insurance agencies. By using this dedicated network (see the following figure) to manage the switches, and by configuring access policies (when routing is enabled), you can manage the switch in a secure fashion.

Figure 132
Dedicated Ethernet management link



You can also use the terminal servers/modems to access the console/modems ports on the switch (see the following figure).

Figure 133
Terminal servers/modem access



When it is an absolute necessity for you to access the switch, Nortel recommends that you use this configuration. The switch is always reachable, even if an issue occurs with the in-band network management interface.

Management access control

The following table shows management access levels. For more information, see *Nortel Ethernet Routing Switch 8600 Security* (NN46205-601) .

Table 33
Ethernet Routing Switch 8600 management access levels

Access level	Description
Read only	Use this level to view the device settings. You cannot change any of the settings.
Layer 1 Read Write	Use this level to view switch configuration and status information and change only physical port parameters.
Layer 2 Read Write	Use this level to view and edit device settings related to Layer 2 (bridging) functionality. The Layer 3 settings (such as OSPF, DHCP) are not accessible. You cannot change the security and password settings.
Layer 3 Read Write	Use this level to view and edit device settings related to Layer 2 (bridging) and Layer 3 (routing). You cannot change the security and password settings.
Read Write	Use this level to view and edit most device settings. You cannot change the security and password settings.
Read Write All	Use this level to do everything. You have all the privileges of read-write access and the ability to change the security settings. The security settings include access passwords and the Web-based management user names and passwords.

Table 33
Ethernet Routing Switch 8600 management access levels (cont'd.)

Access level	Description
	Read-Write-All (RWA) is the only level from which you can modify user-names, passwords, and SNMP community strings, with the exception of the RWA community string, which cannot be changed.
ssladmin	This level lets you logon to connect to and configure the SAM (SSL acceleration module). ssladmin users are granted a broad range of rights that incorporate the Ethernet Routing Switch 8600 read/write access. Users with ssladmin access can also add, delete, or modify all configurations.

High Secure mode

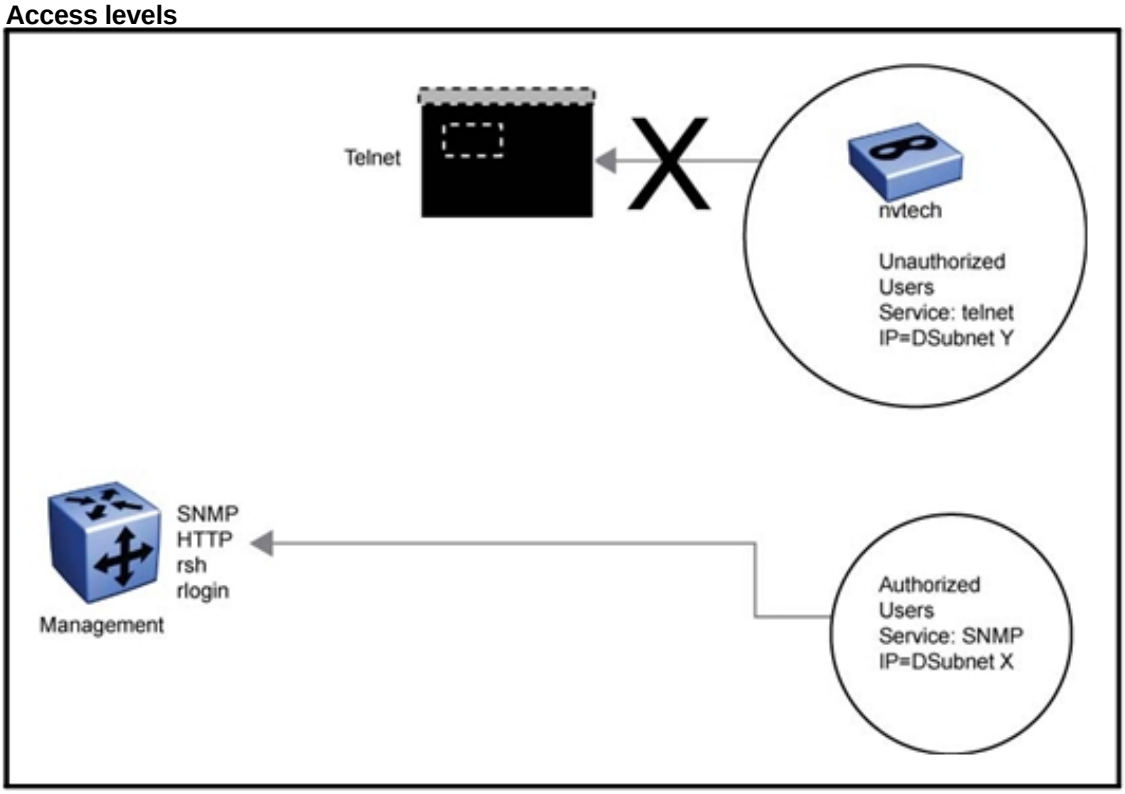
Use High Secure to disable all unsecured application and daemons, such as FTP, TFTP, rlogin, and so on. Nortel strongly recommends that you not use any unsecured protocols. See also "High Secure mode" (page 322).

Use Secure Copy (SCP) rather than FTP or TFTP. For more information, see "SSHv1/v2" (page 327).

Security and access policies

Access policies permit secure switch access by specifying a list of IP addresses or subnets that can manage the switch for a specific daemon, such as Telnet, SNMP, HTTP, SSH, and rlogin. Rather than using a management VLAN that is spread out among all of the switches in the network, you can build a full Layer 3 routed network and securely manage the switch with any of the in-band IP addresses attached to any one of the VLANs (see the following figure).

Figure 134
Access levels



Nortel recommends that you use access policies for in-band management when securing access to the switch. By default, all services are accessible by all networks.

RADIUS authentication

You can enforce access control by utilizing RADIUS (Remote Authentication Dial-in User Service). RADIUS is designed to provide a high degree of security against unauthorized access and to centralize the knowledge of security access based on a client/server architecture. The database within the RADIUS server stores a list of pertinent information about client information, user information, password, and access privileges including the use of the shared secret.

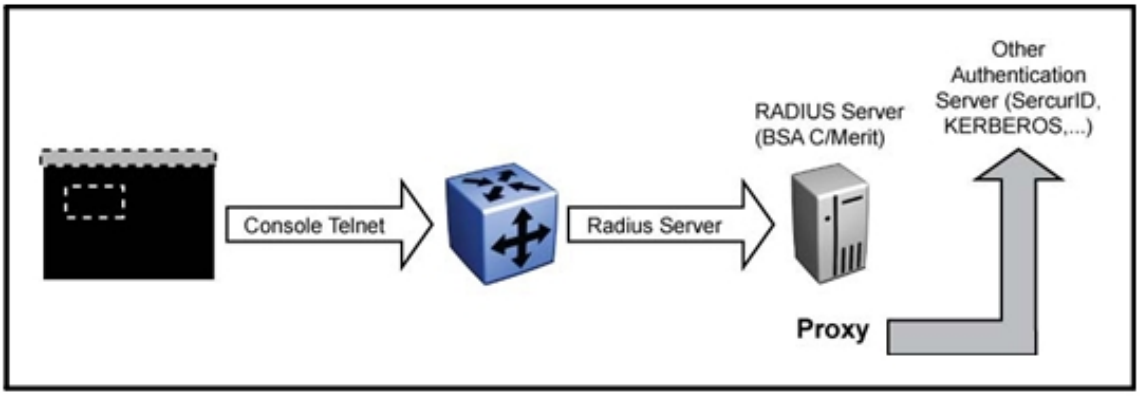
When the switch acts as a Network Access Server, it operates as a RADIUS client. The switch is responsible for passing user information to the designated RADIUS servers. Because the switch operates in a LAN environment, it allows user access through Telnet, rlogin, and Console logon.

You can configure a list of up to 10 RADIUS servers on the client. If the first server is unavailable, the Ethernet Routing Switch 8600 tries the second, and so on, until it establishes a successful connection.

You can use the RADIUS server as a proxy for stronger authentication (see the following figure), such as:

- SecurID cards
- KERBEROS
- other systems like TACACS/TACACS+

Figure 135
RADIUS server as proxy for stronger authentication



You must tell each RADIUS client how to contact its RADIUS server. When you configure a client to work with a RADIUS server, be sure to:

- Enable RADIUS.
- Provide the IP address of the RADIUS server.
- Ensure the shared secret matches what is defined in the RADIUS server.
- Provide the attribute value.
- Indicate the order of priority in which the RADIUS server is used. (Order is essential when more than one RADIUS server exists in the network.)
- Specify the UDP port that is used by the client and the server during the authentication process. The UDP port between the client and the server must have the same or equal value. For example, if you configure the server with UDP 1812, the client must have the same UDP port value.

Other customizable RADIUS parameters require careful planning and consideration on your part, for example, switch timeout and retry. Use the switch timeout to define the number of seconds before the authentication request expires. Use the retry parameter to indicate the number of retries the server accepts before sending an authentication request failure.

Nortel recommends that you use the default value in the attribute-identifier field. If you change the set default value, you must alter the dictionary on the RADIUS server with the new value. To configure the RADIUS feature, you require Read-Write-All access to the switch.

For more information about RADIUS, see *Nortel Ethernet Routing Switch 8600 Security* (NN46205-601) .

TACACS+

Terminal Access Controller Access Control System (TACACS+) is a security application implemented as a client/server-based protocol that provides centralized validation of users attempting to gain access to a router or network access server.

TACACS+ provides management of users wishing to access a device through any of the management channels: Telnet, console, rlogin, SSHv1/v2, and Web management.

TACACS+ also provides management of PPP user connections. PPP provides its own authentication protocols, with no authorization stage. TACACS+ support PPP authentication protocols, but moves the authentication from the local router to the TACACS+ server.

Similar to the RADIUS protocol, TACACS+ provides the ability to centrally manage the users wishing to access remote devices. TACACS+ differs from RADIUS in two important ways:

- TACACS+ is a TCP-based protocol.
- TACACS+ uses full packet encryption, rather than just encrypting the password (RADIUS authentication request).

ATTENTION
TACACS+ encrypts the entire body of the packet but uses a standard TACACS+ header.

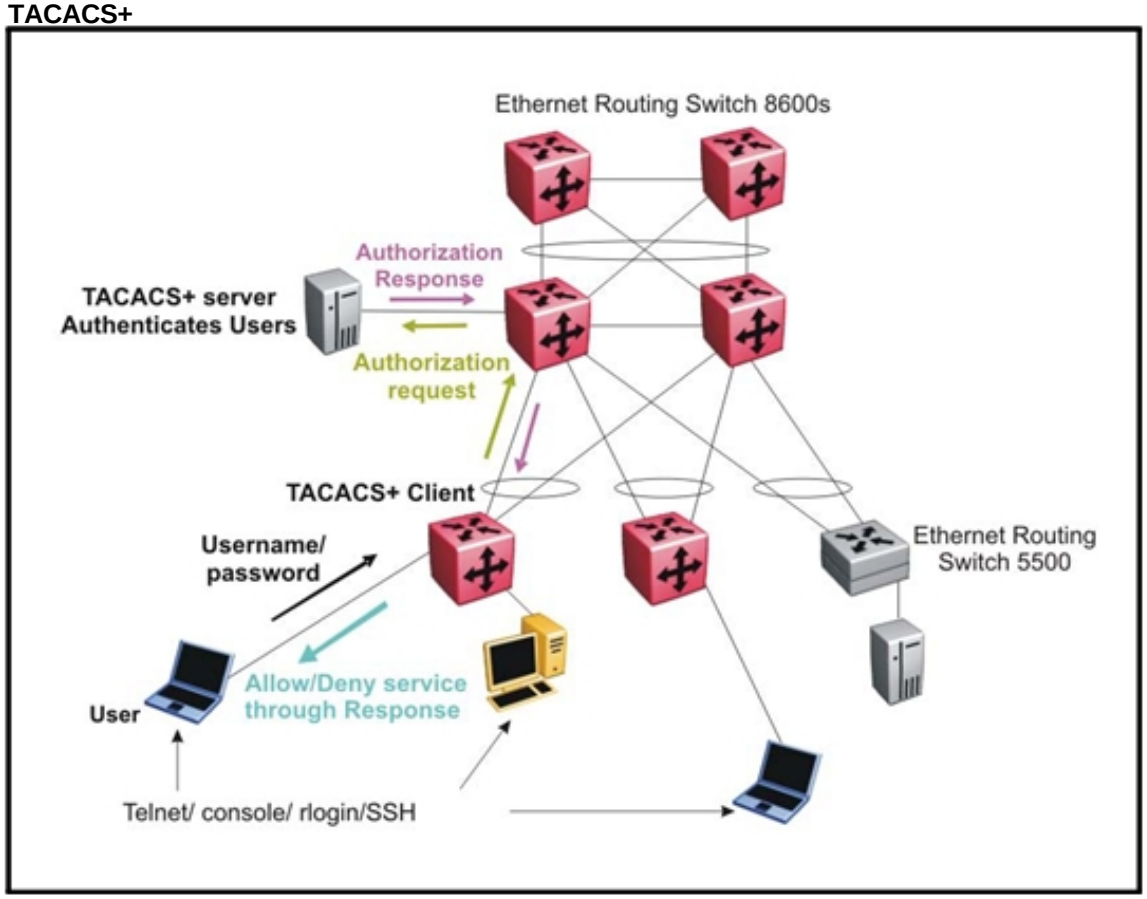
TACACS+ provides separate authentication, authorization and accounting services.

During the log on process, the TACACS+ client initiates the TACACS+ authentication session with the server. The authentication session provides username/password functionality.

After successful authentication, if TACACS+ authorization is enabled, the TACACS+ client initiates the TACACS+ authorization session with the server (see the following figure). The authorization session provides

access level functionality, which enables you to limit the switch commands available to a user. The transition from TACACS+ authentication to the authorization phase is transparent to the user.

Figure 136
TACACS+



After successful authentication, if TACACS+ accounting is enabled, the TACACS+ client sends accounting information to the TACACS+ server. When accounting is enabled, the NAS reports user activity to the TACACS+ server in the form of accounting records. Each accounting record contains accounting AV pairs. The accounting records are stored on the security server. The accounting data can then be analyzed for network management and auditing.

The Ethernet Routing Switch 8600 supports eight users logged in to the chassis simultaneously with TACACS+.

For more information on TACACS+, see *Nortel Ethernet Routing Switch 8600 Security* (NN46205-601) .

Encryption of control plane traffic

Control plane traffic encryption involves SSHv1/v2, SCP, and SNMPv3.

Encryption of control plane traffic navigation

SSHv1/v2

SSH is used to conduct secure communications over a network between a server and a client. The switch supports only the server mode (supply an external client to establish communication). The server mode supports SSHv1 and SSHv2.

The SSH protocol offers:

- Authentication

  SSH determines identities. During the logon process, the SSH client asks for a digital proof of the identity of the user.

- Encryption

  SSH uses encryption algorithms to scramble data. This data is rendered unintelligible except to the intended receiver.

- Integrity

  SSH guarantees that data is transmitted from the sender to the receiver without any alteration. If any third party captures and modifies the traffic, SSH detects this alteration.

The Ethernet Routing Switch 8600 supports:

- SSH version 1, with password and Rivest, Shamir, Adleman (RSA) authentication
- SSH version 2 with password and Digital Signature Algorithm (DSA) authentication
- Triple Digital Encryption Standard (3DES)

SNMP header network address

You can direct an IP header to have the same source address as the management virtual IP address for self-generated UDP packets. If a management virtual IP address is configured and the **`udpsrc-by-vip`** flag is set, the network address in the SNMP header is always the management virtual IP address. This is true for all traps routed out on the I/O ports or on the out-of-band management Ethernet port.

SNMPv3 support

SNMP version 1 and version 2 are not secure because communities are not encrypted.

Nortel strongly recommends that you use SNMP version 3. SNMPv3 provides stronger authentication services and the encryption of data traffic for network management.

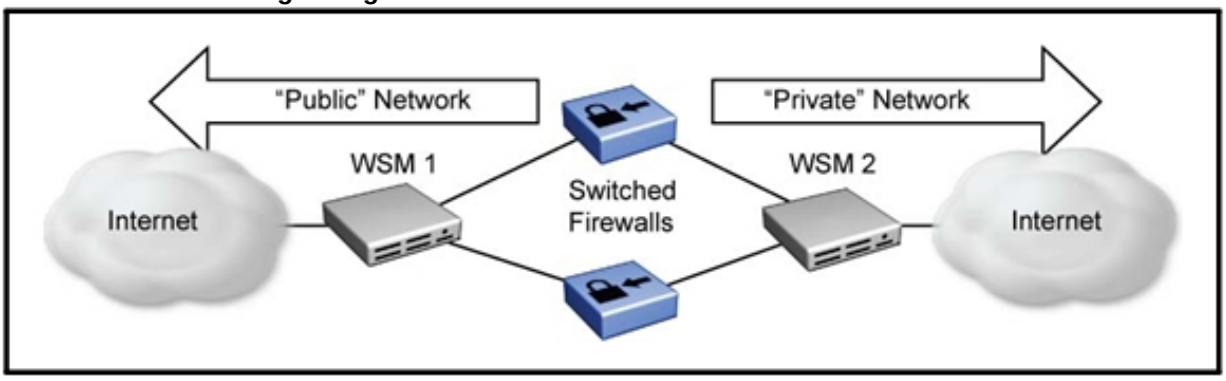
Other security equipment

Nortel offers other devices that increase the security of your network.

For sophisticated state-aware packet filtering (Real Stateful Inspection), you can add an external firewall to the architecture. State-aware firewalls can recognize and track application flows that use not only static TCP and UDP ports, like Telnet or http, but also applications that create and use dynamic ports, such as FTP, and audio and video streaming. For every packet, the state-aware firewall finds a matching flow and conversation.

The following figure shows a typical configuration used in firewall load balancing.

Figure 137
Firewall load balancing configuration



Use this configuration to redirect incoming and outgoing traffic to a group of firewalls and to automatic load balance across multiple firewalls. The WSM can also filter packets at the ingress port so that firewalls see only relevant packets.The benefits of such a configuration are:

- increased firewall performance
- reduced response time
- redundant firewalls ensure Internet access

Virtual private networks (VPN) replace the physical connection between the remote client and access server with an encrypted tunnel over a public network. VPN technology employs IP Security (IPSec) and Secure Sockets Layer (SSL) services.

Several Nortel products support IPSec and SSL. Contivity and the Services Edge Router support IPSEC. Contivity supports up to 5000 IPSEC tunnels, and scales easily to support operational requirements. The Services Edge Router can support up to 30 000 tunnels.

For SSL needs, Nortel offers the Integrated Service Director (iSD) SSL Accelerator Module (SAM). The SAM is used by the Web Switching Module (WSM) to decrypt sessions and to make encrypted cookies and URLs visible to the WSM. The SAM offers:

- secure session content networking at wire speed
- offloading for web servers for better performance
- optimized web traffic for secure Web sites
- cost savings because fewer servers need to be enabled

The Accelerator also terminates each client HTTPS session, performs hardware-assisted key exchange with the client, and establishes an HTTP session to the chosen Web server. On the return path, the SAM encrypts the server response according to the negotiated encryption rules and forewords the response to the requesting client using the established HTTPS session. You can load balance up to 32 iSD-SSL units transparently by using a WSM.

For more information

The following organizations provide the most up-to-date information about network security attacks and recommendations about good practices:

- The Center of Internet Security Expertise (CERT)
- The Research and Education Organization for Network Administrators and Security Professionals (SANS)
- The Computer Security Institute (CSI)

QoS design guidelines

This section provides design guidelines that you can use when you configure your network to provide Quality of Service (QoS) to user traffic.

Quality of Service (QoS) is defined as “the extent to which a service delivery meets user expectations.” In a QoS-aware network, a user can expect the network to meet certain performance expectations. These performance expectations are usually specified in terms of service availability, bandwidth, packet loss, packet delay (latency), and packet delay variation (jitter).

For more information about fundamental QoS mechanisms, and how to configure QoS, see *Nortel Ethernet Routing Switch 8600 Configuration — QoS and IP Filtering for Classic Modules* (NN46205-508) and *Nortel Ethernet Routing Switch 8600 Configuration — QoS and IP Filtering for R and RS Modules* (NN46205-507) .

Navigation

QoS mechanisms

The Ethernet Routing Switch 8600 has a solid, well-defined architecture to handle QoS in an efficient and effective manner. Several QoS mechanisms used by the Ethernet Routing Switch 8600 are briefly described in the sections that follow.

QoS mechanisms navigation

- "QoS classification and mapping" (page 332)
- "QoS and queues" (page 334)
- "QoS and filters" (page 337)
- "Policing and shaping" (page 342)

QoS classification and mapping

The Ethernet Routing Switch 8600 provides a hardware-based Quality of Service platform through hardware packet classification. Packet classification is based on the examination of the QoS fields within the Ethernet packet, primarily the DiffServ Codepoint (DSCP) and the 802.1p fields. Unlike legacy routers that require CPU processing cycles for packet classification, which degrades switch performance, the Ethernet Routing Switch 8600 performs classification in hardware at switching speeds.

You can configure Ingress interfaces in one of two ways. In the first type of configuration, the interface does not classify traffic, but it forwards the traffic based on the packet markings. This mode of operation is applied to trusted interfaces (core port mode) because the DSCP or 802.1p field is trusted to be correct, and the edge switch performs the mapping without any classification.

In the second type of configuration, the interface classifies traffic as it enters the port, and marks the packet for further treatment as it traverses the Ethernet Routing Switch 8600 network. This mode of operation is applied to untrusted interfaces (access port mode) because the DSCP or 802.1p field is not trusted to be correct.

An internal QoS level is assigned to each packet that enters an Ethernet Routing Switch 8600 port. Once the QoS level is set, the egress queue is determined and the packet is transmitted. The mapping of QoS levels to queue is a hard-coded 1-to-1 mapping.

Table 34 "ATM COS, NNSC, DSCP, and 802.1p-bit mappings" (page 333) shows the recommended configuration that a service provider should use for a packet classification scheme. Use the defaults as a starting point because the actual traffic types and flows are unknown. You can change the mapping scheme if the default is not optimal. However, Nortel recommends that you do not change the mappings.

Note the following information with respect to this table:

- If a single Asynchronous Transfer Mode (ATM) virtual circuit (VC) is used with different traffic classes, the ATM Class of Service (COS) is Constant Bit Rate (CBR).
- CBR is used when Voice Activity Detection (VAD) is not used. When VAD is used, Real Time Variable Bit Rate (Rt-VBR) can be used.

Table 34
ATM COS, NNSC, DSCP, and 802.1p-bit mappings

NNSC	DSCP	802.1p	ATM COS
Critical	CS7	7	CBR
Network	CS6	7	nrt-VBR
Premium	EF, CS5	6	CBR
Platinum	AF4x, CS4	5	rt-VBR
Gold	AF3x, CS3	4	rt-VBR
Silver	AF2x, CS2	3	nrt-VBR
Bronze	AF1x, CS1	2	nrt-VBR
Standard	DE, CS0	0	UBR
Custom/best effort	User Defined	1	UBR

In this table, NNSC denotes Nortel Networks Service Class; CS denotes Class Selector; EF denotes Expedited Forwarding; AF denotes Assured Forwarding; DE denotes DEfault forwarding; CBR denotes Constant Bit Rate; nrt denotes nonreal-time; VBR denotes Variable Bit Rate; rt denotes real-time; and UBR denotes Unspecified Bit Rate.

ATTENTION
If you must change the DSCP mappings, ensure that the values are consistent on all other Ethernet Routing Switches and devices in your network. Inconsistent mappings can result in unpredictable service.

The Nortel QoS strategy simplifies QoS implementation by providing a mapping of various traffic types and categories to a Class of Service. These service classes are termed Nortel Networks Service Classes (NNSC). The following table provides a summary of the mappings and their typical traffic types.

Table 35
Traffic categories and NNSC mappings

Traffic category	Application example	NNSC
Network Control	Alarms and heartbeats	Critical
	Routing table updates	Network

Table 35
Traffic categories and NNSC mappings (cont'd.)

Traffic category		Application example	NNSC
Real-Time, Delay Intolerant		IP telephony; interhuman communication	Premium
Real-Time, Delay Tolerant		Video conferencing; interhuman communication.	Platinum
		Audio and video on demand; human-host communication	Gold
NonReal-Time Mission Critical	Interactive	eBusiness (B2B, B2C) transaction processing	Silver
	NonInteractive	Email; store and forward	Bronze
NonReal Time, NonMission Critical		FTP; best effort	Standard
		PointCast; Background/standby	Custom/ best effort

You can select the NNSC for a given device (or a group of devices) and then the network maps the traffic to the appropriate QoS level, marks the DSCP accordingly, sets the 802.1p bits, and sends the traffic to the appropriate egress queue.

QoS and queues

Egress priority and discard priority are used in egress queue traffic management. Egress priority defines the urgency of the traffic, and discard priority defines the importance of the traffic. A packet with high egress priority should be serviced first. Under congestion, apacket with high discard priority is discarded last.

In a communications network, delay-sensitive traffic, such as voice and video, should be classified as high egress priority. Traffic that is sensitive to packet loss, such as financial information, should be classified as high discard priority. The egress priority and discard priority are commonly referred to as latency and drop precedence, respectively.

Each port on the Ethernet Routing Switch 8600 has eight (or 64, depending on the module) egress queues. Each queue is associated with an egress priority. Some queues are designated as Strict Priority queues, which means that they are guaranteed service, and some are designated as Weighted Round Robin (WRR) queues. WRR queues are serviced according to their queue weight after strict priority traffic is serviced.

For more information about queue numbering and priority levels, see *Nortel Ethernet Routing Switch 8600 Configuration — QoS and IP Filtering for R and RS Modules* (NN46205-507) .

The weight of each queue is determined by what is known as its Packet Transmission Opportunity (PTO). The following table shows the default queue configuration for an eight-queue set queue, along with the corresponding packet transmission opportunities (PTO) and queue weights.

Table 36
Eight-queue egress queue set weight, PTO, queue type

NNSC	Egress queue	Type	PTO	Weight
Network	7	Strict priority	2	6%
Premium	6	Strict priority	32	100%
Platinum	5	WRR	10	31%
Gold	4	WRR	8	25%
Silver	3	WRR	6	18%
Bronze	2	WRR	4	12%
Standard	1	WRR	2	6%
Custom	0	WRR	0	0%

The following table shows the relationship between the QoS level and PTO, and the weight and timeslot it receives. In this table, NC denotes not configurable.

Table 37
Packet QoS level and PTO; weight, and timeslot

	QoS level							
	7	**6**	**5**	**4**	**3**	**2**	**1**	**0**
PTO	2	32	10	8	6	4	2	0
Weight (%)	6	100	31	25	18	12	6	0
Time slots								
0	x	x						
1	x	x						

Table 37
Packet QoS level and PTO; weight, and timeslot (cont'd.)

	QoS level							
	7	**6**	**5**	**4**	**3**	**2**	**1**	**0**
2	NC	x	x					
3		x	x					
4		x	x					
5		x		x				
6		x		x				
7		x			x			
8		x	x					
9		x	x					
10		x	x					
11		x		x				
12		x		x				
13		x			x			
14		x			x			
15		x				x		
16		x	x					
17		x	x					
18		x	x					
19		x		x				
20		x		x				
21		x			x			
22		x			x			
23		x				x		
24		x	x					
25		x		x				
26		x		x				
27		x			x			
28		x				x		
29		x				x		
30		x					x	
31		x					x	

32 PTOs exist per round. Queue 7 has two of the 32 transmit opportunities, giving it approximately a 6% weight. Queue weight assignment and the use of WRR prevents the starvation of the lower-priority queues.

The Ethernet Routing Switch 8600 switch offers eight (or sixty-four, depending on the module) egress queues for traffic. These queues are administratively configured so that all the queues are serviced fairly. The weights are assigned through packet transmission opportunities (PTO) to each queue. More PTOs (higher admin weights) are assigned to the high-priority queues so that the time-sensitive transmissions are forwarded with minimum latency. The less time-sensitive (low-priority) traffic goes to the low-priority queues, which are assigned fewer PTOs (lower admin weight). The configured default admin weights allow fair servicing of all low and high priority queues.

Even if an Ethernet Routing Switch 8600 uses a different number of egress queues than other network devices, the DSCP and 802.1p-bit markings are preserved across the network.

QoS and filters

Filters help you provide QoS by permitting or dropping traffic based on the parameters you configure. You can use filters to mark packets for specific treatment.

Ethernet Routing Switch 8600 Classic module filters are hardware-based. Classic filters do not require CPU intervention, so classic filtering can be achieved at wire-speed, resulting in virtually no impact on performance.

Typically, filters act as firewalls or are used for Layer 3 redirection. In more advanced cases, traffic filters can identify Layer 3 and Layer 4 traffic streams. The filters cause the streams to be re-marked and classified to attain a specific QoS level at both Layer 2 (802.1p) and Layer 3 (DSCP).

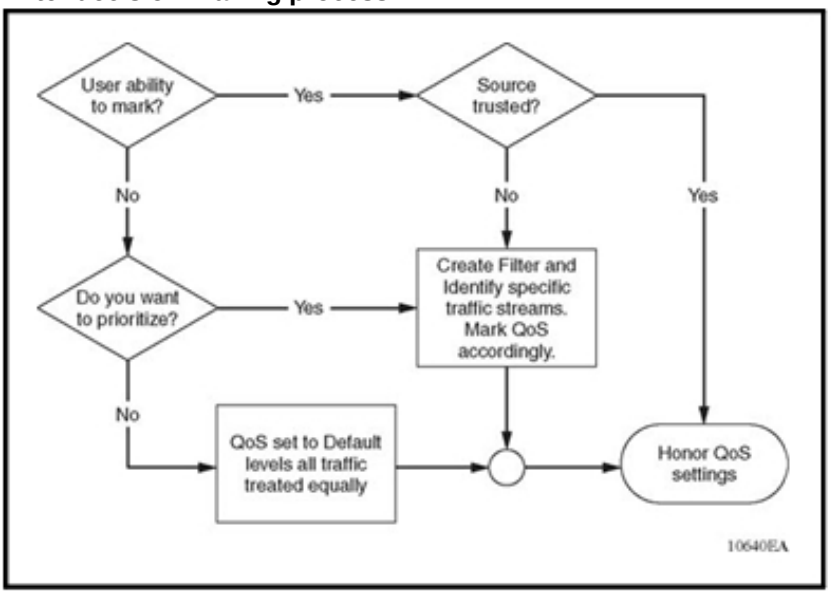
Traffic filtering is a key QoS feature. The Ethernet Routing Switch 8600, by default, determines incoming packet 802.1p or DiffServ markings, and forwards traffic based on their assigned QoS levels. However, situations exist where the markings are incorrect, or the originating user application does not have 802.1p or DiffServ marking capabilities. Also, the administrator may want to give a higher priority to select users (executive class). In any of these situations, use filters to prioritize specific traffic streams.

You can use Classic and Advanced filters to assign QoS levels to devices and applications. To help you decide whether or not to use a filter, key questions include:

1. Does the user or application have the ability to mark QoS information on data packets?
2. Is the traffic source trusted? Are the QoS levels set appropriately for each data source? Users may maliciously set QoS levels on their devices to take advantage of higher priority levels.
3. Do you want to prioritize traffic streams?

This decision-making process is outlined in the following figure.

Figure 138
Filter decision-making process



Be aware that Classic module filters and R series module filters can coexist in the same chassis. You use Classic commands to configure those pre-v4.0 filters that operate only on Classic modules, and use the v4.0 and later commands to configure filters that operate only on R series modules. R series module filters do not interoperate with Classic module filters. In a mixed mode chassis, Classic filters apply only to Classic modules, whereas R series module filters apply only to R series modules.

Classic filters

If you wish to use classic filters for:

- IP routed traffic—use source/destination (SRC/DST) filters.
- IP bridged traffic—use global filters.

Use global filters to remark DSCP and 802.1p bits for IP bridged traffic.

For IP routed traffic, the DSCP field is set to zero at ingress. The QoS level is determined by source and destination filter profiles. The DSCP can be re-marked only by source and destination filters.

For IP routed traffic ingressing a core port, the switch honors the DSCP field and sends traffic to the queues according to the DSCP marking.

Global filters Global filters are executed in the ARU (address resolution unit). This ARU is an Application-specific integrated circuit (ASIC) that makes the forwarding decision without CPU intervention and so does not have an adverse impact on forwarding speeds.

The maximum number of global filters that you can configure per interfaces is:

- eight global filters per group of eight 10BASE-T/100BASE-TX ports
- eight global filters per Gigabit Ethernet port

You can apply global filters to the following types of traffic:

- IP bridged traffic (IP traffic within the same VLAN)
- routed traffic if DiffServ is not used
- multicast traffic with no minimum/maximum mask length

Source/destination filters Source/destination filters are stored in memory associated with the ARU. The time required for the Ethernet Routing Switch 8600 to make a forwarding decision for a given IP routable packet is determined by the following factors:

- the number of source and destination filters configured for and associated with the source/destination IP addresses of this packet
- any IP route that constitutes a less-specific match for one or both of these addresses

You can configure up to 3071 source/destination filters, and each associated IP address must have a minimum mask length of eight bits.

By minimizing the number of source/destination filters associated with IP addresses, you minimize the lookup time necessary for the Ethernet Routing Switch 8600 to complete a forwarding decision for this packet.

When you configure source/destination filters, Nortel recommends the following guidelines:

- In general, minimize the number of source/destination filters in your configuration to avoid multiple filter lookups.
- Design your source/destination filters to be as specific as possible.
- Use the longest possible source/destination masks to avoid multiple filter lookups for different IP traffic flows.

IP filtering and ARP

Classic IP filters only affect the flow of IP traffic that has an Ethertype of 0800; they do not affect traffic from other Ethertypes, such as ARP, which has an Ethertype of 0806. When you configure a physical interface to have a default action of *drop*, it drops all traffic for which no matching forwarding filter exists. NonIP traffic, particularly ARP packets, that ingresses a port with a default action of drop are not answered by the Ethernet Routing Switch 8600.

To ensure that ARP packets traverse the switch when you configure a port in drop mode, Nortel recommends that you do one of the following:

- On the end stations, statically configure the ARP entry related to the gateway.
- Configure a protocol-based VLAN (using the 0x806 type) to capture ARP traffic.

As an alternative to using drop mode, you can configure the port in forward mode, define global filters to specify which traffic must be forwarded, and define a global filter to block all the traffic (a *deny-all* filter). Ensure that the deny-all filter is the last filter and that stop-on-match is set.

R series module filters

Advanced filters are provided for R series modules through the use of Access Control Templates (ACT), Access Control Lists (ACL), and Access Control Entries (ACE), which are implemented in software.

When using ACTs, consider the following:

- For pattern matching filters, three separate patterns per ACT are supported.
- After you configure an ACT, you must activate it. After it is activated, it cannot be modified; only deleted.
- You can only delete an ACT when no ACLs use that ACT. •

4000 ACTs and 4000 ACLs are supported.

- The ACT and ACL IDs 4001 to 4096 are reserved for system-defined ACTs and ACLs. You can use these ACTs and ACLs, but you cannot modify them.

When you configure a new ACT, choose only the attributes you plan to use when setting up the ACEs. For each additional attribute included in an ACT, an additional lookup must be performed. Therefore, to enhance performance, keep the ACT attribute set as small as possible. If too many attributes are defined, you may receive error messages about using up memory. For example, if you plan to filter on source and destination IP addresses and DSCP, only select these attributes. The number of ACEs within an ACL does not impact performance.

For multiple ACEs that perform the same task, for example, deny or allow IP addresses or UDP/TCP-based ports, you can configure one ACE to perform the task with either multiple address entries or address ranges, or a combination of both. This strategy reduces the number of ACEs.

You can configure a maximum of 1000 ACEs per port for ingress and egress. The Ethernet Routing Switch 8600 supports a maximum of 4000 ACEs. For each ACL, a maximum of 500 ACEs are supported.

When you configure R series module filters, keep the following scaling limits in mind.

Table 38
ACT, ACE, ACL scaling

Parameter	Maximum number
ACLs for each switch	4000
ACEs for each switch	4000
ACEs for each ACL	500
ACEs for each port	2000: • 500 inPort • 500 inVLAN • 500 outPort • 500 outVLAN.

The following steps summarize the R series module filter configuration process:

1. Determine your desired match fields.
2. Create your own ACT with the desired match fields.
3. Create an ACL and associate it with the ACT from step 2.
4. Create an ACE within the ACL.
5. Set the desired precedence, traffic type, and action.

The traffic type is determined when you create an ingress or egress ACL.

6. Modify the fields for the ACE.

Policing and shaping

As part of the filtering process, the administrator or service provider can police ingress traffic. Policing is performed according to the traffic filter profile assigned to the traffic flow. For enterprise networks, policing is required to ensure that traffic flows conform to the criteria assigned by network managers.

Both traffic policers and traffic shapers identify traffic using a traffic policy. Traffic that conforms to this policy is guaranteed for transmission, whereas nonconforming traffic is considered to be in violation. Traffic policers drop packets when traffic is excessive, or remark the DSCP or 802.1p markings by using filter actions. With the Ethernet Routing Switch 8600, you can define multiple actions in case of traffic violation.

For service providers, policing at the network edge provides different bandwidth options as part of a Service Level Agreement (SLA). For example, in an enterprise network, you can police the traffic rate from one department to give critical traffic unlimited access to the network. In a service provider network, you can control the amount of traffic customers send to ensure that they comply with their SLA. Policing ensures that users do not exceed their traffic contract for any given QoS level. Policing (or rate metering) gives the administrator the ability to limit the amount of traffic for a specific user in two ways:

- drop out-of-profile traffic
- re-mark out-of-profile traffic to a lower (or higher) QoS level when port congestion occurs

Rate metering can only be performed on a Layer 3 basis.

Traffic shapers buffer and delay violating traffic. These operations occur at the egress queue set level. The Ethernet Routing Switch 8600 supports traffic shaping at the port level and at the per-transmit-queue level for outgoing traffic.

QoS feature availability

The following table lists QoS feature availability as determined by module type and operation mode. In this table, FOQ denotes Feedback Output Queueing.

Table 39
Features supported per operation mode

Chassis configuration	Operation mode	Features supported on respective modules			
		QoS	Filters	Policing	Shaping
All same-type modules	Default	Classic	Classic	Classic	N/A
	M	Classic	Classic	Classic	N/A
	R	Advanced	Advanced	Advanced	Advanced
Mixed modules	Default (32000 records)	Classic Advanced on R and RS modules No FOQ	Classic Advanced on R and RS modules	Classic Advanced on R and RS modules	Advanced on R and RS modules
	M (128000 records)	Classic Advanced on R and RS modules No FOQ	Classic Advanced on R and RS modules	Classic Advanced on R and RS modules	Advanced on R and RS modules
	R (256000 records)	Advanced Supports FOQ	Advanced	Advanced	Advanced

Provisioning QoS networks using R series modules

You can use R series module filters (ACLs) in a mixed or R mode-only chassis but only for R series module ports or VLANs that contain R series module ports. In a mixed-mode chassis, ACLs can only be applied to R series module ports and VLANs. You must use classic filters (src/dst/global) for Classic modules (E and M modules). You can apply an ACL to a VLAN that contains both R and RS module ports and Classic module ports, but the ACL is only applied to the R series module ports within the VLAN.

When you configure Access Control Templates (ACT), only define the attributes on which to match if they are absolutely required. Use as few attributes as possible. The more attributes you configure, the more resource-intensive the filtering action is. If too many attributes are defined, you may receive error messages about using up memory.

Classic IP filtering and DiffServ

You can use Classic module IP filtering with DiffServ features in the following combinations:

- For IP routed traffic on DiffServ access ports, use source/destination filters.
- For IP bridged traffic on DiffServ access ports, use global filters.
- For filtering IP multicast traffic, use global filters.

QoS interface considerations

Four QoS interface types are explained in detail in the following sections. You can configure an interface as trusted or untrusted, and for bridging or routing operations. Use these parameters to properly apply QoS to network traffic.

QoS interface consideration navigation

- "Trusted and untrusted interfaces" (page 344)
- "Bridged and routed traffic" (page 346)
- "802.1p and 802.1Q recommendations" (page 346)

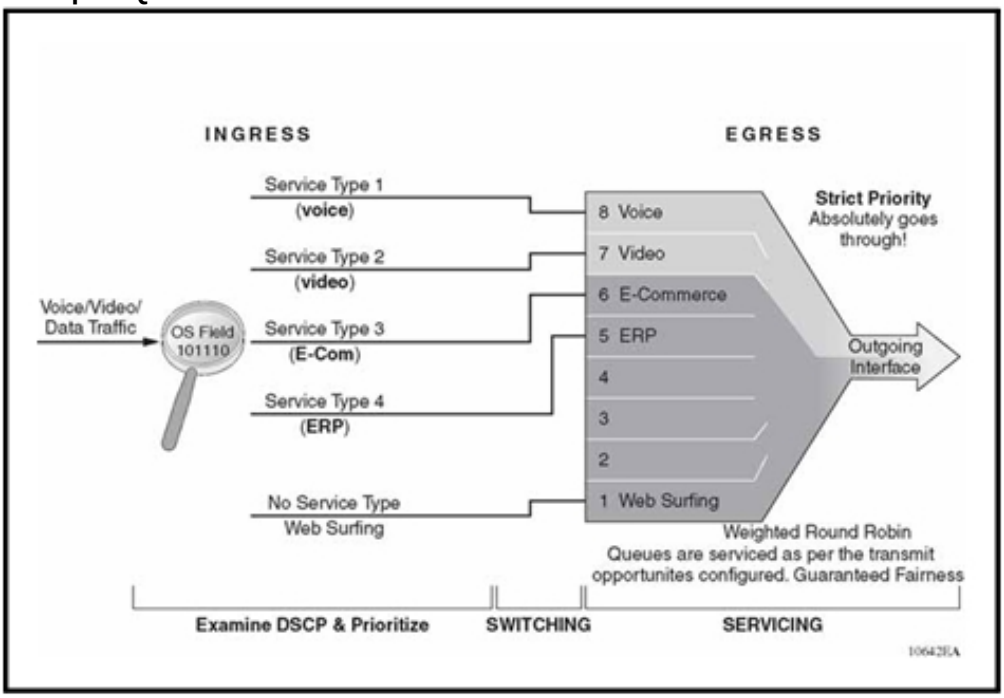
Trusted and untrusted interfaces

You can set an interface as trusted (core) or untrusted (access).

Use a trusted interfaces (core) to mark traffic in a specific way, and to ensure that packets are treated according to the service level of those markings. Use a core setting when control over network traffic prioritization is required. For example, use 802.1p-bits to apply desired CoS attributes to the packets before they are forwarded to the access node. You can also classify other protocol types ahead of IP packets if that is required.

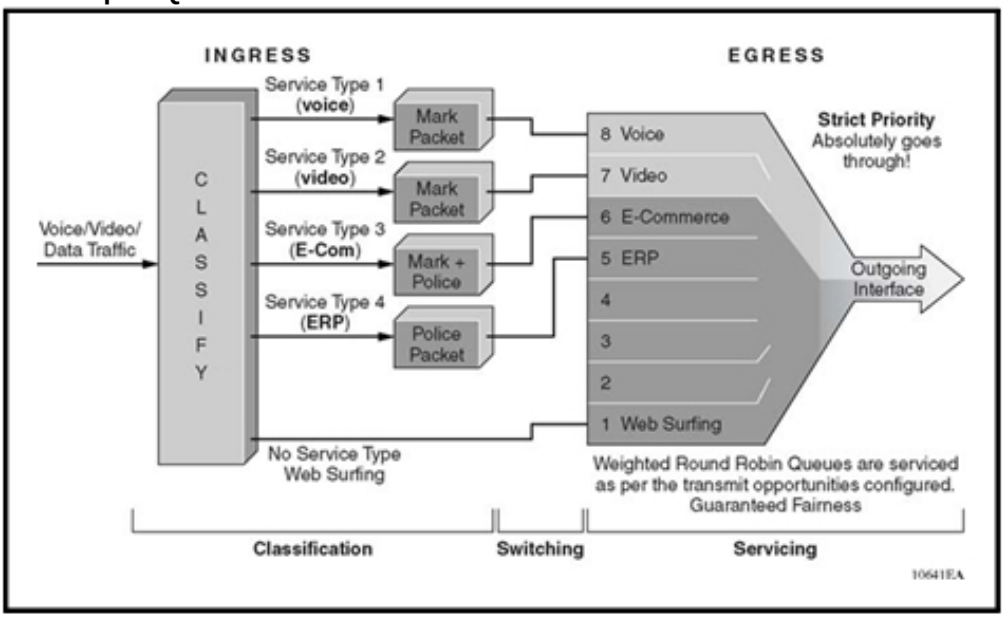
A core port preserves the DSCP and 802.1p-bits markings. The switch uses these values to assign a corresponding QoS level to the packets and sends the packets to the appropriate egress queues for servicing. The following figure illustrates how packets are processed through a core port.

Figure 139
Core port QoS actions



Use the access port setting to control the classification and mapping of traffic for delivery through the network. Untrusted interfaces require you to configure filter sets to classify and re-mark ingress traffic. For untrusted interfaces in the packet forwarding path, the DSCP is mapped to an IEEE 802.1p user priority field in the IEEE 802.1Q frame, and both of these fields are mapped to an IP Layer 2 drop precedence value that determines the forwarding treatment at each network node along the path. Traffic entering an access port is re-marked with the appropriate DSCP and 802.1p markings, and given an internal QoS level. This re-marking is done based on the filters and traffic policies that you configure. The following figure shows access port actions.

Figure 140
Access port QoS actions



Bridged and routed traffic

In a service provider network, access nodes use the Ethernet Routing Switch 8600 configured for bridging. In this case, the Ethernet Routing Switch 8600 uses DiffServ to manage network traffic and resources, but some QoS features are unavailable in the bridging mode of operation. If the Ethernet Routing Switch 8600 is configured for bridging, ingress traffic is mapped from IEEE 802.1p-bits to the appropriate QoS level, and egress traffic is mapped from the QoS level to the appropriate IEEE 802.1p-bits.

In an enterprise network, access nodes use the Ethernet Routing Switch 8600 configured for bridging, and core nodes use the Ethernet Routing Switch 8600 configured for routing. For bridging, ingress, and egress traffic is mapped from the 802.1p-bit marking to a QoS level. For routing, ingress traffic is mapped from the DSCP marking to the appropriate QoS level and egress traffic is mapped from QoS level to the appropriate DSCP as per Table 34 "ATM COS, NNSC, DSCP, and 802.1p-bit mappings" (page 333).

802.1p and 802.1Q recommendations

In a network, to map the 802.1p user priority bits to a queue, 802.1Q-tagged encapsulation must be used on customer premises equipment (CPE). Encapsulation is required because the Ethernet Routing Switch 8600 does not provide classification when it operates in bridging mode. If 802.1Q-tagged encapsulation is not used to connect to the Ethernet Routing Switch 8600, traffic can only be classified based on VLAN membership, port, or MAC address.

To ensure consistent Layer 2 QoS boundaries within the service provider network, you must use 802.1Q encapsulation to connect a CPE directly to an Ethernet Routing Switch 8600 access node. If packet classification is not required, use a Business Policy Switch 2000 to connect to the access node. In this case, the service provider configures the traffic classification functions in the Business Policy Switch 2000.

At the egress access node, packets are examined to determine if their IEEE 802.1p or DSCP values must be re-marked before leaving the network. Upon examination, if the packet is a tagged packet, the IEEE 802.1p tag is set based on the QoS level-to-IEEE 802.1p-bit mapping. For bridged packets, the DSCP is re-marked based on the QoS level.

Network congestion and QoS design

When providing Quality of Service in a network, one of the major elements you must consider is congestion, and the traffic management behavior during congestion. Congestion in a network is caused by many different conditions and events, including node failures, link outages, broadcast storms, and user traffic bursts.

At a high level, three main types or stages of congestion exist:

1. no congestion
2. bursty congestion
3. severe congestion

In a noncongested network, QoS actions ensure that delay-sensitive applications, such as real-time voice and video traffic, are sent before lower-priority traffic. The prioritization of delay-sensitive traffic is essential to minimize delay and reduce or eliminate jitter, which has a detrimental impact on these applications.

A network can experience momentary bursts of congestion for various reasons, such as network failures, rerouting, and broadcast storms. The Ethernet Routing Switch 8600 has sufficient queue capacity and an efficient queue scheduler to handle bursts of congestion in a seamless and transparent manner. Traffic can burst to over 100% within the Weighted Round Robin (WRR) queues, and yet no traffic is dropped: if the burst is not sustained, then the traffic management and buffering process on the switch allows all the traffic to pass without any loss.

Severe congestion is defined as a condition where the network or certain elements of the network experience a prolonged period of sustained congestion. Under such congestion conditions, congestion thresholds are reached, buffers overflow, and a substantial amount of traffic is lost.

When severe congestion is detected, the Ethernet Routing Switch 8600 discards traffic based on drop precedence values. This mode of operation ensures that high-priority traffic is not discarded before lower-priority traffic.

When you perform traffic engineering and link capacity analysis for a network, the standard design rule is to design the network links and trunks for a maximum average-peak utilization of no more than 80%. This means that the network peaks to up to 100% capacity, but the average-peak utilization does not exceed 80%. The network is expected to handle momentary peaks above 100% capacity, as mentioned previously.

QoS examples and recommendations

The sections that follow present QoS network scenarios for bridged and routed traffic over the core network.

Bridged traffic

When you bridge traffic over the core network, you keep customer VLANs separate (similar to a Virtual Private Network). Normally, a service provider implements VLAN bridging (Layer 2) and no routing. In this case, the 802.1p-bit marking determines the QoS level assigned to each packet. When DiffServ is active on core ports, the level of service received is based on the highest of the DiffServ or 802.1p settings.

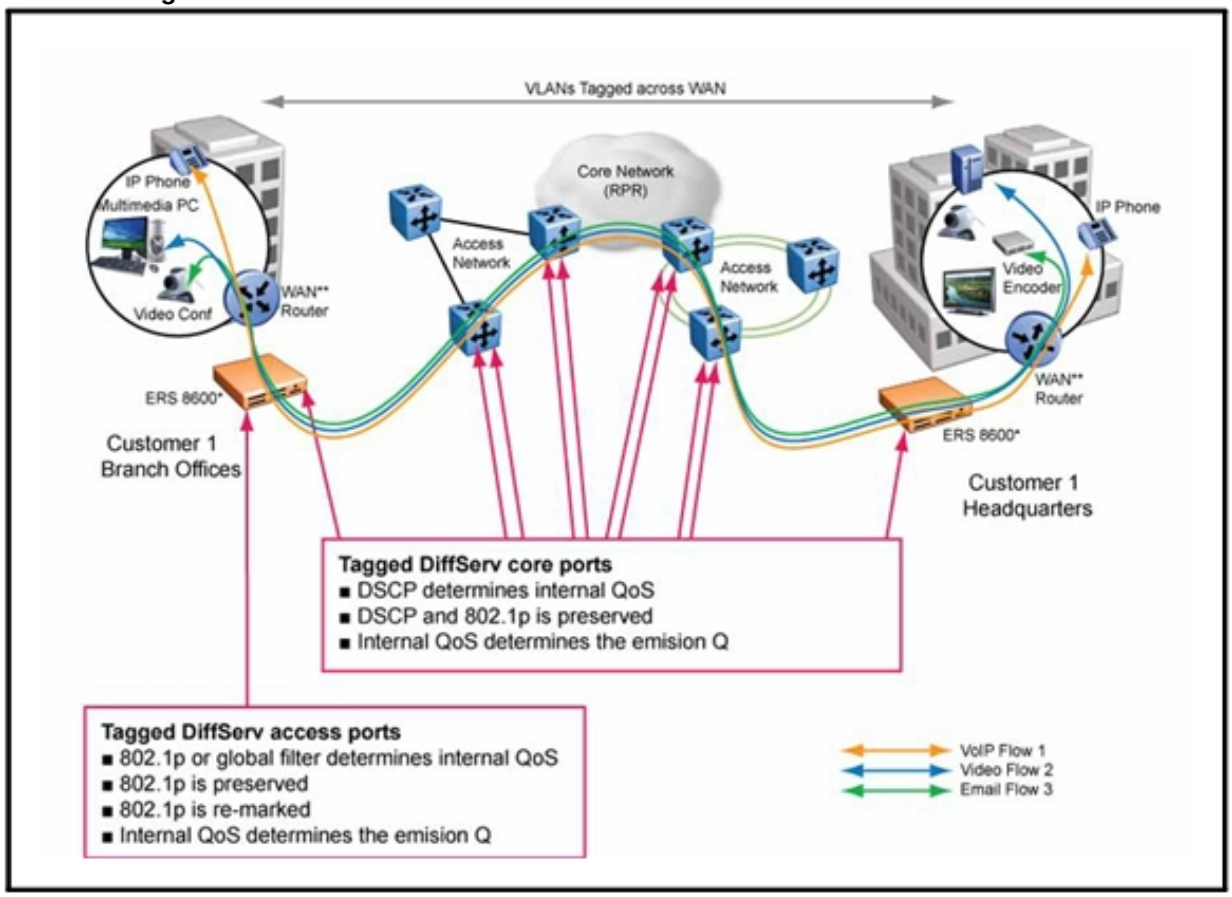
The following cases describe sample QoS design guidelines you can use to provide and maintain high service quality in an Ethernet Routing Switch 8600 network.

Bridged trusted traffic

When you set the port to core, you assume that, for all incoming traffic, the QoS setting is properly marked. All core switch ports simply read and forward packets; they are not re-marked or reclassifiied. All initial QoS markings are performed at the customer device or on the edge devices.

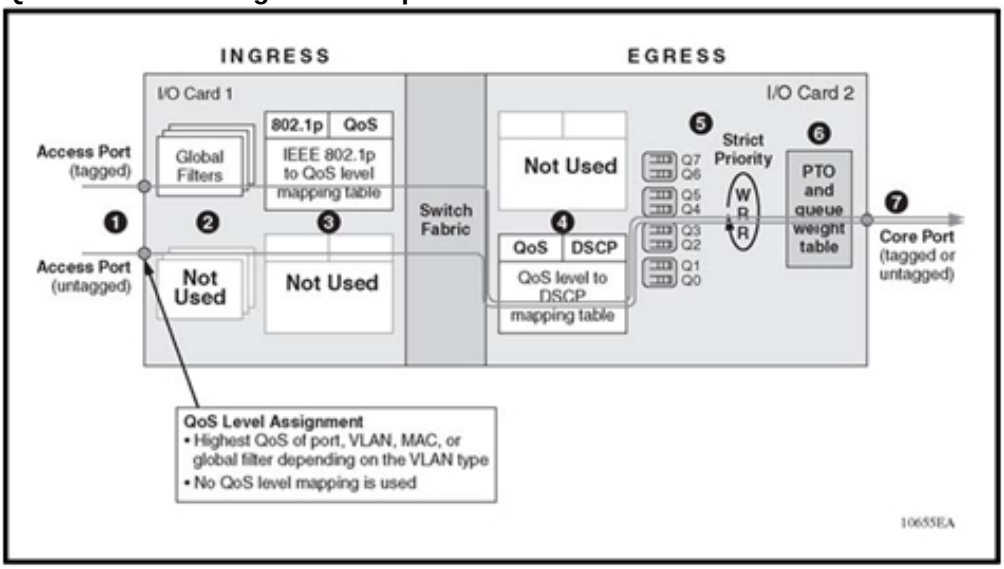
The following figure describes the actions performed on three different bridged traffic flows (that is VoIP, video conference, and e-mail) at access and core ports throughout the network.

Figure 141
Trusted bridged traffic



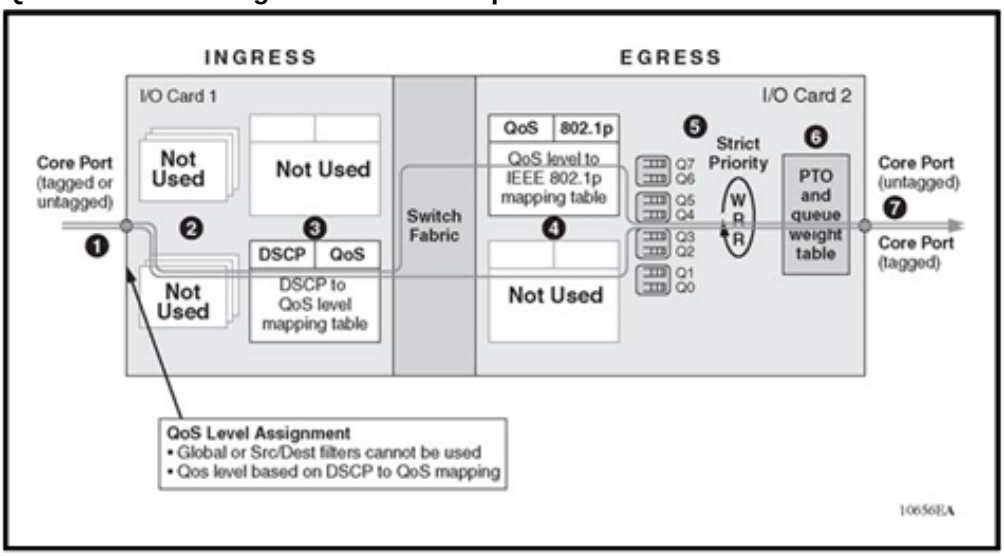
The following figure shows what happens inside an Ethernet Routing Switch 8600 access node. Packets enter through a tagged or untagged access port, and exit through a tagged or untagged core port.

Figure 142
QoS actions on bridged access ports



The following figure shows what happens inside an Ethernet Routing Switch 8600 core node. Packets enter through a tagged or untagged core port, and exit through a tagged or untagged core port.

Figure 143
QoS actions on bridged or routed core ports



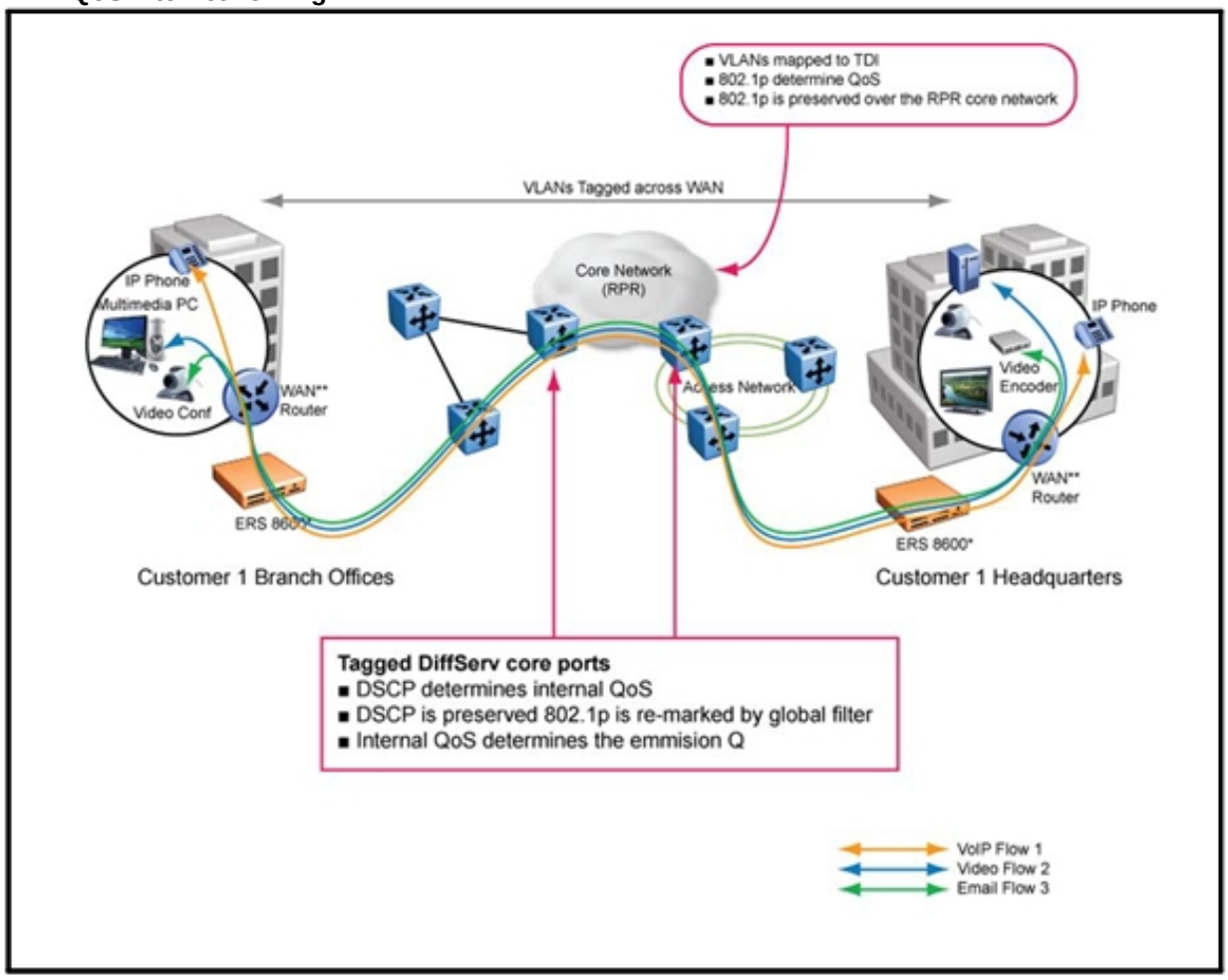
Bridged untrusted traffic

When you set the port to access, mark and prioritize traffic on the access node using global filters. Reclassify the traffic to ensure it complies with the Class of Service specified in the Service Level Agreement (SLA).

Bridged traffic and RPR interworking

For Resilient Packet Ring (RPR) interworking, you can assume that, for all incoming traffic, the QoS setting is properly marked by the access nodes. The RPR interworking is done on the core switch ports that are configured as core/trunk ports. These ports preserve the DSCP marking and re-mark the 802.1p bit to match the 802.1p bit of the RPR. The following figure shows the actions performed on three different traffic flows (VoIP, video conference, and e-mail) over an RPR core network.

Figure 144
RPR QoS internetworking



Routed traffic

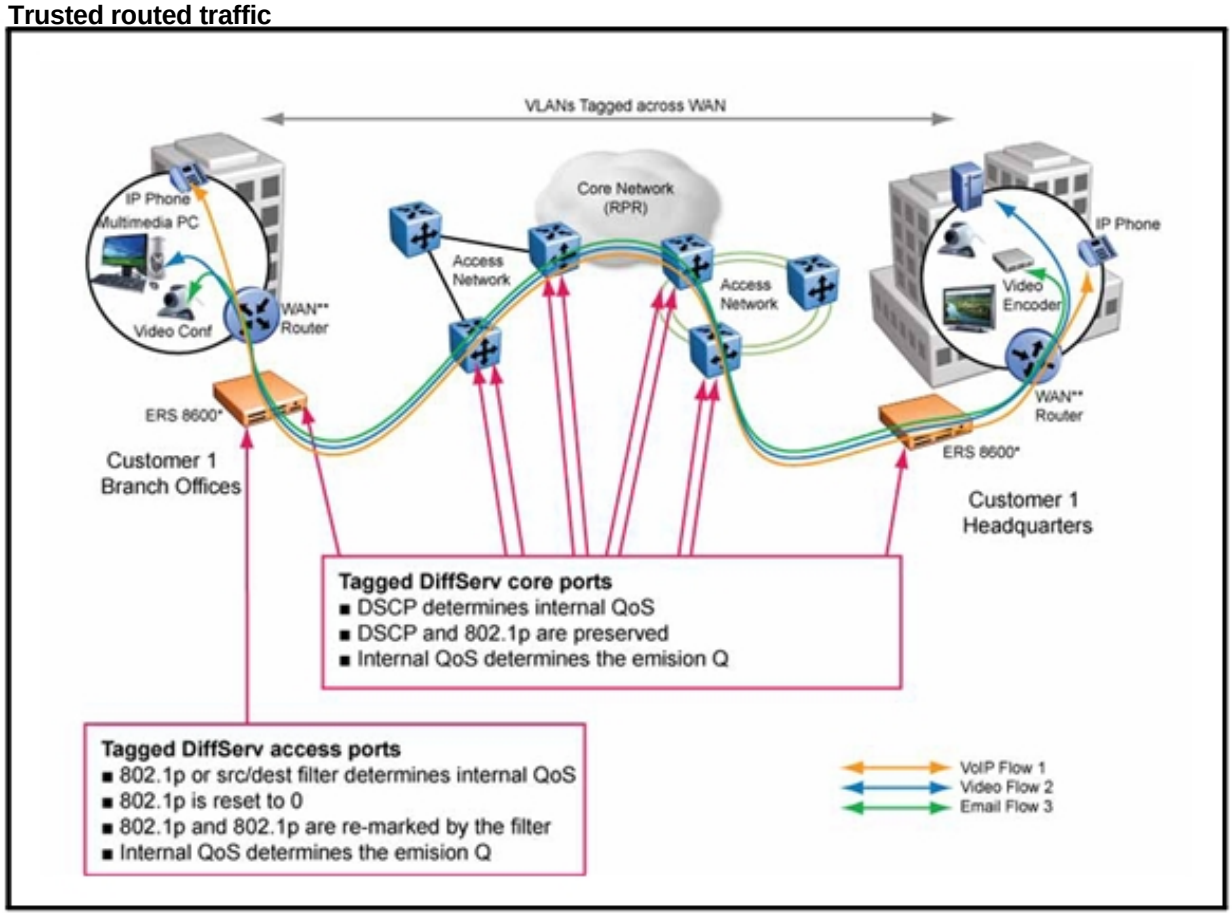
When you route traffic over the core network, VLANs are not kept separate. The following case describes QoS design guidelines you can use to provide and maintain high service quality in an Ethernet Routing Switch 8600 network.

Routed trusted traffic

When you set the port to core, you assume that, for all incoming traffic, the QoS setting is properly marked. All core switch ports simply read and forward packets. The packets are not re-marked or reclassifed from the switch. All initial QoS markings are performed by the customer device or the edge devices, such as the 8003 switch or the Business Policy Switch 2000 (in this case, the 8003 switch treats ingress traffic as trusted).

The following figure shows the actions performed on three different routed traffic flows (that is VoIP, video conference, and e-mail) at access and core ports throughout the network.

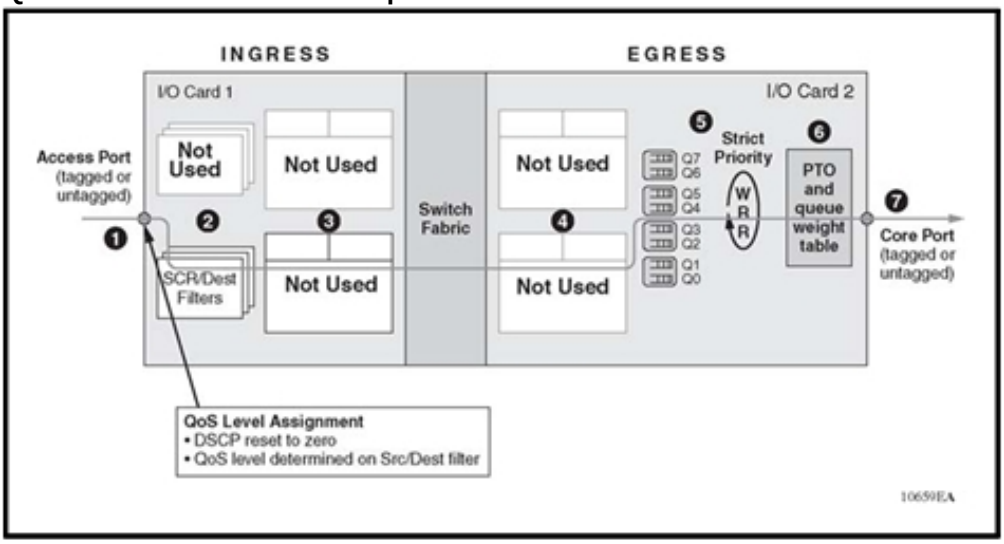
Figure 145
Trusted routed traffic



Routed untrusted traffic

The following figure shows what happens inside an Ethernet Routing Switch 8600 access node. Packets enter through a tagged or untagged access port and exit through a tagged or untagged core port.

Figure 146
QoS actions on routed access ports



Appendix Hardware and supporting software compatibility

The following table describes Ethernet Routing Switch 8600 hardware and the minimum software version required to support the hardware.

EUED RoHS compliancy: Beginning July 1, 2006, products can be ordered with European Union Environmental Directive (EUED) Restriction of Hazardous Substances (RoHS) (EUED RoHS) compliancy. EUED RoHS compliant products are designated with -E5 or -E6, for example, DS1402004-E5.

Table 40
Ethernet Routing Switch 8600 chassis and SF/CPUs

Chassis or switch fabric		Minimum software version	Part number
8010 chassis	10-slot chassis	3.0.0	DS1402001-E5 DS1402001-E5GS
8006 chassis	6-slot chassis	3.0.0	DS1402002-E5 DS1402002-E5GS
8003 chassis	3-slot chassis	3.1.2	DS1402003-E5 DS1402003-E5GS
8010co chassis	10-slot chassis	3.1.2	DS1402004-E5 DS1402004-E5GS
8690 SF/CPU	Switch fabric; discontinued	3.0.0	DS1404001
8691 SF/CPU	Switch fabric	3.1.1	DS1404025
8691 SF/CPU 256 MB	8691 SF/CPU 256 with 256 MB SDRAM installed	3.1.1	DS1404090-E5
8692 SF/CPU	Switch fabric	3.5.6	DS1404065-E5

Table 40
Ethernet Routing Switch 8600 chassis and SF/CPUs (cont'd.)

Chassis or switch fabric		Minimum software version	Part number
8692 SF/CPU Switch Fabric/CPU with factory-installed Enterprise Enhanced CPU Daughter Card (SuperMezz).	Switch fabric	4.1.0	DS1404066-E5
Enterprise Enhanced CPU Daughter Card (SuperMezz)	Optional daughter card for the 8692 SF/CPU	4.1.0	DS1411025-E5
Power supplies			
8001AC	690 Watt (W) AC	3.0.0	DS1405x01
8002DC	780 W DC	3.0.0	DS1405002
8003AC	500 W AC (8003 chassis only)	3.1.2	DS1405x03
8004AC	850 W AC	3.1.2	DS1405x08
8004DC	850 W DC	3.1.2	DS1405007
8005AC	1462 W AC	4.0.0	DS1405012
8005DI AC	1492 W Dual input AC	5.0	DS1405016-E6
8005DC	1462 W DC	4.0.x	DS1405011
Upgrade kits			
256 MB CPU upgrade kit	Nortel recommends that you upgrade the 8691 SF/CPU to 256 MB. This memory upgrade may be required for the 3.5 software to run properly.	3.5	DS1411016
MAC upgrade kit	Use this kit to add Media Access Control (MAC) addresses to your system.	3.5	DS1404015

Table 41
Ethernet Routing Switch 8600 modules and components

Module or component	Minimum software version	Part number
Service Delivery Module		

Table 41
Ethernet Routing Switch 8600 modules and components (cont'd.)

Module or component		Minimum software version	Part number
8660 Service Delivery Module with one Firewall iSD module	Firewall iSD module	3.7.6 (SDM) 4.5 (SourceFire)	DS1404104
8660 Service Delivery Module with four Firewall iSD modules	Firewall iSD module	3.7.6 (SDM) 4.5 (SourceFire)	DS1404080
8660 Service Delivery Module with two Firewall iSD modules	Firewall iSD module	3.7.6 (SDM) 4.5 (SourceFire)	DS1404081
8660 Service Delivery Module with four Intrusion Sensor iSD modules	Firewall iSD module connects to Nortel Threat Protection System	3.7.6 (SDM) 4.5 (SourceFire)	DS1404082-E5
8660 Service Delivery Module COMBO with two Firewall iSD modules and two Intrusion Sensor iSD modules	Firewall iSD module connects to Nortel Threat Protection System	3.7.6 (SDM) 4.5 (SourceFire)	DS1404086-E5
8660 Service Delivery Module COMBO with one Firewall iSD module and one Intrusion Sensor iSD module	Firewall iSD module connects to Nortel Threat Protection System	3.7.6 (SDM) 4.5 (SourceFire)	DS1404087-E5
8660 Service Delivery Module spare disk drive	Spare Field Replaceable Hard Disk Drive with FW SW (20GB)	3.7.6 (SDM) 4.5 (SourceFire)	DS1411023
8660 Service Delivery Module spare PrPMC	Spare Field Replaceable PrPMC module	-	DS1411024
Security module			

Table 41
Ethernet Routing Switch 8600 modules and components (cont'd.)

Module or component		Minimum software version	Part number
8661 SSL Acceleration Module (SAM)	The 8661 SAM and WSM security solution also requires WebOS version 10.0.27.3 or newer. Ethernet Routing Switch 8600 Software Release 3.3.1 was specifically designed to introduce the 8661 SAM module. Release 3.3.2 does not support this module.	3.3.1	DS1404070
Layer 4 to 7 module			
Web Switching Module (WSM)	1000BASE-SX or 10BASE-T/100BASE-TX	3.1.3, 3.2.1, 3.3.0	DS1404045-E5
Ethernet modules			
8608GB module	Discontinued	3.0.0	DS1404015
8608GT module	Discontinued	3.1.0	DS1404012
8608SX module	Discontinued	3.0.0	DS1404003
8624FX module	Discontinued	3.0.0	DS1404005
8648TX module	Discontinued	3.0.0	DS1404002
10 Gigabit Ethernet modules			
8681XLW module	Discontinued	3.3.0	DS1404052
8681XLR module	Discontinued	3.3.0	DS1404053
Ethernet E modules			
8608GBE	8-port Gigabit Ethernet GBIC	3.1.1	DS1404038-E5
8608GTE	8-port 1000BASE-T	3.1.1	DS1404044-E5
8608SXE	8-port 1000BASE-SX	3.1.1	DS1404036-E5
8616SXE	16-port 1000BASE-SX	3.1.0	DS1404011-E5
8616GTE	16-port 1000BASE-T	3.3.0	DS1404034-E5
8624FXE	24-port 100BASE-FX	3.1.1	DS1404037-E5
8648TXE	48-port 10BASE-T/100BASE-TX	3.1.1	DS1404035-E5
8632TXE	32-port 10BASE-T/100BASE-TX plus 2-port GBIC	3.1.2	DS1404024-E5
Ethernet M modules			

Table 41
Ethernet Routing Switch 8600 modules and components (cont'd.)

Module or component		Minimum software version	Part number
8608GBM	8-port Gigabit Ethernet GBIC	3.3.0	DS1404059-E5
8608GTM	8-port 1000BASE-T	3.3.0	DS1404061-E5
8632TXM	32-port 10BASE-T/100BASE-TX plus 2-port GBIC	3.3.0	DS1404055-E5
8648TXM	48-port 10BASE-T/100BASE-TX	3.3.0	DS1404056-E5
Ethernet R modules			
8630GBR	30-port Gigabit Ethernet SFP GBIC baseboard	4.0.0	DS1404063-E5
8648GTR	48-port 10BASE-T/ 100BASE-TX/1000Base-T	4.0.0	DS1404092-E5
8683XLR	3-port 10 Gbit/s LAN XFP baseboard	4.0.0	DS1404101-E5
8683XZR	3-port 10 Gbit/s LAN/WAN XFP baseboard	4.1.0	DS1404064-E5
Ethernet RS modules			
8612XLRS	12-port 10 GbE LAN module	5.0.0	DS1404097-E6
8634XGRS	Combination 2-port 10 GbE; 24-port SFP; 8-port RJ-45	5.0.0	DS1404109-E6
8648GBRS	48-port SFP baseboard	5.0.0	DS1404102-E6
8648GTRS	48-port 10BASE-T/100BASE-TX/1000BASE-T	5.0.0	DS1404110-E6
ATM/ATME/ATMM modules			
8672ATM module	Discontinued	3.1.0	DS1304001
8672ATME module	2-slot MDA baseboard ATME module	3.1.1	DS1304008-E5
8672ATMM module	2-slot MDA baseboard ATMM module-expanded memory	3.3.0	DS1304009-E5
ATM/ATME/ATMM MDAs			
DS-3 MDA	2-port 75 ohm coaxial	3.3.0	DS1304002-E5
OC-12c/STM-4 MDA	1-port MMF	3.1.0, 3.1.1, 3.3.0	DS1304004-E5
OC-12c/STM-4 MDA	1-port SMF	3.1.0, 3.1.1, 3.3.0	DS1304005-E5

Table 41
Ethernet Routing Switch 8600 modules and components (cont'd.)

Module or component		Minimum software version	Part number
OC-3c/STM-1 MDA	4-port MMF	3.1.0, 3.1.1, 3.3.0	DS1304006-E5
OC-3c/STM-1 MDA	4-port SMF	3.1.0, 3.1.1, 3.3.0	DS1304007-E5
POS/POSE/POSM modules			
8683POS module	Discontinued	3.1.0	DS1404016
8683POSE module	Discontinued	3.1.1	DS1404043
8683POSM module	POSM 3-slot MDA baseboard	3.3.0	DS1404060-E5
POS/POSE/POSM MDAs			
OC-3c/STM-1 MDA	2-port MMF	3.1.0, 3.1.1, 3.3	DS1333003-E5
OC-3c/STM-1 MDA	2-port SMF	3.1.0, 3.1.1, 3.3	DS1333004-E5
OC-12c/STM-4 MDA	1-port MMF	3.1.0, 3.1.1, 3.3	DS1333001-E5
OC-12c/STM-4 MDA	1-port SMF	3.1.0, 3.1.1, 3.3	DS1333002-E5
GBICs			
1000BASE-SX	850 nm	3.0.0	AA1419001-E5
1000BASE-LX	1300 nm	3.0.0	AA1419002-E5
1000BASE-XD	Up to 50 km, SC duplex SMF	3.0.0	AA1419003-E5
1000BASE-ZX	Up to 70 km, SC duplex SMF	3.0.0	AA1419004-E5
CWDM GBICs	Discontinued	3.1.2	AA1419005 to AA1419012
1000BASE-EX CWDM APD	1470 nm to 1610 nm	3.1.4	AA1419017-E5 to AA1419024-E5
1000BASE-T	Category 5 copper unshielded twisted pair (UTP)	3.5.0	AA1419041-E5
SFPs			
1000BASE-XD CWDM	1470 nm to 1610 nm		AA1419025-E5 to AA1419032-E5
1000BASE-ZX CWDM	1470 nm to 1610 nm		AA1419033-E5 to AA1419040-E5
1000BASE-T	CAT 5 UTP PAM-5		AA1419043-E6

Table 41
Ethernet Routing Switch 8600 modules and components (cont'd.)

Module or component		Minimum software version	Part number
1000BASE-SX	Up to 550 m, 850 nm DDI	DDI requires 5.0.0	AA1419048-E6
1000BASE-LX	Up to 10 km, 1310 nm DDI	DDI requires 5.0.0	AA1419049-E6
1000BASE-XD	Up to 40 km, 1310 nm DDI	DDI requires 5.0.0	AA1419050-E6
1000BASE-XD	Up to 40 km, 1550 nm DDI	DDI requires 5.0.0	AA1419051-E6
1000BASE-ZX	Up to 70 km, 1550 nm DDI	DDI requires 5.0.0	AA1419052-E6
1000BASE-XD CWDM	Up to 40 km, 1470 nm to 1610 nm DDI	DDI requires 5.0.0	AA1419053-E6 to AA1419060-E6
1000BASE-ZX CWDM	Up to 70 km, 1470 nm to 1610 nm DDI	DDI requires 5.0.0	AA1419061-E6 to AA1419068-E6
1000BASE-BX	Up to 10 km, 1310 nm DDI	4.1.0	AA1419069-E6
1000BASE-BX	Up to 10 km, 1490 nm DDI	4.1.0	AA1419070-E6
1000BASE-EX	Up to 120 km, 1550 nm DDI	5.0.0	AA1419071-E6
XFPs			
10GBASE-LR	1310 nm LAN/WAN	DDI requires 5.0	AA1403001-E5
10GBASE-ER	1550 nm LAN/WAN	DDI requires 5.0	AA1403003-E5
10GBASE-SR	850 nm LAN	DDI requires 5.0	AA1403005-E5
10GBASE-ZR	1550 nm LAN/WAN	4.1.0; DDI requires 5.0	AA1403006-E5
10GBASE-LRM	Up to 300 m	5.0.0; DDI requires 5.0	AA1403007-E6
DWDM XFPs			
10GBASE DWDM	1530.33 nm (195.90 Terahertz [THz])	5.1.0	NTK587AEE5
10GBASE DWDM	1531.12 nm (195.80 THz)	5.1.0	NTK587AGE5
10GBASE DWDM	1531.90 nm (195.70 THz)	5.1.0	NTK587AJE5
10GBASE DWDM	1532.68 nm (195.60 THz)	5.1.0	NTK587ALE5
10GBASE DWDM	1533.47 nm (195.50 THz)	5.1.0	NTK587ANE5
10GBASE DWDM	1534.25 nm (195.40 THz)	5.1.0	NTK587AQE5

Table 41
Ethernet Routing Switch 8600 modules and components (cont'd.)

Module or component		Minimum software version	Part number
10GBASE DWDM	1535.04 nm (195.30 THz)	5.1.0	NTK587ASE5
10GBASE DWDM	1535.82 nm (195.20 THz)	5.1.0	NTK587AUE5
10GBASE DWDM	1536.61 nm (195.10 THZ)	5.1.0	NTK587AWE5
10GBASE DWDM	1537.40 nm (195.0 THz)	5.1.0	NTK587AYE5
10GBASE DWDM	1538.19 nm (194.9 THz)	5.1.0	NTK587BAE5
10GBASE DWDM	1538.98 nm (194.8 THz)	5.1.0	NTK587BCE5
10GBASE DWDM	1539.77 nm (194.7 THz)	5.1.0	NTK587BEE5
10GBASE DWDM	1540.56 nm (194.6 THz)	5.1.0	NTK587BGE5
10GBASE DWDM	1541.35 nm (194.5 THz)	5.1.0	NTK587BJE5
10GBASE DWDM	1542.14 nm (194.4 THz)	5.1.0	NTK587BLE5
10GBASE DWDM	1542.94 nm (194.3 THz)	5.1.0	NTK587BNE5
10GBASE DWDM	1543.73 nm (194.2 THz)	5.1.0	NTK587BQE5
10GBASE DWDM	1544.53 nm (194.1 THz)	5.1.0	NTK587BSE5
10GBASE DWDM	1545.32 nm (194.0 THz)	5.1.0	NTK587BUE5

The following information pertains to the information shown in the preceding table:

- Ethernet Routing Switch 8600 Software Release 3.1.3 is the first and only release in the 3.1.x software branch that supports the Web Switching Module.
- Ethernet Routing Switch 8600 Software Release 3.2.1 (and later) supports the Web Switching Module.
- Ethernet Routing Switch 8600 Software Release 3.3.0 introduced support for WebOS 10.0 on the Web Switching Module.
- The 8624FX and the 8624FXE modules support only full-duplex mode and cannot connect to half-duplex devices.
- ATM MDAs inserted into an 8672ATME module require Ethernet Routing Switch 8600 Software Release 3.1.1 or later. ATM MDAs inserted into a 8672ATMM module require Ethernet Routing Switch 8600 Software Release 3.3.0 or later.
- POS MDAs inserted into a 8683POSE module require Ethernet Routing Switch 8600 Software Release 3.1.1 or later, and POS MDAs

inserted into a 8683POSM module require Ethernet Routing Switch 8600 Software Release 3.3.0 or later.

- Unsupported GBICs are indicated by the CLI and Device Manager as GBIC-other.

Appendix Supported standards, RFCs, and MIBs

This section identifies the IEEE standards, RFCs, and network management MIBs supported in this release.

IEEE standards

The following table lists supported IEEE standards.

Table 42
Supported IEEE standards

Supported standard	Description
EEE 802.1D	Spanning Tree Protocol
IEEE 802.1p	Priority Queues
IEEE 802.1Q	VLAN Tagging
IEEE 802.1s	Multiple Spanning Tree Protocol (MSTP)
IEEE 802.1w	Rapid Spanning Tree Protocol (RSTP)
IEEE 802.1v	VLAN Classification by Protocol and Port
IEEE 802.1x	Ethernet Authentication Protocol
IEEE 802.3	CSMA/CD Ethernet(ISO/IEC 8802-3)
IEEE 802.3ab	1000BASE-T Ethernet
IEEE 802.3ab	1000BASE-LX Ethernet
IEEE 802.3ab	1000BASE-ZX Ethernet
IEEE 802.3ab	1000BASE-CWDM Ethernet
IEEE 802.3ab	1000BASE-SX Ethernet
IEEE 802.3ab	1000BASE-XD Ethernet
IEEE 802.3ab	1000BASE-BX Ethernet
IEEE 802.3ad	Link Aggregation Control Protocol (LACP)
IEEE 802.3ae	10GBASE-X XFP
IEEE 802.3i	10BASE-T - Autonegotiation

Supported standard	Description
IEEE 802.3	10BASE-T Ethernet
IEEE 802.3u	100BASE-TX Fast Ethernet (ISO/IEC 8802-3,Clause 25)
IEEE 802.3u	100BASE-FX
IEEE 802.3u	Auto-negotiation on Twisted Pair (ISO/IEC 8802-3,Clause 28)
IEEE 802.3x	Flow Control on the Gigabit Uplink port
IEEE 802.3z	Gigabit Ethernet 1000BASE-SX and LX

IETF RFCs

This section identifies the supported IETF RFCs.

Layer 2 features ATM/POS

The following table describes the supported ATM/POS IETF RFCs for Layer 2 features.

Table 43
Supported ATM/POS RFCs

Supported standard	Description
RFC 1332	IPCP (POS module)
RFC 1471	LCP (POS module)
RFC 1473	NCP (POS module)
RFC 1474	Bridge NCP (POS module)
RFC 1552	IPXCP (POS module)
RFC 1638	BCP (POS module)
RFC 1661	PPP (POS module)
RFC 1989	PPP Link Quality Monitoring (POS module)
RFC 2558	Sonet / SDH (POS module)
RFC 2615	PPP over Sonet / SDH (POS module)

IPv4 Layer 3/Layer 4 Intelligence

The following table describes the supported IETF RFCs for IPv4 Layer 3/Layer 4 Intelligence.

Table 44
IPv4 Layer 3/Layer 4 Intelligence RFCs

Supported standard	Description
RFC 768	UDP Protocol
RFC 783	TFTP Protocol
RFC 791	IP Protocol

Supported standard	Description
RFC 792	ICMP Protocol
RFC 793	TCP Protocol
RFC 826	ARP Protocol
RFC 854	Telnet Protocol
RFC 894	A standard for the Transmission of IP Datagrams over Ethernet Networks
RFC 896	Congestion control in IP/TCP internetworks
RFC 903	Reverse ARP Protocol
RFC 906	Bootstrap loading using TFTP
RFC 950	Internet Standard Subnetting Procedure
RFC 951 / RFC 2131	BootP / DHCP
RFC 1027	Using ARP to implement transparent subnet gateways/ Nortel Subnet based VLAN
RFC 1058	RIPv1 Protocol
RFC 1112	IGMPv1
RFC 1253	OSPF
RFC 1256	ICMP Router Discovery
RFC 1305	Network Time Protocol v3 Specification, Implementation and Analysis3
RFC 1332	The PPP Internet Protocol Control Protocol (IPCP)
RFC 1340	Assigned Numbers
RFC 1541	Dynamic Host Configuration Protocol1
RFC 1542	Clarifications and Extensions for the Bootstrap Protocol
RFC 1583	OSPFv2
RFC 1587	The OSPF NSSA Option
RFC 1591	DNS Client
RFC 1631	NAT (Network Address Translation) — only with WSM
RFC 1695	Definitions of Managed Objects for ATM Management v8.0 using SMIv2
RFC 1723	RIP v2 - Carrying Additional Information
RFC 1745	BGP / OSPF Interaction
RFC 1771 / RFC 1772	BGP-4
RFC 1812	Router Requirements
RFC 1866	HTMLv2 Protocol
RFC 1965	BGP-4 Confederations

Supported standard	Description
RFC 1966	BGP-4 Route Reflectors
RFC 1998	An Application of the BGP Community Attribute in Multi-home Routing
RFC 1997	BGP-4 Community Attributes
RFC 2068	Hypertext Transfer Protocol
RFC 2131	Dynamic Host Control Protocol (DHCP)
RFC 2138	RADIUS Authentication
RFC 2139	RADIUS Accounting
RFC 2178	OSPF MD5 cryptographic authentication/ OSPFv2
RFC 2205	Resource ReSerVation Protocol (RSVP) — v1 Functional Specification
RFC 2210	The Use of RSVP with IETF Integrated Services
RFC 2211	Specification of the Controlled-Load Network Element Service
RFC 2236	IGMPv2 for snooping
RFC 2270	BGP-4 Dedicated AS for sites/single provide
RFC 2283	Multiprotocol Extensions for BGP-4
RFC 2328	OSPFv2
RFC 2338	VRRP: Virtual Redundancy Router Protocol
RFC 2362	PIM-SM
RFC 2385	BGP-4 MD5 authentication
RFC 2439	BGP-4 Route Flap Dampening
RFC 2453	RIPv2 Protocol
RFC 2475	An Architecture for Differentiated Service
RFC 2547	BGP/MPLS VPNs
RFC 2597	Assured Forwarding PHB Group
RFC 2598	An Expedited Forwarding PHB
RFC 2702	Requirements for Traffic Engineering Over MPLS
RFC 2765	Stateless IP/ICMP Translation Algorithm (SIIT)
RFC 2796	BGP Route Reflection — An Alternative to Full Mesh IBGP
RFC 2819	Remote Monitoring (RMON)
RFC 2858	Multiprotocol Extensions for BGP-4
RFC 2918	Route Refresh Capability for BGP-4
RFC 2961	RSVP Refresh Overhead Reduction Extensions
RFC 2992	Analysis of an Equal-Cost Multi-Path Algorithm
RFC 3031	Multiprotocol Label Switching Architecture

Supported standard	Description
RFC 3032	MPLS Label Stack Encoding
RFC 3036	LDP Specification
RFC 3037	LDP Applicability
RFC 3065	Autonomous System Confederations for BGP
RFC 3210	Applicability Statement for Extensions to RSVP for
RFC 3215	LDP State Machine
RFC 3270	Multi-Protocol Label Switching (MPLS) Support of Differentiated Services
RFC 3376	Internet Group Management Protocol, v3
RFC 3392	Capabilities Advertisement with BGP-4 LSP-Tunnels
RFC 3443	Time To Live (TTL) Processing in Multi-Protocol Label Switching (MPLS) Networks
RFC 3569	An overview of Source-Specific Multicast (SSM)
RFC 3917	Requirements for IP Flow Information Export (IPFIX)
RFC 4364	BGP/MPLS IP Virtual Private Networks (VPNs)
RFC 4379	Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures
draft-holbrook-idmr-igmpv3-ssm -02.txt	IGMPv3 for SSM
draft-ietf-bfd-v4v6-1hop-06	IETF draft Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)

IPv4 Multicast

The following table describes the supported IETF RFCs for IPv4 Multicast.

Table 45
IPv4 Multicast RFCs

Supported standard	Description
RFC 1075	DVMRP Protocol
RFC 1112	IGMP v1 for routing / snooping
RFC 1519	Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy
RFC 2236	IGMP v2 for routing / snooping
RFC 2362	+ some PIM-SM v2 extensions (PIM-SM)
RFC 3446	Anycast Rendevous Point (RP) mechanism using Protocol Independent Multicast (PIM) and Multicast Source Discovery Protocol (MSDP)

Supported standard	Description
RFC 3618	Multicast Source Discovery Protocol (MSDP)
RFC 3768	Virtual Router Redundancy Protocol (VRRP)

IPv6

The following table describes the supported IETF RFCs for IPv6.

Table 46
IPv6 RFCs

Supported standard	Description
RFC 1881	IPv6 Address Allocation Management
RFC 1886	DNS Extensions to support IP version 6
RFC 1887	An Architecture for IPv6 Unicast Address Allocation
RFC 1981	Path MTU Discovery for IP v6
RFC 2030	Simple Network Time Protocol (SNTP) v4 for IPv4, IPv6 & OSI
RFC 2373	IPv6 Addressing Architecture
RFC 2375	IPv6 Multicast Address Assignments
RFC 2460	Internet Protocol, v6 (IPv6) Specification
RFC 2461	Neighbor Discovery
RFC 2462	IPv6 Stateless Address Autoconfiguration
RFC 2463	Internet Control Message Protocol (ICMPv6) for the Internet Protocol v6 (IPv6) Specification
RFC 2464	Transmission of IPv6 Packets over Ethernet Networks
RFC 2474	Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers
RFC 2526	Reserved IPv6 Subnet Anycast Addresses
RFC 2710	Multicast Listener Discovery (MLD) for IPv6
RFC 2740	OSPF for IPv6
RFC 2893	Configured Tunnels and Dual Stack Routing per port
RFC 2893	Transition Mechanisms for IPv6 Hosts and Routers
RFC 3056	Connection of IPv6 Domains via IPv4 Clouds
RFC 3363	Representing Internet Protocol Version 6 Addresses in DNS3
RFC 3484	Default Address Selection for IPv6
RFC 3513	Internet Protocol Version 6 (IPv6) Addressing Architecture
RFC 3587	IPv6 Global Unicast Address Format
RFC 3596	DNS Extensions to Support IP v6
RFC 3587	IPv6 Global Unicast Address Format

Supported standard	Description
RFC 3590	Source Address Selection for the Multicast Listener Discovery (MLD) Protocol
RFC 3596	DNS Extensions to support IP version 6
RFC 3810	IPv6 Multicast capabilities SSH/SCP, Telnet, Ping, CLI, JDM support for IPv6

Platform

The following table describes the supported IETF platform RFCs.

Table 47
Platform RFCs

Supported standard	Description
RFC 1305	(NTP client / unicast mode only)
RFC 1340	Assigned Numbers
RFC 1350	The TFTP Protocol (Revision 2)

Quality of Service (QoS)

The following table describes the supported IETF RFCs for Quality of Service (QoS).

Table 48
QoS RFCs

Supported standard	Description
RFC 2474 / RFC 2475	DiffServ Support
RFC 2597 / RFC 2598	DiffServ per Hop Behavior

Network Management

The following table describes the supported IETF RFCs for Network Management.

Table 49
Network Management RFCs

Supported standard	Description
RFC 1155	SMI
RFC 1157	SNMP
RFC 1215	Convention for defining traps for use with the SNMP
RFC 1269	Definitions of Managed Objects for the Border Gateway Protocol: v3
RFC 1271	Remote Network Monitoring Management Information Base

Supported standard	Description
RFC 1304	Definitions of Managed Objects for the SIP Interface Type
RFC 1354	IP Forwarding Table MIB
RFC 1389	RIP v2 MIB Extensions
RFC 1565	Network Services Monitoring MIB
RFC 1757 / RFC 2819	RMON
RFC 1907	SNMPv2
RFC 1908	Coexistence between v1 & v2 of the Internet-standard Network Management Framework
RFC 1930	Guidelines for creation, selection, and registration of an Autonomous System (AS)
RFC 2571	An Architecture for Describing SNMP Management Frameworks
RFC 2572	Message Processing and Dispatching for the Simple Network Management Protocol (SNMP)
RFC2573	SNMP Applications
RFC 2574	User-based Security Model (USM) for v3 of the Simple Network Management Protocol (SNMPv3)
RFC 2575	View-based Access Control Model (VACM) for the Simple Network Management Protocol (SNMP)
RFC 2576	Coexistence between v1, v2, & v3 of the Internetstandard Network Management Framework

Supported network management MIBs

The Ethernet Routing Switch 8600 includes an SNMPv1/v2/v2c/v3 agent with Industry Standard MIBs, as well as private MIB extensions, which ensure compatibility with existing network management tools.

All these MIBs are included with any software version that supports them. Consult the Nortel Web site for a file called mib.zip, which contains all MIBs, and a special file called *manifest*.

The following tables list the network management MIBs and standards that this release supports.

Table 50
Standard IEEE MIBs

Protocol	IEEE standard	File name
LACP	802.3ad	ieee802-lag.mib
EAPoL	802.1x	ieee8021x.mib

Table 51
Standard MIBs (RFC)

RFC number	MIB name
RFC 1212	Concise MIB definitions
RFC 1213	TCP/IP Management Information Base
RFC 1213	MIB II
RFC 1354	IP Forwarding Table MIB
RFC 1389 / RFC 1724	RIPv2 MIB extensions
RFC 1398	Definitions of Managed Objects for the Ethernet-Like Interface Types
RFC 1406	Definitions of Managed Objects for the DS1 and E1 Interface Types
RFC 1414	Identification MIB
RFC 1442	Structure of Management Information for version 2 of the Simple Network Management Protocol (SNMPv2)
RFC 1447	Party MIB for v2 of the Simple Network Management Protocol bytes)
RFC 1450	Management Information Base for v2 of the Simple Network Management Protocol (SNMPv2)
RFC 1472	The Definitions of Managed Objects for the Security Protocols of the Point-to-Point Protocol
RFC 1483	Multiprotocol Encapsulation over ATM Adaptation Layer 5
RFC 1493	Bridge MIB
RFC 1525	Definitions of Managed Objects for Source Routing Bridges
RFC 1565	Network Services Monitoring MIB
RFC 1573	Interface MIB
RFC 1643	Ethernet MIB
RFC 1650	Definitions of Managed Objects for the Ethernet-like Interface Types using SMIv2
RFC 1657	BGP-4 MIB using SMIv2
RFC 1658	Definitions of Managed Objects for Character Stream Devices using SMIv2.)
RFC 1695	Definitions of Managed Objects for ATM Management v8.0 using SMIv2
RFC 1696	Modem Management Information Base (MIB) using SMIv2
RFC 1724	RIP v2 MIB Extension
RFC 1850	OSPF MIB
RFC 2021	RMON MIB using SMIv2
RFC 2037	Entity MIB using SMIv2
RFC 2096	IP Forwarding Table MIB
RFC 2233	Interfaces Group MIB using SMIv2

RFC number	MIB name
RFC 2452	IPv6 MIB: TCP MIB
RFC 2454	IPv6 MIB: UDP MIB
RFC 2465	IPv6 MIB: IPv6 General group and textual conventions
RFC 2466	IPv6 MIB: ICMPv6 Group
RFC 2578	Structure of Management Information v2 (SMIv2)
RFC 2613	Remote Network Monitoring MIB Extensions for Switched Networks v1.0
RFC 2665	Definitions of Managed Objects for the Ethernet-like Interface Types
RFC 2668	Definitions of Managed Objects for IEEE 802.3 Medium Attachment Units (MAUs)
RFC 2674	Bridges with Traffic MIB
RFC 2787	Definitions of Managed Objects for the Virtual Router Redundancy Protocol
RFC 2863	Interface Group MIB
RFC 2925	Remote Ping, Traceroute & Lookup Operations MIB
RFC 2932	IPv4 Multicast Routing MIB
RFC 2933	IGMP MIB
RFC 2934	PIM MIB
RFC 3019	IPv6 MIB: MLD Protocol
RFC 3411	An Architecture for Describing Simple Network Management Protocol (SNMP) Management Frameworks
RFC 3412	Message Processing and Dispatching for the Simple Network Management Protocol (SNMP)
RFC 3416	v2 of the Protocol Operations for the Simple Network Management Protocol (SNMP)
RFC 3635	Definitions of Managed Objects for the Ethernet-like Interface Types
RFC 3636	Definitions of Managed Objects for IEEE 802.3 Medium Attachment Units (MAUs)
RFC 3810	Multicast Listener Discovery v2 (MLDv2) for IPv6
RFC 3811	Definitions of Textual Conventions (TCs) for Multiprotocol Label Switching (MPLS) Management
RFC 3812	Multiprotocol Label Switching (MPLS) Traffic Engineering (TE) Management Information Base (MIB)
RFC 3813	Multiprotocol Label Switching (MPLS) Label Switching Router (LSR) Management Information Base (MIB)
RFC 3815	Definitions of Managed Objects for the Multiprotocol Label Switching (MPLS), Label Distribution Protocol (LDP)

RFC number	MIB name
RFC 4022	Management Information Base for the Transmission Control Protocol (TCP) 4087 IP Tunnel MIB
RFC 4113	Management Information Base for the User Datagram Protocol (UDP)
RFC 4624	Multicast Source Discovery Protocol (MSDP) MIB

Table 52
Proprietary MIBs

Proprietary MIB name	File name
Rapid City MIB	rapid_city.mib
SynOptics Root MIB	synro.mib
Other SynOptics definitions	s5114roo.mib
Other SynOptics definitions	s5tcs112.mib
Other SynOptics definitions	s5emt103.mib
Nortel RSTP/MSTP proprietary MIBs	nnrst000.mib, nnmst000.mib
Nortel IPX MIBs	ipx_rcc.mib, ipxripsap_rcc.mib
Nortel IGMP MIB	rfc_igmp.mib
Nortel IP Multicast MIB	ipmroute_rcc.mib
Nortel DVMRP MIB	dvmrp_rcc.mib
Nortel PIM MIB	pim-rcc.mib
Nortel ATM MIB	atm_tc.mib
Nortel MIB definitions	wf_com.mib
Nortel PGM MIB	wf_pgm.mib
The Definitions of Managed Objects for the Link Control Protocol of the Point-to-Point Protocol - Nortel Proprietary	rfc1471rcc.mib
The Definitions of Managed Objects for the IP Network Control Protocol of the Point-to-Point Protocol - Nortel Proprietary	rfc1473rcc.mib
The Definitions of Managed Objects for the Bridge Network Control Protocol of the Point-to-Point Protocol	rfc1474rcc.mib
Definitions of Managed Objects for the SONET/SDH Interface Type - Nortel Proprietary	rfc1595rcc.mib

Table 52
Proprietary MIBs (cont'd.)

Proprietary MIB name	File name
OSPF Version 2 Management Information Base - Nortel proprietary extensions	rfc1850t_rcc.mib
Nortel IPv6 proprietary MIB definitions	rfc_ipv6_tc.mib, inet_address_tc.mib, ipv6_flow_label.mib

Index

D

E

F

G

H

I

L

M

N

O

S

T

V

W

Nortel Ethernet Routing Switch 8600

Planning and Engineering — Network Design

Release: 5.1
Publication: NN46205-200
Document revision: 02.02
Document release date: 27 August 2010

To provide feedback or to report a problem in this document, go to www.nortel.com/documentfeedback.

www.nortel.com
LEGAL NOTICE