Planning and Engineering — Network Design
Avaya Ethernet Routing Switch 8800/8600

7.1.3
NN46205-200, 06.02
July 2012

Notice

While reasonable efforts have been made to ensure that the information in this document is complete and accurate at the time of printing, Avaya assumes no liability for any errors. Avaya reserves the right to make changes and corrections to the information in this document without the obligation to notify any person or organization of such changes.

Documentation disclaimer

“Documentation” means information published by Avaya in varying mediums which may include product information, operating instructions and performance specifications that Avaya generally makes available to users of its products. Documentation does not include marketing materials. Avaya shall not be responsible for any modifications, additions, or deletions to the original published version of documentation unless such modifications, additions, or deletions were performed by Avaya. End User agrees to indemnify and hold harmless Avaya, Avaya's agents, servants and employees against all claims, lawsuits, demands and judgments arising out of, or in connection with, subsequent modifications, additions or deletions to this documentation, to the extent made by End User.

Link disclaimer

Avaya is not responsible for the contents or reliability of any linked Web sites referenced within this site or documentation provided by Avaya. Avaya is not responsible for the accuracy of any information, statement or content provided on these sites and does not necessarily endorse the products, services, or information described or offered within them. Avaya does not guarantee that these links will work all the time and has no control over the availability of the linked pages.

Warranty

Avaya provides a limited warranty on its Hardware and Software (“Product(s)”). Refer to your sales agreement to establish the terms of the limited warranty. In addition, Avaya’s standard warranty language, as well as information regarding support for this Product while under warranty is available to Avaya customers and other parties through the Avaya Support Web site: http://support.avaya.com. Please note that if you acquired the Product(s) from an authorized Avaya reseller outside of the United States and Canada, the warranty is provided to you by said Avaya reseller and not by Avaya.

Licenses

THE SOFTWARE LICENSE TERMS AVAILABLE ON THE AVAYA WEBSITE, HTTP://SUPPORT.AVAYA.COM/LICENSEINFO/ ARE APPLICABLE TO ANYONE WHO DOWNLOADS, USES AND/OR INSTALLS AVAYA SOFTWARE, PURCHASED FROM AVAYA INC., ANY AVAYA AFFILIATE, OR AN AUTHORIZED AVAYA RESELLER (AS APPLICABLE) UNDER A COMMERCIAL AGREEMENT WITH AVAYA OR AN AUTHORIZED AVAYA RESELLER. UNLESS OTHERWISE AGREED TO BY AVAYA IN WRITING, AVAYA DOES NOT EXTEND THIS LICENSE IF THE SOFTWARE WAS OBTAINED FROM ANYONE OTHER THAN AVAYA, AN AVAYA AFFILIATE OR AN AVAYA AUTHORIZED RESELLER; AVAYA RESERVES THE RIGHT TO TAKE LEGAL ACTION AGAINST YOU AND ANYONE ELSE USING OR SELLING THE SOFTWARE WITHOUT A LICENSE. BY INSTALLING, DOWNLOADING OR USING THE SOFTWARE, OR AUTHORIZING OTHERS TO DO SO, YOU, ON BEHALF OF YOURSELF AND THE ENTITY FOR WHOM YOU ARE INSTALLING, DOWNLOADING OR USING THE SOFTWARE (HEREINAFTER REFERRED TO INTERCHANGEABLY AS “YOU” AND “END USER”), AGREE TO THESE TERMS AND CONDITIONS AND CREATE A BINDING CONTRACT BETWEEN YOU AND AVAYA INC. OR THE APPLICABLE AVAYA AFFILIATE (“AVAYA”).

Copyright

Except where expressly stated otherwise, no use should be made of materials on this site, the Documentation, Software, or Hardware provided by Avaya. All content on this site, the documentation and the Product provided by Avaya including the selection, arrangement and design of the content is owned either by Avaya or its licensors and is protected by copyright and other intellectual property laws including the sui generis rights relating to the protection of databases. You may not modify, copy, reproduce, republish, upload, post, transmit or distribute in any way any content, in whole or in part, including any code and software unless expressly authorized by Avaya. Unauthorized reproduction, transmission, dissemination, storage, and or use without the express written consent of Avaya can be a criminal, as well as a civil offense under the applicable law.

Third-party components

Certain software programs or portions thereof included in the Product may contain software distributed under third party agreements (“Third Party Components”), which may contain terms that expand or limit rights to use certain portions of the Product (“Third Party Terms”). Information regarding distributed Linux OS source code (for those Products that have distributed the Linux OS source code), and identifying the copyright holders of the Third Party Components and the Third Party Terms that apply to them is available on the Avaya Support Web site: http://support.avaya.com/Copyright.

Preventing Toll Fraud

“Toll fraud” is the unauthorized use of your telecommunications system by an unauthorized party (for example, a person who is not a corporate employee, agent, subcontractor, or is not working on your company's behalf). Be aware that there can be a risk of Toll Fraud associated with your system and that, if Toll Fraud occurs, it can result in substantial additional charges for your telecommunications services.

Avaya Toll Fraud Intervention

If you suspect that you are being victimized by Toll Fraud and you need technical assistance or support, call Technical Service Center Toll Fraud Intervention Hotline at +1-800-643-2353 for the United States and Canada. For additional support telephone numbers, see the Avaya Support Web site: http://support.avaya.com. Suspected security vulnerabilities with Avaya products should be reported to Avaya by sending mail to: securityalerts@avaya.com.

Trademarks

The trademarks, logos and service marks (“Marks”) displayed in this site, the Documentation and Product(s) provided by Avaya are the registered or unregistered Marks of Avaya, its affiliates, or other third parties. Users are not permitted to use such Marks without prior written consent from Avaya or such third party which may own the Mark. Nothing contained in this site, the Documentation and Product(s) should be construed as granting, by implication, estoppel, or otherwise, any license or right in and to the Marks without the express written permission of Avaya or the applicable third party.

Avaya is a registered trademark of Avaya Inc.

All non-Avaya trademarks are the property of their respective owners, and “Linux” is a registered trademark of Linus Torvalds.

Downloading Documentation

For the most current versions of Documentation, see the Avaya Support Web site: http://support.avaya.com.

Contact Avaya Support

Avaya provides a telephone number for you to use to report problems or to ask questions about your Product. The support telephone number is 1-800-242-2121 in the United States. For additional support telephone numbers, see the Avaya Web site: http://support.avaya.com.

Contents

Chapter 1: 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 Avaya Ethernet Routing Switch 8800/8600.

Notices

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

Attention notice

Important:

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

Caution ESD notice

Electrostatic alert:

ESD

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

Electrostatic alert:

ESD (décharge électrostatique)

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

Electrostatic alert:

ACHTUNG ESD

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

⚠ Electrostatic alert:

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 Avaya.

⚠ Electrostatic alert:

CUIDADO ESD

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

⚠ Electrostatic alert:

ATTENZIONE ESD

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

Caution notice

⚠ Caution:

Caution notices provide information about how to avoid possible service disruption or damage to Avaya 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 Avaya.

⚠ Caution:

ACHTUNG

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

⚠ Caution:

PRECAUCIÓN

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

⚠ Caution:

CUIDADO

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

⚠ Caution:

ATTENZIONE

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

Chapter 2: Purpose of this document

This document describes a range of design considerations and related information that helps you to optimize the performance and stability of your Avaya Ethernet Routing Switch 8800/8600 network.

Important:

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

Chapter 3: New in this release

The following section details what's new in *Avaya Ethernet Routing Switch 8800/8600 Planning and Engineering — Network Design,* NN46205-200 for Release 7.1.3.

Features

See the following section for information about feature changes.

8812XL SFP+ I/O module

Release 7.1.3 introduces a new Ethernet Routing Switch 8800 interface module — the 8812XL SFP+ I/O module. This module supports 12 SFP+ ports at 10Gbps and provides the same functionality as its RS module equivalent, the 8612XLRS.

The 8812XL SFP+ I/O module, like all 8800 modules, uses the new enhanced network processor-the RSP 2.7. It requires a minimum software version of 7.1.3 to operate.

Other changes

See the following section for information about changes in release 7.1.3 that are not feature-related.

SPBM implementation option name change

The Shortest Path Bridging MAC (SPBM) implementation option called Global Routing Table (GRT) Shortcuts is now changed to Internet Protocol (IP) Shortcuts. This document has been updated to reflect the change.

Chapter 4: Network design fundamentals

To efficiently and cost-effectively use your Avaya 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 the following:

- 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.

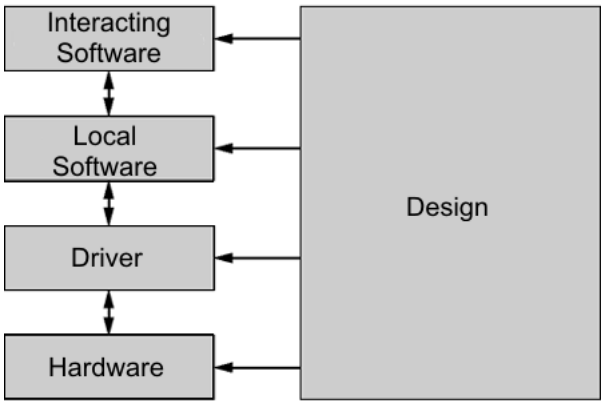


Figure 1: Hardware and software interaction

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 the same device. No external interaction is needed, so you can easily test its function.

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 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, one goal of network design is to off-load the interacting software level as much as possible to the other levels, especially to the hardware level. Therefore, Avaya recommends that you follow these generic rules when you design networks:

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

Chapter 5: Hardware fundamentals and guidelines

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

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 Avaya Ethernet Routing Switch 8800/8600 chassis provides redundant power options, depending on the chassis and the number of modules installed.

The 8003-R chassis supports up to two power supplies, and the 8006 and 8010 chassis support up to three power supplies. You must install at least one power supply for each chassis.

To determine the number of power supplies required for your switch configuration, use the *Power Supply Calculator for Avaya ERS 8800/8600, NN48500-519*. This is available at www.avaya.com/support with the rest of the ERS 8800/8600 documentation.To support a full configuration of RS modules, you require an 8004 or 8005 power supply. Do not mix 8004 and 8005 power supplies in the same chassis.

Power supply circuit requirements

The Avaya Ethernet Routing Switch 8800/8600 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.

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

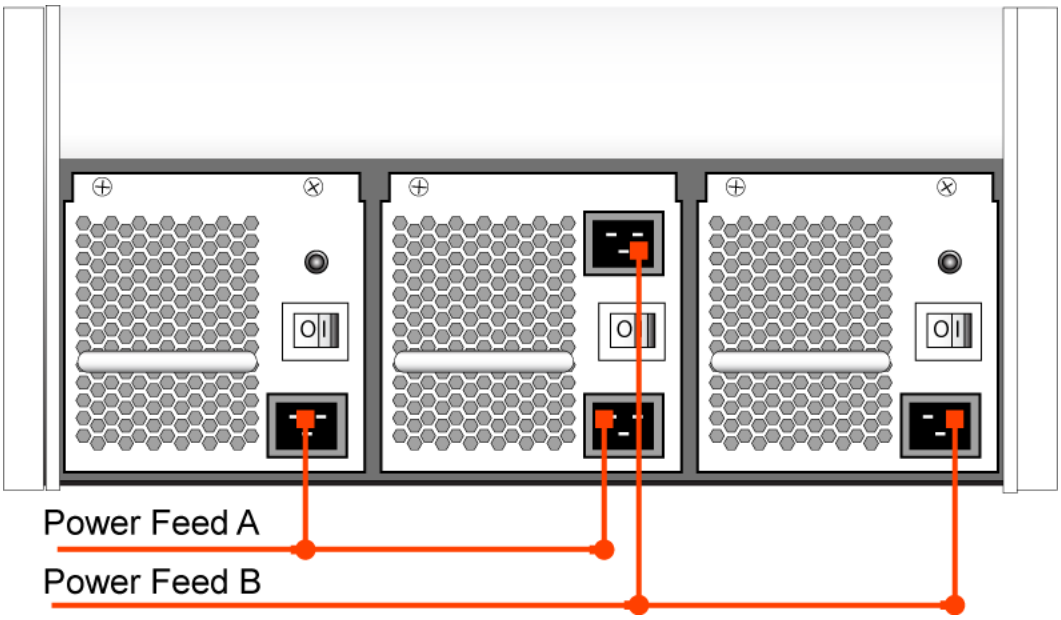


Figure 2: Dual-input power supply connections

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

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

You can use the Avaya 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 also change.

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 watts (W). For 220 AC volts (VAC), the draw is 1462 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 more information about these specifications, see *Avaya Ethernet Routing Switch 8800/8600 Installation — AC Power Supply, NN46205-306* and *Avaya Ethernet Routing Switch 8800/8600 Installation — DC Power Supply, NN46205-307*. Add these values to the cooling capacity calculation. For a multiple 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 or 8800 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 module. For information about installing high speed cooling modules, see *Avaya Ethernet Routing Switch 8800/8600 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 Avaya Ethernet Routing Switch 8800/8600 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.

Release 7.0 supports only the 8895 SF/CPU and the 8692 SF/CPU with SuperMezz. The 8692 SF/CPU must be equipped with the Enterprise Enhanced CPU Daughter Card (SuperMezz) for proper functioning with Release 7.0 and later. The 8895 SF/CPU has SuperMezz capabilities built into the module, and so does not support a SuperMezz card.

The use of dual 8692 SF/CPU or 8895 SF/CPU modules enables a maximum switch bandwidth of 512 Gbit/s. Dual modules provide redundancy and load sharing between the modules. Split

MultiLink Trunking (SMLT) in the core in a resilient cluster configuration (redundant switch with two 8692 or 8895 SF/CPU modules) can provide over 1 terabit per second (Tbit/s) of core switching capacity.

You can install the 8692 or 8895 SF/CPU module in slots 5 or 6 of the 8006, 8010, or 8010co chassis. The 8692 and 8895 SF/CPU modules are supported in slot 3 of the 8003-R chassis.

With dual SF/CPUs, you must install the same type of SF/CPU in both slots. You cannot install one type of SF/CPU in one slot, and a different type in the other slot. For example, you cannot install the 8692 SF/CPU in one slot and the 8895 SF/CPU in the other.

8800 series I/O modules

All 8800 series modules provide the same functionality as the RS module equivalents. The 8800 modules however use an enhanced network processor — the RSP 2.7.

Important:

Support for 8800 series I/O modules start with Release 7.1. They are not backwards compatible with older Ethernet Routing Switch 8800/8600 releases.

The following table displays the 8800 series modules and their RS module equivalents.

RS module	8800 series module
8648GTRS	8848GT
8648GBRS	8848GB
8634XGRS	8834XG
8612XLRS	8812XL

Important:

Ensure that you are running software release 7.1 or later for the 8800 modules to operate properly.

R and RS modules continue to be supported and you can install a mix of R/RS and 8800 modules in the same chassis.

Important:

You can only replace a module with another module of the same type. For example, you cannot replace a 48– port copper module with a fiber module. You can replace a copper module with another copper module, or a fiber module with another fiber module.

Note:

RS and 8800 I/O modules require a High Speed Cooling Module. If the High Speed Cooling Module is not installed in the chassis, these I/O modules will not power on.

To help you configure Avaya Ethernet Routing Switch 8800/8600 Ethernet modules, see *Avaya Ethernet Routing Switch 8800/8600 Configuration — Ethernet Modules*, (NN46205–503).

For module specifications and installation procedures, see *Avaya Ethernet Routing Switch 8800/8600 Installation — Modules*, (NN46205–304).

For optical transceiver specifications and installation procedures, see *Avaya Ethernet Routing Switch 8800/8600 Installation — SFP, SFP+, XFP, and OADM Hardware Components*, (NN46205–320).

8812XL information and recommendations

Release 7.1.3 introduces a new 8800 interface module, the 8812XL SFP+ I/O module. This module supports 12 SFP+ ports at 10Gbps. For this module to operate properly, ensure that you are running a minimum software version of 7.1.3.

Note:

The 8812XL supports only Avaya-qualified 10Gbps SFP+ pluggable Ethernet transceivers. 1Gbps transceivers are not supported, even if Avaya-qualified.

If you plug a 1Gbps SFP onto the 8812XL I/O interface, the port remains offline and the system displays the following message in the system logs and on the console.

```
CPU# [date time] COP-SW INFO Slot #: 1G SFP detected in 10G SFP+
port #. Port offline.
```

Passive DACs are not supported but may be recognized. If so, they will display as `10GbOther`. However, they will not pass data.

RS modules

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 8895 SF/CPU or 8692 SF/CPU with SuperMezz.

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.

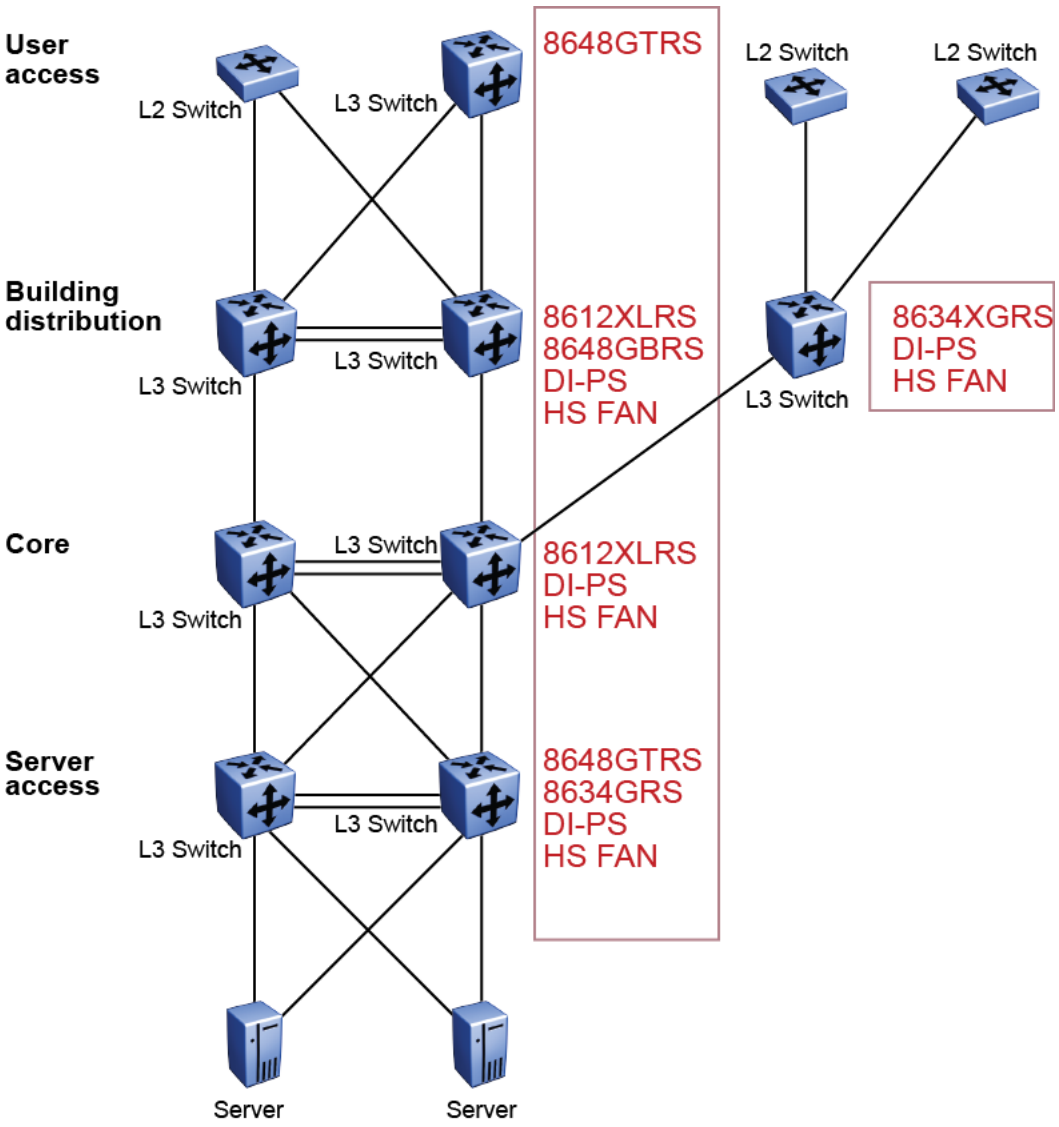


Figure 3: RS module usage

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)

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 more information about XFP specifications, see *Avaya Ethernet Routing Switch 8800/8600 Installation — SFP, SFP+, XFP, 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 Avaya Ethernet Routing Switch 8800/8600 supports the following R modules, which require the use of the 8895 SF/CPU or 8692 SF/CPU with SuperMezz:

- 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 8010, 8010co, 8006, and 8003-R 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.

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 by 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
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, Avaya 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 R modules and 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 ACLI 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. Chassis Revision A indicates that the chassis is not a high performance chassis and must be upgraded.

R and RS series modules and global FDB filters

The Avaya Ethernet Routing Switch 8800/8600 provides global forwarding database filter (FDB) operations for R and RS series modules. The global FDB filter command for R and RS series modules is **`config fdb fdb-filter add <mac-address>`**.

For more information about the FDB filters, see *Avaya Ethernet Routing Switch 8800/8600 Configuration — VLANs and Spanning Tree, NN46205-517*.

8648GTR recommendations

Avaya supports the 8648GTR module in a high-performance slot only. Avaya 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.

The 10 GbE modules support only full-duplex mode. In accordance with the IEEE 802.3ae standard, autonegotiation 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 autonegotiation
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 more information about XFP specifications, see *Avaya Ethernet Routing Switch 8800/8600 Installation — SFP, SFP+, XFP, 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 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, Avaya 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

Features and scaling

The following tables show scaling information and features available on the Avaya Ethernet Routing Switch 8800/8600. For the most recent scaling information, always consult the latest version of the Release Notes.

Table 5: Supported scaling capabilities

	Maximum supported 8692SF with SuperMezz or 8895SF (R or RS series modules)
Layer 2	

	Maximum supported 8692SF with SuperMezz or 8895SF (R or RS series modules)
MAC address table entries	64 000 32 000 when SMLT is used
VLANs (port- protocol-, and IEEE 802.1Q-based)	4000
IP subnet-based VLANs	800
Ports in Link Aggregation Group (LAG, MLT)	8
Aggregation groups 802.3ad aggregation groups Multi Link Trunking (MLT) group	128
SMLT links	128
SLT (single link SMLT)	382
VLANs on SMLT/IST link	with Max VLAN feature enabled: 2000
RSMLT per VLAN	32 SMLT links with RSMLT-enabled VLANs
RSTP/MSTP (number of ports)	384, with 224 active. Configure the remaining interfaces with Edge mode
MSTP instances	32
Advanced Filters	
ACLs for each system	4000
ACEs for each system	1000
ACEs for each ACL	1000
ACEs for each port	2000: 500 inPort 500 inVLAN 500 outPort 500 outVLAN
IP, IP VPN/MPLS, IP VPN Lite, VRF Lite	
IP interfaces (VLAN- and brouter-based)	1972
VRF instances	255
ECMP routes	5000
VRRP interfaces	255
IP forwarding table (Hardware)	250 000
BGP/mBGP peers	250
iBGP instances	on GRT
eBGP instances	on 256 VRFs (including GRT)
BGP forwarding routes BGP routing information base (RIB) BGP forwarding information base (FIB)	BGP FIB 250 000 BGP RIB 500 000

	Maximum supported 8692SF with SuperMezz or 8895SF (R or RS series modules)
IP VPN routes (total routes for each system)	180 000
IP VPN VRF instances	255
Static ARP entries	2048 in a VRF 10 000 in the system
Dynamic ARP entries	32 000
DHCP Relay instances (total for all VRFs)	512
Static route entries	2000 in a VRF 10 000 in the system
OSPF instances for each switch	on 64 VRFs (including GRT)
OSPF areas for each switch	5 in a VRF 24 in the system
OSPF adjacencies for each switch	80 200 in the system
OSPF routes	20 000 in a VRF 50 000 in the system
OSPF interfaces	238 500 in the system
OSPF LSA packet maximum size	3000 bytes
RIP instances	64
RIP interfaces	200
RIP routes	2500 in a VRF 10 000 in the system
Multiprotocol Label Switching	
MPLS LDP sessions	200
MPLS LDP LSPs	16 000
MPLS RSVP static LSPs	200
Tunnels	2500
IP Multicast	
DVMRP passive interfaces	1200
DVMRP active interfaces/neighbors	80
DVMRP routes	2500
PIM instances	64
PIM active interfaces	500 (200 for all VRFs)
PIM passive interfaces	1972 (2000 for all VRFs)
PIM neighbors	80 (200 for all VRFs)
Multicast streams: with SMLT/ without SMLT	2000/4000

	Maximum supported 8692SF with SuperMezz or 8895SF (R or RS series modules)
Multicast streams per port	1000
IGMP reports/sec	250
IPv6	
IPv6 interfaces	250
IPv6 tunnels	350
IPv6 static routes	2000
OSPFv3 areas	5
OSPFv3 adjacencies	80
OSPFv3 routes	5000
Operations, Administration, and Maintenance	
IPFIX	384 000 flows per chassis
RMON alarms with 4000K memory	2630
RMON events with 250K memory	324
RMON events with 4000K memory	5206
RMON Ethernet statistics with 250K memory	230
RMON Ethernet statistics with 4000K memory	4590

Avaya supports only 25 spanning tree groups (STG). Although you can configure up to 64 STGs, configurations of more than 25 STGs are not supported. If you need to configure more than 25 STGs, contact your Avaya Customer Support representative for more information.

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 more information about the Avaya optical routing system (Coarse Wavelength Division Multiplexing system) information, see Optical routing design on page 41.

Optical device guideline navigation

Optical power considerations

When you connect 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 three to five 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 *Avaya Ethernet Routing Switch 8800/8600 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 Avaya OM5200 10G optical transponder utilizes a 1310 nm client-side transmitter.

1000BASE-X and 10GBASE-X reach

Various SFP (1 Gbit/s), SFP+(10Gbit/s), and XFP (10 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 on page 47.

For more information about these devices, including compatible fiber type, see *Avaya Ethernet Routing Switch 8800/8600 Installation — SFP, SFP+, XFP, and OADM Hardware Components* (NN46205-320).

Table 6: Optical devices and maximum reach

Optical device (SFP/ SFP+/XFP)	Maximum reach
SFPs	
1000BASE-SX	Up to 275 or 550 m reach (fiber-dependent) over a fiber pair.
1000BASE-LX	Up to 10 km reach over a single mode fiber (SMF) pair. Up to 550 m reach over a multimode fiber (MMF) pair.
1000BASE-XD	Up to 40 km reach over a single mode fiber pair.
1000BASE-ZX	Up to 70 km reach over a single mode fiber pair.
1000BASE-BX	Up to 40 km reach. Bidirectional over one single mode fiber.
1000BASE-EX	Up to 120 km reach over a single mode fiber pair.
SFP+	
10GBASE-LR/LW	Up to 10 km.
10GBASE-ER/EW	Up to 40 km.
10GBASE-SR/SW	Using 62.5 µm MMF optic cable: • 160 MHz-km fiber: 2 to 26 m • 200 MHz-km fiber: 2 to 33 m Using 50 µm MMF optic cable: • 400 MHz-km fiber: 2 to 66 m • 500 MHz-km fiber: 2 to 82 m • 2000 MHz-km fiber: 2 to 300 m
10GBASE-LRM	Up to 220 m.
10GBASE-CX	4-pair direct attach twinaxial copper cable to connect 10 Gb ports. Supported ranges are 3 m, 5m and 10 m.
XFP	
10GBASE-SR	Using 62.5 µm MMF optic cable: • 160 MHz-km fiber: 2 to 26 m • 200 MHz-km fiber: 2 to 33 m Using 50 µm MMF optic cable: • 400 MHz-km fiber: 2 to 66 m • 500 MHz-km fiber: 2 to 82 m • 2000 MHz-km fiber: 2 to 300 m
10GBASE-LR/LW	Up to 10 km
10GBASE-LRM	Up to 220 m

Optical device (SFP/ SFP+/XFP)	Maximum reach
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 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 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. With the availability of 10 Gbit/s and 40 Gbit/s devices, you must consider dispersion.

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 7: 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 8: 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, measure the PMD of the fiber proposed to carry 10 Gbit/s.

The following table shows the PMD limits.

Table 9: 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.

If no PMD compensating devices are available and the proposed fiber is over the PMD limit, 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 10: 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
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, 4 pair: 100 m
STP/Coaxial cable distance	500 m	100 m	N/A	

10/100BASE-TX Autonegotiation recommendations

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

The autonegotiation 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 autonegotiation 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 autonegotiation. In this case, only the link speed is sensed; not the duplex mode. Avaya recommends the autonegotiation settings as shown in the following table, where A and B are two Ethernet devices.

Table 11: Recommended autonegotiation setting on 10/100BASE-TX ports

Port on A	Port on B	Remarks	Recommendations
Autonegotiation enabled	Autonegotiation enabled	Ports negotiate on highest supported mode on both sides.	Recommended setting if both ports support autonegotiation mode.

Port on A	Port on B	Remarks	Recommendations
Full-duplex	Full-duplex	Both sides require the same mode.	Recommended setting if full-duplex is required, but autonegotiation is not supported.

Autonegotiation 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 autonegotiation 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 autonegotiation 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, you must enable autonegotiation.

Important:

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 you must apply the same configuration 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 use CANA 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 on page 57. For information on remote fault indication for Gigabit Ethernet, see Gigabit Ethernet and remote fault indication on page 58.

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 noncontrol multicast traffic and noncontrol broadcast traffic on the physical port that is not sent to the CPU (for example, IP subnet broadcast). The CP-Limit feature is configured by port within the chassis.

The CP-Limit default settings are as follows:

- 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 can impair network traffic flow, as this is a critical port for SMLT configurations. Avaya recommends that an IST MLT contain at least two physical ports, although this is not a requirement. Avaya also recommends that you disable CP-Limit 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 if a protection condition (CP-Limit) arises. Be aware that, although one of the SMLT MLT ports (risers) is likely to be disabled in such a condition, traffic continues 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 protects the CPU from any traffic hitting the CPU by shutting down the ports that 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 to 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 monitors the most highly utilized ports in the

chassis. The number of highly utilized ports monitored is configured in the MaxPorts parameter.

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

- MinCongTime (Minimum Congestion Time) sets the minimum time, in milliseconds, during which the CPU frame buffers can be oversubscribed 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, during which the bandwidth utilization on a port 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 after 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 is 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.

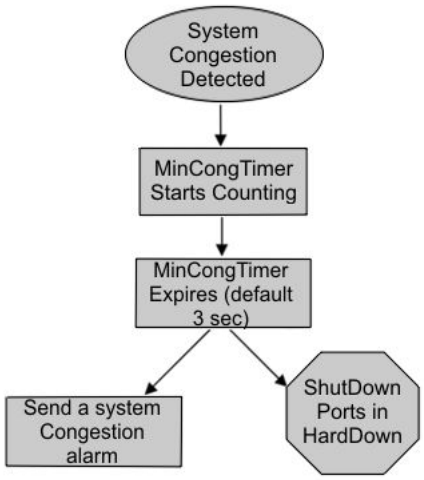


Figure 4: Extended CP-Limit HardDown operation

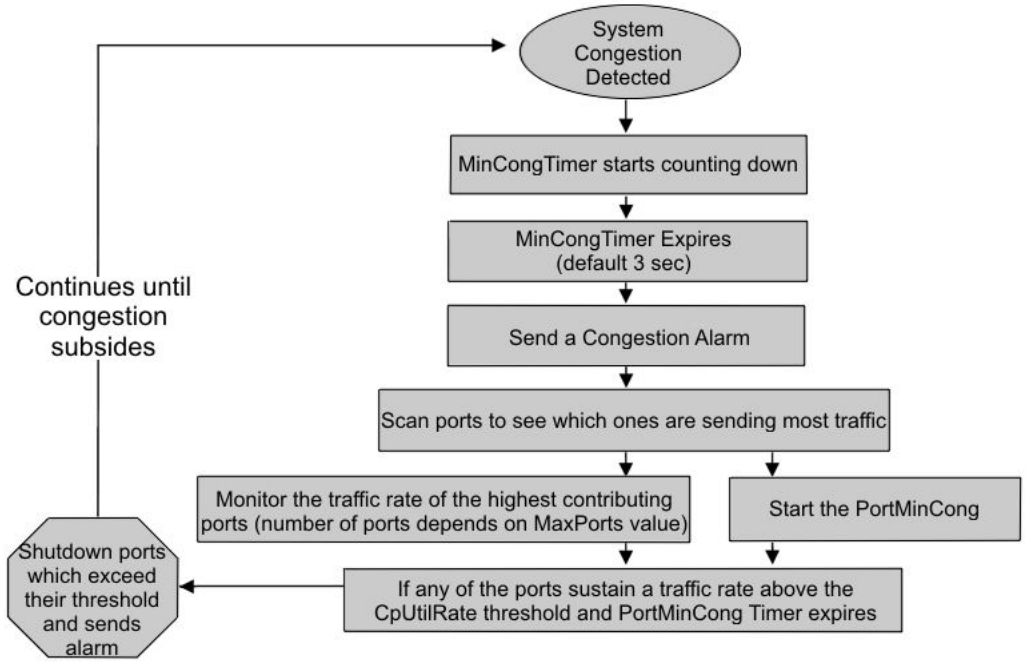


Figure 5: Extended CP-Limit SoftDown operation

For information about using CP-Limit and Extended CP-Limit with SLPP and VLACP, see SLPP, Loop Detect, and Extended CP-Limit on page 109.

For more information about CP-Limit and Extended CP-Limit, see *Avaya Ethernet Routing Switch 8800/8600 Administration, NN46205-605*.

Chapter 6: Optical routing design

Use the Avaya 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.

Optical routing system components

The Avaya optical routing system uses coarse wavelength division multiplexing (CWDM) in a grid of eight optical wavelengths. CWDM Small Form Factor Pluggable (SFP) and Small Form Factor Pluggable Plus (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.

Important:

The Avaya Ethernet Routing Switch 8800/8600 switch no longer supports CWDM Gigabit Interface Converters (GBIC).

The following figure shows multiplexer and demultiplexer operations.

Important:

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.

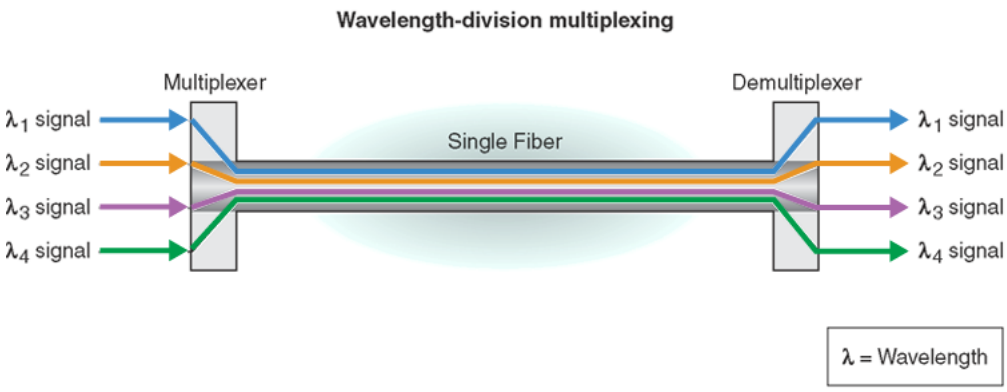


Figure 6: Wavelength division multiplexing

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

- CWDM SFPs
- CWDM SFP+s
- 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 12: Parts of the optical routing system

Wavelength	SFP/SFP+ part numbers	Multiplexer part number		
		OADM	OMUX-4	OMUX-8
1470 nanometers (nm) Gray	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
1490 nm Violet	AA1419026-E5, up to 40 km SFP AA1419034-E5, up to 70 km SFP AA1419054-E6, up to 40 km DDI SFP	AA1402003-E5	AA1402009-E5	

Wavelength	SFP/SFP+ part numbers	Multiplexer part number		
		OADM	OMUX-4	OMUX-8
	AA1419062-E6, up to 70 km DDI SFP			
1510 nm Blue	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	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	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 AA1403013-E6, up to 40 km SFP+	AA1402006-E5		
1570 nm Orange	AA1419030-E5, up to 40 km SFP AA1419038-E5, up to 70 km SFP AA1419058-E6, up to 40 km DDI SFP AA1419066-E6, up to 70 km DDI SFP	AA1402007-E5	AA1402009-E5	
1590 nm Red	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		

Wavelength	SFP/SFP+ part numbers	Multiplexer part number		
		OADM	OMUX-4	OMUX-8
1610 nm Brown	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 SFP+s, including technical specifications and installation instructions, see *Avaya Ethernet Routing Switch 8800/8600 Installation — SFP, SFP+, XFP, 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.

Navigation

- OADM ring on page 44
- Optical multiplexer in a point-to-point application on page 45
- OMUX in a ring on page 46

OADM ring

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

Important:

The wavelength of the OADM and the corresponding transceiver 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.

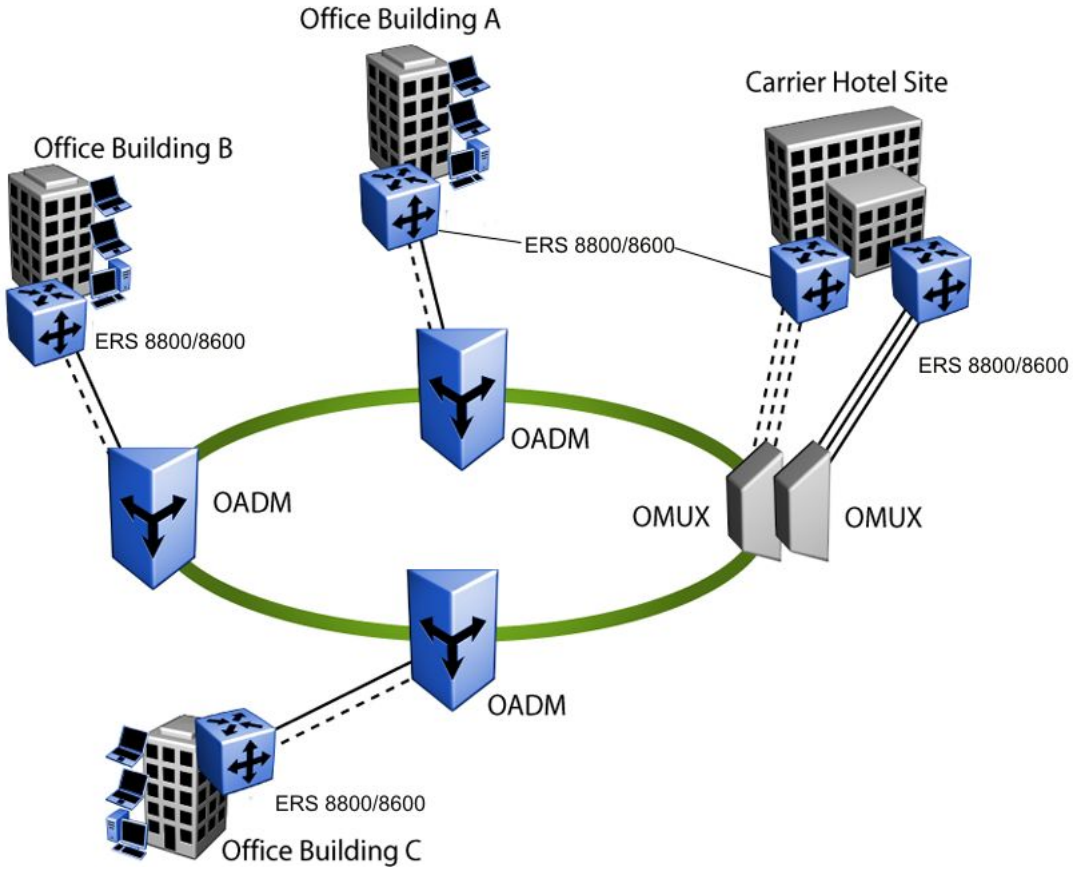


Figure 7: OADM ring configuration example

For information about calculating network transmission distance, see Transmission distance on page 47.

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. Point-to-Point connections (see the following figure) are made between mux/demuxs at each end. Point-to-Point 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 Point-to-Point 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 transceivers (SFPs/SFP+s). 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.

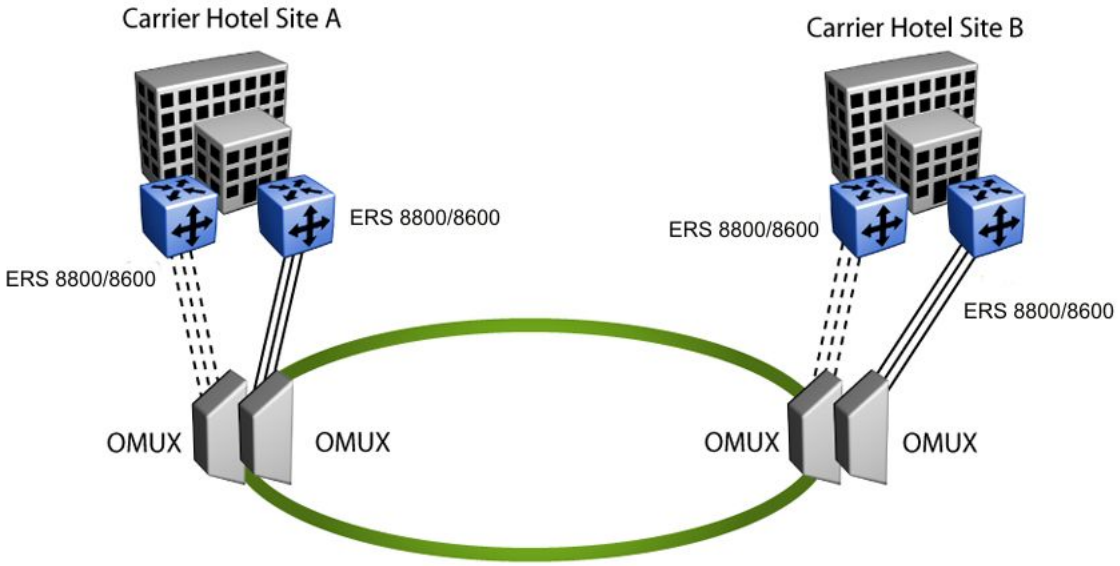


Figure 8: OMUX point-to-point configuration example

OMUX in a ring

OMUXs are also used as the hub site in OMUX-based ring applications. (For more information, see Figure 7: OADM ring configuration example on page 45.) 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.

Navigation

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 the following:

- 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).

Important:

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 SFP, SFP+, XFP, and

multiplexer specifications, see *Avaya Ethernet Routing Switch 8800/8600 Installation — SFP, SFP+, XFP, 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. Ensure that you use the appropriate value for your network.

Table 13: 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	0.25 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

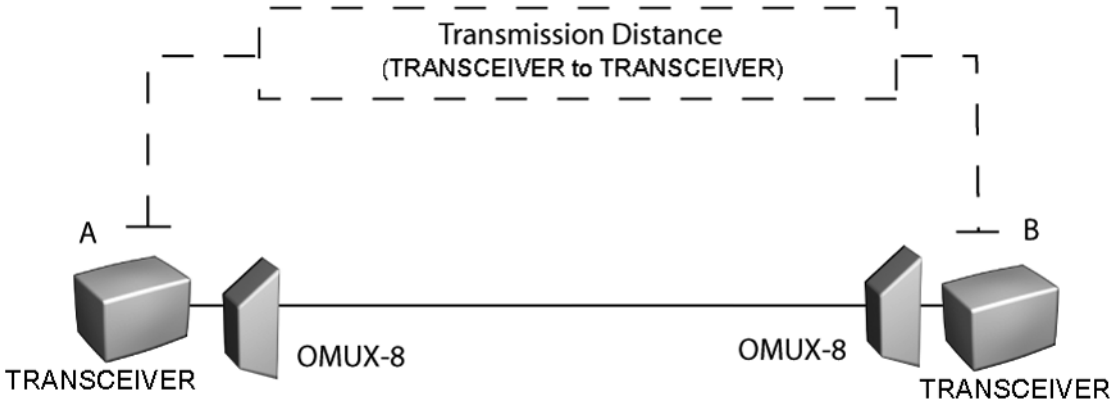


Figure 9: Point-to-point network configuration example

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 transceiver to transceiver.

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

Table 14: 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
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.

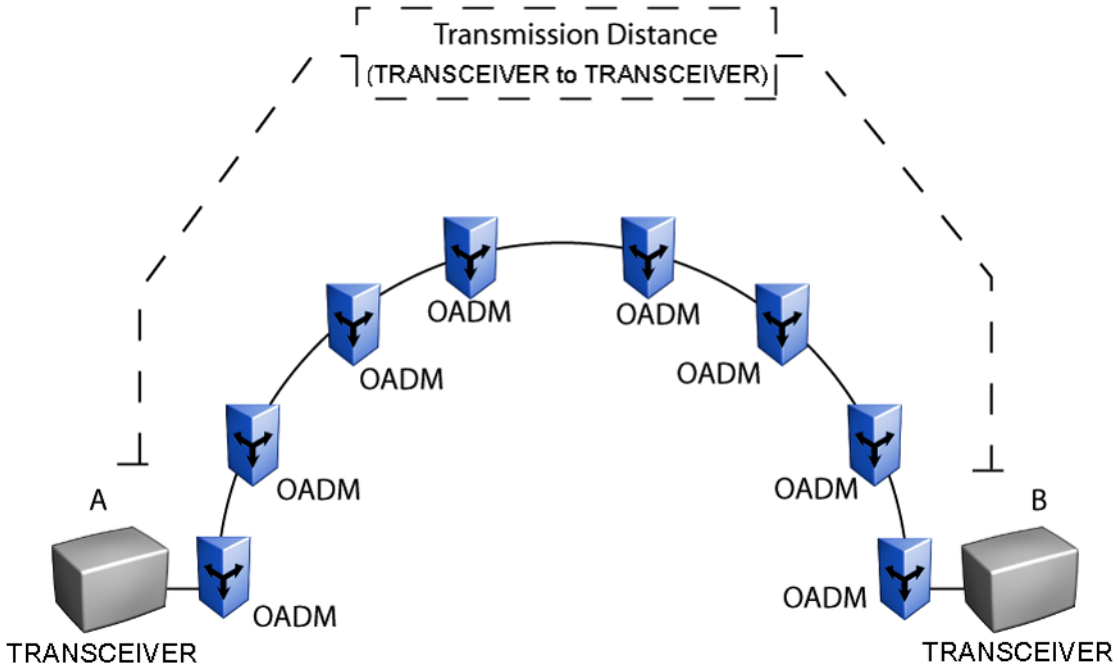


Figure 10: Mesh ring network configuration

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 transceivers.

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 transceiver to transceiver.

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 15: 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	0.25 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

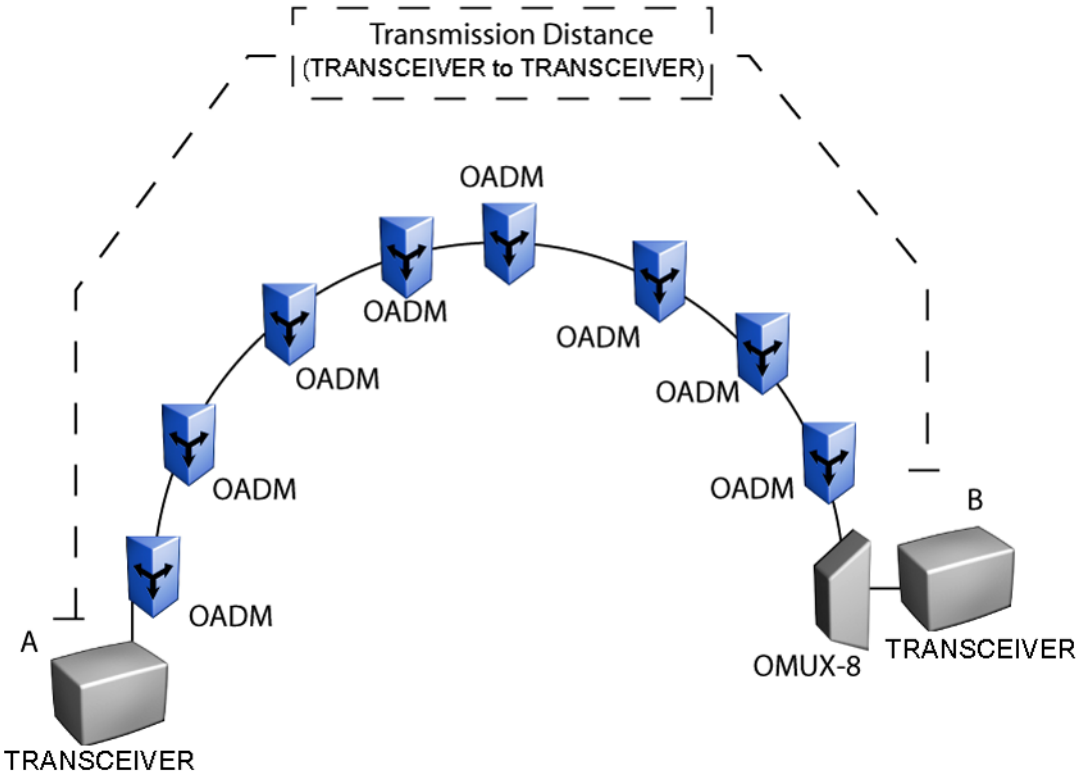


Figure 11: Hub and spoke network configuration

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 transceivers. 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 16: 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
OMUX-8 demux loss	4.5 dB
System margin	3.0 dB
Fiber attenuation	0.25 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

Chapter 7: 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.

Operational modes

With Release 7.0 and later, the Avaya Ethernet Routing Switch 8800/8600 operates in R mode only. You cannot configure the switch to run in M mode.

Similarly, enhanced operational mode configuration is not applicable as the system always operates in enhanced mode.

R mode supports up to 256 000 IP routes, 64 000 MAC entries, and 32 000 Address Routing Protocol (ARP) entries. This mode supports R and RS modules only.

With Software Release 7.0 and later, R and RS modules support up to 128 MLT groups and up to eight ECMP routing paths.

Release 5.0 and later supports up to 4000 VLANs with a default of 1972. VLAN scaling is reduced if you use multicast MAC filters.

Chapter 8: 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.

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 Avaya Ethernet Routing Switch 8800/8600 supports Far End Fault Indication (FEFI). FEFI ensures that link failures are reported to the switch. FEFI is enabled when the autonegotiation 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 12: 100BASE-FX FEFI on page 58). 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.

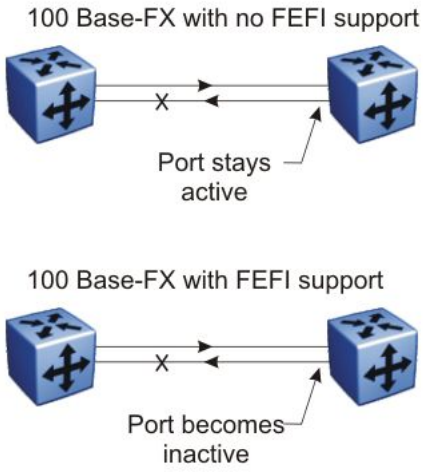


Figure 12: 100BASE-FX FEFI

With Avaya-to-Avaya 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 on page 59.

Gigabit Ethernet and remote fault indication

The 802.3z Gigabit Ethernet standard defines remote fault indication (RFI) as part of the autonegotiation 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 autonegotiation function, if autonegotiation is disabled, RFI is automatically disabled. Therefore, Avaya] recommends that autonegotiation be enabled on Gigabit Ethernet links when autonegotiation is supported by the devices on both ends of a fiber link.

For information about autonegotiation for 10 and 100 Mbit/s links, see 10/100BASE-TX Autonegotiation recommendations on page 36.

SFFD recommendations

The Ethernet switching devices listed in the following table do not support autonegotiation on fiber-based Gigabit Ethernet ports. These devices are unable to participate in remote fault indication (RFI), which is a part of the autonegotiation 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 17: Ethernet switching devices that do not support autonegotiation

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 autonegotiation, 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 autonegotiation disabled.

A better alternative to SFFD is VLACP (see End-to-end fault detection and VLACP on page 59).

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, Avaya has developed a feature called Virtual LACP (VLACP), which 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, but rather 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 an Avaya 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 13: Problem description (1 of 2) on page 60 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 13: Problem description (1 of 2) on page 60 shows a MLT running with VLACP. VLACP can operate end-to-end, but can be used in a point-to-point link.

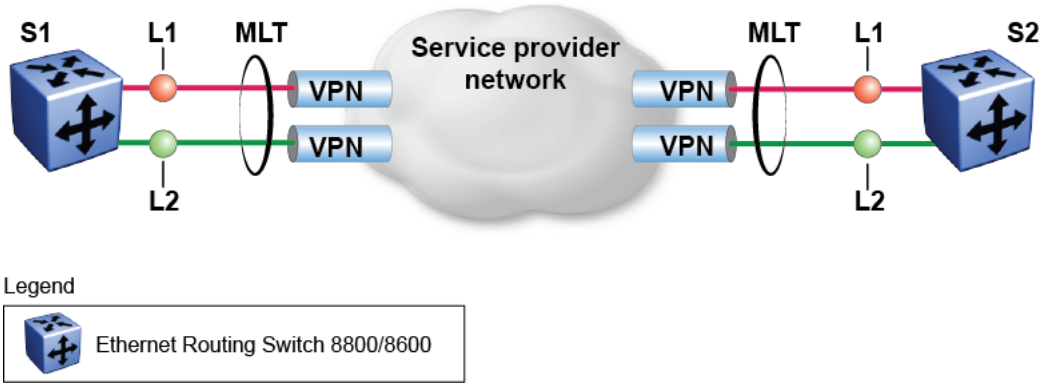


Figure 13: Problem description (1 of 2)

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.

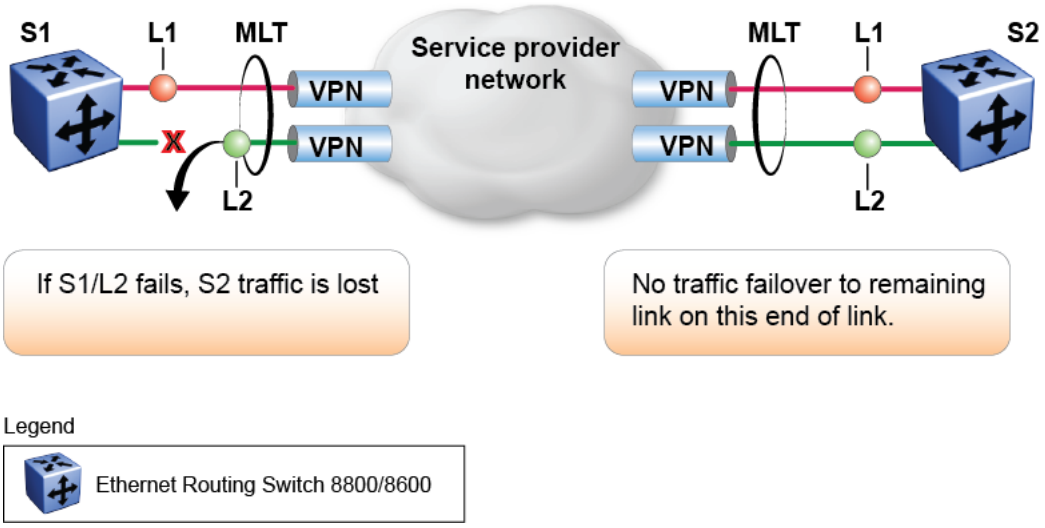


Figure 14: Problem description (2 of 2)

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

The Avaya Ethernet Routing Switch 8800/8600 can provide sub-100 millisecond failover using short timers on the 8692 SF/CPU with SuperMezz or on the 8895 SF/CPU. The target scenario, as shown in Figure 15: VLACP sub-100 millisecond convergence on page 62, is a core network of at least two Ethernet Routing Switch 8800/8600s (this feature works only between at least two Ethernet Routing Switch 8800/8600s equipped with the 8895 SF/CPU or with the 8692 SF/CPU with SuperMezz).

The Ethernet Routing Switch 8800/8600 supports sub-100 millisecond failover, but not as a best practice general recommendation. This functionality is only supported between two Ethernet Routing Switch 8800/8600 switches, generally across the core of a square of a full-mesh multiple cluster design. As an environment is scaled, sub-100 millisecond failover may not be stable. Therefore, if you enable this feature, minimize the number of links running sub-100 millisecond operation. Upon implementing sub-100 millisecond links or timers, if any VLACP instability is seen, increase the timers.

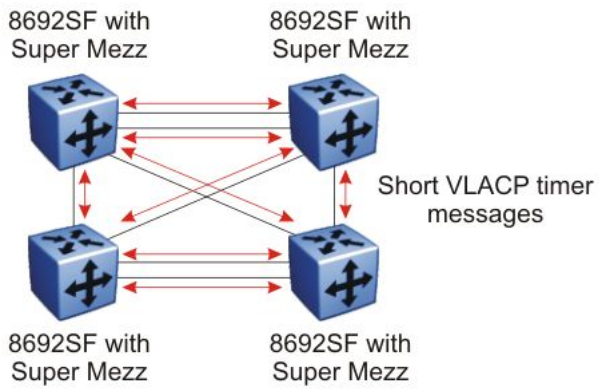


Figure 15: VLACP sub-100 millisecond convergence

VLACP recommendations and considerations

Avaya recommends the following:

- When connecting to a stackable switch, Avaya recommends settings VLACP to use a time-out scale of 5, short timers, and a timeout value of 500 milliseconds. Both faster timers and lower time-out scales are supported, but if any VLACP flapping occurs, increase the time-out scale and the short timer to their recommended values. Both the 8895 SF/CPU and 8692 SF/CPU with SuperMezz can support a fast periodic timer value as low as 30 milliseconds (ms).
- Do not use VLACP on configured LACP MLTs because LACP provides the same functionality as VLACP for link failure. VLACP and LACP running on the same link is not recommended.
- 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 millisecond failover (seeVLACP sub-100 ms convergence on page 61). 30 ms timers are not supported in High Availability (HA) mode, and may not be stable in scaled networks.
- For interswitch trunk (IST) links, Avaya recommends using a time-out scale of 5, long timers, and slow-periodic-time of 10000. For IST MLTs, Avaya 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. Some more complex configurations require higher timer values.
- When a VLACP-enabled port does not receive a VLACPDU, it normally enters 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.

- VLACP is configured by port. 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 is 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 stay in a disabled state. When VLACP is enabled at the IST peer, the VLACPDU is received and the ports are reenabled. 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 stays 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.

Avaya 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

  Avaya 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 on page 64.

- 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, Avaya 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, Avaya 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 non-HA 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 release 7.1.

Table 18: Feature support for HA

Feature	Release 7.1
Modules	R and RS only
Platform	Yes
Layer 2	Yes
Layer 3	Yes (see Note 1)
Multicast	Yes but no PGM (see Note 1)
IPv6	Yes, Restart
Security	Yes TACACS+ DHCP snooping ARP Inspection IP Source Guard
Note 1 — For Release 7.1, HA-CPU supports the following: • Hot Standby mode: - Shortest Path Bridging MAC (SPBM) - platform configuration - Layer 2 protocols: IGMP, STP, MLT, SMLT, ARP, LACP, VLACP - Layer 3 protocols: RIP, OSPF, VRRP, RSMLT, VRF Lite • Warm Standby mode: - DVMRP, PIM-SM, and PIM-SSM - BGP - MPLS - BFD - IPv6 and all associated IPv6 protocols	

Table 19: Synchronization capabilities in HA mode

Synchronization of:	Release 7.0
Layer 1	
Port configuration parameters	Yes

Synchronization of:	Release 7.0
Layer 2	
VLAN parameters	Yes
STP parameters	Yes
RSTP/MSTP parameters	Yes
SMLT parameters	Yes
QoS parameters	Yes
Layer 3	
Virtual IP (VLANs)	Yes
ARP entries	Yes
Static and default routes	Yes
VRRP	Yes
RIP	Yes
OSPF	Yes
Layer 3 Filters; ACE/ACLs	Yes
BGP	Yes
DVMRP/PIM	Yes
IGMP, PIM-SM, and PIM-SSM virtualization	Yes
BFD	Yes
IPv6	Partial HA

For more information about configuring HA, see *Avaya Ethernet Routing Switch 8800/8600 Administration, NN46205-605*.

High availability mode limitations and considerations

This section describes the limitations and considerations of the High Availability (HA) 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.

HA-CPU is not compatible with the PacketCapture (PCAP) Tool. Be sure to disable HA-CPU prior to using PCAP.

A restart 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 *Avaya Ethernet Routing Switch 8800/8600 Configuration — IP Routing,*

NN46205-523 and *Avaya Ethernet Routing Switch 8800/8600 Configuration — OSPF and RIP, NN46205-522*.

HA does not currently support the following protocols:

- PGM

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, Avaya 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 7.0 and later:

- All HA configurations support protocols with short timers and fast 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 on page 68
- 802.3ad-based link aggregation on page 71
- Bidirectional Forwarding Detection on page 74
- Multihoming on page 75

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 7.0 supports 128 MLT aggregation groups with up to 8 ports.
- Up to eight same-type ports can belong to a single MLT group. The same port type means that the ports operate on the same physical media, at the same speed, and in the same duplex mode.
- For Release 7.0, 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

Ensure that all ports that belong to the same MLT/LACP group use the same port speed, for example, 1 Gbit/s, even if autonegotiation is used. The software does not enforce this requirement. Avaya 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 can restart autonegotiation and the link can 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 autonegotiation and CANA are not used, the same speed and duplex mode settings must be used on all ports of the MLT.

Switch-to-switch MLT link recommendations

Avaya 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 Avaya Ethernet Routing Switch 8800/8600, 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 for each 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 can 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), Avaya 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 non-MLT ports, or the MLT with the least active links, has a higher value than the MLT link with the most active ports.

Withthe implementation of 802.1w (Rapid Spanning Tree Protocol—RSTP) and802.1s (Multiple Spanning Tree Protocol—MSTP), a new path cost calculationmethod is implemented. The following table describes the new pathcosts associated with each interface type:

Table 20: New path cost for RSTP or MSTP mode

Link speed	Recommended path cost
Less than or equal 100 Kbit/s	200 000 000
1 Mbit/s	20 000 000
10 Mbit/s	2 000 000
100 Mbit/s	200 000
1 Gbit/s	20 000
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

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 Avaya Ethernet Routing Switch 8800/8600 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 8800/8600 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 Avaya Ethernet Routing Switch 8800/8600 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 Avaya 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. Avaya 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 ACLI 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 8800, 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 ACLI, 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 Avaya Ethernet Routing Switch 8800/8600 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, you can 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.

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.
- A maximum of 128 link aggregation groups are supported.
- A maximum of eight active links are supported per LAG.

For LACP fundamentals and configuration information, see *Avaya Ethernet Routing Switch 8800/8600 Configuration — Link Aggregation, MLT, and SMLT, NN46205-518*.

Bidirectional Forwarding Detection

The Avaya Ethernet Routing Switch 8800/8600 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 8800/8600 BFD implementation complies with IETF drafts draft-ietf-bfd-base-06 and draft-ietf-bfd-v4v6-1hop-06.

Operation

The Avaya Ethernet Routing Switch 8800/8600 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.

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

The Ethernet Routing Switch 8800/8600 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 Avaya Ethernet Routing Switch 8800/8600 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 8800/8600 does not support the following IETF BFD options:

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

The Ethernet Routing Switch 8800/8600 does not support:

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

The Ethernet Routing Switch 8800/8600 supports partial HA for BFD.

The Ethernet Routing Switch 8800/8600 also supports the modification of transmit and receive intervals during an active BFD session.

Multihoming

Multihoming enables the Avaya Ethernet Routing Switch 8800/8600 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 8800/8600 supports multihomed hosts with up to 16 IP addresses per MAC address.

For more information about multihoming, see *Avaya Ethernet Routing Switch 8800/8600 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 on page 76
- Network edge redundancy on page 79
- Split Multi-Link Trunking on page 80
- Routed SMLT on page 92
- Switch clustering topologies and interoperability with other products on page 101

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, Avaya 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.

Important:

Avaya 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 followingfigure shows a three-tiered campus architecture with edge, distribution,and core layers. In addition, a server access layer is directly connectedto the core, representing a two-layer data center architecture.

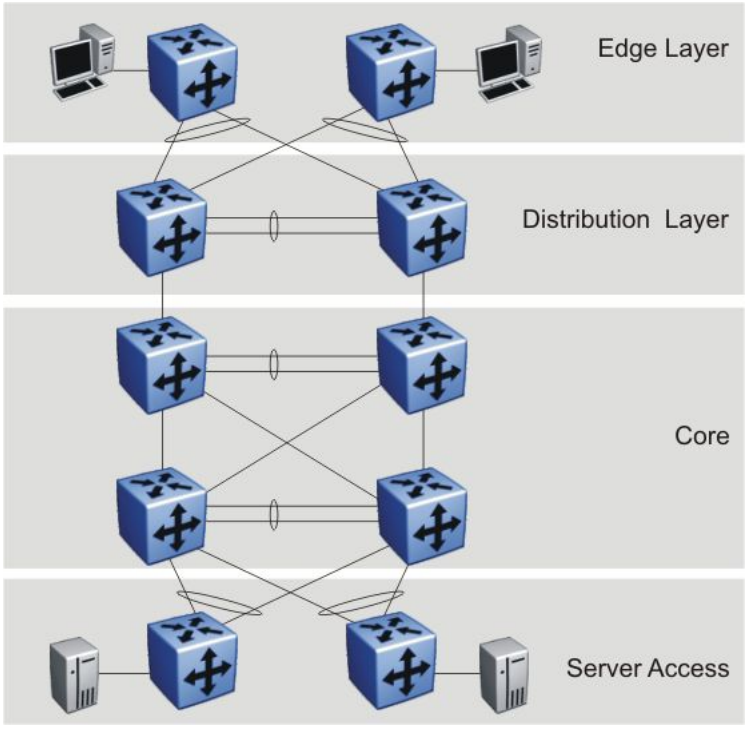


Figure 16: Three-tiered architecture plus data center

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

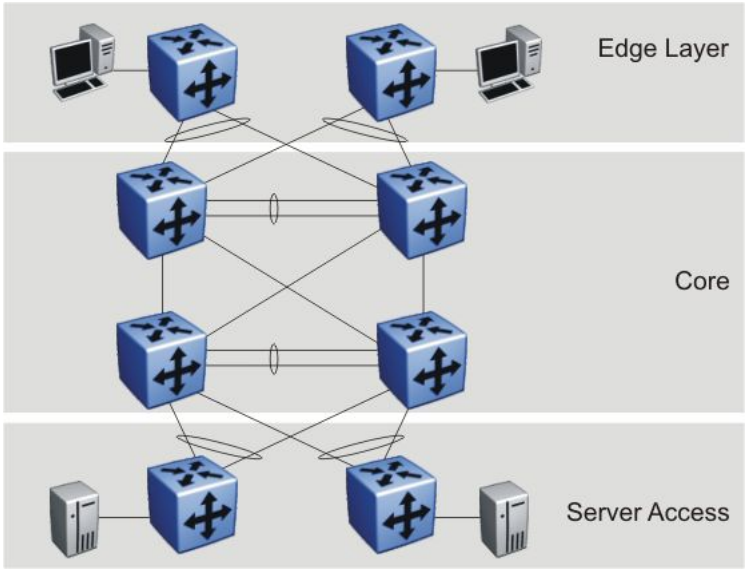


Figure 17: Two-tiered architecture with four-switch core plus data center

Thefollowing figure shows a two-tiered architecture with a two-switchcore.

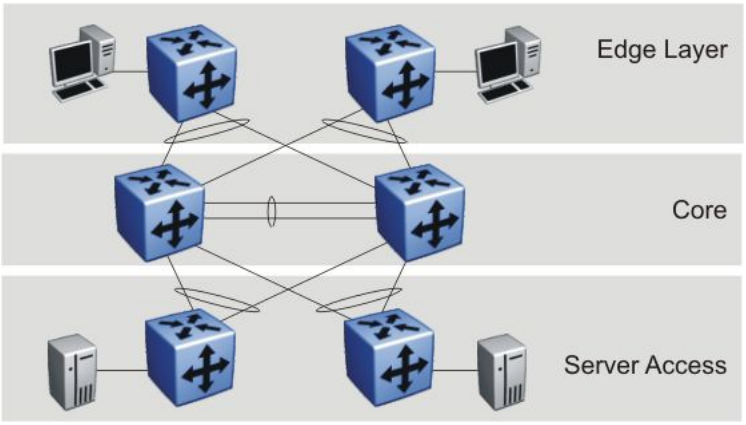


Figure 18: Two-tiered architecture with two-switch core plus data center

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.

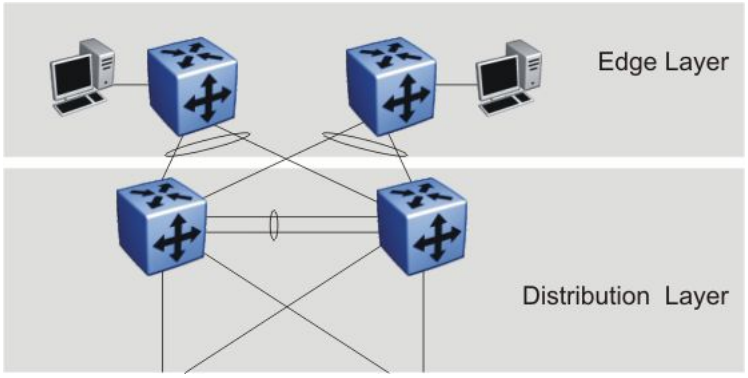


Figure 19: Redundant network edge diagram

Avaya recommends the network edge design shown in Figure 20: Recommended network edge design on page 80. This setup is simple to implement and maintain, yet still provides redundancy if one of the edge or distribution layer switches fails.

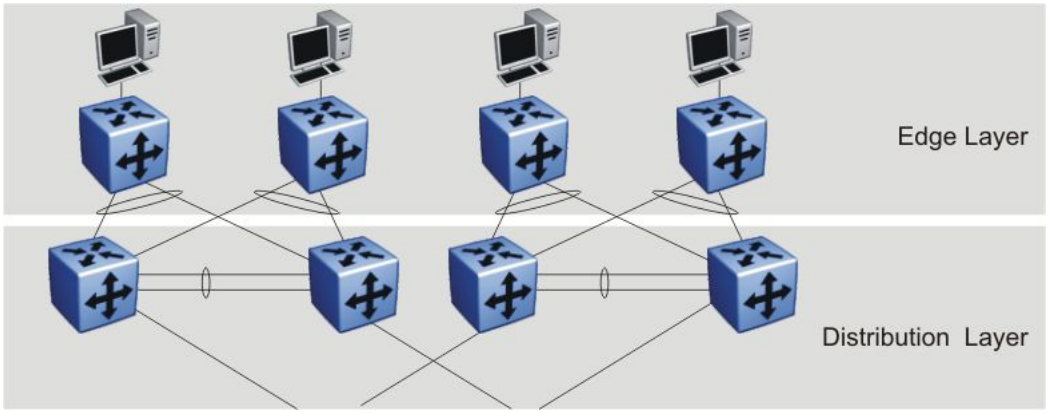


Figure 20: Recommended network edge design

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 Avaya Split Multi-Link Trunking feature for the Ethernet Routing Switch 8800/8600, 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).

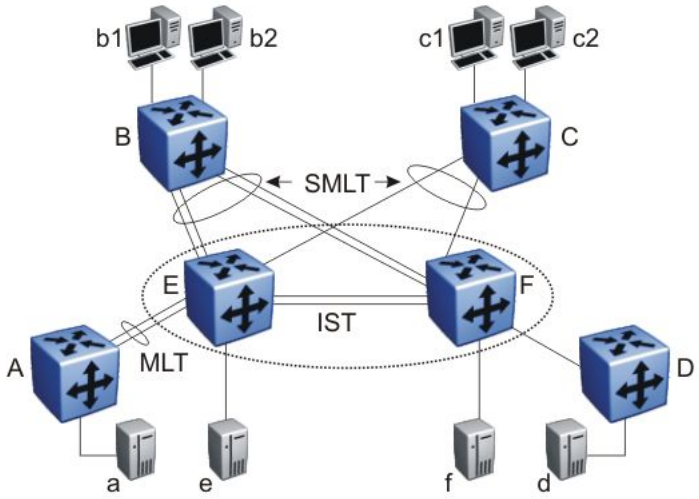


Figure 21: SMLT configuration with switches as aggregation switches

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 21: SMLT configuration with switches as aggregation switches on page 82 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).

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 an Avaya switch.

For more information about VLACP, see End-to-end fault detection and VLACP on page 59.

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 on page 109.

Interswitch Trunking recommendations

Figure 21: SMLT configuration with switches as aggregation switches on page 82 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 non-SMLT 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, Avaya recommends that the interswitch trunk be a multigigabit 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. Avaya 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.

Avaya recommends that you not enable any dynamic routing protocols on the IST VLAN. The purpose of the IST VLAN is to support adjacent switches; do not use the IST as a next-hop

route for non-IST 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.

Avaya 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.

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

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

Avaya 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.

IPv4 IST with IPv6 RSMLT

Avaya Ethernet Routing Switch 8800/8600 supports IPv6 RSMLT. However, messaging between IST peers is supported over IPv4 only.

Dual MLTs in SMLT

Dual MLTs in SMLT designs are supported, as long as only one is configured as an IST MLT (the system does not allow misconfiguration), and as long as any use of any form of spanning tree, and the VLANs/ports associated with this form of spanning tree, remain solely on the non-IST MLT; there can be no association or interaction with the IST MLT.

SMLT and client/server applications

Do not use unbalanced client-server configuration, where core switches have directly-connected servers 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 22: SMLT square configuration on page 91 and Figure 23: SMLT full-mesh configuration on page 91.

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

- 1 to 128 for MLT-based SMLTs
- 1 to 512 for SLTs

Important:

Avaya 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, on each chassis, an SMLT ID can belong to either an MLT-SMLT or to an SLT. 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, and POS 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.
- A port cannot be configured as MLT-based SMLT and as SLT at the same time.

For information about configuring SLT, see *Avaya Ethernet Routing Switch 8800/8600 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 *Avaya Ethernet Routing Switch 8800/8600 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. Avaya 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, Avaya 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.

Important:

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. Avaya 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, Avaya recommends that, for routing purposes, you configure per-VLAN IP addresses on both SMLT aggregation switches. Avaya 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 Avaya Ethernet Routing Switch 8800/8600 switch fully supports the IEEE 802.3adLink Aggregation Control Protocol (LACP) on MLT and distributed MLTlinks, and on a pair of SMLT switches. Be aware of 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. Avaya 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 restarts, 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 restarted 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 *Avaya Ethernet Routing Switch 8800/8600 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 with R series modules, use the following formula:

(number of VLANs on regular ports or MLT ports) + (2 * number of VLANs on SMLT ports) = 1972

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 on a 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).

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 on page 183.

For more information about VLAN scalability, see *Avaya Ethernet Routing Switch 8800/8600 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 127 SMLT client switch connections, and up to 512 SLT connections. When you use the square design (Figure 22: SMLT square configuration on page 91), keep in mind that all links facing each other (denoted by the MLT ring on an aggregation pair) must use the same SMLT IDs.

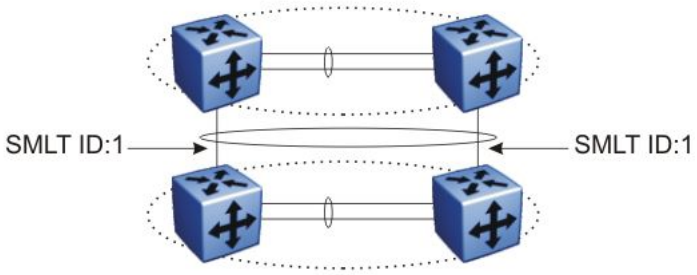


Figure 22: SMLT square configuration

You can configure an SMLT full-mesh configuration as shown in Figure 23: SMLT full-mesh configuration on page 91. In this configuration, all SMLT ports use the same SmltId (denoted by the MLT ring). The SMLT ID is of local significance only and must be the same on a cluster. For example, the top cluster could use SMLT ID 1 while the bottom cluster can use SMLT ID 2.

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.

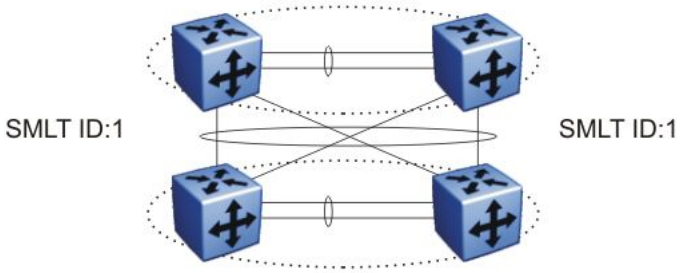


Figure 23: SMLT full-mesh configuration

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.

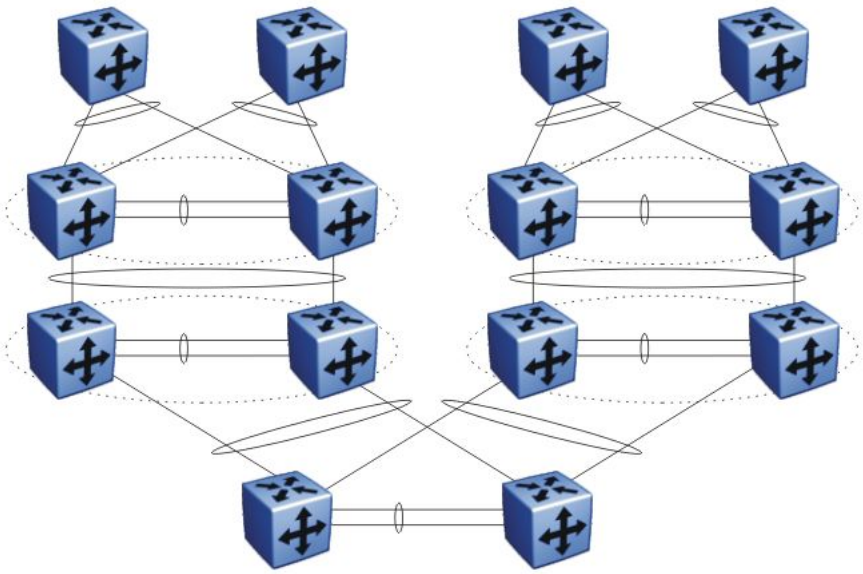


Figure 24: SMLT scaling

For more information about the SMLT triangle, square, and full-mesh designs, see *Avaya Ethernet Routing Switch 8800/8600 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), Avaya 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 configuration is recommended because CP-generated traffic is always sent out on the lower numbered MLT ports when active. This configuration keeps 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

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.

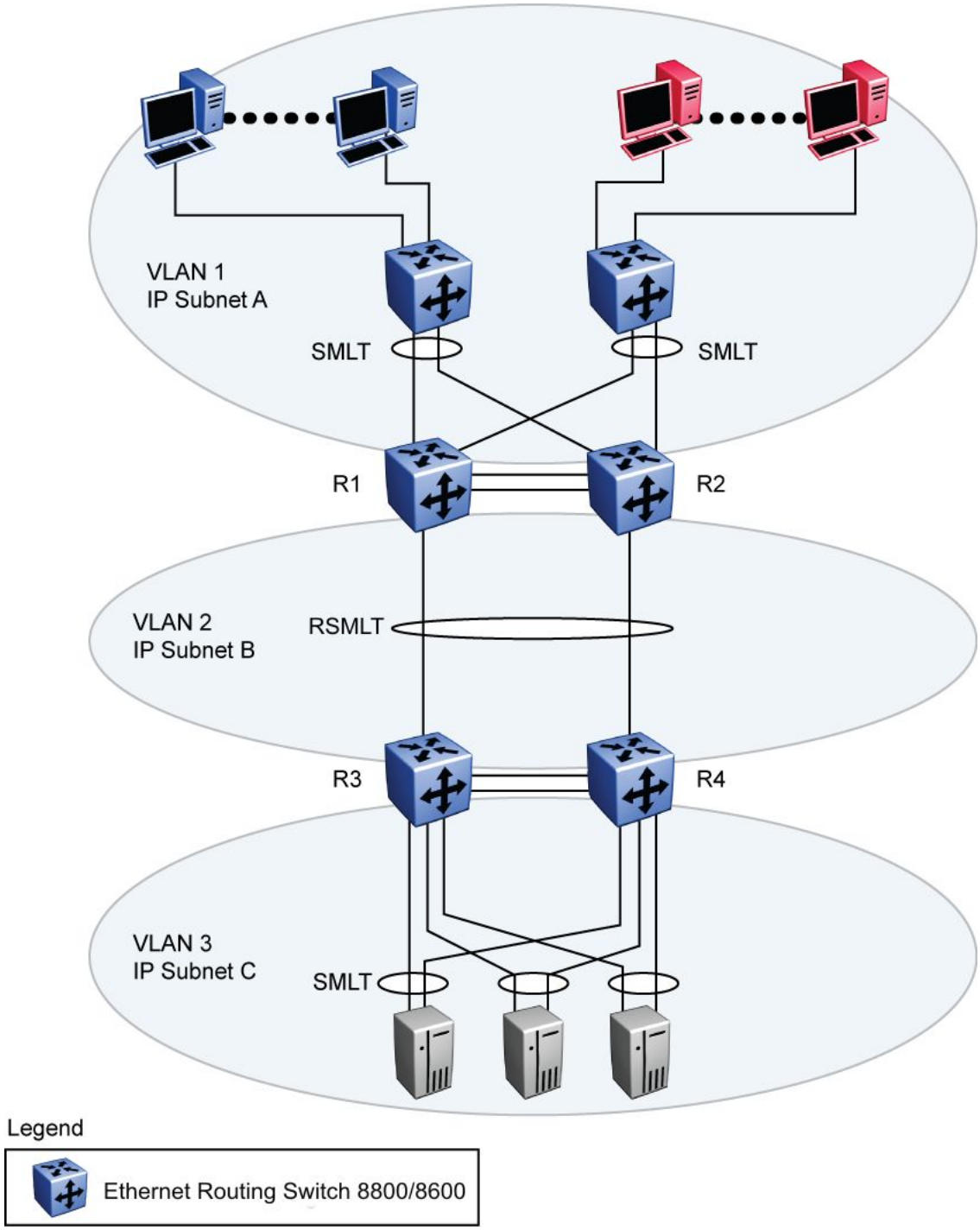


Figure 25: SMLT and RSMLT in Layer 3 environments

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 25: SMLT and RSMLT in Layer 3 environments on page 94.

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). Avaya 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, Avaya 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, Avaya 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 on page 282.

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.

Important:

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). Avaya 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.

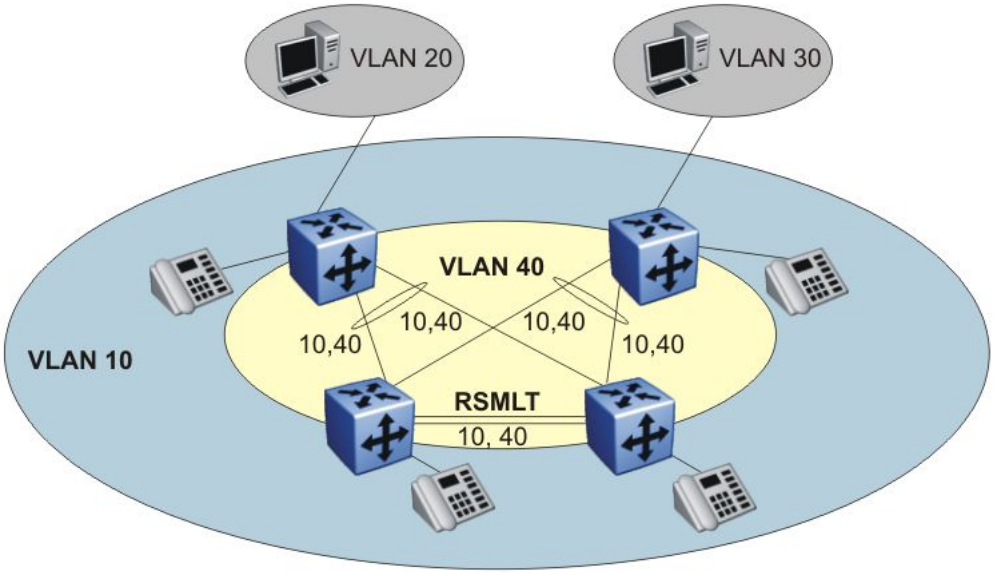


Figure 26: VLAN with all IP telephones as members

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.

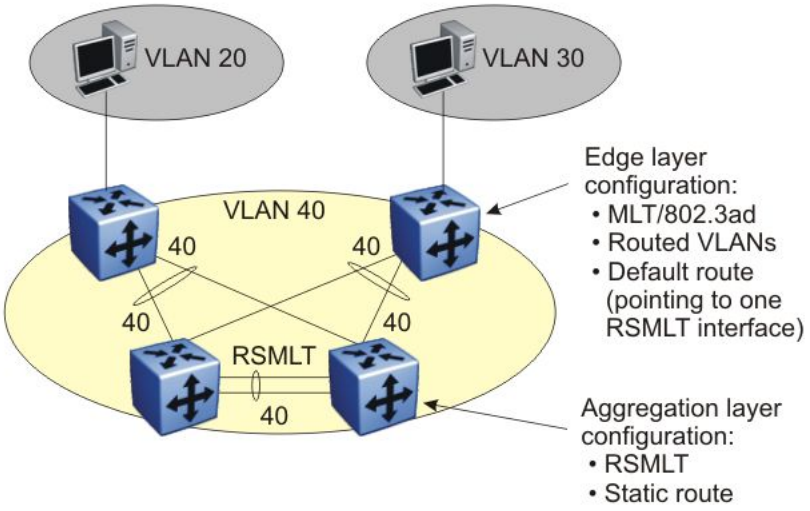


Figure 27: VLAN edge configuration

IPv6 RSMLT

While Avaya's Routed Split MultiLink Trunk (RSMLT) functionality originally provided sub-second failover for IPv4 forwarding only, the Avaya Ethernet Routing Switch 8800/8600

extends RSMLT functionality to IPv6. The overall model for IPv6 RSMLT is essentially identical to that of IPv4 RSMLT. In short, RSMLT peers exchange their IPv6 configuration and track each other’s state by means of IST messages. An RSMLT node always performs IPv6 forwarding on the IPv6 packets destined to the peer’s MAC addresses – thus preventing IPv6 data traffic from being sent over the IST. When an RSMLT node detects that its RSMLT peer is down, the node also begins terminating IPv6 traffic destined to the peer's IPv6 addresses.

With RSMLT enabled, an SMLT switch performs IP forwarding on behalf of its SMLT peer – thus preventing IP traffic from being sent over the IST.

IPv6 RSMLT supports the full set of topologies and features supported by IPv4 RSMLT, including SMLT triangles, squares, and SMLT full-mesh topologies, with routing enabled on the core VLANs.

With IPv6, you must configure the RSMLT peers using the same set of IPv6 prefixes.

Supported routing protocols include the following:

- IPv6 Static Routes
- OSPFv3

IPv4 IST with IPv6 RSMLT

The Avaya Ethernet Routing Switch 8800/8600 does not support the configuration of an IST over IPv6. IST is supported over IPv4 only.

Example network

The following figure shows a sample IPv6 RSMLT topology. It shows a typical redundant network example with user aggregation, core, and server access layers. To minimize the creation of many IPv6 prefixes, one VLAN (VLAN 1, IP prefix A) spans all wiring closets.

RSMLT provides the loop-free topology. The aggregation layer switches are configured with routing enabled and provide active-active default gateway functionality through RSMLT.

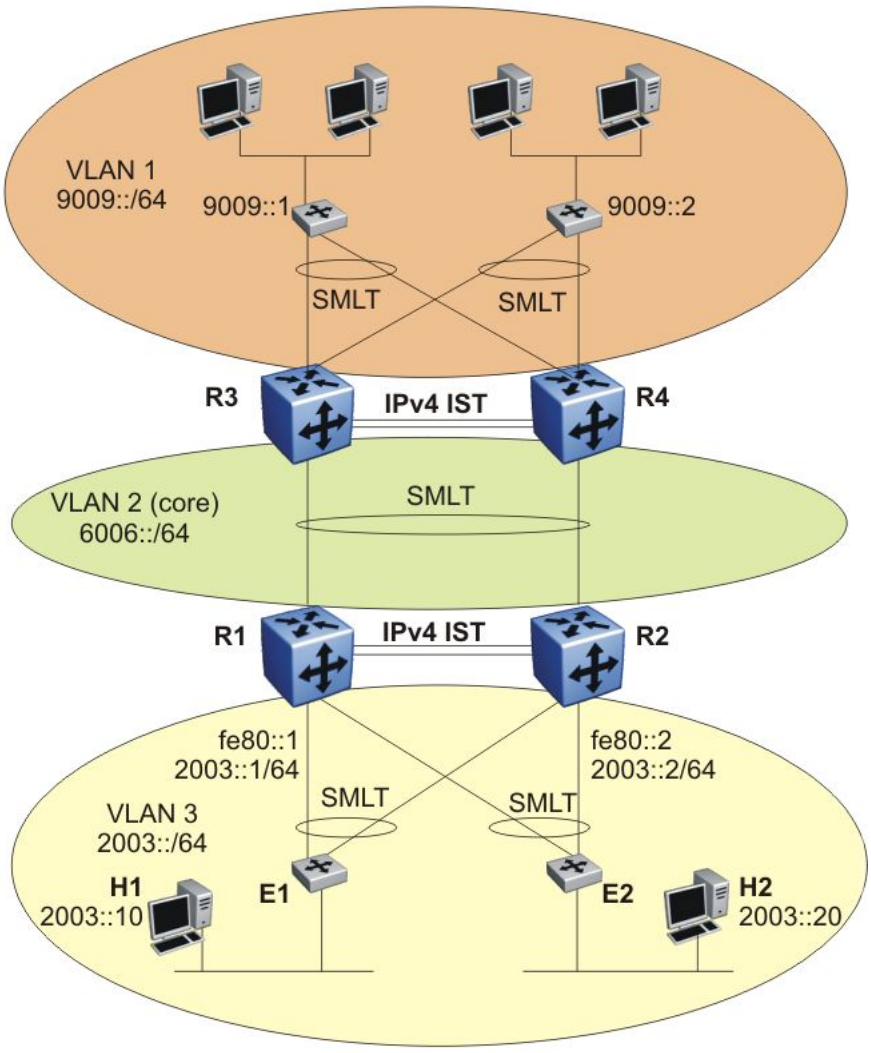


Figure 28: IPv6 RSMLT network example

In the VLAN 3 portion of the network shown in the preceding figure, routers R1 and R2 provide RSMLT-enabled IPv6 service to hosts H1 and H2. Router R1 can be configured as the default IPv6 router for H1 and R2 can be the default router for H2. R1 is configured with the link-local address of fe80::1, the global unicast address 2003::1, and the routing prefix of 2003::/64 (as a shorthand, the last two items are referred to as 2003::1/64). R2 is configured with fe80::2 and 2003::2/64.

Host H1 sends its IPv6 traffic destined to VLAN 1 to R1's MAC address (after resolving the default router address fe80::1 to R1's MAC). H2 sends its traffic to R2's MAC. When an IPv6 packet destined to R1's MAC address is received at R2 on its SMLT links (which is the expected MLT behavior), R2 performs IPv6 forwarding on the packet and does not bridge it over the IST. The same behavior occurs on R1.

At startup, R1 and R2 use the IST link to exchange full configuration information including MAC address for the IPv6 interfaces residing on SMLT VLAN 3.

When R2 detects that the RSMLT in R1 transitions to the DOWN state (for example, if R1 itself is down, or its SMLT links are down, or the IST link is down) R2 takes over IPv6 termination and IPv6 Neighbor Discovery functionality on behalf or R1's IPv6 SMLT interface. Specifically:

- When the above event is detected, R2 transmits an unsolicited IPv6 Neighbor Advertisement for each IPv6 address configured on R1's SMLT link using R1's MAC address (fe80::1 and 2003::1 in this example).
- R2 also transmits an unsolicited Router Advertisement for each of R1's routing prefixes (unless R1's prefixes are configured as "not advertised").
- R2 responds to Neighbor Solicitations and (if configuration allows) Router Advertisements on behalf of R1
- R2 terminates IPv6 traffic (such as pings) destined to R1's SMLT IPv6 addresses

When R1's RSMLT transitions back into the UP state and the HoldDown timer expires it resumes IPv6 forwarding and R2 ceases to terminate IPv6 traffic on R1's behalf.

Note that IPv6 allows a rich set of configuration options for advertising IPv6 routing prefixes (equivalent to IPv4 subnets) and configuring hosts on a link. A prefix can be configured to be or not to be advertised, to carry various flags or lifetime. These parameters affect how hosts can (auto)configure their IPv6 addresses and select their default routers. Most relevant from the RSMLT perspective is that an RSMLT node fully impersonates its peer's IPv6 configuration and behavior on the SMLT link – whatever its configuration happens to be. The above network example illustrates one of the many possible deployment schemes for IPv6 routers and hosts on a VLAN.

RSMLT provides both router failover and link failover. For example, if the Split MultiLink Trunk link between R2 and R4 is broken, the traffic fails over to R1 as well.

Router R1 recovery

After R1 reboots after a failure, it becomes active as a VLAN bridge first. Packets destined to R1 are switched, using the bridging forwarding table, to R2. R1 operates as a VLAN bridge for a period defined by the hold-down timer.

After the hold-down time expires and the routing tables converge, R1 starts routing packets for itself and also for R2. Therefore, it does not matter which of the two routers is used as the next hop from R3 and R4 to reach IPv6 prefix 2003::/64.

When an IPV6 RSMLT peer recovers, the peer installs a temporary default route in the IPv6 routing table to point all the IPv6 traffic to the IST peer IP address for the hold down time. (This is the same behavior as in IPv4 RSMLT.)

Coexistence with IPv4 RSMLT

The IPv6 RSMLT feature introduces no changes to the existing IPv4 RSMLT state machine including RSMLT configuration, definitions of events, logic of state transitions, or timer operations. A single instance of state and configuration parameter set controls both IPv4 and IPv6 RSMLT logic. With the introduction of this feature, RSMLT is best thought of as a property of the VLAN layer as opposed to the IP (v4 or v6) layer above it. RSMLT configuration and states affect IPv4 and IPv6 operation simultaneously.

For a given SMLT VLAN RSMLT is supported for any of the following scenarios:

- IPv4 Only: IPv4 is configured on the VLAN and IPv6 is not. RSMLT operation and logic remains unchanged from the current implementation.
- IPv6 Only: IPv6 is configured on the VLAN and IPv4 is not. IPv6 RSMLT operation follows that of IPv4 as described in this document.
- IPv4 and IPv6: Both IPv4 and IPv6 are configured on the VLAN. IPv4 RSMLT operation and logic remains unchanged from the current implementation and unaffected by IPv6. IPv6 operation follows that of IPv4 as described in this document.

Switch clustering topologies and interoperability with other products

When the Avaya Ethernet Routing Switch 8800/8600 is used with other Avaya Ethernet Routing Switch products, the switch clustering bridging, unicast routing, and multicast routing configurations vary with switch type. Avaya 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 8800 / 8600 / 5500 / 8300 / 1600, NN48500-555*.

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*.

Chapter 9: 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.

Spanning tree

Spanning Tree prevents loops in switched networks. The Avaya Ethernet Routing Switch 8800/8600 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 *Avaya Ethernet Routing Switch 8800/8600 Configuration — VLANs and Spanning Tree, NN46205-517*.

Spanning tree navigation

- Spanning Tree Protocol on page 103
- Per-VLAN Spanning Tree Plus on page 108
- MSTP and RSTP considerations on page 108

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.

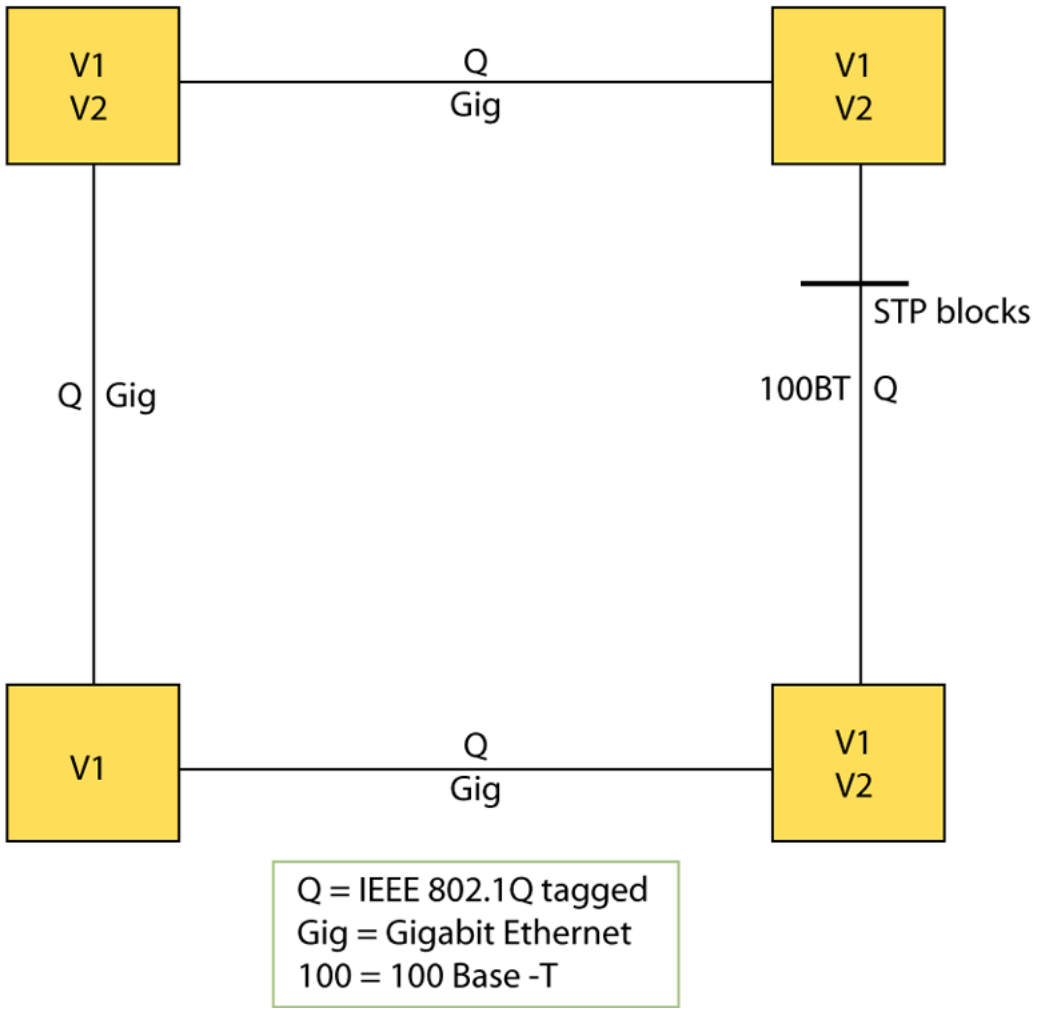


Figure 29: VLAN isolation

Multiple STG interoperability with single STG devices

Avaya provides multiple spanning tree group (STG) interoperability with single STG devices. When you connect the Avaya Ethernet Routing Switch 8800/8600 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 Avaya Ethernet Routing Switch 8800/8600 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.

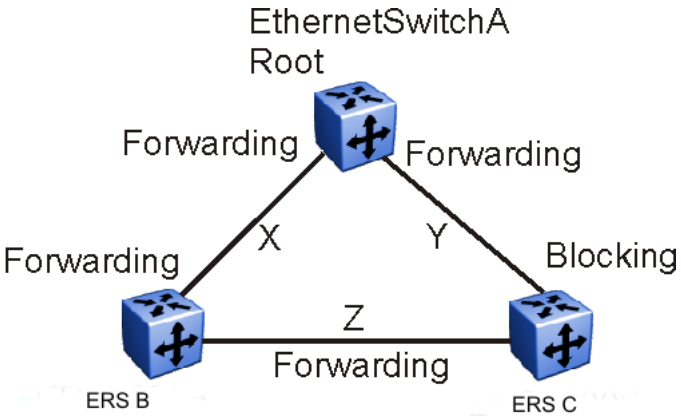


Figure 30: One STG between two Layer 3 devices and one Layer 2 device

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 31: Alternative configuration for STG and Layer 2 devices on page 106, 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 8800/8600s, 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.

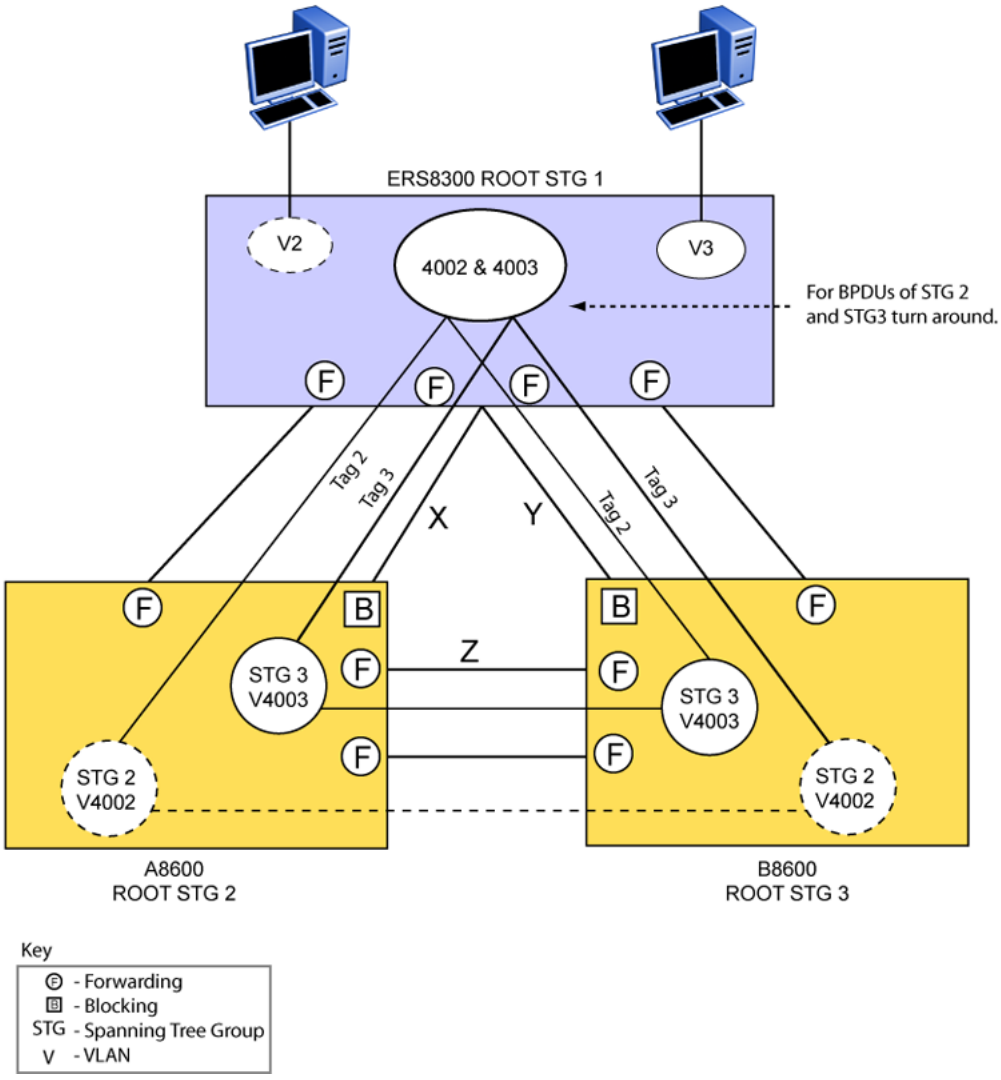


Figure 31: Alternative configuration for STG and Layer 2 devices

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

On the Ethernet Routing Switch 8800/8600s (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 8800/8600. 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 32: VLANs on the Layer 2 switch on page 107 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 8800/8600s.

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.

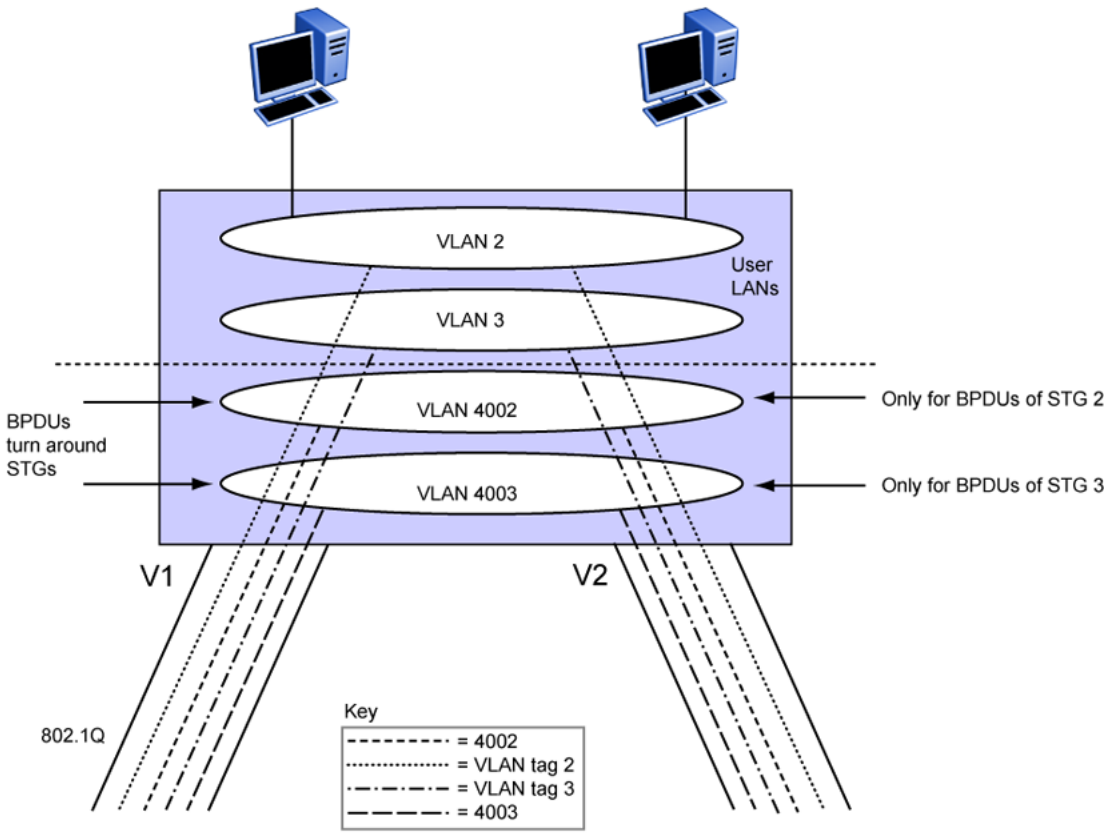


Figure 32: VLANs on the Layer 2 switch

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 *Avaya Ethernet Routing Switch 8800/8600 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 Avaya proprietary MSTP, except that the Avaya 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 Avaya 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. Avaya 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 *Avaya Ethernet Routing Switch 8800/8600 Administration, NN46205-605*. For more information about loop detection, see *Avaya Ethernet Routing Switch 8800/8600 Configuration — VLANs and Spanning Tree, NN46205-517*.

Simple Loop Prevention Protocol (SLPP)

Beginning with Software Release 4.1, Avaya 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 Recovery to reenable the port after a predefined interval. For more information on Auto Recovery, see *Administration* (NN46205–605).

SLPP prevents loops in an SMLT network, but it 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.

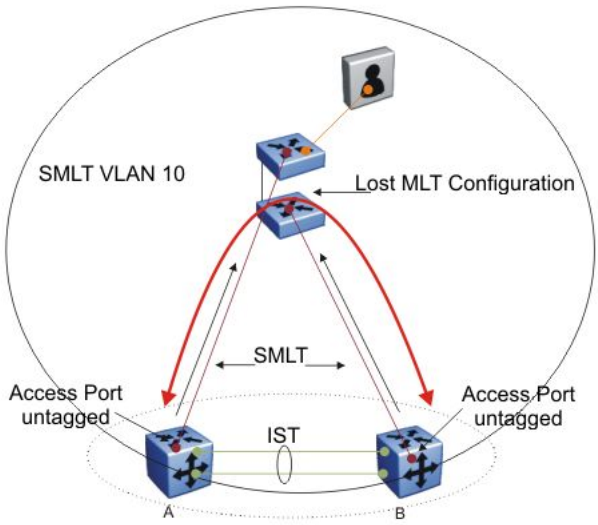


Figure 33: Untagged SMLT links

For the network shown in Figure 33: Untagged SMLT links on page 110, 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:

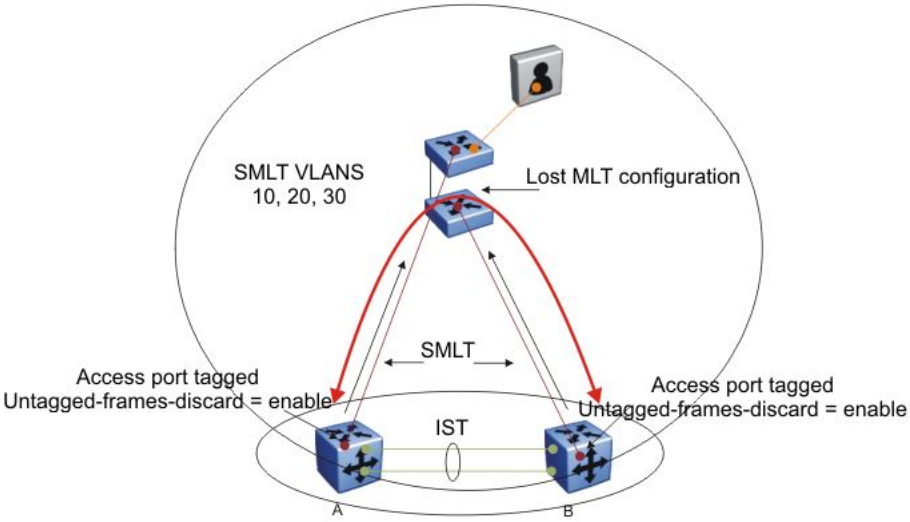


Figure 34: Tagged SMLT links

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, Avaya recommends that you enable the untagged-frames-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, Avaya 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 Avaya recommended SLPP values.

Table 21: 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)

	Setting
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.

Important:

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 Avaya recommended Extended CP-Limit values.

Table 22: 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 = 4000 ms S = 70000 ms T = 140 000 ms Q = 210 000 ms

Setting	Value
Port congestion time	P = 4 seconds S = 70 seconds T = 140 seconds Q = 210 seconds
Primary (P) – primary target for convergence Secondary (S) – secondary target for convergence Tertiary (T) – third target for convergence Quarternary (Q) – fourth target for convergence Avaya 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 can 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 logically disables that port and no traffic can pass. This insures that even if there is a link on the port at the other end, if it is not processing VLACP PDUs correctly, no traffic is sent. This alleviates potential black hole situations by only sending traffic to ports that are functioning properly.

Loop prevention recommendations

The following table describes the loop prevention features available for release 4.1.x and later. For best loop prevention, Avaya recommends that you use SLPP.

Table 23: Loop prevention by release

Software release	CP-Limit	Loop detect	Ext-CP-Limit	SLPP
4.1.x and on	Yes (see Note 1)	No	Yes (soft down) (see Notes 1 and 2)	Yes (see Note 3)
Note 1: SF/CPU protection mechanism; do not enable on IST links. Note 2: With Release 4.1.1.0 and later, Avaya recommends that you use the Soft Down option versus Hard Down. Note 3: Do not enable SLPP on IST links.				

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

Table 24: 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	10 000	10 000

SF/CPU protection and loop prevention compatibility

Avaya recommends several best-practice methods for loop prevention, especially in any Avaya Ethernet Routing Switch 8800/8600 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*.

Chapter 10: Layer 3 network design

This section describes some Layer 3 design considerations that you need to be aware of to properly design an efficient and robust network.

VRF Lite

Release 7.0 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 install 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 *Avaya Ethernet Routing Switch 8800/8600 Configuration — IP Routing, NN46205-523*.

VRF Lite route redistribution

Using VRF Lite, the Avaya Ethernet Routing Switch 8800/8600 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.

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 performs the following:

- 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 do the following:

- 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.

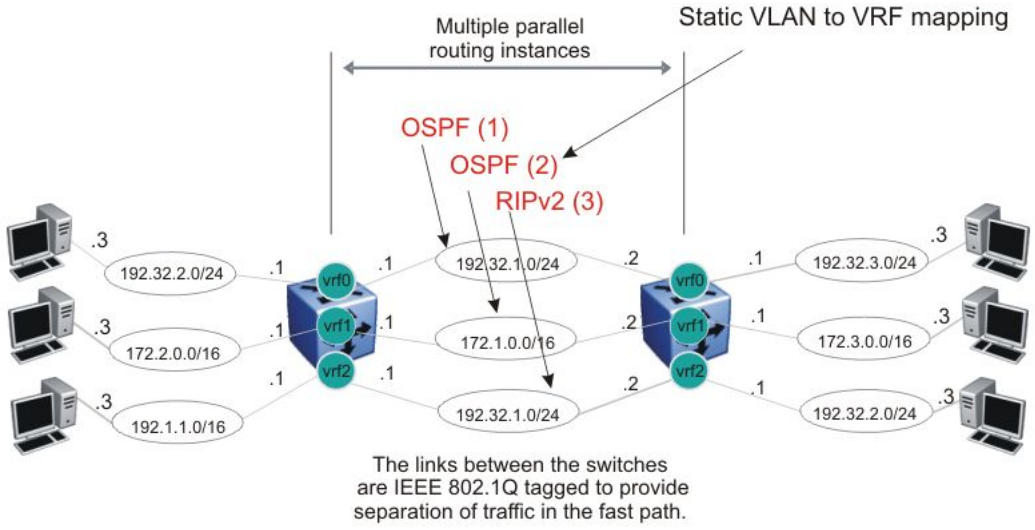


Figure 35: VRF Lite example

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

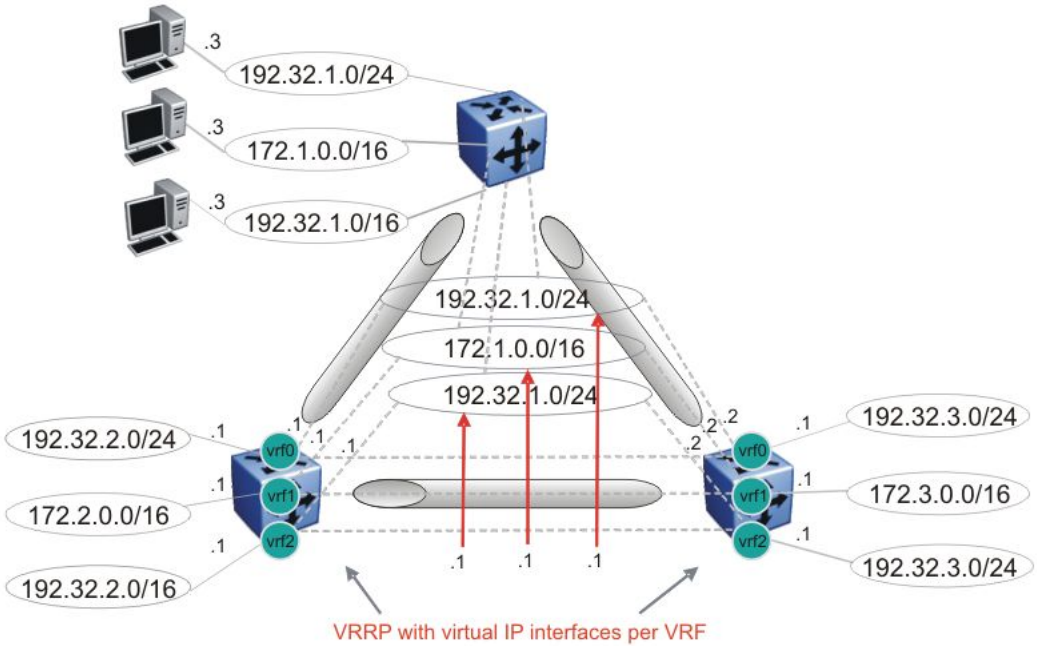


Figure 36: VRRP and VRF in SMLT topology

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

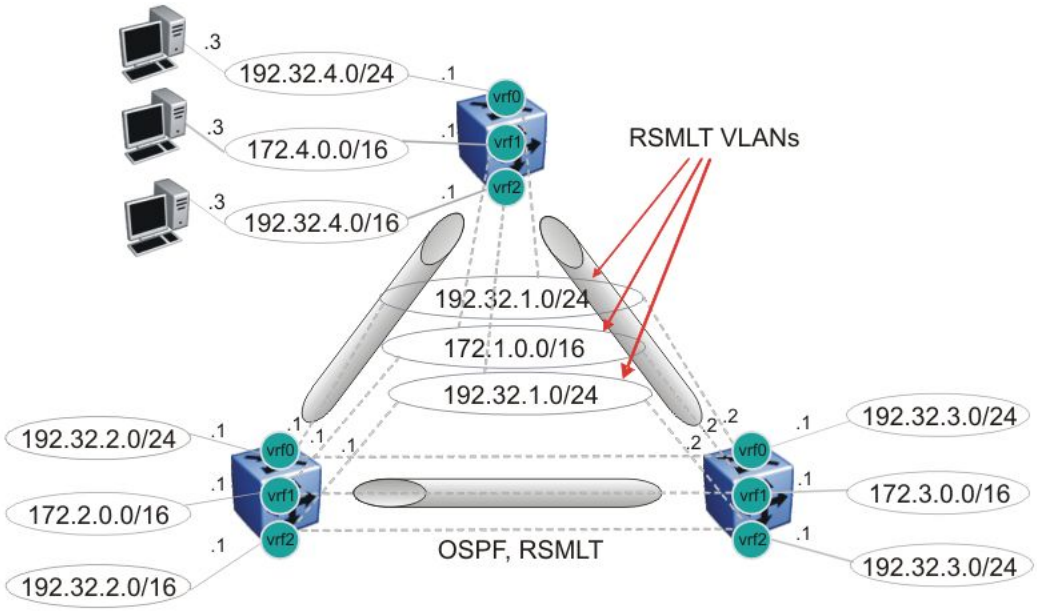


Figure 37: Router redundancy for multiple routing instances (using RSMLT)

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

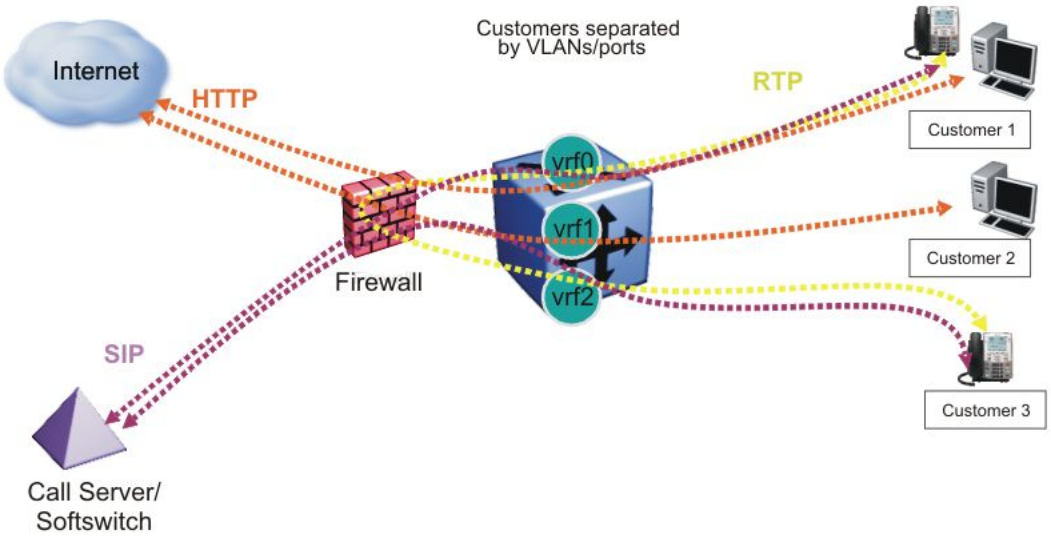


Figure 38: Inter-VRF forwarding based on external firewall

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 inter-VRF 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.

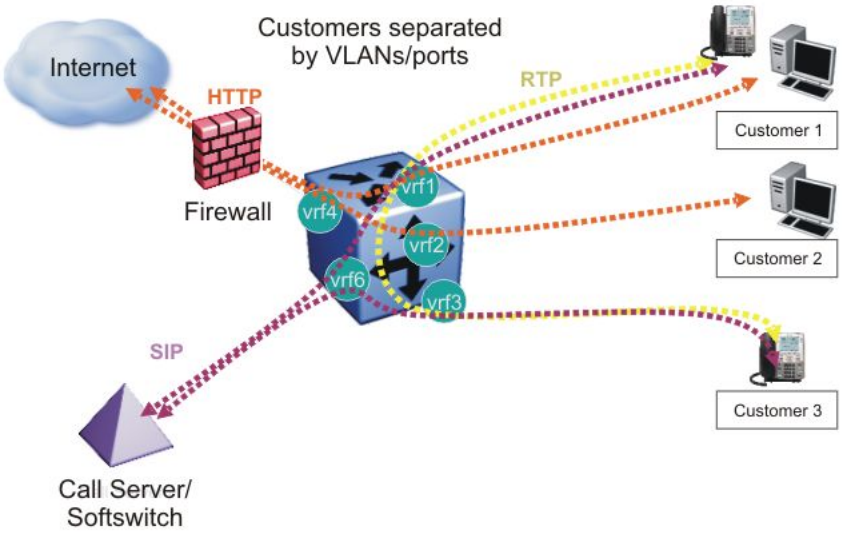


Figure 39: Inter VRF communication, internal inter-VRF forwarding

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, Avaya 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.

The following figure shows a sample network topology that uses VRRP with BackupMaster.

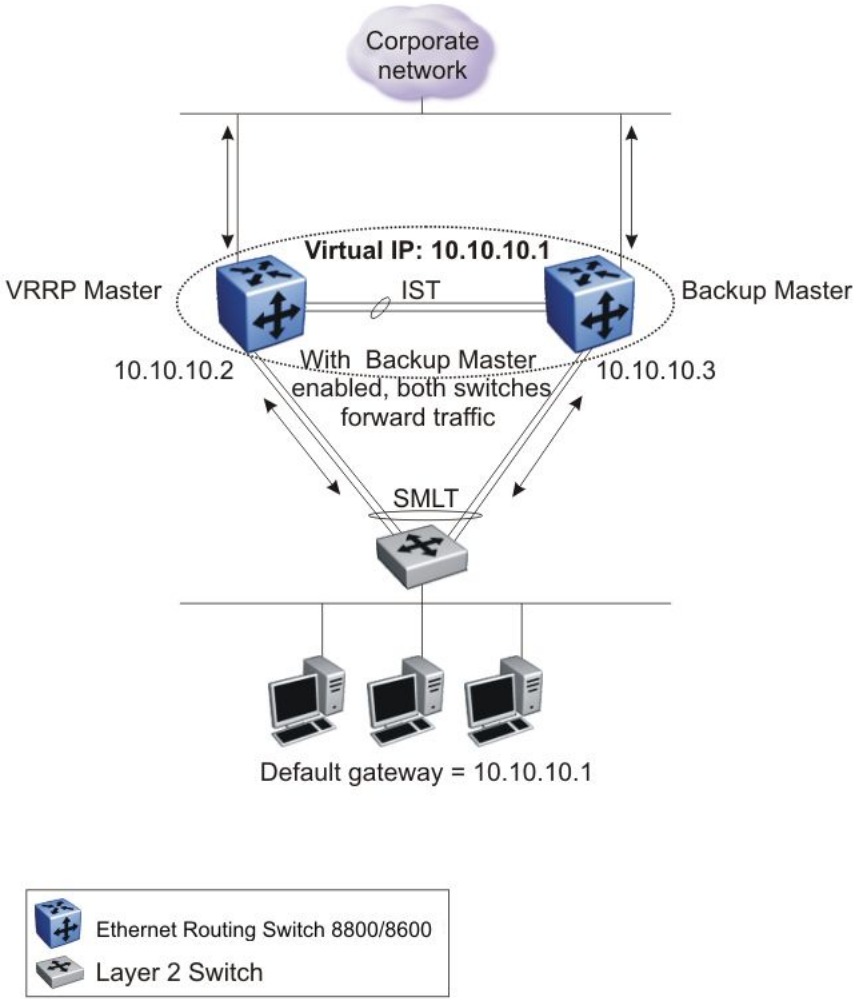


Figure 40: VRRP with BackupMaster

Avaya 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 Avaya 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, Avaya recommends that you 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-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.)
- 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, Avaya 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.

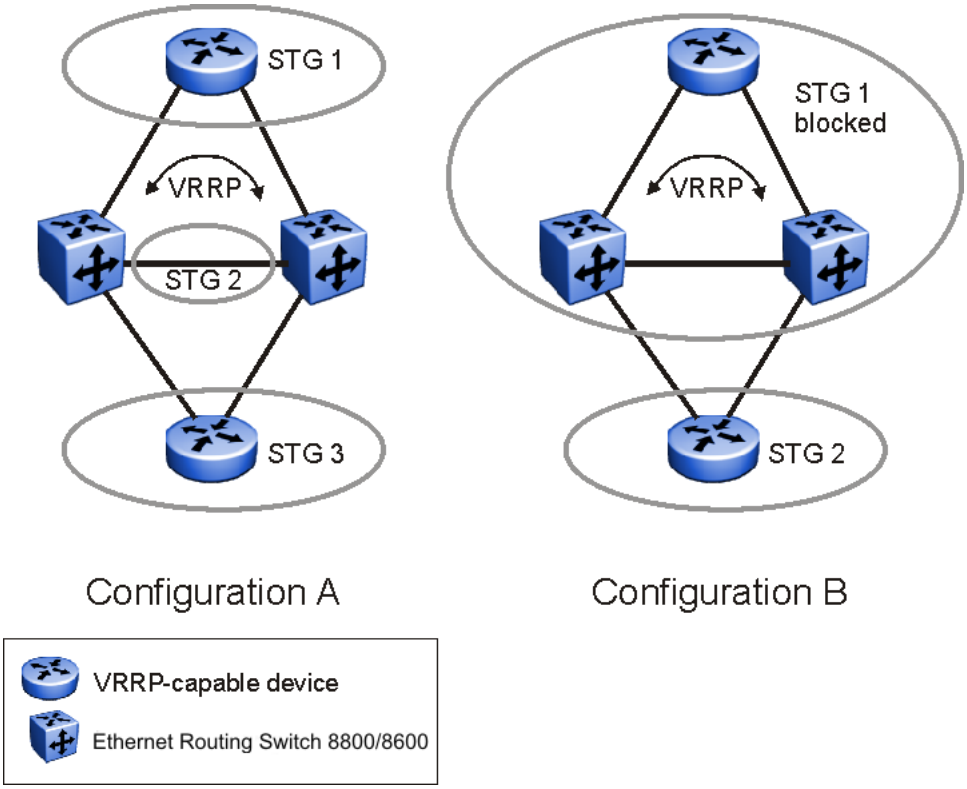


Figure 41: VRRP and STG configurations

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, Avaya recommends that you enable SMLT with VRRP to simplify the network configuration and reduce the failover time. For more information about VRRP and SMLT, see SMLT and Layer 3 traffic Redundant Default Gateway: VRRP on page 86.

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.

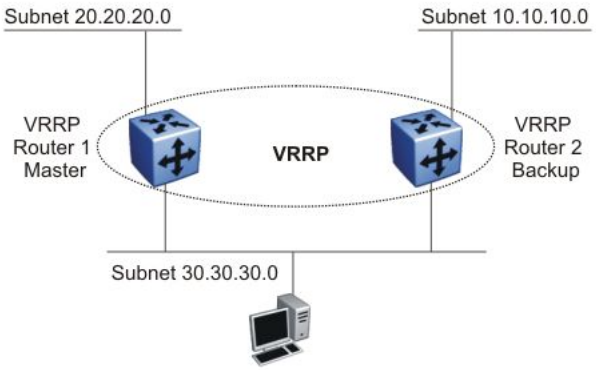


Figure 42: ICMP redirect messages diagram

To avoid excessive ICMP redirect messages if network clients do not recognize ICMP redirect messages, Avaya 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 on page 57.

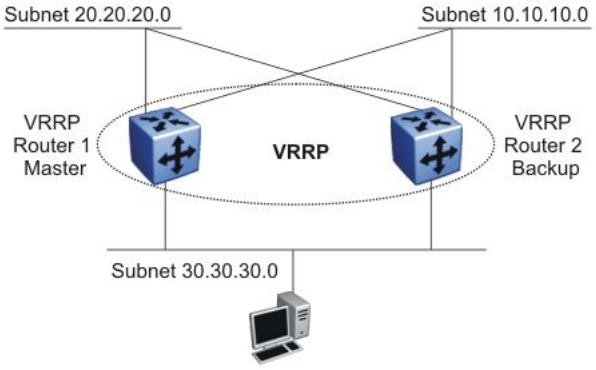


Figure 43: Avoiding excessive ICMP redirect messages

IPv6 VRRP

For IPv6 hosts on a LAN to learn about one or more default routers, IPv6-enabled routers send Router Advertisements using the IPv6 Neighbor Discovery (ND) protocol. The routers multicast these Router Advertisements every few minutes.

The ND protocol includes a mechanism called Neighbor Unreachability Detection to detect the failure of a neighbor node (router or host) or the failure of the forwarding path to a neighbor. Nodes can monitor the health of a forwarding path by sending unicast ND Neighbor Solicitation messages to the neighbor node. To reduce traffic, nodes only send Neighbor Solicitations to neighbors to which they are actively sending traffic and only after the node receives no positive indication that the neighbors are up for a period of time. Using the default ND parameters, it takes a host approximately 38 seconds to learn that a router is unreachable before it switches to another default router. This delay is very noticeable to users and causes some transport protocol implementations to timeout.

While you can decrease the ND unreachability detection period by modifying the ND parameters, the current lower limit that can be achieved is five seconds, with the added downside of significantly increasing ND traffic. This is especially so when there are many hosts all trying to determine the reachability of one of more routers.

To provide fast failover of a default router for IPv6 LAN hosts, the Avaya Ethernet Routing Switch 8800/8600 supports the Virtual Router Redundancy Protocol (VRRP v3) for IPv6 (defined in draft-ietf-vrrp-ipv6-spec-08.txt).

VRRPv3 for IPv6 provides a faster switchover to an alternate default router than is possible using the ND protocol. With VRRPv3, a backup router can take over for a failed default router in approximately three seconds (using VRRPv3 default parameters). This is accomplished without any interaction with the hosts and with a minimum amount of VRRPv3 traffic.

The operation of Avaya's IPv6 VRRP implementation is similar to the IPv4 VRRP operation, including support for hold-down timer, critical IP, fast advertisements, and backup master. With backup master enabled, the backup switch routes all traffic according to its routing table. It does not Layer 2-switch the traffic to the VRRP master.

New to the IPv6 implementation of VRRP, you must specify a link-local address to associate with the virtual router. Optionally, you can also assign global unicast IPv6 addresses to associate with the virtual router. Network prefixes for the virtual router are derived from the global IPv6 addresses assigned to the virtual router.

With the current implementation of VRRP, one active master switch exists for each IPv6 network prefix. All other VRRP interfaces in a network are in backup mode.

The following figure shows a sample IPv6 VRRP configuration with SMLT. Because the backup router is configured as the backup master, routing traffic is load-shared between the two devices.

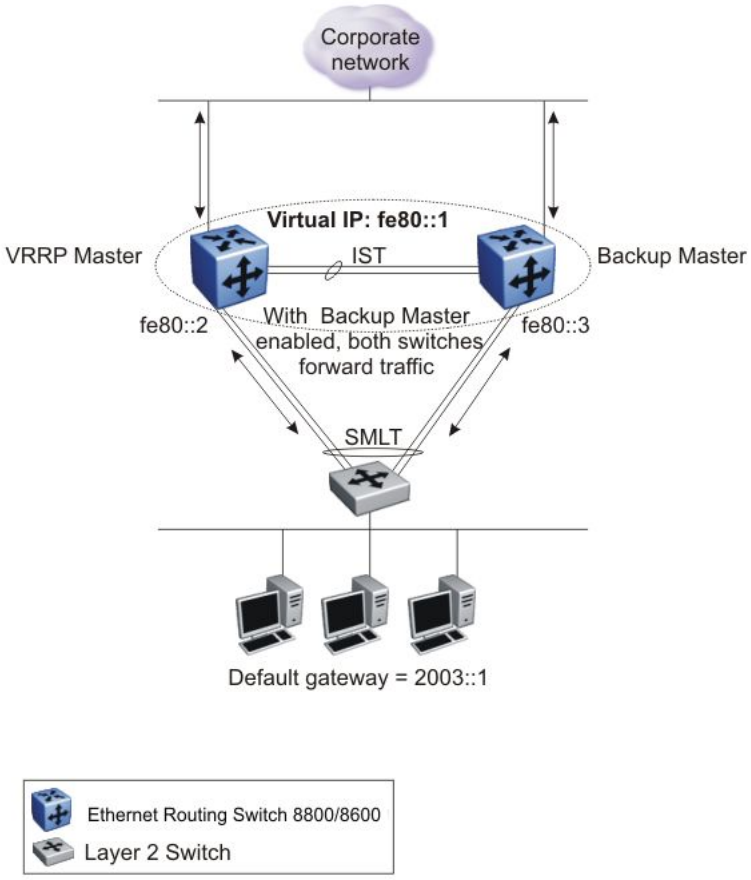


Figure 44: IPv6 VRRP configuration with SMLT

The backup master feature only supports the triangular SMLT topology.

Important:

Do not use VRRP backup master and critical IP at the same time. Use one or the other.

IPv6 VRRP and ICMP redirects

In IPv6 networks, do not enable ICMP redirects on VRRP VLANs. If you enable this option (using the **`config ipv6 icmp redirect-msg`** command), VRRP cannot function. The option is disabled by default.

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, Avaya 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 default gateway resiliency for end stations. The configurations of these features are different, but both provide the same end result and are transparent to the end station.

For more information about RSMLT, see Routed SMLT on page 92.

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 non-IP 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.

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. For more information, 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 an 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 Avaya Ethernet Routing Switch 8800/8600 (when users are attached directly to the 8800/8600).
- Each user is assigned their own VLAN.
- Each user has two VLANs when directly connected to the Avaya Ethernet Routing Switch 8800/8600: one for IP traffic and the other for PPPoE traffic.
- PPPoE bridged traffic preserves user VLANs.

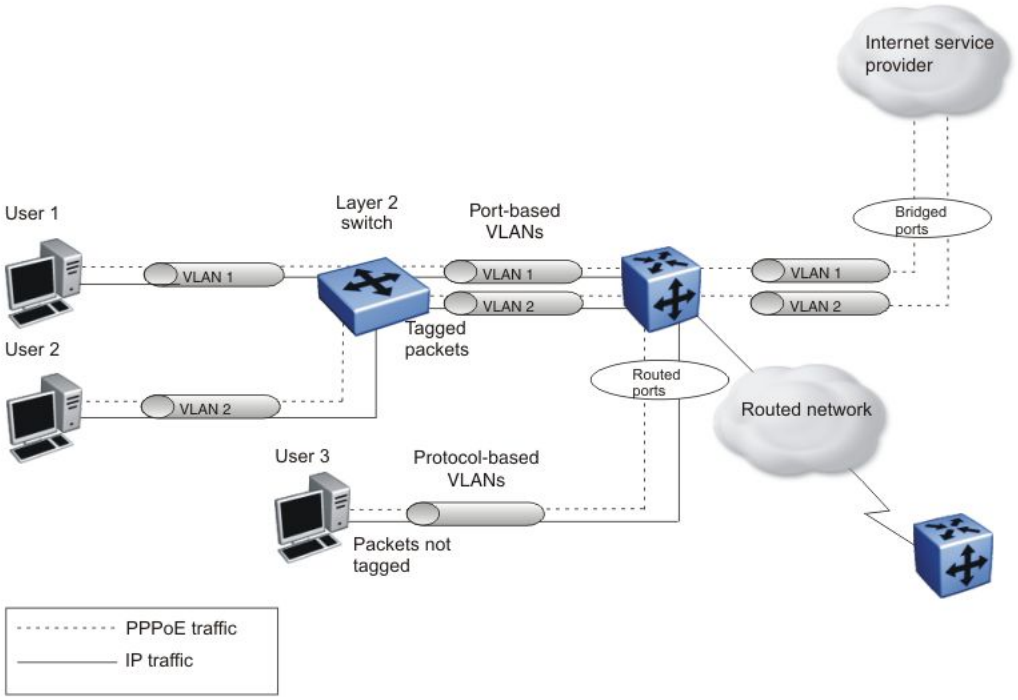


Figure 45: PPPoE and IP traffic separation

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 8800/8600). 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 8800/8600 are bridged, whereas packets between the Ethernet Routing Switch 8800/8600 and the routed network are routed.
- VLANs between the Layer 2 switch and the 8800/8600 are port-based.
- VLANS from the directly connected user (User 3) are protocol-based.
- The connection between the Ethernet Routing Switch 8800/8600 and the ISP is a single port connection.
- The connection between the Layer 2 switch and the Ethernet Routing Switch 8800/8600 can be a single port connection or a MultiLink Trunk (MLT) connection.
- Ethernet Routing Switch 8800/8600 ports connected to the user side (Users 1, 2, and 3) and the routed network are routed ports.
- Ethernet Routing Switch 8800/8600 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.

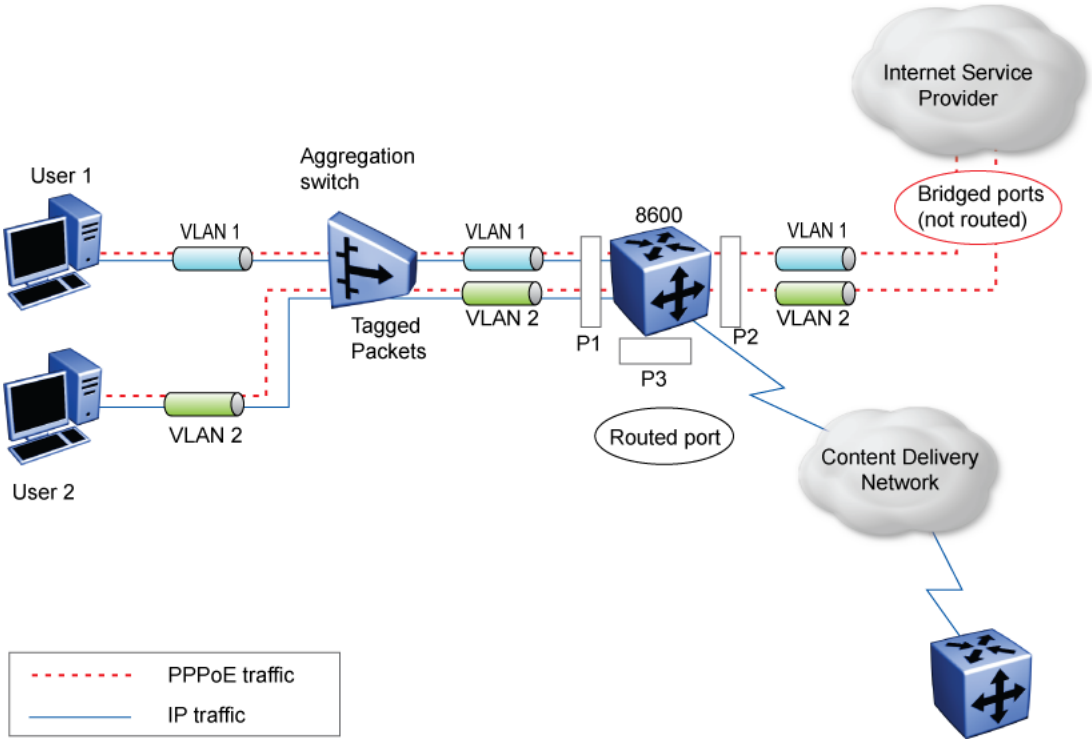


Figure 46: Indirect PPPoE and IP configuration

Direct connections

To directly connect to the Avaya Ethernet Routing Switch 8800/8600, 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).

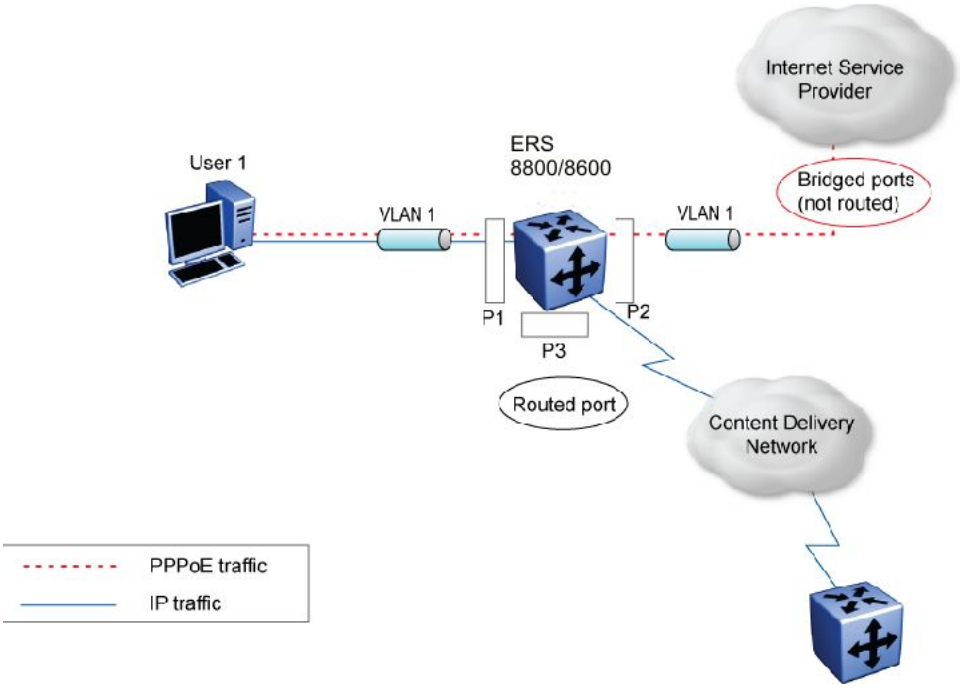


Figure 47: Direct PPPoE and IP configuration

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 8800/8600 software version 3.3 or later installed. BGP is supported on all interface modules. For large BGP environments, Avaya 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 *Avaya Ethernet Routing Switch 8800/8600 Configuration — BGP Services, NN46205-510*.

BGP navigation

BGP scaling

For information about BGP scaling numbers, see Table 5: Supported scaling capabilities on page 28 and *Avaya Ethernet Routing Switch 8800/8600 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), Avaya 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 Avaya Ethernet Routing Switch 8800/8600 and products from other vendors, including Cisco and Juniper.

For more information about BGP and BGP commands, see *Avaya Ethernet Routing Switch 8800/8600 Configuration — BGP Services, NN46205-510*. For configuration examples, see *Border Gateway Protocol (BGP-4) Technical Configuration Guide, NN48500-538*.

BGP design examples

The following design examples describe typical Avaya Ethernet Routing Switch 8800/8600 BGP applications.

BGP and Internet peering

By using BGP, you can perform Internet peering directly between the Avaya Ethernet Routing Switch 8800/8600 and another edge router. In such a scenario, you can use each Avaya Ethernet Routing Switch 8800/8600 for aggregation and peer it with a Layer 3 edge router, as shown in the following figure.

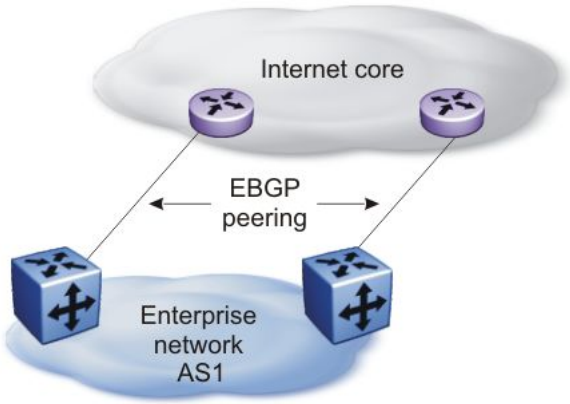


Figure 48: BGP and Internet peering

In cases where the Internet connection is single-homed, to reduce the size of the routing table, Avaya 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).

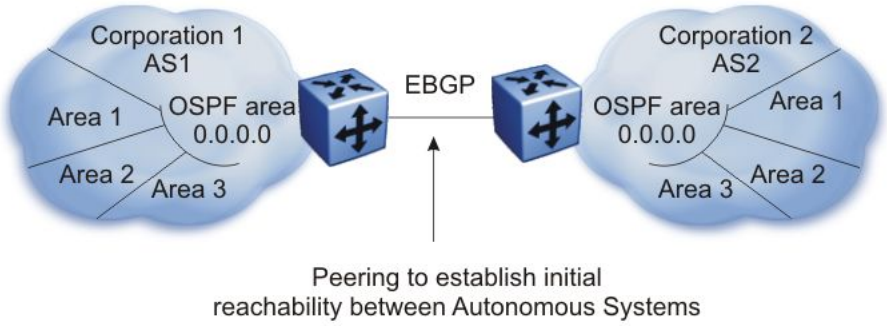


Figure 49: Routing domain interconnection with BGP

BGP and edge aggregation

You can perform edge aggregation with multiple point of presence/edge concentrations. The Avaya Ethernet Routing Switch 8800/8600 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).

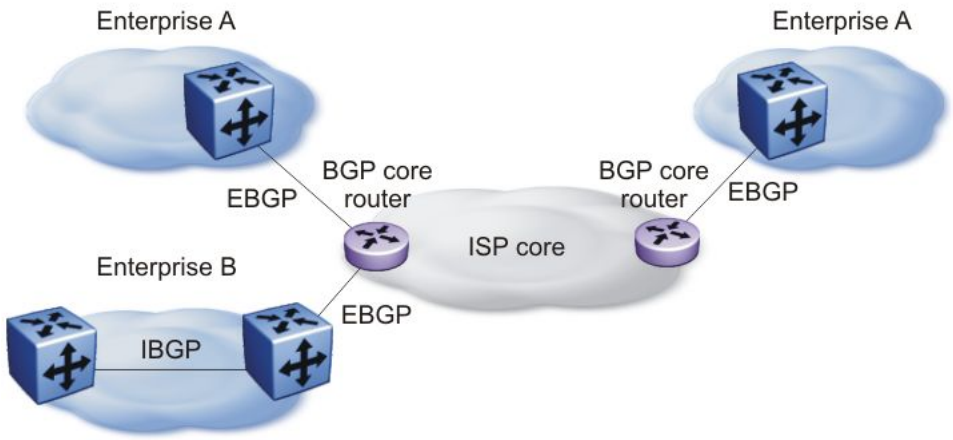


Figure 50: BGP and edge aggregation

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).

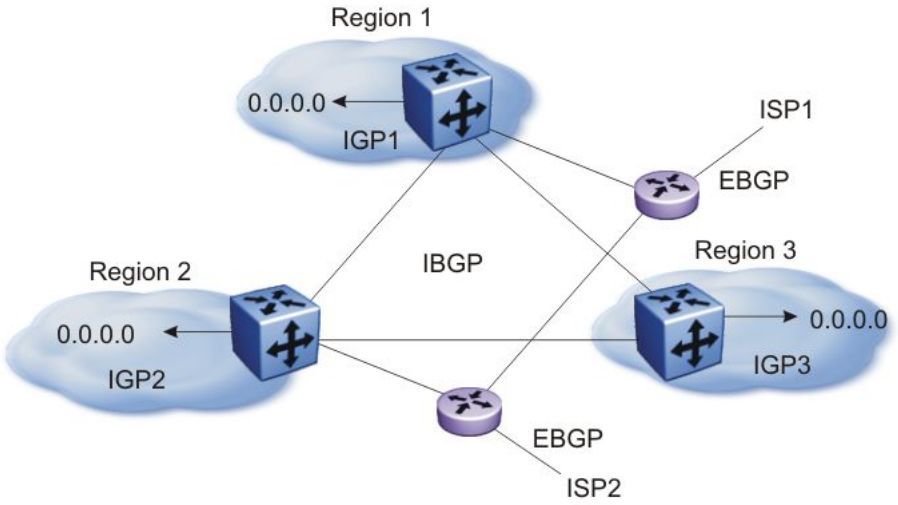


Figure 51: Multiple regions separated by IBGP

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.

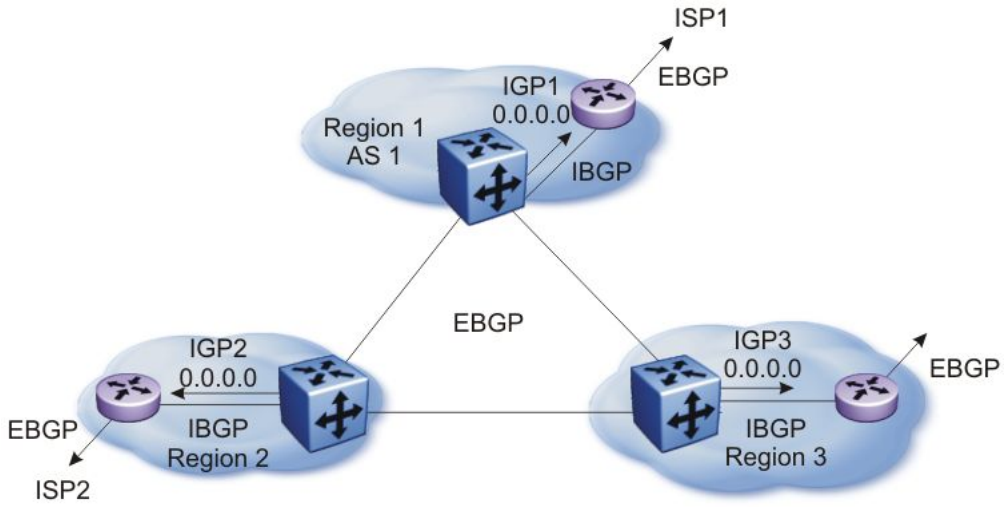


Figure 52: Multiple regions separated by EBGP

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.

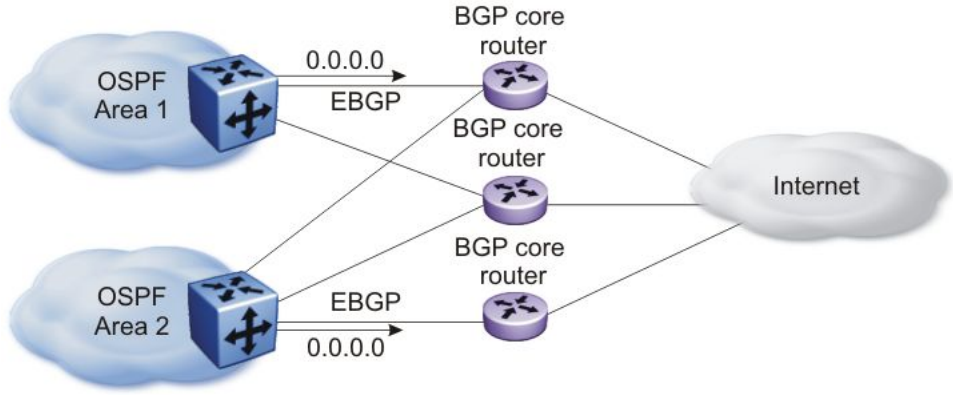


Figure 53: Multiple OSPF regions peering with the Internet

IPv6 BGP+

The Avaya Ethernet Routing Switch 8800/8600 extends the BGPv4 process to support the exchange of IPv6 routes using BGPv4 peering. BGP+ is an extension of BGPv4 for IPV6.

Note that the Ethernet Routing Switch 8800/8600 BGP+ support is not an implementation of BGPv6. Native BGPv6 peering uses the IPv6 Transport layer (TCPv6) for establishing the

BGPv6 peering, route exchanges, and data traffic. Native BGPv6 peering is not supported in Release 7.0.

Ethernet Routing Switch 8800/8600 supports the exchange of BGP+ reachability information over IPv4 transport. To support BGP+, the Ethernet Routing Switch supports two BGP protocol extensions, standards RFC 4760 (multi-protocol extensions to BGP) and RFC 2545 (MP-BGP for IPv6). These extensions allow BGPv4 peering to be enabled with IPv6 address family capabilities.

The Ethernet Routing Switch 8800/8600 implementation of BGP+ uses an existing TCPv4 stack to establish a BGPv4 connection. Optional, nontransitive BGP properties are used to transfer IPv6 routes over the BGPv4 connection. Any BGP+ speaker has to maintain at least one IPv4 address to establish a BGPv4 connection.

Different from IPv4, IPv6 introduces scoped unicast addresses, identifying whether the address is global or link-local. When BGP+ is used to convey IPv6 reachability information for inter-domain routing, it is sometimes necessary to announce a next hop attribute that consists of a global address and a link-local address. For BGP+, no distinction is made between global and site-local addresses.

The BGP+ implementation includes support for BGPv6 policies, including redistributing BGPv6 into OSPFv3, and advertising OSPFv3, static, and local routes into BGPv6 (through BGP+). It also supports the aggregation of global unicast IPv6 addresses and partial HA.

BGP+ does not support confederations. In this release, you can configure confederations for IPv4 routes only.

The basic configuration of BGP+ is the same as BGPv4 with one additional parameter added and some existing commands altered to support IPv6 capabilities. You can enable and disable IPv6 route exchange by specifying the address family attribute as IPv6. Note that an IPv6 tunnel is required for the flow of IPv6 data traffic.

BGP+ is only supported on the global VRF instance.

Limitations

IPv6 BGP convergence in case of SMLT scenarios cannot be guaranteed. Avaya does not recommend to configure BGP peers between SMLT core routers or in between the core router and any switch connecting through SMLT links for the failover scenarios.

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 *Avaya Ethernet Routing Switch 8800/8600 Configuration — OSPF and RIP, NN46205-522*.

OSPF navigation

OSPF scaling guidelines

For information about OSPF scaling numbers, see Table 5: Supported scaling capabilities on page 28 and *Avaya Ethernet Routing Switch 8800/8600 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 = 1 to the number of areas per switch

   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, Avaya 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.

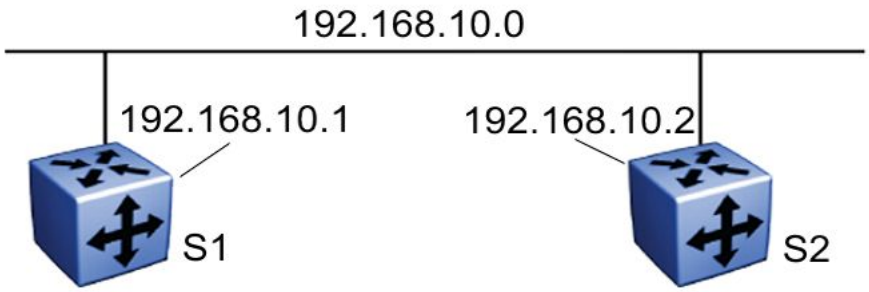


Figure 54: Example 1: OSPF on one subnet in one area

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 as follows:

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.

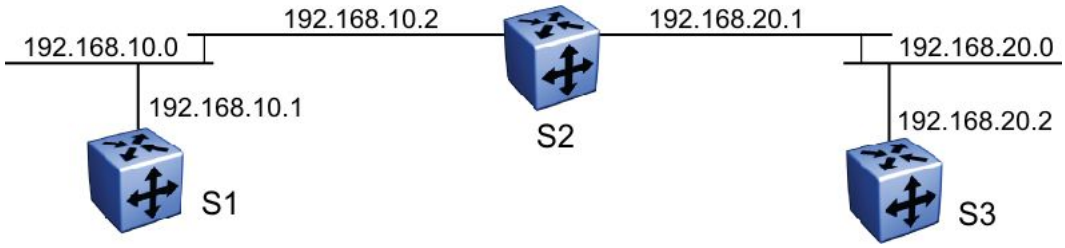


Figure 55: Example 2: OSPF on two subnets in one area

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.

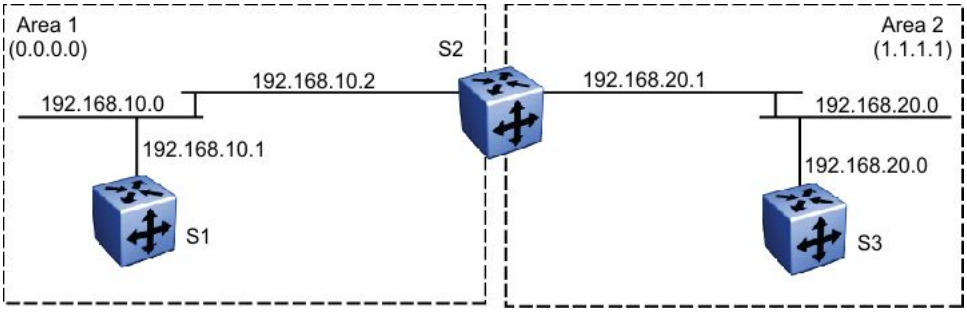


Figure 56: Example 3: OSPF on two subnets in two areas

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 Avaya switches/routers, you may need to manually modify the OSPF parameter RtrDeadInterval to 40 seconds.

IP routed interface scaling

The Avaya Ethernet Routing Switch 8800/8600 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 more information, see *Avaya Ethernet Routing Switch 8800/8600 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. Avaya 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 *Avaya Ethernet Routing Switch 8800/8600 Configuration — IPv6 Routing Operations, NN46205-504*.

IPv6 navigation

- IPv6 requirements on page 147
- IPv6 design recommendations on page 147
- Transition mechanisms for IPv6 on page 147
- Dual-stack tunnels on page 147

IPv6 requirements

To use IPv6, the switch requires at least one 8895 SF/CPU or 8692 SF/CPU module with Enterprise Enhanced CPU daughter card (SuperMezz)

IPv6 design recommendations

Avaya Layer 2 and Layer 3 Ethernet switches support protocol-based IPv6 VLANs. To simplify network configuration with IPv6, Avaya 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 on page 28.

Transition mechanisms for IPv6

The Avaya Ethernet Routing Switch 8800/8600 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 information, see *Avaya Ethernet Routing Switch 8800/8600 Configuration — IPv6 Routing Operations, NN46205-504*.

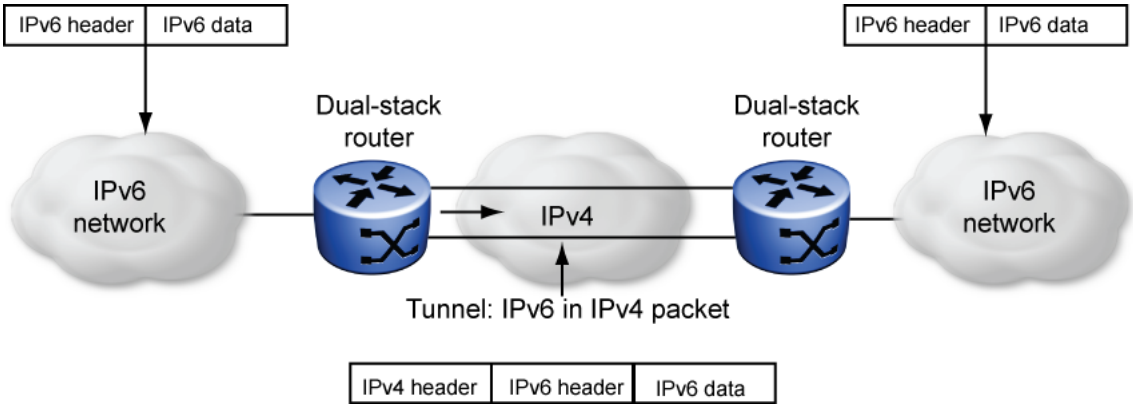


Figure 57: IPv6 tunnels

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.

Chapter 11: SPBM design guidelines

Shortest Path Bridging MAC (SPBM) is a next generation virtualization technology that revolutionizes the design, deployment and operations of Enterprise Campus core networks and Data Centers. The benefits of the technology are clearly evident in its ability to provide massive scalability while at the same time reducing the complexity of the network. SPBM makes network virtualization a much easier paradigm to deploy within the Enterprise environment than other technologies.

This section provides design guidelines that illustrate the operational simplicity of SPBM. This section also lists best practices for configuring SPBM in your network. For more information about SPBM, see the following documents:

- For information on fundamental concepts, command structure and basic configurations, see *Avaya Ethernet Routing Switch 8800/8600 Configuration — Shortest Path Bridging MAC (SPBM)* (NN46205-525).
- For information on advanced configurations, see *Shortest Path Bridging (802.1aq) for ERS 8800/8600 Technical Configuration Guide* (NN48500-617).

SPBM IEEE 802.1aq standards compliance

Release 7.1 introduced a pre-standard implementation of the IEEE 802.1aq standard for Shortest Path Bridging MAC (SPBM) because the standard was not yet ratified. The standard is now ratified and Release 7.1.3 supports it.

Avaya continues to support the pre-standard (or draft) SPBM for previous releases, but all future releases will support standard SPBM only. Release 7.1.3 is a bridge release that supports both draft and standard SPBM. For migration purposes, it is very important to understand the following upgrade considerations:

- Releases prior to 7.1.3 support draft SPBM only.
- Release 7.1.3 supports both draft and standard SPBM.
- Future releases (after 7.1.3) will support standard SPBM only.

Important:

To upgrade to standard SPBM and to use future releases, you must first upgrade to 7.1.3. as an intermediate upgrade step. You can use the CLI or the ACLI to migrate to standard SPBM, but EDM does not support this feature. For more information on migrating and configuring this feature, see *Configuration — Shortest Path Bridging MAC (SPBM)* (NN46205–525). For SPBM deployments, future ERS 8800/8600 releases cannot interoperate with releases prior to 7.1.3.

SPBM 802.1aq standard

The Ethernet Routing Switch 8800/8600 supports the IEEE 802.1aq standard of Shortest Path Bridging MAC (SPBM). SPBM makes *network virtualization* easy to deploy within the Enterprise environment by reducing the complexity of the network while at the same time providing greater scalability. This technology provides all the features and benefits required by Carrier-grade deployments to the Enterprise market without the complexity of alternative technologies traditionally used in Carrier deployments (typically MPLS). SPBM integrates into a single control plane all the functions that MPLS requires multiple layers and protocols to support.

IS-IS

SPBM eliminates the need for multiple overlay protocols in the core of the network by reducing the core to a single Ethernet-based, link-state protocol (IS-IS). IS-IS provides virtualization services, both layer 2 and layer 3, using a pure Ethernet technology base. SPBM also uses IS-IS to discover and advertise the network topology, which enables it to compute the shortest path to all nodes in the SPBM network.

Spanning Tree is a topology protocol that prevents loops but does not scale very well. Because SPBM uses IS-IS, which has its own mechanisms to prevent loops, SPBM does not have to use Spanning Tree to provide a loop free Layer 2 domain.

SPBM uses the IS-IS shortest path trees to populate forwarding tables for each participating node's individual Backbone MAC (B-MAC) addresses. Depending on the topology, SPBM supports as many equal cost multi path trees as there are Backbone VLAN IDs (B-VIDs) provisioned (with a maximum of 16 B-VIDs allowed by the standard and 2 allowed in ERS 8800 release 7.1) per IS-IS instance. IS-IS interfaces operate in point-to-point mode only, which means that for any given Ethernet or MLT interface where IS-IS has been enabled, there can be only one IS-IS adjacency across that interface.

B-MAC

An SPBM backbone includes Backbone Edge Bridges (BEB) and Backbone Core Bridges (BCB). A BEB performs the same functionality as a BCB, but it also terminates one or more Virtual Service Networks (VSN). A BCB does not terminate any VSNs and is unaware of the VSN traffic it transports. A BCB simply knows how to reach any other BEB in the SPBM backbone.

To forward customer traffic across the service provider backbone, the BEB for the VSN encapsulates the customer Ethernet packet received at the edge into a Backbone MAC header using the 802.1ah MAC-in-MAC encapsulation. This encapsulation hides the Customer MAC (C-MAC) address in a Backbone MAC (B-MAC) address pair. MAC-in-MAC encapsulation defines a BMAC-DA and BMAC-SA to identify the backbone source and destination addresses. The originating node creates a MAC header that is used for delivery from end to end. Intermediate BCB nodes within the SPBM backbone perform packet forwarding using BMAC-DA alone. When the packet reaches the intended egress BEB, the Backbone MAC header is removed and the original customer packet is forwarded onwards.

I-SID

SPBM introduces a service instance identifier called I-SID. SPBM uses I-SIDs to separate services from the infrastructure. Once you create an SPBM infrastructure, you can add additional services (such as VLAN extensions or VRF extensions) by provisioning the endpoints only. The SPBM endpoints are called Backbone Edge Bridges (BEBs), which mark the boundary between the core MAC-in-MAC SPBM domain and the edge customer 802.1Q domain. I-SIDs are provisioned on the BEBs to be associated with a particular service instance. In the SPBM core, the bridges are referred to as Backbone Core Bridges (BCBs). BCBs forward encapsulated traffic based on the BMAC-DA.

The SPBM B-MAC header includes an I-SID with a length of 24 bits. I-SIDs identify and transmit virtualized traffic in an encapsulated SPBM frame. These I-SIDs are used in a Virtual Service Network (VSN) for VLANs or VRFs across the MAC-in-MAC backbone.

- With L2 VSN, the I-SID is associated with a customer VLAN, which is then virtualized across the backbone. L2 VSNs offer an any-any LAN service type. L2 VSNs associate one VLAN per I-SID.
- With L3 VSN, the I-SID is associated with a customer VRF, which is also virtualized across the backbone. L3 VSNs are always full-mesh topologies. L3 VSNs associate one VRF per I-SID.

Encapsulating customer MAC addresses in backbone MAC addresses greatly improves network scalability (no end-user C-MAC learning required in the core) and also significantly improves network robustness (loops have no effect on the backbone infrastructure).

The following figure shows the components of a basic SPBM architecture.

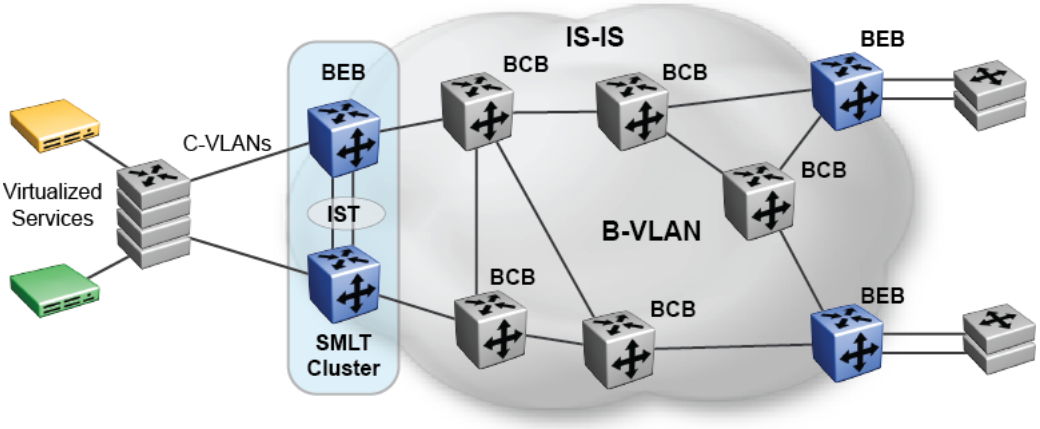


Figure 58: SPBM basic architecture

SPBM provisioning

This section summarizes how to provision SPBM. For information on specific configuration commands, see *Avaya Ethernet Routing Switch 8800/8600 Configuration — Shortest Path Bridging MAC (SPBM)* (NN46205–525).

Infrastructure provisioning

Provisioning an SPBM core is as simple as enabling SPBM and IS-IS globally and on all the IS-IS core Ethernet links on all the BCB and BEB nodes. The IS-IS protocol operates at layer 2 so it does not need IP addresses configured on the links to form IS-IS adjacencies with neighboring switches (like OSPF does). Hence there is no need to configure any IP addresses on any of the core links. The encapsulation of customer MAC addresses in backbone MAC addresses greatly improves network scalability.

There is no flooding and learning of end-user MACs in the backbone. This SPBM provisioning significantly improves network robustness, as customer-introduced network loops have no effect on the backbone infrastructure.

Service provisioning

Provision I-SIDs on a BEB to associate that BEB with a particular service instance. After you map the customer VLAN or VRF into an I-SID, any BEB that has the same I-SID configured can participate in the same L2 or L3 virtual services network (VSN). This same simplicity extends to provisioning the services to run above the SPBM backbone.

- To create an L2 VSN, associate an I-SID number with an edge VLAN.
- To create an L3 VSN, associate an I-SID number with a VRF and configure the desired IS-IS IP route redistribution within the newly created L3 VSN.

Note:

There is no service provisioning needed on the core BCB SPBM switches. This provides a robust carrier grade architecture where configuration on the core switches never needs to be touched when adding new services.

SPBM implementation options

The SPBM architecture is architecturally simple and easy to provision, but it is *not* just for simple networks. SPBM supports multiple implementation options within the same network to meet the demands of the most complex network configurations. The following figure shows how SPBM supports multiple campus networks as well as multiple data centers.

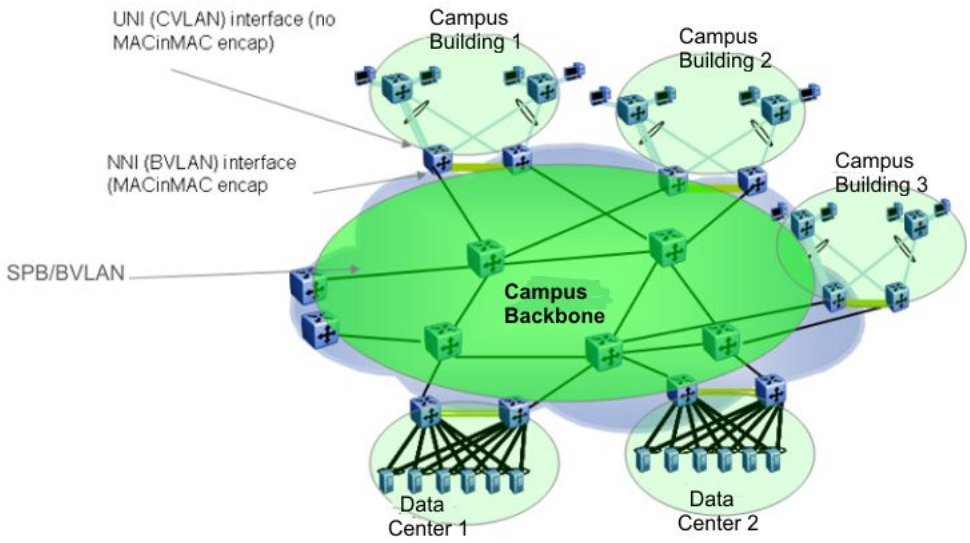


Figure 59: SPBM support for campus and data center architecture

Within the SPBM architecture, you can implement multiple options. The following figure shows all the options that SPBM supports.

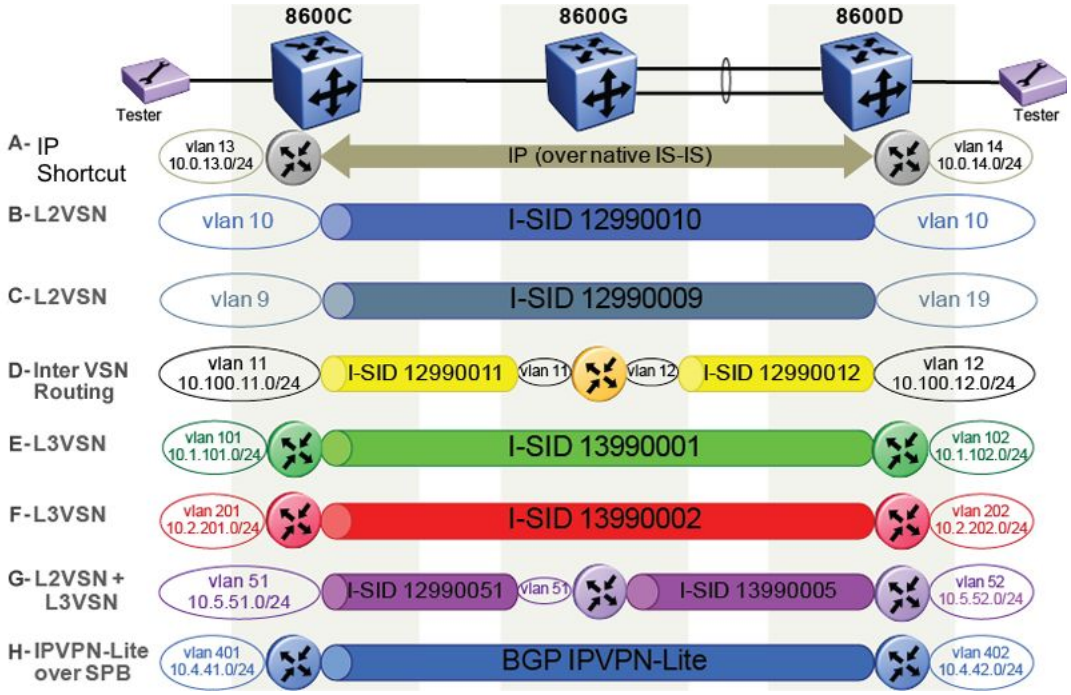


Figure 60: SPBM implementation options

A—IP Shortcut

IP Shortcuts forward standard IP packets over IS-IS. This option enables you to forward IP over the SPBM core, which is a simpler method than traditional IP routing or MPLS. SPBM nodes propagate Layer 3 reachability as “leaf” information in the IS-IS LSPs using Extended IP reachability TLV 135, which contains routing information such as neighbors and locally

configured subnets. SPBM nodes receiving the reachability information use this information to populate the routes to the announcing nodes. All TLVs announced in the IS-IS LSPs are grafted onto the shortest path tree (SPT) as leaf nodes.

An IP route lookup is only required once where the source BEB uses the GRT to identify the BEB closest to the destination subnet. All other nodes perform standard Ethernet switching based on the existing SPT. This allows for end to end IP-over-Ethernet forwarding without the need for ARP, flooding, or reverse learning. Because BCB SPBM nodes only forward on the MAC addresses that comprise the B-MAC header, and since unknown TLVs in IS-IS are relayed to the next hop but ignored locally, SPBM BCB nodes need not be aware of IP subnets to forward IP traffic. Only BEBs generate and receive Extended IP reachability TLV to build routing table in GRT; BCBs just relay the TLV to the next hop based on SPT. In fact, the Extended IP reachabilty TLV is ignored on BCBs.

With IP Shortcuts there are only two IP routing hops (ingress BEB and egress BEB) as the SPBM backbone acts as a virtualized switching backplane.

IP Shortcuts do not require any I-SID configuration. However, IP must be enabled on IS-IS and the IS-IS source address is configured to matched a circuitless/loopback IP address.

In the figure above, node 8600G is acting as a BCB for the service, and has no IP configuration whatsoever.

B—L2 VSN

An L2 Virtual Services Network (VSN) bridges customer VLANs (C-VLANs) over the SPBM core infrastructure. An L2 VSN associates a C-VLAN with an I-SID, which is then virtualized across the backbone. All VLANs in the network that share the same I-SID will be able to participate in the same VSN. If SMLT clusters are used or if you want IS-IS to distribute traffic across two equal cost paths then two backbone VLANs (B-VLAN) are required with a primary B-VLAN and a secondary B-VLAN. Otherwise, only a single B-VLAN is required.

One of the key advantages of the SPBM L2 VSN is that network virtualization provisioning is achieved by configuring the edge of the network (BEBs) *only*. The intrusive core provisioning that other Layer 2 virtualization technologies require is not needed when new connectivity services are added to the SPBM network. For example, when new virtual server instances are created and need their own VLAN instances, they are provisioned at the network edge only and do not need to be configured throughout the rest of the network infrastructure.

Based on its I-SID scalability, this solution can scale much higher than any 802.1Q tagging based solution. Also, due to the fact that there is no need for Spanning Tree in the core, this solution does not need any core link provisioning for normal operation. Redundant connectivity between the C-VLAN domain and the SPBM infrastructure can be achieved by operating two SPBM switches in Switch Clustering (SMLT) mode. This allows the dual homing of any traditional link aggregation capable device into an SPBM network

In the figure above, nodes 8600C & 8600D act as BEBs for the VSN. Only these nodes have a MAC table/FDB for C-VLAN 10.

C—L2 VSN with VLAN translation

L2 VSNs with VLAN translation are basically the same as the L2 VSNs, except that the BEBs on either end of the SPBM network belong to different VLANs. This option enables you to connect one VLAN to another VLAN. In the figure above, VLAN 9 is connected to VLAN 19. The mechanism used to connect them is that they are using the same I-SID (12990009).

D—Inter-VSN Routing

Inter-VSN Routing allows routing between Layer 2 VLANs with different I-SIDs. You can use Inter-VSN Routing to redistribute routes between L2 VLANs. This option allows effective networking of multiple Virtual Service Networks. Where **L2 VSN with VLAN translation** enabled you to interconnect VLANs. This option takes that concept one step further and allows you to interconnect VSNs. It also provides the ability to route IP traffic on L2-VSNs ingressing on NNI interfaces, which is especially useful for L2 edge solutions.

As illustrated in the figure above, routing between VLANs 11 and 12 occurs on the SPBM core switch 8600G shown in the middle of the figure. With Inter-VSN Routing enabled, 8600G transmits traffic between VLAN 11 (I-SID 12990011) and VLAN 12 (I-SID 12990012) on the VRF instance configured. Note that for these VSNs, node 8600G is acting as a BEB.

E—L3 VSN

L3 VSNs are very similar to L2 VSNs. The difference is that L2 VSNs associate I-SIDs with VLANs; L3 VSNs associate I-SIDs with VRFs. With the L3 VSN option, all VRFs in the network that share the same I-SID will be able to participate in the same VSN by advertising into IS-IS their reachable IP routes and install IP routes learnt from IS-IS. Suitable IP redistribution policies need to be defined to determine what IP routes a BEB will advertise to IS-IS.

As illustrated in the figure above, the green VRF on 8600C is configured to advertise its local/direct IP routes into IS-IS within I-SID 13990001; the VRF on node 8600D, which is also a member of the same I-SID, installs these IP routes in its VRF IP routing table with a next-hop B-MAC address of 8600C. Therefore, when the VRF on node 8600D needs to IP route traffic to the IP subnet off 8600C, it performs a lookup in its IP routing table and applies a MAC-in-MAC encapsulation with B-MAC DA of 8600C. The SPBM core ensures delivery to the egress BEB 8600C where the encapsulation is removed and the packet is IP routed onwards.

Note:

Like the IP Shortcut service, there are only two IP routing hops (ingress BEB and egress BEB) as the SPBM backbone acts as a virtualized switching backplane.

F—L3 VSN

The figure above shows two VRFs (green and red) to illustrate that the BEBs can associate I-SIDs with multiple VRFs. The L3 VSN option provides IP connectivity over SPBM for all of your VRFs.

G—L2 VSN and L3 VSN

The figure above shows both an L2 VSN and an L3 VSN to show that you can configure both options on the same BEBs. This topology is simply made up of a number of BEBs that terminate VSNs of both types. What this example shows is the flexibility to extend one or more edge VLANs (using one or more L2 VSNs) to use a default gateway that is deeper into the SPBM core. From here, traffic can then be IP routed onwards as either non-virtualized with IP Shortcuts or, as shown in this example, with a virtualized L3 VSN. Note that in this example the central node 8600G is now also acting as BEB for both service types as it now maintains both a MAC table for the L2 VSN it terminates and an ARP cache and IP routing table for the L3 VSN it also terminates.

H—IP VPN-Lite over SPBM

IP VPN-Lite over SPBM switches IP-in-IP packets in the SPBM core. This option is similar to the traditional 2547 VPN with BGP running between the GRTs to exchange routes between the VRFs. The VRFs themselves are identified by the BGP communities configured for the VRFs. However, unlike the 2547 VPNs, the VPN-Lite model does not use MPLS labels to identify the edge node or the VRF, nor does it use the MPLS transport. Instead it maps a service label, which is an IP address, to each VRF and uses IP-in-IP encapsulation with the outer IP being the service label for the VRF. SPBM switches the IP-in-IP packet in the core by looking up the service label and applying the B-MAC header corresponding to the particular IP service label. IP VPN-Lite can work over any IP routed backbone. In this case it is simply running above the native IP Shortcuts that SPBM provides.

IP VPN-Lite over SPBM also enables you to use hub and spoke configurations by manipulating the import and export Route Target (RT) values. For example, this allows a server frame in a central site to have connectivity to all spokes but no connectivity between the spoke sites. You only have to configure BGP on the BEBs. The BCBs have no knowledge of any Layer 3 VPN IP addresses or routes.

Multiple tenants using different SPBM services

The following figure shows multiple tenants using different services within an SPBM metro network. In this network, you can use some or all of the SPBM implementation options to meet the needs of the community while maintaining the security of information within VLAN members.

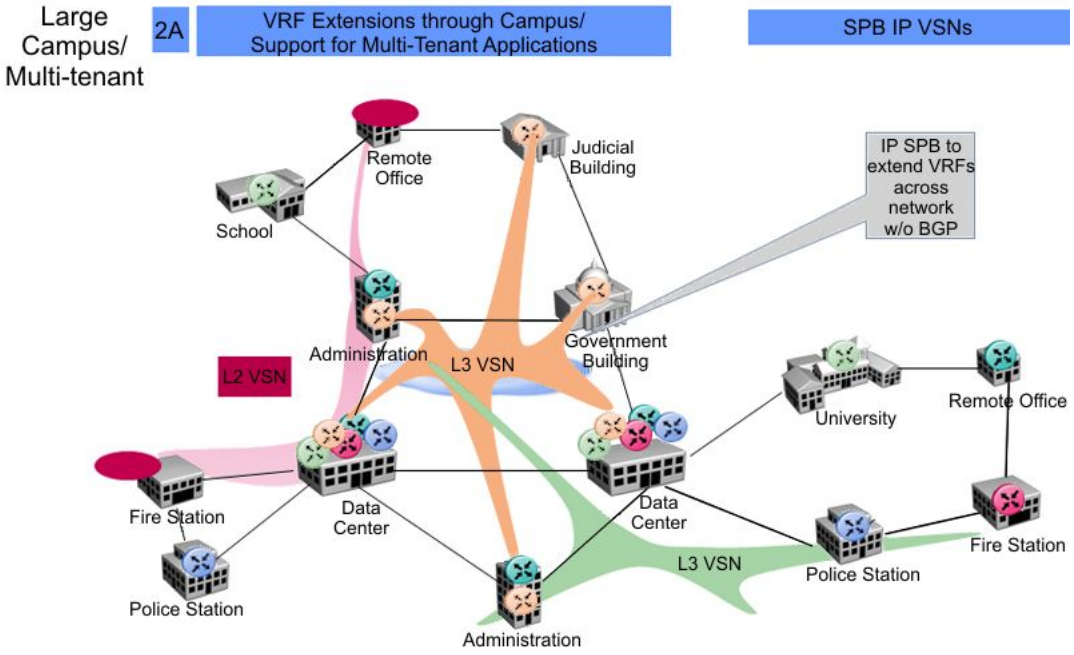


Figure 61: Multi-tenant SPBM metro network

To illustrate the versatility and robustness of SPBM even further, the following figure shows a logical view of multiple tenants in a ring topology. In this architecture, each tenant has their own domain where some users have VLAN requirements and are using L2 VSNs and others

have VRF requirements and are using L3 VSNs. In all three domains, they can share data center resources across the SPBM network.

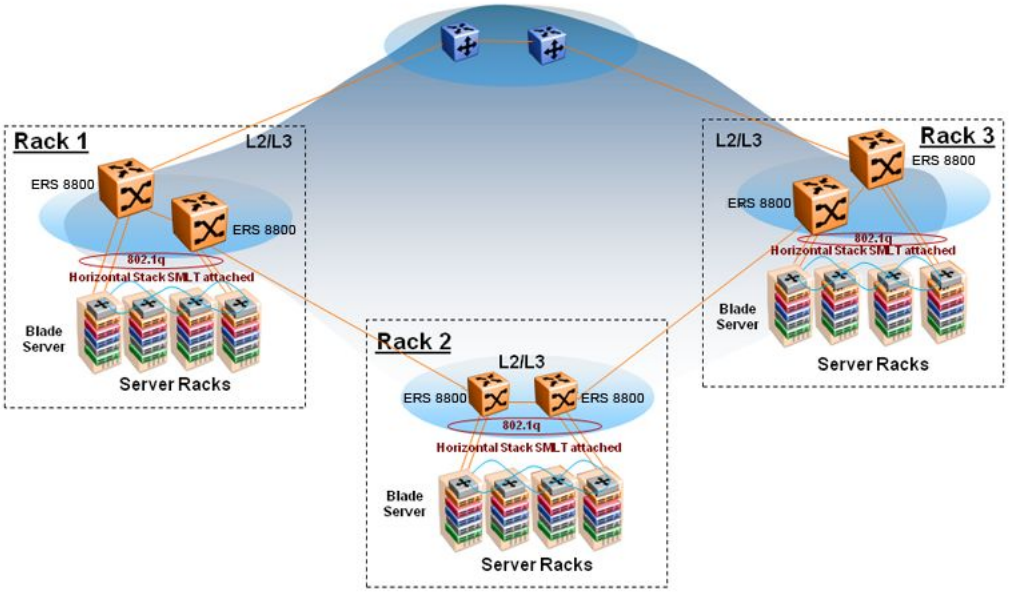


Figure 62: SPBM ring topology with shared data centers

SPBM reference architectures

SPBM has a straightforward architecture that simply forwards encapsulated C-MACs across the backbone. Because the B-MAC header stays the same across the network, there is no need to swap a label or do a route lookup at each node. This allows the frame to follow the most efficient forwarding path from end to end.

The following figure shows the MAC-in-MAC SPBM domain with Backbone Edge Bridges (BEBs) on the boundary and Backbone Core Bridges (BCBs) in the core.

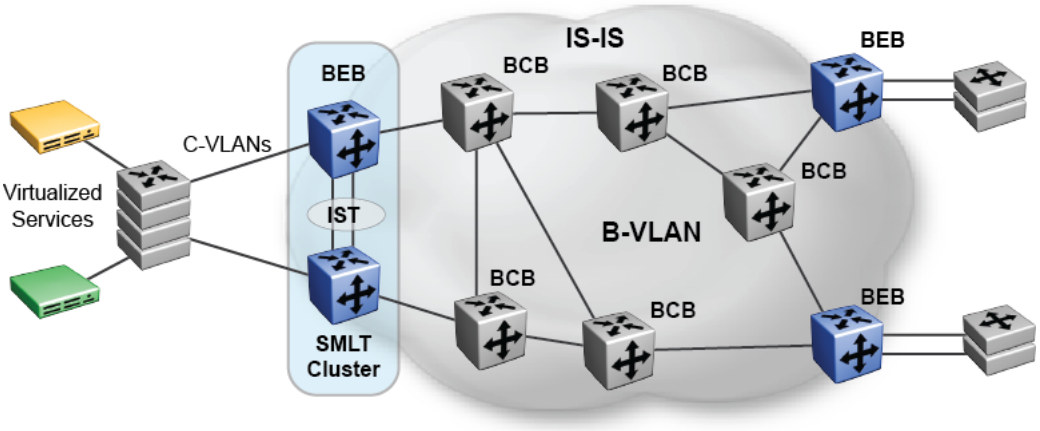


Figure 63: SPBM architecture

Provisioning an SPBM core is as simple as enabling SPBM and IS-IS globally on all the nodes and on the core facing links. To migrate an existing edge configuration into an SPBM network is just as simple.

The boundary between the MAC-in-MAC SPBM domain and the 802.1Q domain is handled by the Backbone Edge Bridges (BEBs). At the BEBs, VLANs or VRFs are mapped into I-SIDs based on the local service provisioning. Services (whether L2 or L3 VSNs) only need to be configured at the edge of the SPBM backbone (on the BEBs). There is no provisioning needed on the core SPBM nodes.

The following figure illustrates an existing Edge connecting to an SPBM Core.

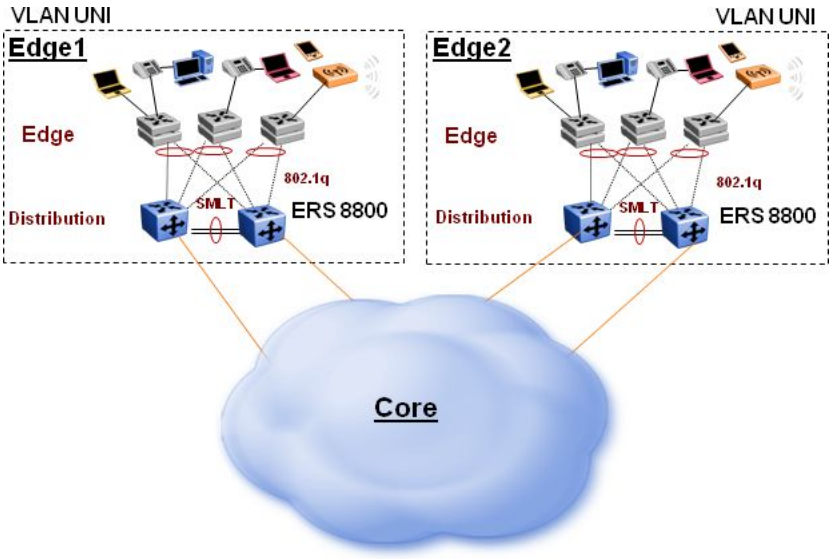


Figure 64: Migrating an existing configuration into SPBM

- For Layer 2 virtualized bridging (L2 VSN), identify all the VLANs that you want to migrate into SPBM and assign them to an I-SID on the BEB.
- For Layer 3 virtualized routing (L3 VSN), map IPv4-enabled VLANs to VRFs, create an IPVPN instance on the VRF, assign an I-SID to it, and then configure the desired IP redistribution of IP routes into IS-IS.

All BEBs that have the same I-SID configured can participate in the same VSN. That completes the configuration part of the migration and all the traffic flows should be back to normal.

The following kinds of traffic are supported by SPBM in the Release 7.1:

- Layer-2 bridged traffic (L2 VSN)
- IPv4 unicast routed traffic on the Global Router (IP Shortcuts)
- IPv4 unicast routed traffic using a VRF (L3 VSN)
- IPv4 Unicast routed traffic using an IPVPN-Lite over SPBM

If your existing edge configuration uses SMLT, you can maintain that SMLT-based resiliency for services configured on the IST peer switches. SPBM requires that you upgrade both IST peer to 7.1 and identify two VLANs to be used as B-VLANs. SPBM then automatically creates a virtual backbone MAC for the IST pair and advertises it with IS-IS. By operating two SPBM switches in Switch Clustering (SMLT) mode, you can achieve redundant connectivity between the C-VLAN domain and the SPBM infrastructure. This allows the dual homing of any traditional link aggregation capable device into an SPBM network.

Related topics:

Campus architecture on page 160
Multicast architecture on page 163
Large data center architecture on page 163

Campus architecture

For migration purposes, you can add SPBM to an existing network that has SMLT configured. In fact, if there are other protocols already running in the network such as OSPF, you can leave them in place too. SPBM uses IS-IS, and operates independently from other protocols. However, Avaya recommends that you eventually eliminate SMLT in the core and other unnecessary protocols. This reduces the complexity of the network and makes it much simpler to maintain and troubleshoot.

Whether you configure SMLT in the core or not, the main point to remember is that SPBM separates services from the infrastructure. For example, in a large campus, a user may need access to other sites or data centers. SPBM enables you to grant that access by associating the user to a specific I-SID. This mechanism enables the user to do his work without getting access to another department's confidential information.

The following figure depicts a topology where the BEBs in the Edge and Data Center Distribution nodes (blue icons) are configured in SMLT Clusters. Prior to implementing SPBM, the core nodes (yellow icons) would also have been configured as SMLT Clusters. When migrating SPBM onto this network design, it is important to note that you can deploy SPBM over the existing SMLT topology without any network interruption. Once the SPBM infrastructure is in place, you can create VSN services over SPBM (or migrate them from the previous end to end SMLT-based design).

After migrating all services to SPBM, the customer VLANs (C-VLANs) will exist **only** on the BEB SMLT Clusters at the edge of the SPBM network (blue icons). The C-VLANs will be assigned to an I-SID instance and then associated with either a VLAN in an L2 VSN or terminated into a VRF in an L3 VSN. You can also terminate the C-VLAN into the default Global Routing Table (GRT), which uses IP Shortcuts to route over the SPBM core.

In an SPBM network design, the only nodes where it makes sense to have an SMLT Cluster configuration is on the BEB nodes where VSN services terminate. These are the SPBM nodes where C-VLANs exist and these C-VLANs need to be redundantly extended to non-SPBM devices such as L2 edge stackable switches. On the BCB core nodes where no VSNs are terminated and no L2 edge stackables are connected, there is no longer any use for the SMLT Clustering functionality. Therefore, in the depicted SPBM design, the SMLT/IST configuration can be removed from the core nodes because they now act as pure BCBs that have no knowledge of the VSN they transport and the only control plane protocol they need to run is IS-IS.

Since SMLT BEB nodes exist in this design (the edge BEBs) and it is desirable to use equal cost paths to load balance VSN traffic across the SPBM core, all SPBM nodes in the network are configured with the same two B-VIDs.

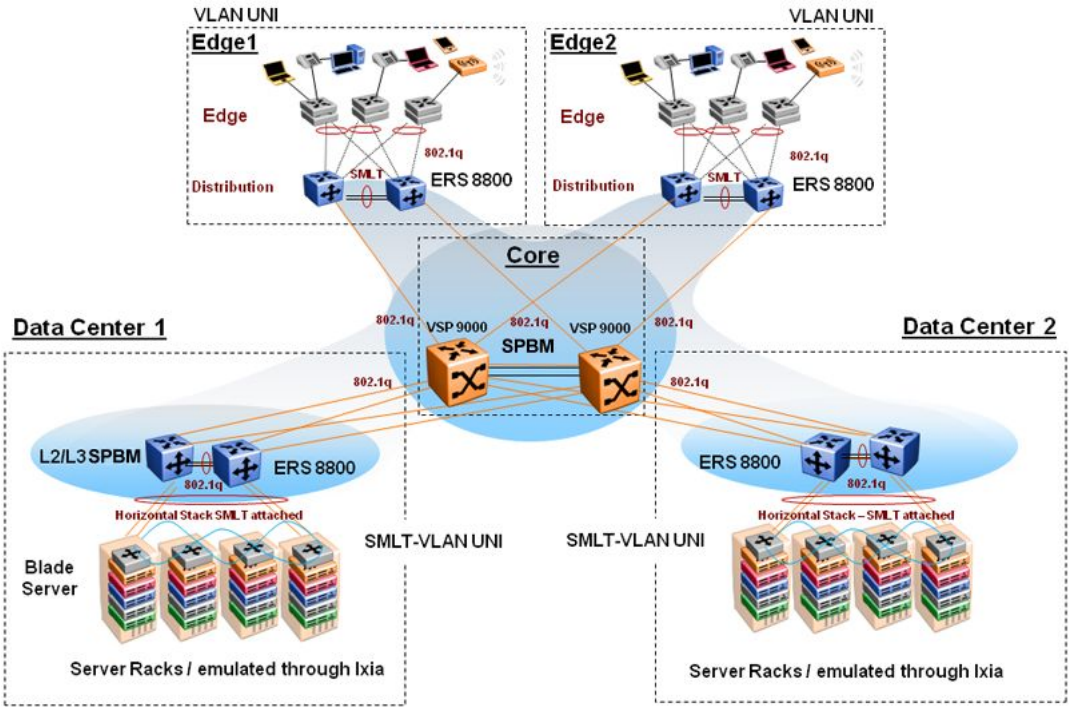


Figure 65: SPBM campus without SMLT

Where the above figure shows the physical topology, the following two figures illustrate a “logical” rendition of the same topology. In both of the following figures, you can see that the core is almost identical. Why? Because the SPBM core just serves as a transport mechanism that transmits traffic to the destination BEB. All the provisioning is done at the edge.

In the data center, VLANs are attached to Inter-VSNs that transmit the traffic across the SPBM core between the data center on the left and the data center on the right. A common application of this service is VMotion moving VMs from one data center to another.

The first figure below uses IP Shortcuts that route VLANs in the GRT. There is no I-SID configuration and no Layer 3 virtualization between the Edge Distribution and the Core. This is normal IP forwarding to the BEB.

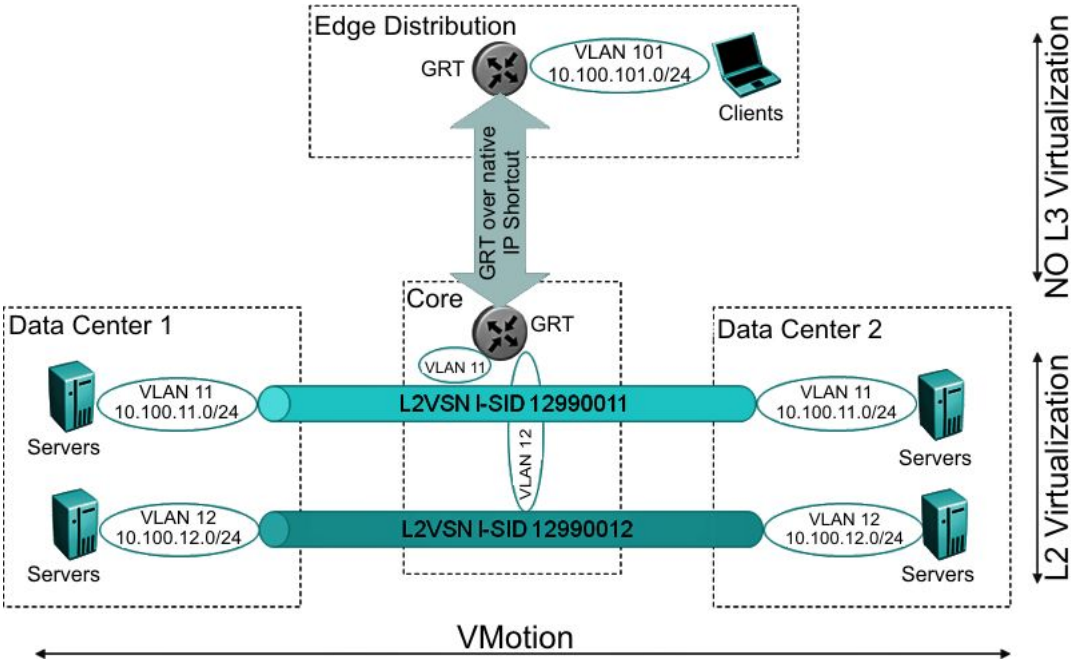


Figure 66: IP Shortcut scenario to move traffic between data centers

The figure below uses L3 VSNs to route VRFs between the Edge Distribution and the Core. The VRFs are attached to I-SIDs and use Layer 3 virtualization.

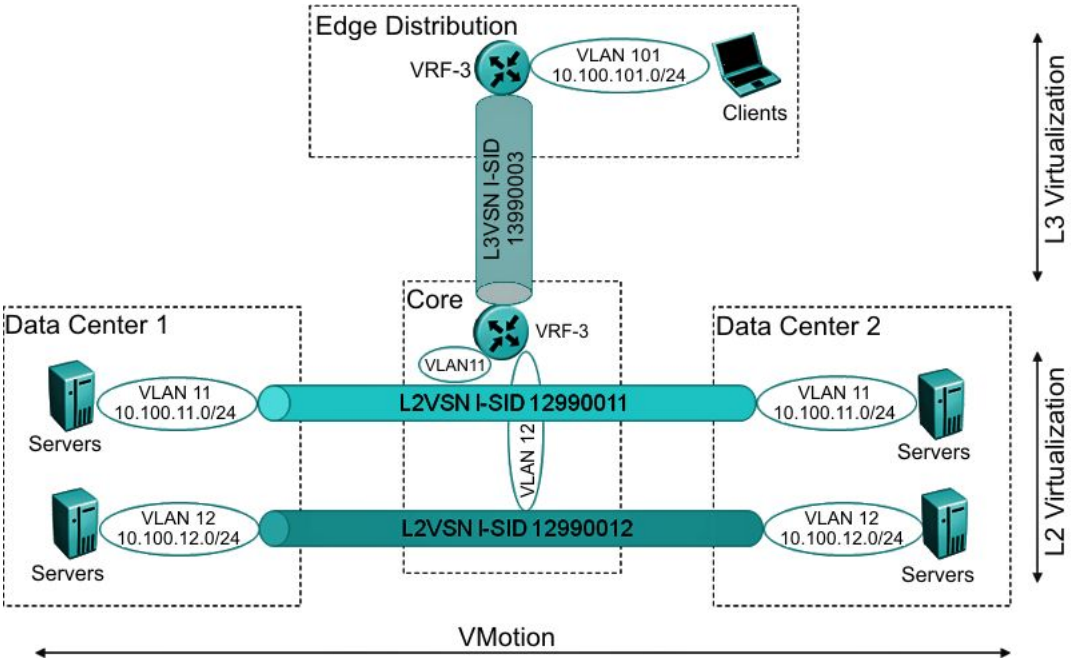


Figure 67: VRF scenario to move traffic between data centers

Multicast architecture

SPBM transports multicast streams, but SPBM does not support multicast routing in Release 7.1. However, you can keep a traditional SMLT/OSPF/PIM design operating in parallel to SPBM.

SPBM uses L2 VSNs to tunnel multicast traffic between multicast routers. This means that SPBM transports multicast streams across the core, but there is no multicast **routing** until traffic reaches the edge switches.

Large data center architecture

SPBM supports data centers with IP Shortcuts, L2 VSNs, or L3 VSNs. If you are using vMotion, then Layer 2 must be used between data centers (L2 VSN). With L2 VSNs, you can simply add an IP addresses to the VLAN on both data centers and run VRRP between them to allow the ESX server to route to the rest of the network.

The following figure shows an SPBM topology of a large data center. This figure represents a full-mesh Virtual Enterprise Network Architecture (VENA) data center fabric using SPBM for storage over Ethernet. This topology is optimized for storage transport because traffic never travels more than two hops.

Note:

Avaya recommends a two-tier, full-mesh topology for large data centers.

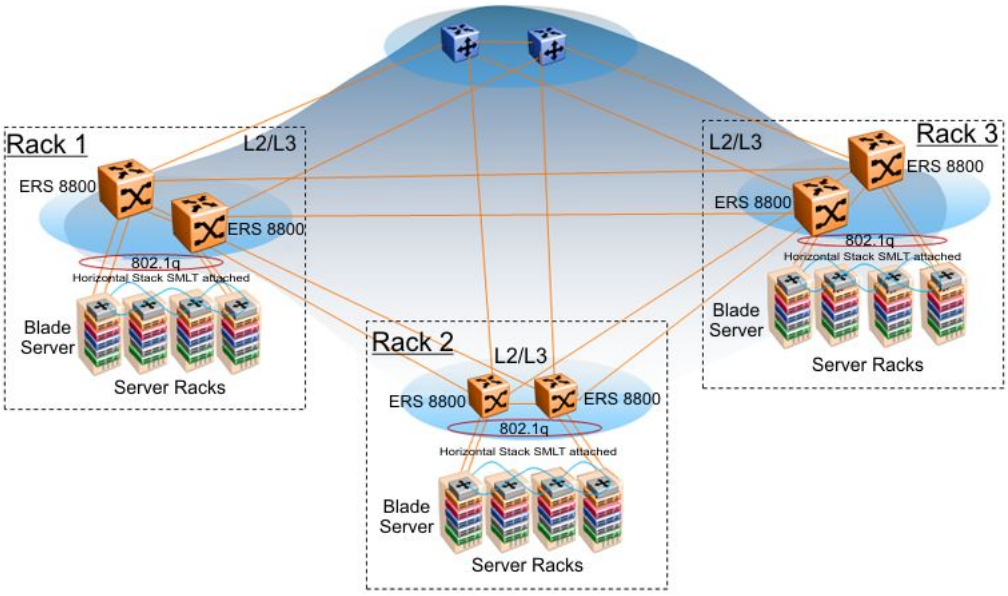


Figure 68: SPBM data center — full-mesh

Traditional data center routing of VMs

In a traditional data center configuration, the traffic flows into the network to a VM and out of the network in almost a direct path. (The red device in the following figures represent the VM.)

The figure below shows an example of a traditional data center with Virtual Router Redundancy Protocol (VRRP) configured. Because end stations are often configured with a static default gateway IP address, a loss of the default gateway router causes a loss of connectivity to the remote networks. VRRP eliminates the single point of failure that can occur when the single static default gateway router for an end station is lost.

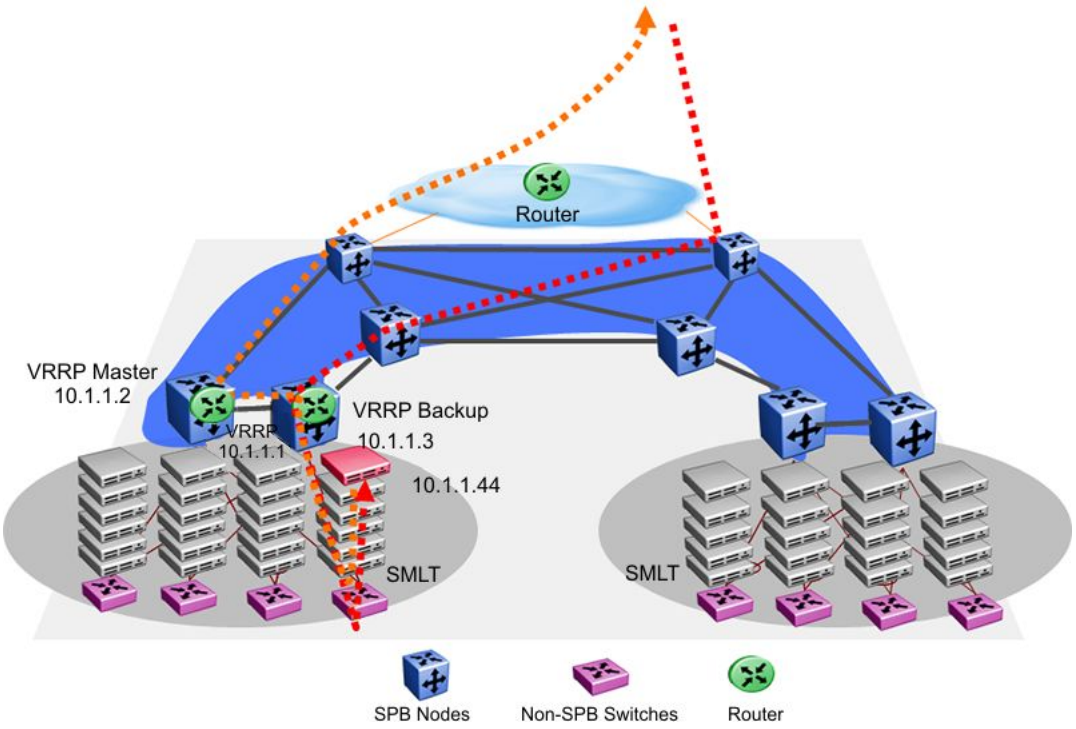


Figure 69: Traditional routing before moving VMs

A VM is a virtual server. When a VM is moved, the virtual server is moved as is. This means that the IP addresses of that server remain the same when the server is moved from one data center to the other. This in turn dictates that the same IP subnet (and hence VLAN) be present in both data centers.

In the following figure, the VM (red device) moved from the data center on the left to the data center on the right. To ensure a seamless transition that is transparent to the user, the VM retains its network connections through the default gateway. This method works, but it adds more hops to all traffic. As you can see in the figure below, one VM move results in a convoluted traffic path. Multiply this with many moves and soon the network look like a tangled mess that is very inefficient, difficult to maintain, and almost impossible to troubleshoot.

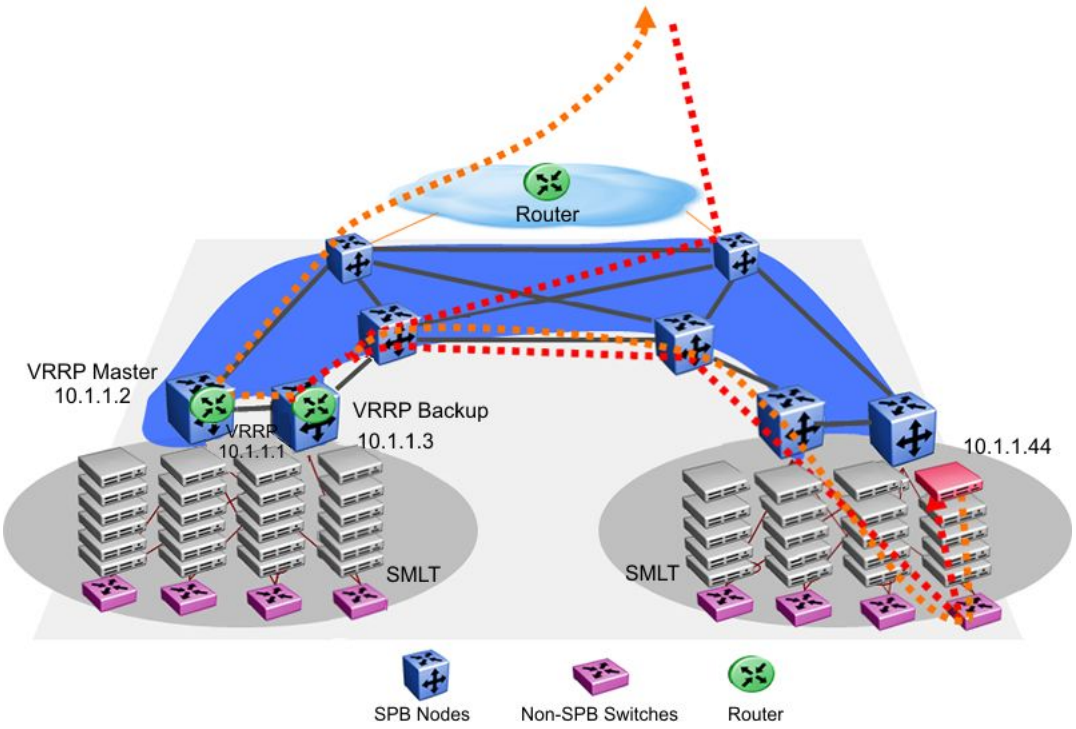


Figure 70: Traditional routing after moving VMs

Optimized data center routing of VMs

There are two main features that make a data center *optimized*:

- VLAN Routers in the Layer 2 domain (green icons)
- VRRP BackupMaster

The VLAN Routers use lookup tables to determine the best path to route incoming traffic (red dots) to the destination VM.

VRRP BackupMaster solves the problem with traffic congestion on the IST. Because there can only be one VRRP Master, all other interfaces are in backup mode. In this case, all traffic is forwarded over the IST link towards the primary VRRP switch. All traffic that arrives at the VRRP backup interface is forwarded, so there is not enough bandwidth on the IST link to carry all the aggregated riser traffic. VRRP BackupMaster overcomes this issue by ensuring that the IST trunk is not used in such a case for primary data forwarding. The VRRP BackupMaster acts as an IP router for packets destined for the logical VRRP IP address. All traffic is directly routed to the destined subnetwork and not through Layer 2 switches to the VRRP Master. This avoids potential limitation in the available interswitch trunk bandwidth.

The following figure shows a solution that optimizes your network for bidirectional traffic flows. However, this solution turns two SPBM BCB nodes into BEBs where MAC and ARP learning will be enabled on the Inter-VSN routing interfaces. If you do not care about top-down traffic flows, you can omit the Inter-VSN routing interfaces on the SPBM BCB nodes. This makes the IP routed paths top-down less optimal, but the BCBs will remain pure BCBs, thus simplifying core switch configurations.

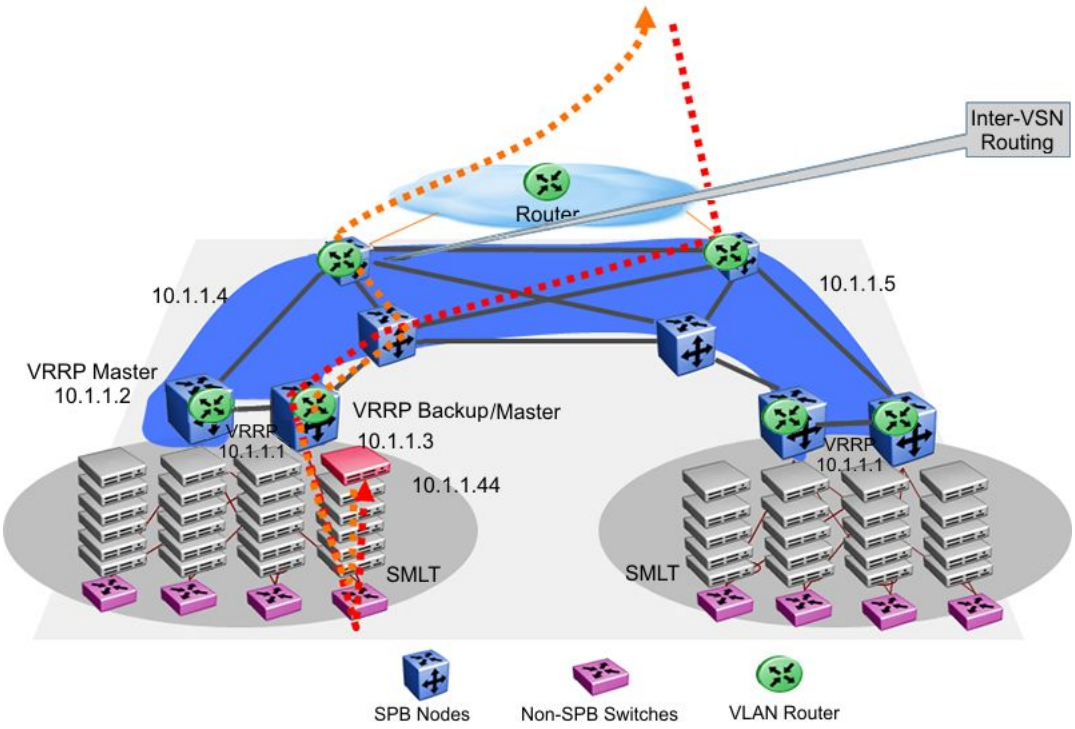


Figure 71: Optimized routing before moving VMs

In the traditional data center, we saw the chaos that resulted when a lot of VMs were moved. In an optimized data center as shown below, the incoming traffic enters the Layer 2 domain where an edge switch uses Inter-VSN Routing to attach an I-SID to a VLAN. The I-SID bridges traffic directly to the destination. With VRRP Backup Master, the traffic no longer goes through the default gateway; it takes the most direct route in and out of the network.

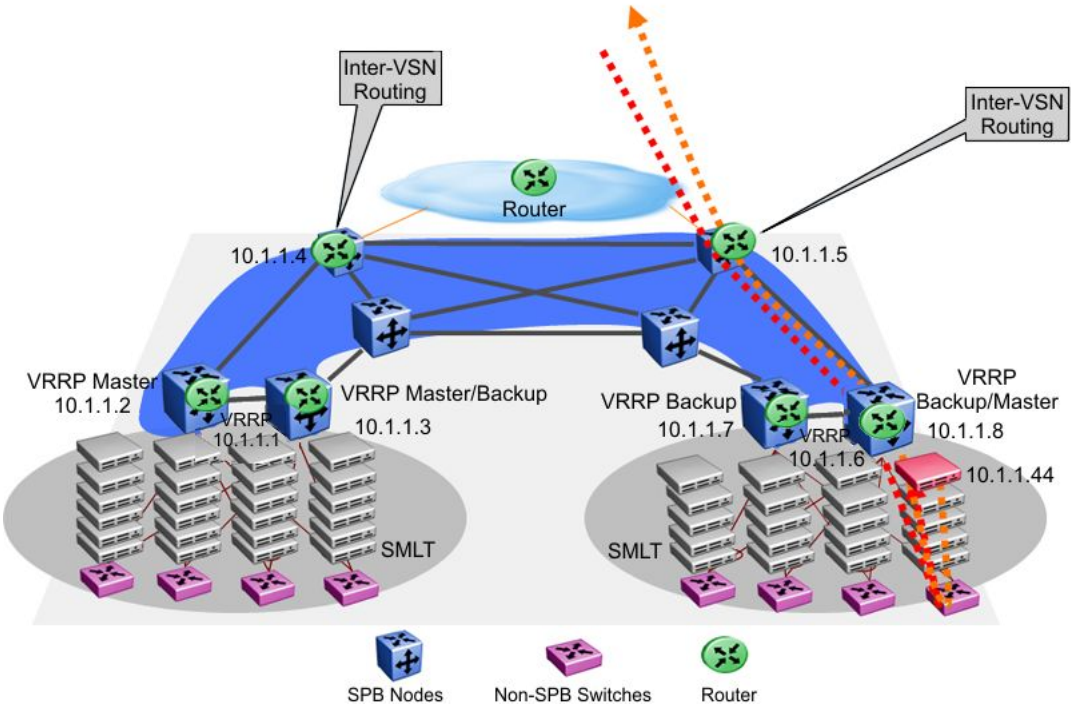


Figure 72: Optimized routing after moving VMs

SPBM scaling and performance capabilities

The test scenarios in this section simulate deployments of SPBM in Enterprise networks. Each figure represents an actual configuration that was tested in the Avaya Data Solutions Test Lab. The table accompanying each figure shows the scaling results of the tests performed in that test bed.

Upgrade scenarios

The figures in this section illustrate how two networks with SMLT switch clusters can be upgraded to SPBM. The accompanying test results in the table below prove that SMLT can be successfully upgraded to SPBM with no loss of functionality. The Figure 73: Test bed for upgrading a data center to STP/SPBM on page 169 figure shows a network with an SMLT switch cluster configuration upgrading to an SPBM configuration. When SPBM is used with SMLT, a virtual backbone MAC is automatically created for the IST pair and is advertised by IS-IS. The gray-dashed line outlines all the devices included in the SMLT and STP network. The red-dashed line outlines all the devices included in the SPBM network.

When you complete the configuration steps, this upgrade reduces the complexity of the network. For more information and a configuration example, see *Shortest Path Bridging (802.1aq) for ERS 8600 / 8800 Technical Configuration Guide* (NN48500-617).

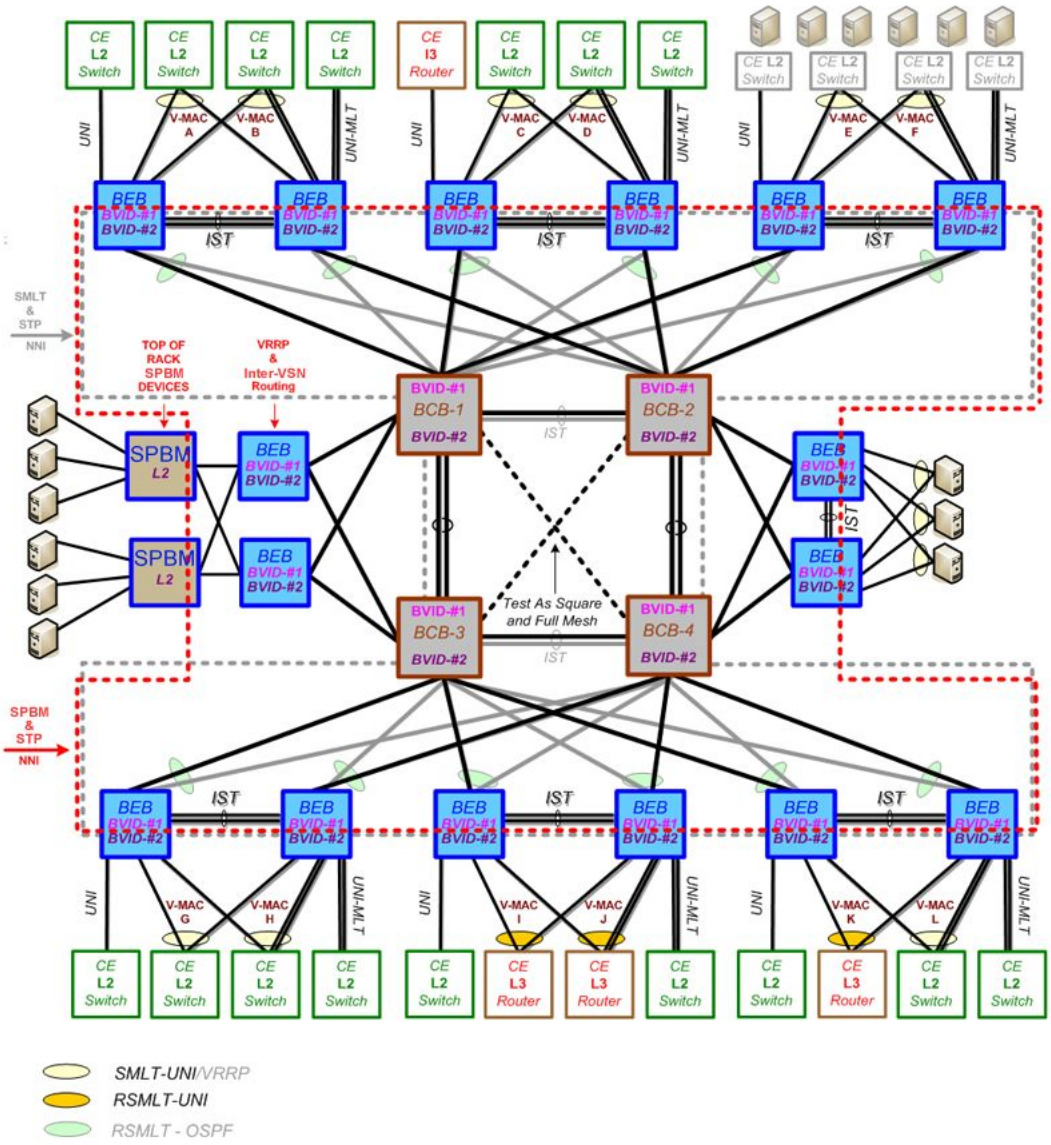


Figure 73: Test bed for upgrading a data center to STP/SPBM

The following table shows how SPBM scales in a data center with an SMLT switch cluster.

Table 25: Scaling capability for the STP/SPBM data center test bed

Service	Maximum number tested in this scenario	Maximum number supported in all test scenarios
Multicast streams on non-SPBM VLANs when SPBM is enabled on the switch	500	1500

Service	Maximum number tested in this scenario	Maximum number supported in all test scenarios
ARP entries with SPBM enabled on the switch	2000	6000
VLAN entries with SPBM enabled on the switch	20	100
VRF instances	64	256 (including GRT)
GRT routes	2000	8000
VRF routes total	2000	8000
L2 VPNs	500	1000
Number of MACs on VPNs	2000	30 000

The Test bed for upgrading a network to STP/SPBM on page 171 figure is basically a subset of the figure above except this scenario has different scaling requirements. Again, the gray-dashed line outlines all the devices included in the SMLT and STP network and the red-dashed line outlines all the devices included in the SPBM network. In this figure, SPBM aggregates the local and remote C-MACs on the edge of the network and associates them to an I-SID. To transport them across the IS-IS network, remote C-MACs are referenced to an I-SID/NextHop pair. The NextHop will be either an I-SID/IS-IS SystemId-MAC or an I-SID/IS-IS Virtual-Mac.

IS-IS advertises a unique Virtual MAC to support SMLT-UNI so that LSP re-advertising and remote C-MAC re-learning is unnecessary when SMLT-UNI link failures occur. IS-IS also supports two B-MAC instances.

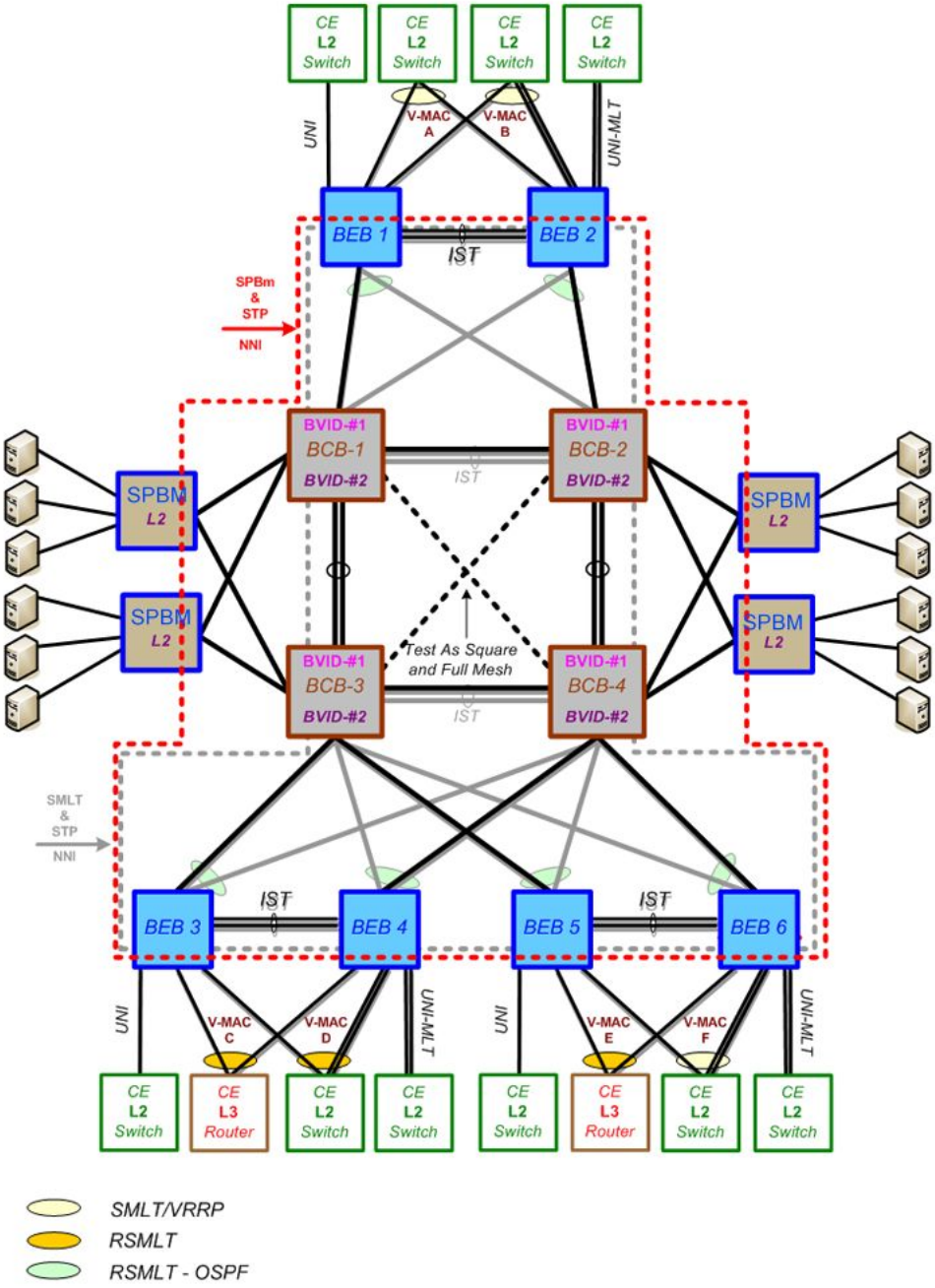


Figure 74: Test bed for upgrading a network to STP/SPBM

The following table shows how SPBM scales in a smaller configuration with SMLT.

Table 26: STP/SPBM scaling capability for the small test bed

Service	Maximum number tested in this scenario	Maximum number supported in all test scenarios
Multicast groups on non-SPBM VLANs when SPBM is enabled on the switch	1500	1500
ARP entries with SPBM enabled on the switch	6000	6000
VLAN entries with SPBM enabled on the switch	100	100
VRF instances	64	256 (including GRT)
GRT routes	8000	8000
VRF routes total	8000	8000
L2 VPNs	300	1000
Number of MACs on VPNs	10 000	30 000

Greenfield scenarios

The two figures in this section show greenfield installations of SPBM. To streamline and simplify the networks, there are no SMLT switch clusters in the core of these configurations. The scenario shown in the Test bed for a greenfield MSTP/SPBM configuration on page 173 figure tests a square topology in the core and the red-dashed line outlines all the devices included in the SPBM network. This figure shows a network with an MSTP/SPBM configuration

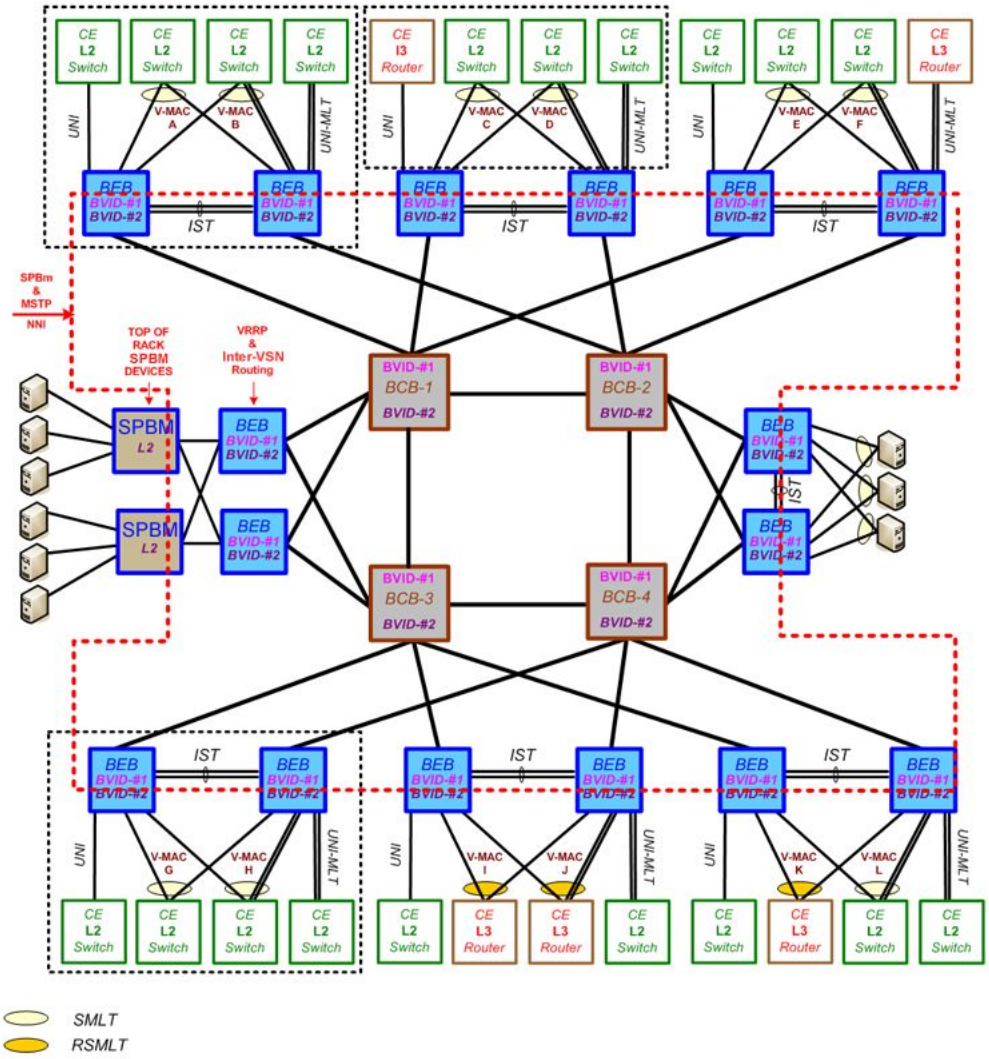


Figure 75: Test bed for a greenfield MSTP/SPBM configuration

The following table shows how SPBM scales in a data center with MSTP.

Table 27: Greenfield MSTP/SPBM scaling capability

Service	Maximum number tested in this scenario	Maximum number supported in all test scenarios
Multicast streams on non-SPBM VLANs when SPBM is enabled on the switch	500	1500
ARP entries with SPBM enabled on the switch	2000	6000

Service	Maximum number tested in this scenario	Maximum number supported in all test scenarios
VLAN entries with SPBM enabled on the switch	20	100
VRF instances	64	256 (including GRT)
GRT routes	2000	8000
VRF routes total	2000	8000
L2 VPNs	500	1000
Number of MACs on VPNs	2000	30 000

The scenario shown in the Test bed for a greenfield MSTP/SPBM ring configuration on page 175 figure tests a ring topology and the red-dashed line outlines all the devices included in the SPBM network.

The focus of this test scenario is to verify the multicast scaling requirements for some customers. In this configuration, there are up to 1,500 multicast groups streaming data across the SPBM network.

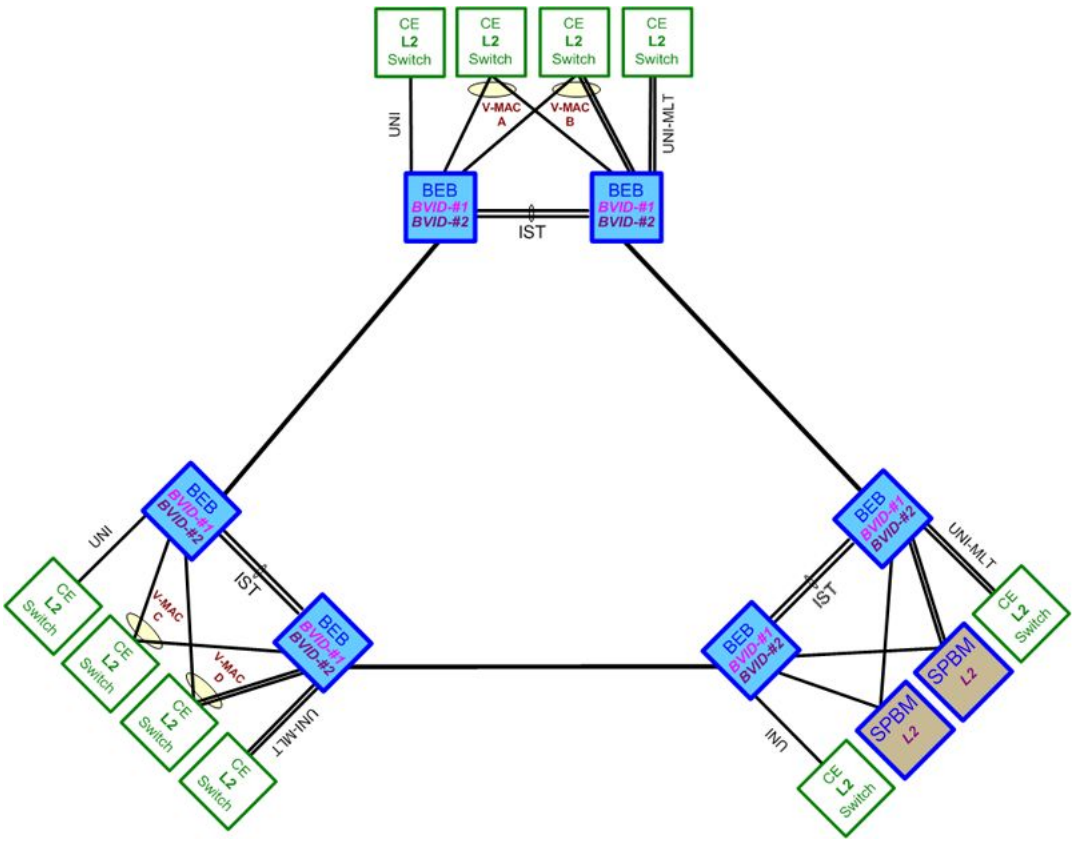


Figure 76: Test bed for a greenfield MSTP/SPBM ring configuration

The following table shows how SPBM scales in a ring with MSTP.

Table 28: Greenfield MSTP/SPBM ring scaling capability

Service	Maximum number tested in this scenario	Maximum number supported in all test scenarios
Multicast groups on non-SPBM VLANs when SPBM is enabled on the switch	100	1500
ARP entries with SPBM enabled on the switch	2000	6000
VLAN entries with SPBM enabled on the switch	20	100
VRF instances	10	256 (including GRT)
GRT routes	500	8000
VRF routes total	500	8000
L2 VPNs	1000	1000

Service	Maximum number tested in this scenario	Maximum number supported in all test scenarios
Number of MACs on VPNs	10 000	30 000

SPBM best practices

This section describes best practices when setting up SPBM networks.

IS-IS

- Avaya recommends that you change the IS-IS system ID from the default B-MAC value to a recognizable address to easily identify a switch. This helps to recognize source and destination addresses for troubleshooting purposes.
  - If you leave the system ID as the default value (safe practice as it ensures no duplication in the network), it can be difficult to recognize the source and destination B-MAC for troubleshooting purposes.
  - If you do manually change the system ID, take the necessary steps to ensure there is no duplication in the network.
- Create two B-VLANs to allow load distribution over both B-VLANs. This is required when using SMLT. Even if SMLT is not used in the network, this is still good practice as adding a second B-VLAN to an existing configuration allows SPBM to load balance traffic across two equal cost multipaths if the physical topology grants it.
- In a ring topology with OSPF and IS-IS configured in the core, a core link break causes slow convergence that may lead to SPBM L2 traffic loss. If the last member link of an OSPF VLAN fails, it takes down the IP interface and OSPF has to reconverge. Keep in mind that while OSPF is reconverging, SPBM cannot get any CPU time so there is some traffic loss.

SPBM

- Use a different, easily recognizable IS-IS nickname on each switch.
- If IP Shortcuts is enabled, you must configure an IS-IS IP source address on the switch.

IST

If the switch will be part of an SMLT cluster, you must create the IST prior to enabling IS-IS.

SMLT

- Each switch in the cluster must be configured to peer with its neighbor. If the IS-IS interface between them is the IST MLT (recommended, but not strictly required), traffic paths will generally be similar to existing behavior.
- A virtual B-MAC is automatically created based on the lowest system ID in the cluster plus one.

  Important:

  The virtual B-MAC or any System ID created must not conflict with any other System ID or virtual B-MAC in the network. Verify that there is no duplication of System IDs or virtual B-MACs anywhere in the network .

  A safe practice is to leave the lowest byte in the system ID as all zeroes.
- There is a consistency check in place to ensure that L2 VSN VLANs cannot be added to the IST.

Physical or MLT links between IS-IS switches

Only a single port or single MLT is supported between a pair of IS-IS switches. For example, if there are two individual ports between a pair of IS-IS switches, IS-IS can only be configured on one of the two ports.

If a single MLT is configured between a pair of IS-IS switches, all ports (1-8) in the MLT are utilized. Note that you must configure the MLT before you enable IS-IS on the MLT.

CFM

- The CFM Domain name must be the same on all switches in an IS-IS area.
- The Maintenance Association must be the same on all switches in an IS-IS area.
  - To allow CFM testing over both B-VLANs, create two Maintenance Associations, one for each B-VLAN.
- The MIP can be configured the same on all switches in an IS-IS area or uniquely defined per switch.

Important:

The MIP must be configured at the same level as the MEP on all switches in the SPBM network.

Example of a configuration using best practices

```
- spbm-id               : 1
- BVID #1 & BVID #2     : 4040, 4041 (ignore warning message when configuring)
- nick-name             : b:b0:<node-id>
- MEP-id                : md.ma.<node-id>
- BMAC                  : 00:bb:00:00:<node-id>:00
- VirtBMAC              : 00:bb:00:00:<node-id>:ff
- MD                    : spbm (level 4)
- MA                    : 4040 & 4041
- mep                   : <node-id>
- mip                   : (level 4)
- isis manual area      : 49.0001
```

Migration best practices

Before you migrate your network to SPBM, perform an audit to determine if the desired configuration and traffic is supported by SPBM in the 7.1 release. If you determine that the migration is supported, follow the high level steps in the following sections to ensure a successful migration.

Supported traffic on SPBM networks

The following kinds of traffic are supported by SPBM in the 7.1 release:

- Layer-2 bridged traffic
- IPv4 unicast routed traffic on the Global Router
- IPv4 unicast routed traffic using a VRF
- IPv4 Unicast routed traffic using an IP VPN

The following traffic can only be bridged on an L2 VSN and not routed in an L3 VSN:

- IPv6 routed traffic (unicast or multicast)
- IPv4 multicast routed traffic
- IGMP snooping

Common Procedures and Exclusions on Migration

- Make sure that the SPBM infrastructure is enabled and functional in the network.
- If the VLAN has multicast enabled or has IPv6 routing configured, then it can only be migrated to an SPBM L2 VSN and routed with an external router.
- Identify the UNI and NNI ports that are currently port members of the VLAN on all the switches in the network. If STP/MSTP/RSTP is being used on the UNI, and the relevant domain touches more than one SPBM-enabled switch in the network (without having to go through any NNI ports), then the VLAN cannot be migrated to SPBM.

Migrating a VLAN to be an L2 VSN (C-VLAN)

The following procedure can be used to provide L2 connectivity for a VLAN across the SPBM core.

- Follow the pre-migration procedures checks described in the preceding section, “Common Procedures and Exclusions on Migration”.
- Identify the UNI and NNI ports that are currently port members of the VLAN on all the switches in the network.
- On all the switches in the network that are currently connected by the VLAN, remove the NNI ports from the membership list of the VLAN.

⚠ Warning:

This step will cause service interruption.

- Make the VLAN an L2 VSN using the **`config vlan vlanid i-sid <isidvalue>`** CLI command or the **`vlan i-sid vlanid isidvalue`** ACLI command. Use the same value of I-SID on all the switches. This step should restore service.

Migrating to Inter-ISID Routing

Inter-ISID routing provides the ability to route traffic between extended VLANs where the VLANs have different I-SIDs. All of the traditional IPv4 unicast routing and gateway redundancy protocols (including OSPF, RIP, VRRP, and RSMLT) are supported on top of any VLAN that is mapped to an I-SID.

The high level procedure to migrate a configuration to use Inter-ISID routing is described below.

- Follow the pre-migration procedures checks described in the preceding section, “Common Procedures and Exclusions on Migration”.
- For each VLAN in the SPBM core:
  - On all the switches where the VLAN is configured, remove all NNI ports

    ⚠ Warning:

    This step will cause service interruption.
  - On all the switches where the VLAN is configured, map the VLAN to an I-SID. This will restore L2 connectivity (the l2tracetree command can be used to validate L2 connectivity within the VLAN at this point). L3 will be restored once the routing protocols configured on top of the VLAN converge.
  - If an IGP (OSPF) is being used on the VLAN, the impact on traffic during the migration can be reduced by using two VLANs in each routed segment and configuring the interface cost to select the VLAN used for the routed next hop. Migrating the two VLANs one after the other will reduce the duration for which there is a loss of IGP adjacency.
  - Once all the VLANs identified for migration have been assigned an I-SID, the configuration part of the migration is completed. At this point all the traffic flows should be back to normal.

Restrictions and limitations

This section describes the restrictions and limitations associated with SPBM on the Avaya Ethernet Routing Switch 8800/8600.

STP/RSTP/MSTP

- There is no SPBM support in RSTP mode.
- A C-VLAN-level loop across SPBM NNI ports cannot be detected and needs to be resolved at the provisional level.
- SPBM NNI ports are not part of the L2 VSN C-VLAN, and BPDUs are not transmitted over the SPBM tunnel. SPBM can only guarantee loop-free topologies consisting of the NNI ports. Avaya recommends that you always use SLPP in any SMLT environment.

  Note:

  Avaya recommends deploying SLPP on C-VLANs to detect loops created by customers in their access networks. However, SLPP is not required on B-VLANs, and it is not supported. The B-VLAN active topology is controlled by IS-IS that has loop mitigation and prevention capabilities built into the protocol.
- SPBM uses STG 63 (with Avaya STP enabled) or MSTI 62 (with MSTP enabled) for internal use. So STG 63 or MSTI 62 cannot be used by other VLAN/MSTI.

SBPM IS-IS

- The current release supports IP over IS-IS using IP Shortcuts. However, the current release does not support RFC 1195.
- SPBM only uses level 1 IS-IS. Level 2 IS-IS is not support in this release.
- The IS-IS standard defines wide (32bit) metrics and narrow (8 bits) metrics. Only the wide metric is supported in this release.
- SPBM supports full High Availability (HA). The SPBM and IS-IS configuration and dynamic information (such as adjacencies and LSPs) are all HA synced to the standby CPU to ensure seamless switchover. Since the Ethernet Routing Switch 8800/8600 HA framework cannot guarantee seamless switchover, there is a 6 to 7 seconds gap between the active CPU going down and the standby CPU coming up. To avoid IS-IS adjacencies bouncing during the switchover, the recommended hello interval is 9 seconds and the hello multiple is 3.

  Important:

  If the switch has a large number of MACs in the VLAN FDB entry table and the primary SF/CPU fails, it can take longer than 27 seconds for the secondary SF/CPU to become the new primary. In this scenario, the IS-IS adjacency goes down because the default hold time (27 seconds) is too short. To prevent this from happening, increase the Hello multiplier to allow more time for the HA failover.

SBPM NNI SMLT

For NNI-facing MLT splitting towards two IST peers, only one link is supported in each direction towards each of the IST switches.

Multicast

In this release, SPBM does not support enabling PIM or IGMP snooping on a C-VLAN.

VLACP

VLACP is generally used when a repeater or switch exists between connected Ethernet Routing Switch 8800/8600 switches to detect when a connection is down even when the link

light is up. If you have VLACP configured on an SPBM link that is also an IST link, during a connection fail over (where the link lights stay up) the IS-IS hellos time out first (after 27 seconds, using default values) and take down the IS-IS adjacency. IS-IS then calculates the new shortest path and fails over the SPBM traffic. Then, 90 seconds after the connection failure (using default values), VLACP goes down but the IST link was already taken down by IS-IS.

In this scenario, there is no data traffic impact since IS-IS can find another path in the SPBM network before VLACP goes down.

SNMP traps

On each SPBM peer, if the SPBM B-VLANs are configured using different VLAN IDs (for example, VLAN 10 and 20 on one switch and VLAN 30 and 40 on the second), no trap message is generated alerting of the mismatch because the two switches cannot receive control packets from one another. Be sure to configure the SPBM B-VLANs using matching VLAN IDs.

Others

- C-VLAN and B-VLAN cannot be enabled on the same port.
- Filters are not supported on the B-VLAN NNI port.
- Filters are not supported on the C-VLAN port for egress traffic.

Legacy IS-IS

An SPBM node can form an adjacency with a legacy IS-IS router. This allows SPBM to be introduced into existing networks and provide for easy migration.

Chapter 12: 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 *Avaya Ethernet Routing Switch 8800/8600 Configuration — IP Multicast Routing Protocols, NN46205-501*.

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 Avaya Ethernet Routing Switch 8800/8600 must be equipped with the following:

- Release 5.1 (or later) software
- Premier Software License
- R/RS/8800 modules
- 8692 SF/CPU with SuperMezz CPU-Daughter card or 8895 SF/CPU

Multicast virtualization network scenarios

The following figure shows an example of multicast virtualization in an RSMLT topology.

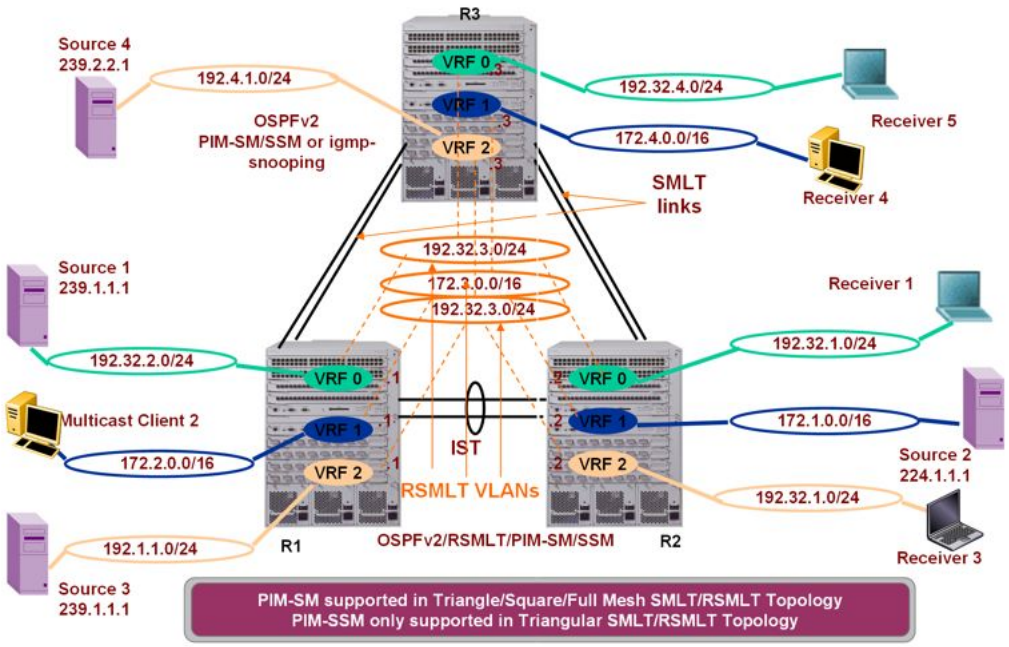


Figure 77: Multicast virtualization in RSMLT topology

The following figure shows an example of multicast virtualization in an Enterprise/Metro network.

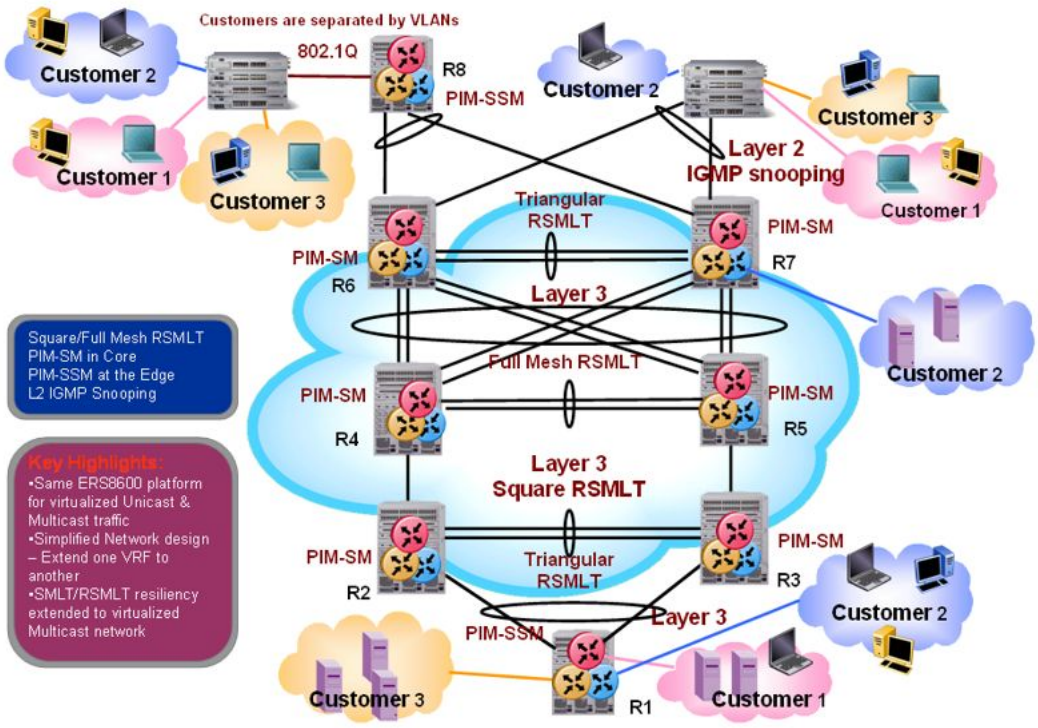


Figure 78: Multicast virtualization for Enterprise/Metro network

The following figure shows an example of multicast virtualization supporting an end-to-end triple play solution for an MSO/Large Enterprise.

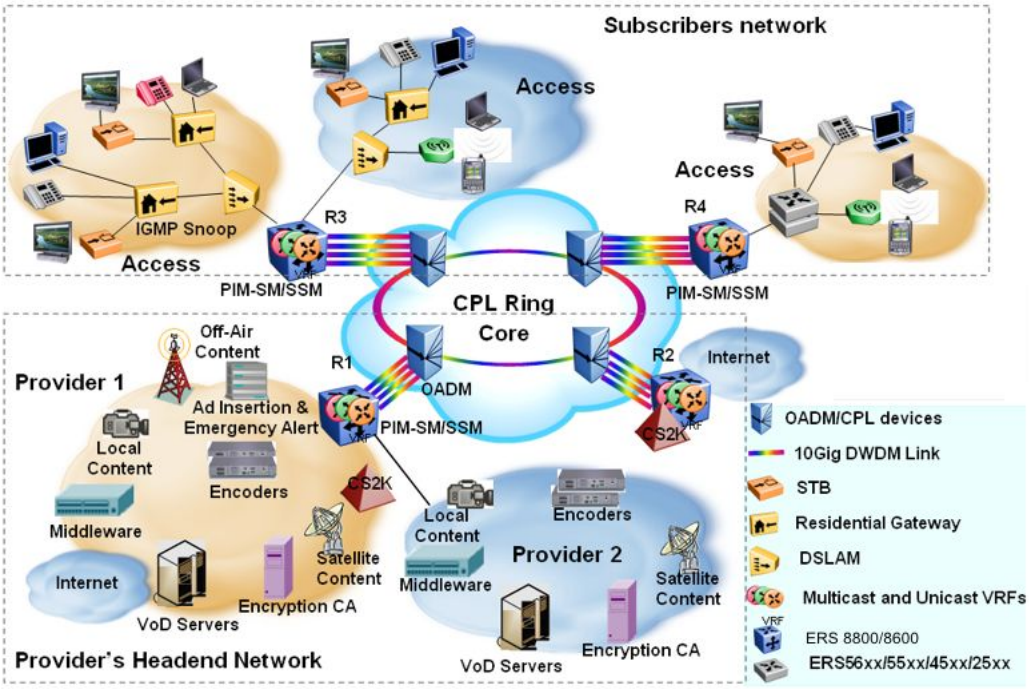


Figure 79: End-to-end triple play solution for MSO/Large Enterprise

The following figure shows an example of multicast virtualization in a data center.

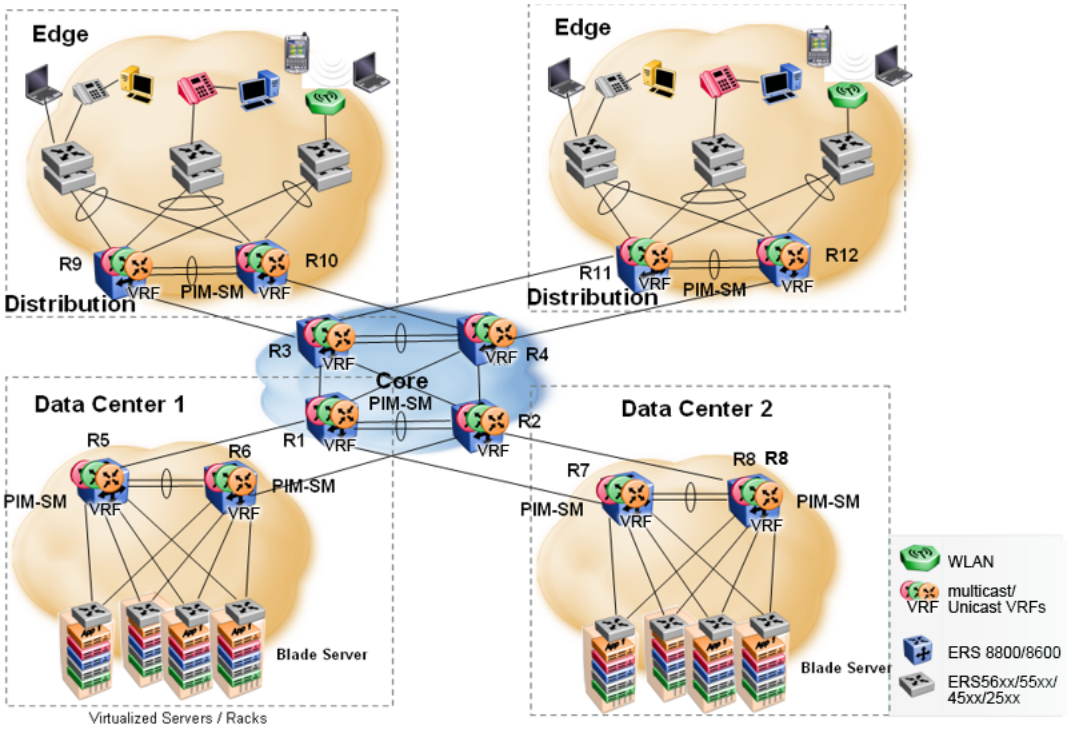


Figure 80: Data center solution with multicast virtualization

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 Avaya Ethernet Routing Switch 8800/8600 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 on page 188
- Multicast flow distribution over MLT on page 189

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.

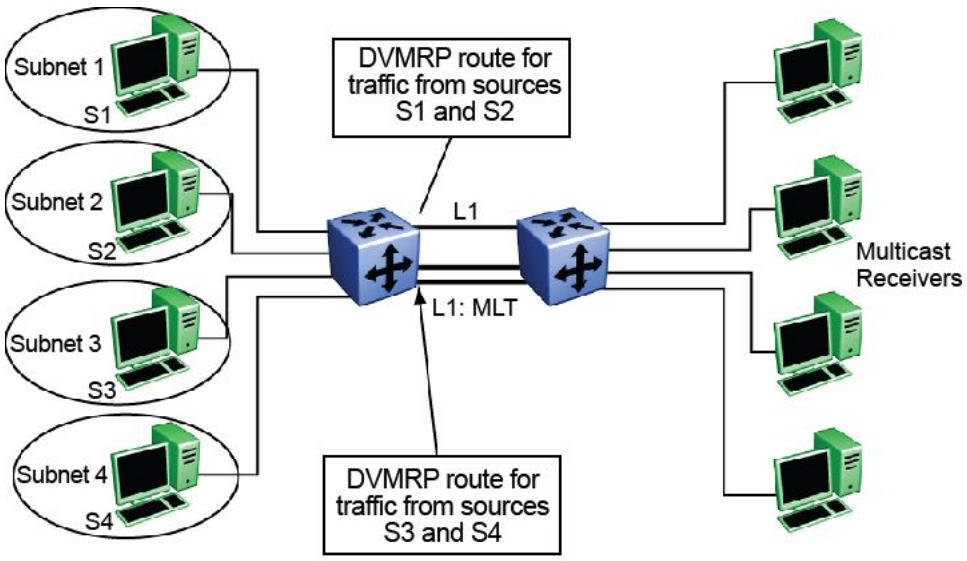


Figure 81: Traffic distribution for multicast data

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.

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.

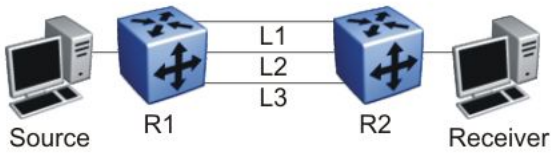


Figure 82: Multicast flow distribution over MLT

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, the activity check is set to 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:

- Whenever possible, use simple network designs that do not use VLANs that span several switches. Instead, use routed links to connect switches.
- Whenever possible, group sources should send to the same group in the same subnet. The Avaya Ethernet Routing Switch 8800/8600 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 ACLI, the command is **`show ip mroute hw-group-trace`**.
- 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.
- 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.
- 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 can 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.

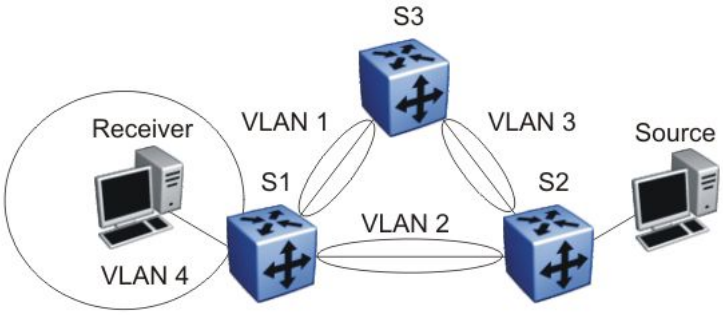


Figure 83: IP multicast sources and receivers on interconnected VLANs

- 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 on page 203.
- For faster convergence, configure the Bootstrap and Rendezvous Point routers on a circuitless IP. For more information, see Circuitless IP for PIM-SM on page 217.
- For faster convergence, Avaya 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. (For more information, 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 must 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 can use announce and accept policies as described in Announce and accept policy examples on page 203.

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.

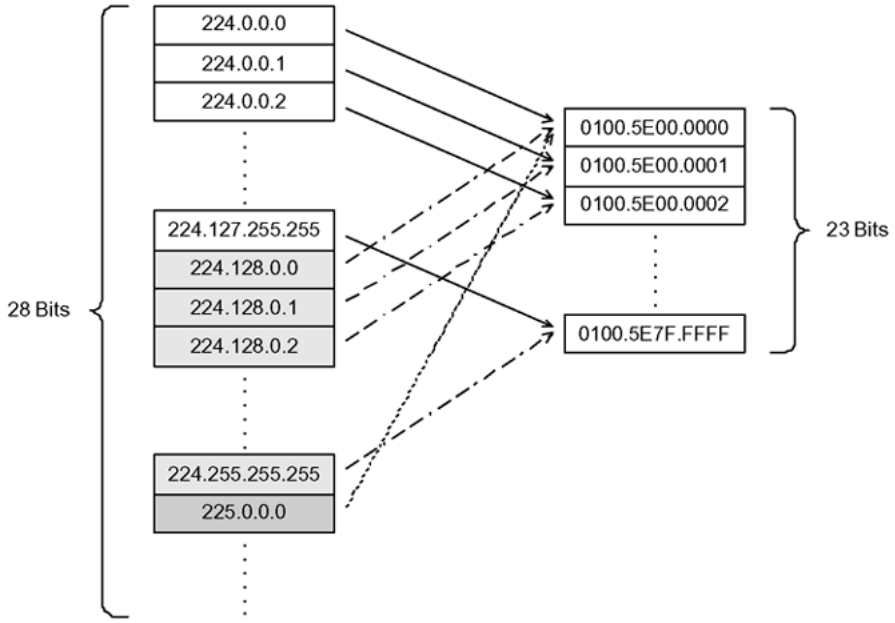


Figure 84: Multicast IP address to MAC address mapping

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 Avaya Ethernet Routing Switch 8800/8600 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 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, Avaya 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 Avaya Ethernet Routing Switch 8800/8600 does not suffer from the problem described in the previous example, other switches in the network, particularly pure Layer 2 switches, can.

In a network that includes non-Ethernet Routing Switch 8800/8600 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

Avaya 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, Avaya 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 Avaya Ethernet Routing Switch8800/ 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 Avaya Ethernet Routing Switch 8800/8600 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. Avaya 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 8800/8600 (Switch 1) to drop multicast traffic with a TTL of less than 2 (see Figure 85: IP multicast traffic with low TTL on page 195). In this configuration, all streams that egress the switch (Switch 1) with a TTL of 1 are dropped.

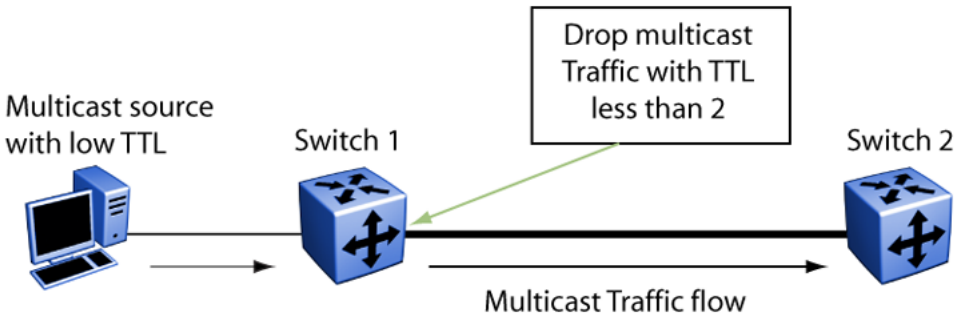


Figure 85: IP multicast traffic with low TTL

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 8800/8600 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 Avaya Ethernet Routing Switch 8800/8600, 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.

Avaya 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 86: Applying IP multicast access policies for DVMRP on page 197 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.

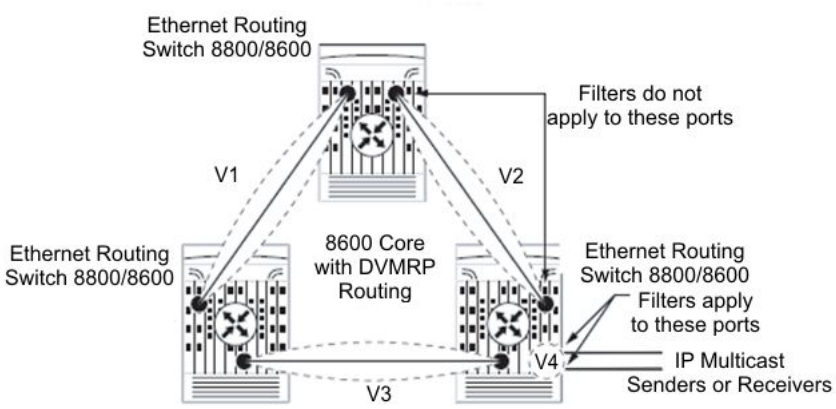


Figure 86: Applying IP multicast access policies for DVMRP

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.

Layer 2 multicast features

On Layer 2 VLANs, the Avaya Ethernet Routing Switch 8800/8600 provides the following features to support multicast traffic.

- IGMP snoop and proxy on page 198
- Multicast VLAN Registration (MVR) on page 199
- IGMP Layer 2 querier on page 199

IGMP snoop and proxy

On a Layer 2 VLAN, if at least one host on the VLAN specifies that it is a member of a multicast group, by default, the Avaya Ethernet Routing Switch 8800/8600 forwards to that VLAN all datagrams bearing the multicast address of that group. All ports on the VLAN receive the traffic for that group.

With IGMP snoop enabled on a VLAN, the Avaya Ethernet Routing Switch 8800/8600 forwards the multicast group data to only those ports that are members of the multicast group.

The switch identifies multicast group members by listening to IGMP packets (IGMP reports, leaves, and queries) from each port. Using the information gathered from the reports, the switch builds a list of group members. After the group members are identified, the switch blocks the IP Multicast stream from exiting any port that does not connect to a group member, conserving bandwidth.

With IGMP snoop enabled, the switch can receive multiple reports for the same multicast group. Rather than forward each report upstream, the Ethernet Routing Switch 8800/8600 can consolidate these multiple reports using the IGMP proxy feature. With IGMP proxy enabled, if the switch receives multiple reports for the same multicast group, it does not transmit each report to the upstream multicast router. Instead, the switch forwards the first report to the querier and suppresses the rest. If new information emerges, for example if the switch adds another multicast group or receives a query since the last report is transmitted upstream, then the switch forwards a new report to the multicast router ports.

For more information about IGMP snoop and proxy, see Avaya Ethernet Routing Switch 8800/8600 Configuration — IP Multicast Routing Protocols (NN46205-501).

Multicast VLAN Registration (MVR)

On Layer 2 VLANs, the Avaya Ethernet Routing Switch 8800/8600 uses IGMP Snoop to listen for report, leave and query packets, and then creates or deletes multicast groups for receiver ports to receive multicast data streams. In IGMP Snoop, all the ports, including receiver and source ports, are members of the same VLAN. When users in different VLANs join the same group, the multicast router replicates one stream into multiple streams that are sent to these VLANs. Multiple streams waste bandwidth and decrease the performance of the multicast router.

The Multicast VLAN Registration (MVR) Protocol solves this problem. With MVR, the receiver ports remain in the IGMP Snoop VLAN, but one VLAN is designated as the MVR VLAN. Each VRF on the Ethernet Routing Switch 8800/8600 supports only one MVR VLAN.

The MVR VLAN has a source port, which connects to the multicast router. After you bind several IGMP Snoop VLANs to the MVR VLAN, and a multicast data packet arrives from the source port, the switch replicates this packet and forwards it to all the IGMP Snoop VLANs that are bound to the MVR VLAN.

MVR is based on IGMP Snoop, but the two features work independently. The MVR VLAN controls only the VLANs that are bound to it; other IGMP Snoop VLANs operate as usual.

After you enable MVR globally, all IGMP control packages that are received from IGMP Snoop VLANs that are bound to the MVR VLAN (including report, leave, and query) are processed by MVR.

If an IGMPv3 snoop interface is used for MVR, both the MVR VLAN and the IGMP Snoop VLANs must be set to version 3.

For more information about MVR, see Avaya Ethernet Routing Switch 8800/8600 Configuration — IP Multicast Routing Protocols (NN46205-501).

IGMP Layer 2 querier

In a multicast network, if the multicast traffic only needs to be Layer 2 switched, no multicast routing is required. However, for multicast traffic to flow from sources to receivers, an IGMP querier must exist on the network, a function that is normally provided by a multicast router.

To provide a querier on a Layer 2 network without a multicast router, you can use IGMP Layer 2 querier.

The IGMP Layer 2 querier provides the querier functions of a multicast router on the Layer 2 multicast network. The Layer 2 querier forwards queries for multicast traffic, and processes the responses accordingly. On the connected Layer 2 VLANs, IGMP snoop continues to

provide services as normal, responding to queries and identifying receivers for the multicast traffic.

To enable Layer 2 querier, you must configure an IP address for the querier, in order for it to receive forwarded report and leave messages.

If a multicast router is present on the network, the Layer 2 querier is automatically disabled.

In the Layer 2 multicast network, enable Layer 2 querier on one of the switches in the VLAN. Only one Layer 2 querier is supported in the same Layer 2 multicast domain. No querier election is available.

You cannot enable MVR and Layer 2 querier on the same VLAN.

For more information about IGMP Layer 2 querier, see Avaya Ethernet Routing Switch 8800/8600 Configuration — IP Multicast Routing Protocols (NN46205-501).

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 Avaya Ethernet Routing Switch 8800/8600 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 *Avaya Ethernet Routing Switch 8800/8600 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, Avaya 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 more information about using policies, see DVMRP policies on page 203.

The recommended maximum number of active multicast source/group pairs (S,G) is 2000.

Avaya recommends that the number of source subnets multiplied by 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 on page 188.

Important:

In some DVMRP scaled configurations with more than one thousand streams, to avoid multicast traffic loss, you need to increase routing protocol timeouts (for example, the dead interval for OSPF).

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 86: Applying IP multicast access policies for DVMRP on page 197, 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 Avaya Ethernet Routing Switch 8800/8600 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, Avaya 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) with respect to 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.

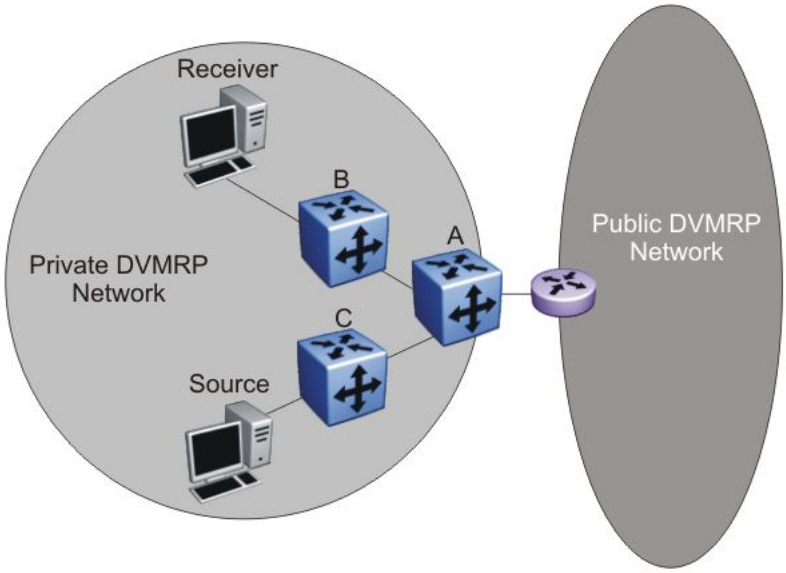


Figure 87: Announce policy on a border router

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.

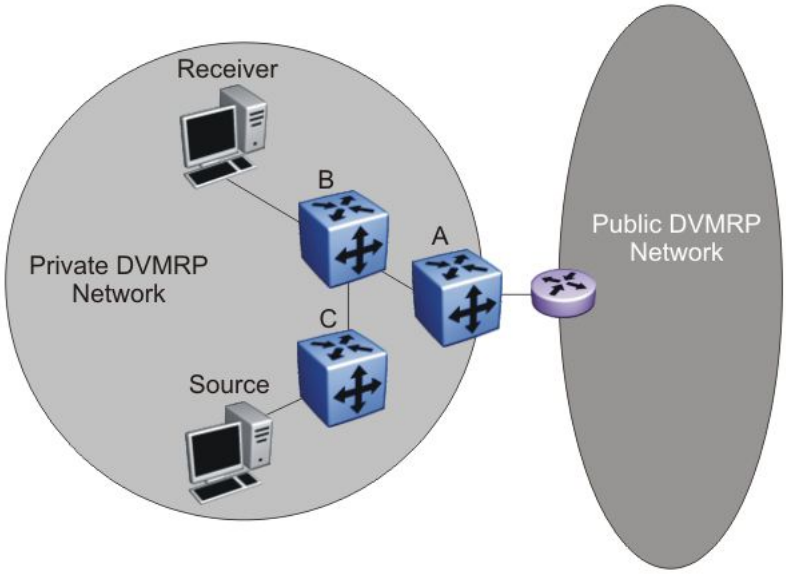


Figure 88: Accept policy on a border router

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.

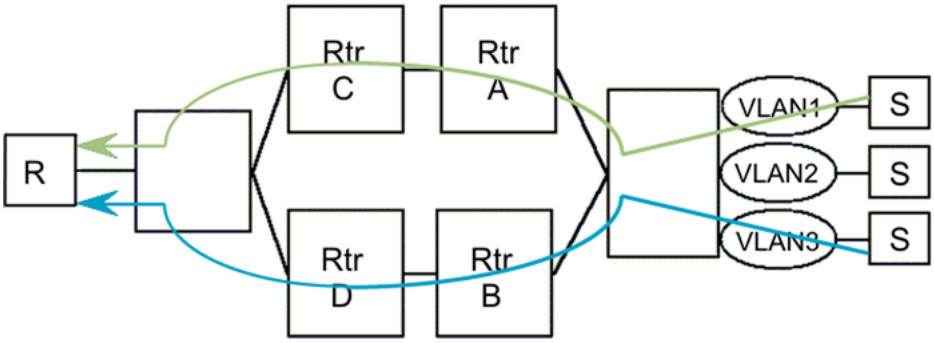


Figure 89: Load balancing with announce policies

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.

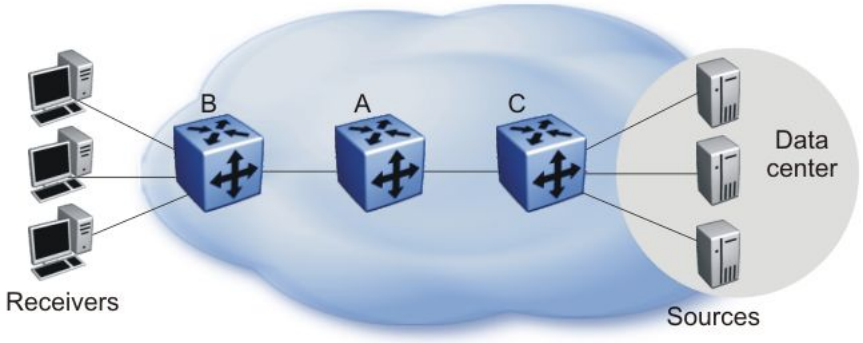


Figure 90: Do not advertise local route policies

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. When 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.

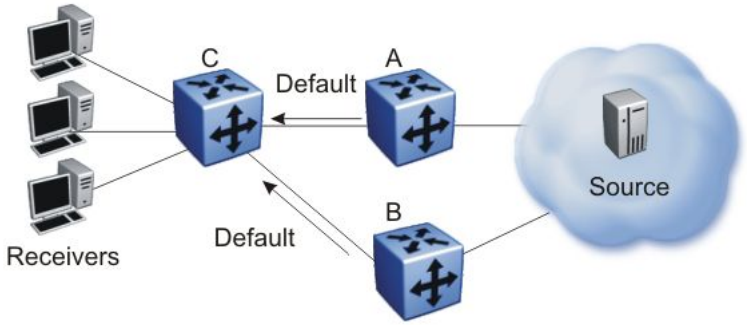


Figure 91: Default route

Avaya 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, Avaya 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 receivers and VLANs on page 223
- PIM network with non-PIM interfaces on page 224

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.

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 are not 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. Avaya 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 information, see DVMRP passive interfaces on page 207.

Important:

Avaya 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, Avaya 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 non-PIM interfaces on page 224.

Avaya 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.

PIM general requirements

Avaya 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.

Avaya 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.

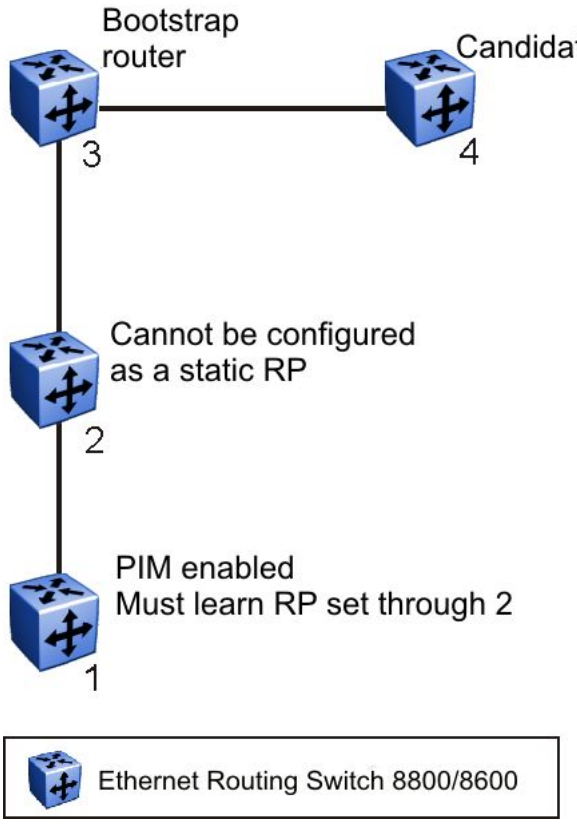


Figure 92: Example 1

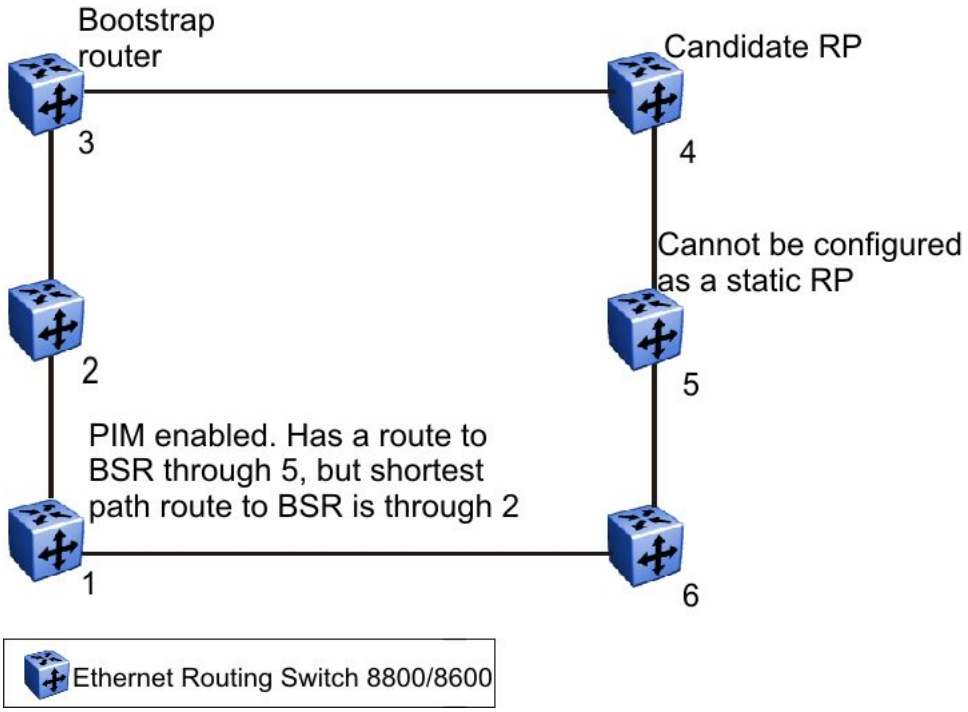


Figure 93: Example 2

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 Avaya Ethernet Routing Switch 8800/8600 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 workaroundis to configure static RP on the peer SMLT switches. This avoids theprocess of selecting an active RP router from the list of candidateRPs, and also of dynamically learning about RPs through the BSR mechanism.Then, when the Designated Router comes back, traffic resumes as soonas OSPF converges. This workaround reduces the traffic delay.

PIM-SM to DVMRP connection: MBR

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.

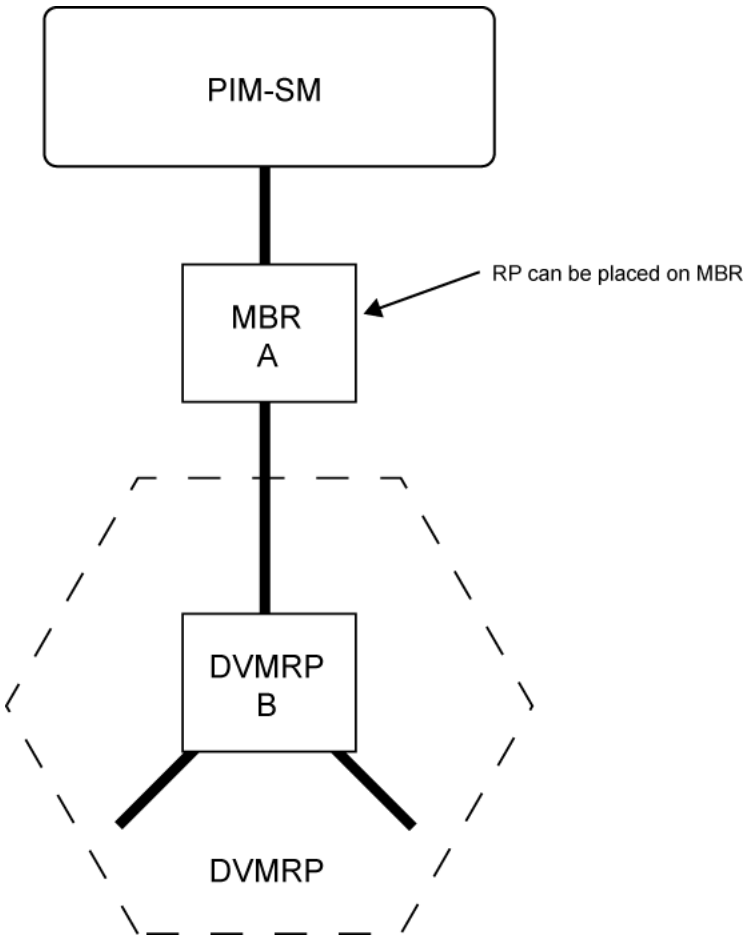


Figure 94: MBR configuration

With the Avaya Ethernet Routing Switch 8800/8600 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.

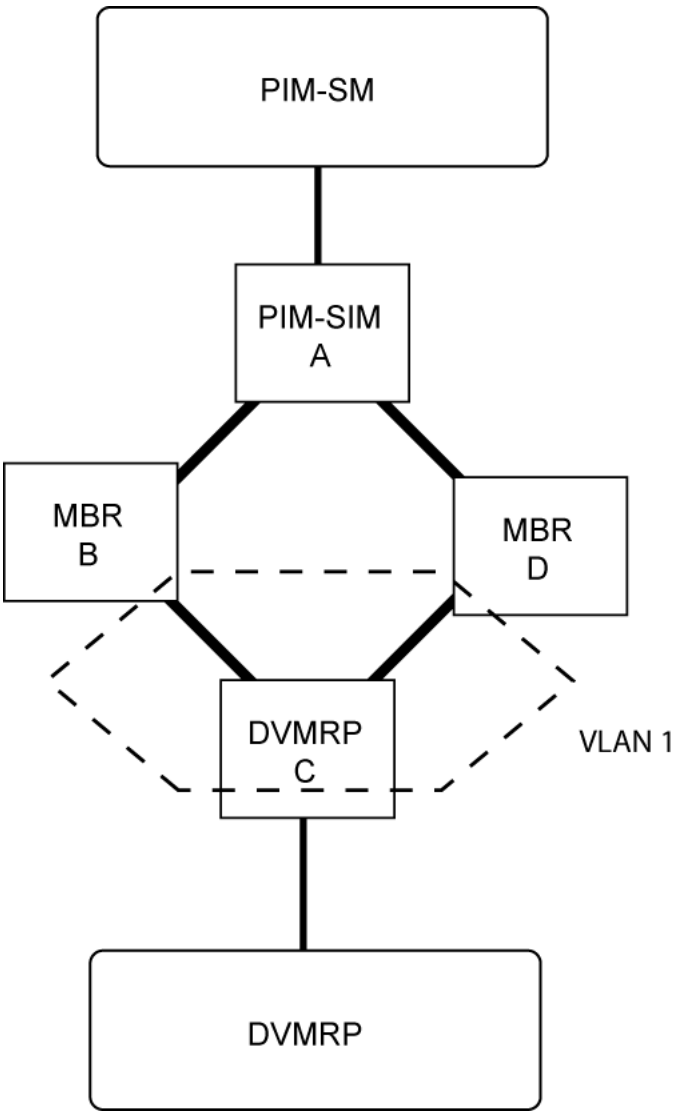


Figure 95: Redundant MBR configuration

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.

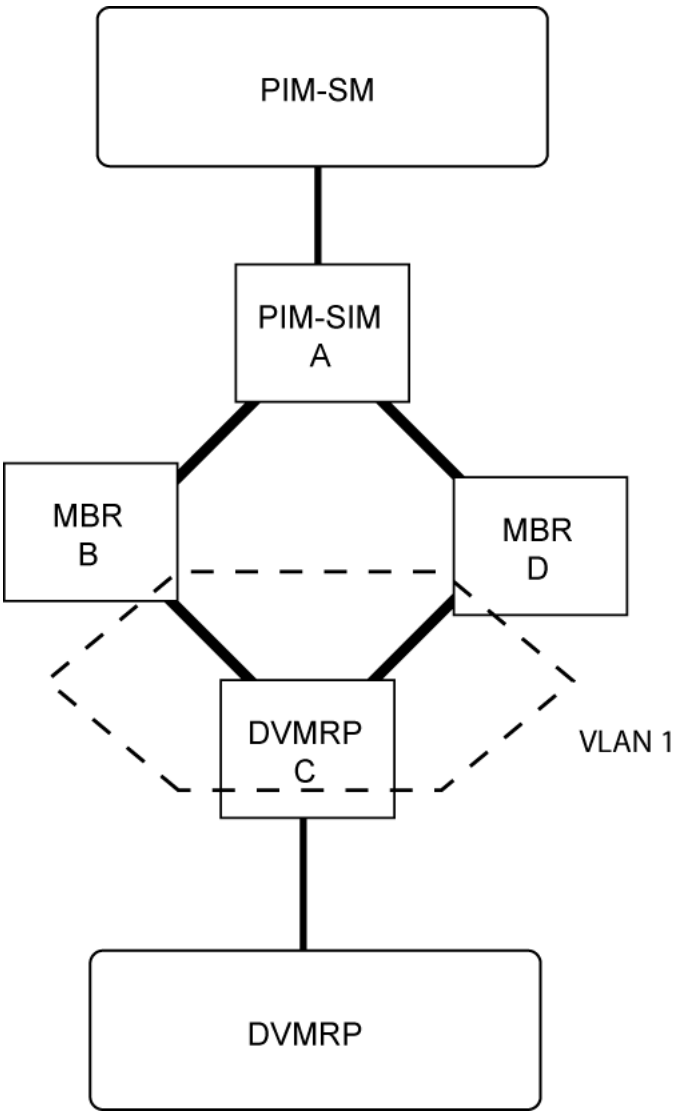


Figure 96: Redundant MBR configuration with two separate VLANs

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, Avaya recommends that you use the simple design proposed in PIM-SM to DVMRP connection: MBR.

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.

Avaya 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 Avaya Ethernet Routing Switch 8800/8600 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 Avaya Ethernet Routing Switch 8800/8600 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 8800/8600 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.

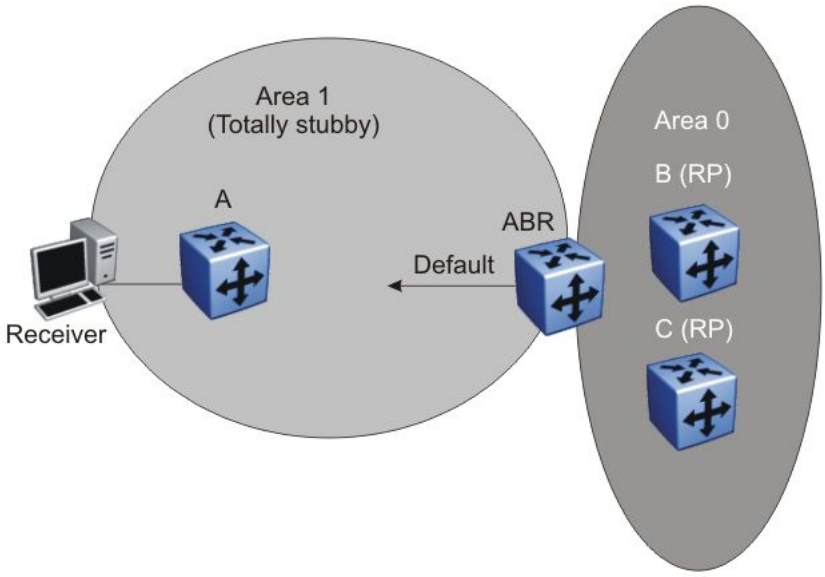


Figure 97: RP failover with default unicast routes

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, Avaya 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.

Specific route for static RP

With static RP enabled, the Avaya Ethernet Routing Switch 8800/8600 detects RP failure based on IGP convergence and, more specifically, on the removal of the route to the RP from the routing table. With the route to the failed RP removed, the Avaya Ethernet Routing Switch 8800/8600 can fail over to an alternate static RP.

If a default route is injected into the routing table, that default route still appears as an active route to the failed RP. Therefore, in this case, the switch does not fail over to the alternate RP.

A similar situation exists with SMLT-based configurations, where an internal-only default static route is used during IST failover and recovery. In this case, the internal default route appears as an active route to the failed RP, and therefore does not failover to the alternate RP.

To resolve the preceding situations, you can configure the lookup for static RP to be chosen from the specific route rather than the best route. In this case, when the route to the active RP fails, the switch no longer interprets the default route as a valid route for RP purposes, and therefore fails over to the alternate RP.

Avaya recommends that you always enable the specific route option for any SMLT/RSMLT cluster running PIM-SM with static RPs because of the implementation of the internal-only default static route on the IST.

This resolution applies only to static RP configurations, not to C-RP configurations.

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.

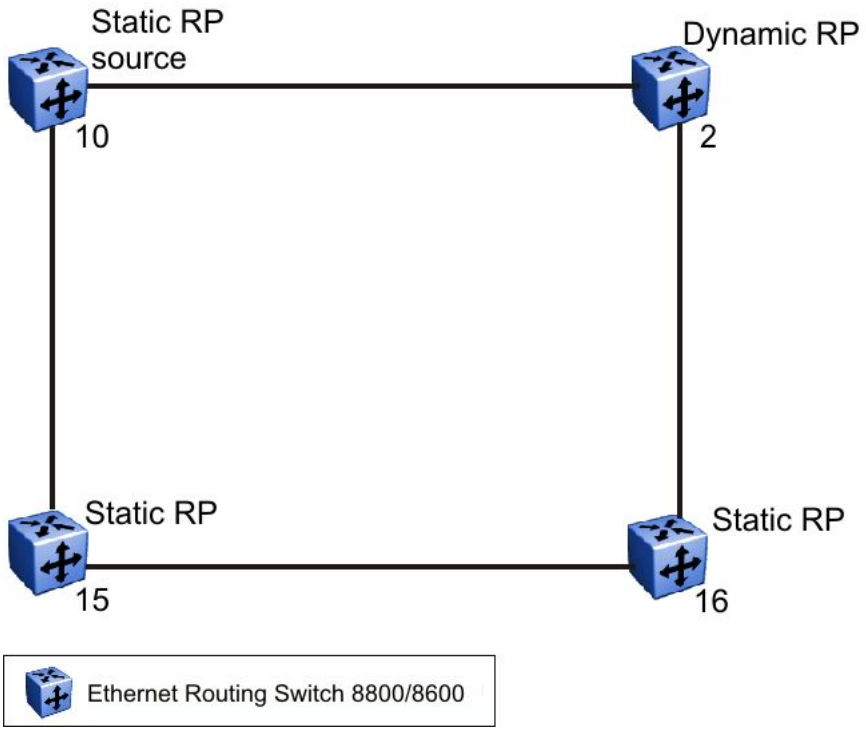


Figure 98: Nonsupported static RP configuration

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, Avaya 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.

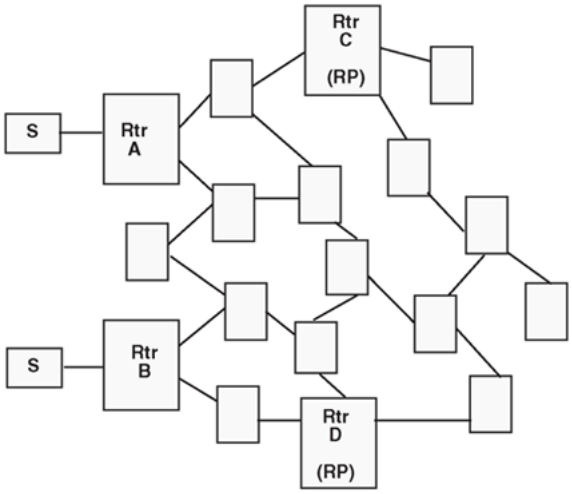


Figure 99: Example multicast network

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.

- 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.

- 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.

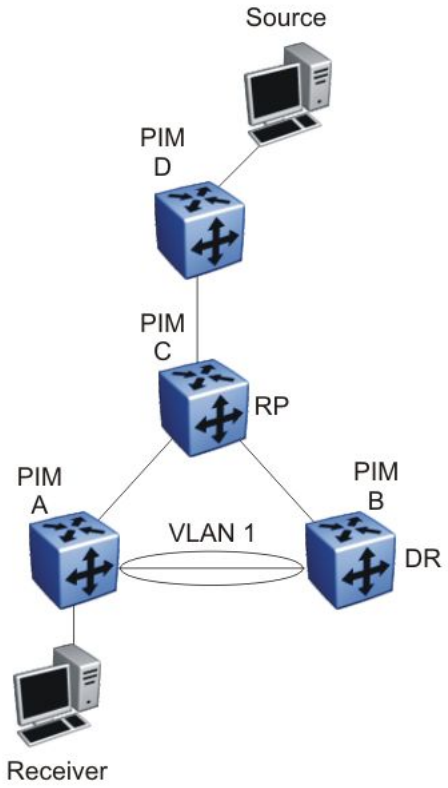


Figure 100: Receivers on interconnected VLANs

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, Avaya 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 non-PIM 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 101: PIM network with non-PIM interfaces on page 225 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.

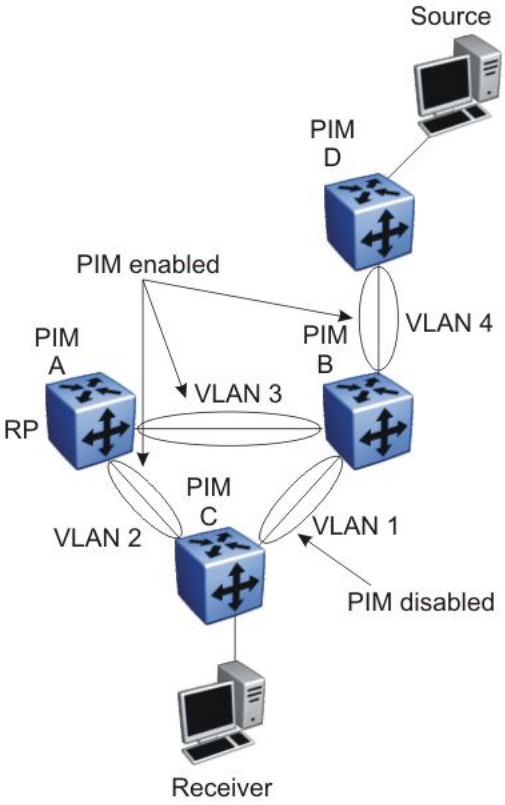


Figure 101: PIM network with non-PIM interfaces

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 SSM-only implementation is not a full IGMPv3 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 can 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.

RP Set configuration considerations

When you configure RP sets (C-RPs or static RPs), Avaya recommends as best practice not to configure multiple entries that each specify a unique group, but instead specify a range of groups when possible, thereby decreasing the number of entries required.

PIM-SSM design considerations

Consider 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 more information about PIM-SSM scaling, see PIM-SM and PIM-SSM scalability on page 209.

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.

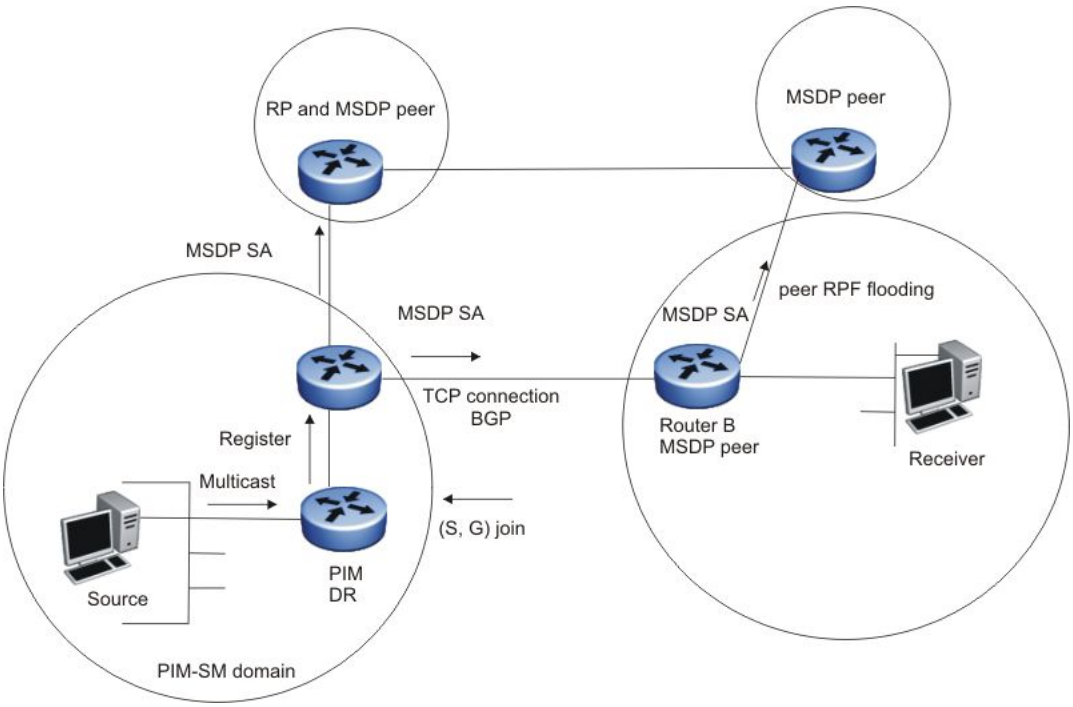


Figure 102: MSDP operation between peers

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

Avaya 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.

Avaya Ethernet Routing Switch 8800/8600 supports the MSDP management information base (MIB) as described in RFC 4624.

Static mroute

The Avaya Ethernet Routing Switch 8800/8600 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.

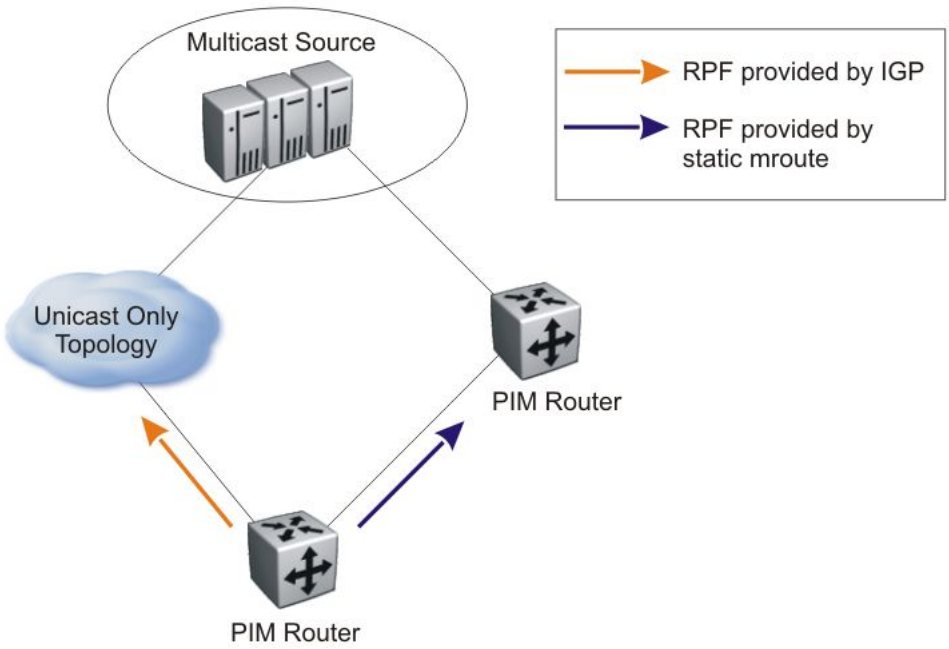


Figure 103: Static mroute

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

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 101: PIM network with non-PIM interfaces on page 225, 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

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.

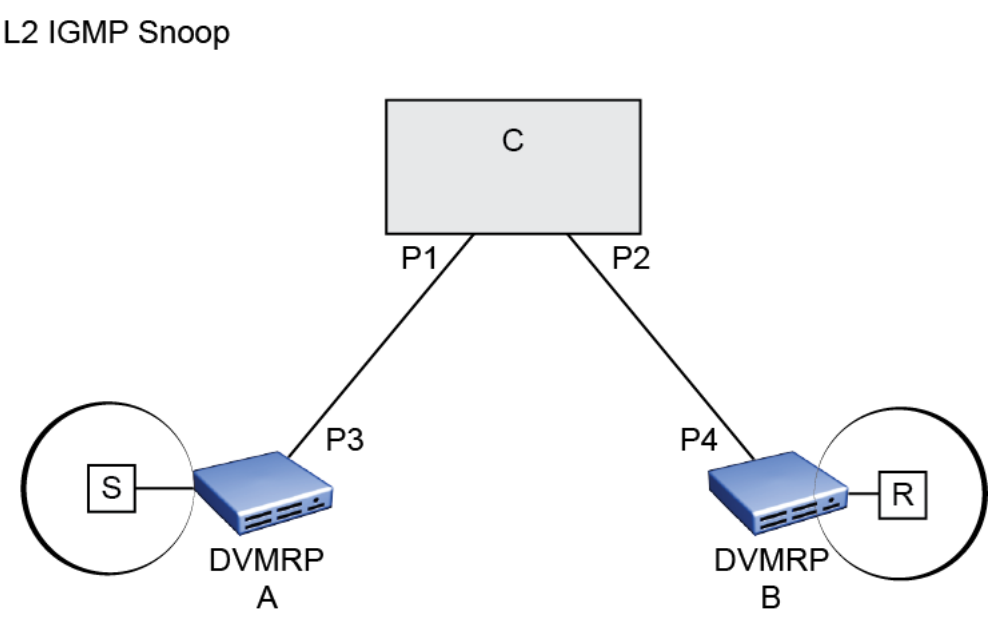


Figure 104: IGMP interaction with DVMRP

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

The following sections provide configuration guidelines for multicast SMLT networks.

For more information about SMLT topologies, see SMLT topologies on page 90 or *Avaya Ethernet Routing Switch 8800/8600 Configuration — Link Aggregation, MLT, and SMLT, NN46205-518*.

Multicast and SMLT guidelines navigation

- PIM-SSM over SMLT/RSMLT on page 239
- Static-RP in SMLT using the same CLIP address on page 244

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:

- a configuration with Layer 3 PIM-SM routing on both the edge and aggregation switches
- a configuration with Layer 2 snooping on the client switches and Layer 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, Avaya 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.

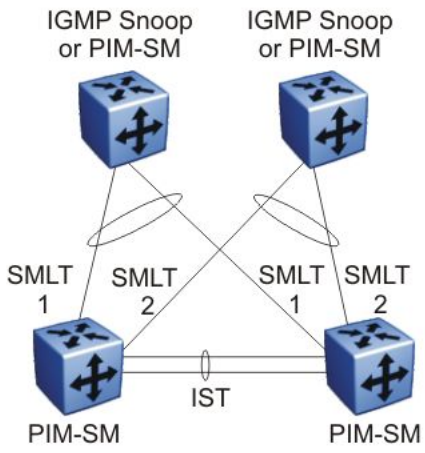


Figure 105: Multicast routing using PIM-SM

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 Avaya Ethernet Routing Switch 8800/8600 does not support DVMRP in SMLT full-mesh designs.

In a square design, you must configure all switches with PIM-SM. Avaya 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 protocolsare used in an SMLT environment, the protocol should be enabled onthe IST. Although, in general, routing protocols should not run overan IST, multicast routing protocols are an exception.

In a single PIM domain with an MBR (Multicast Border Router), Avaya 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 Avaya Ethernet Routing Switch 8800/8600s have OSPF enabled, and PIM is enabled and active on VLAN 101.
- Both Avaya Ethernet Routing Switch 8800/8600s have OSPF enabled, and PIM is either disabled or passive on VLAN 102.

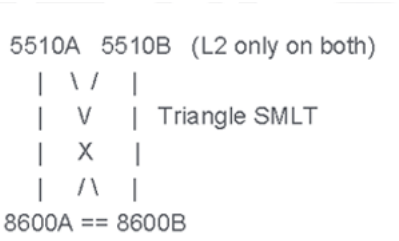


Figure 106: Unicast route example

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.

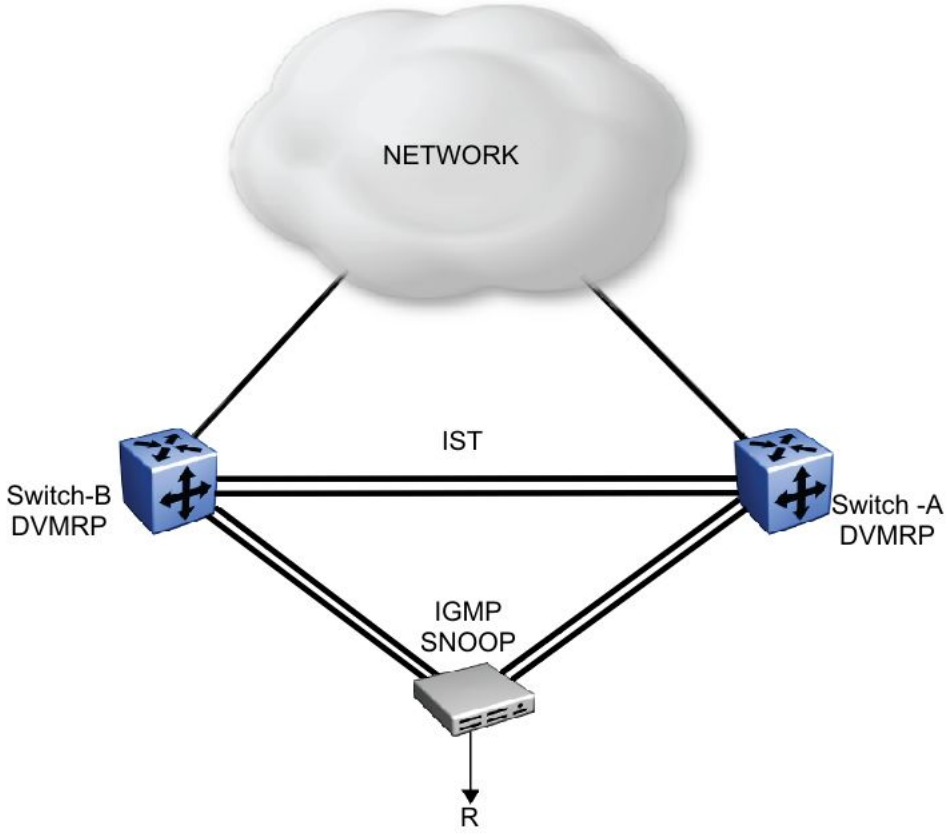


Figure 107: Multicast and SMLT design that avoids duplicate traffic

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.

- A has a DVMRP route 10.10.10.0 with a metric of 10 and an upstream neighbor through the interface connecting to NETWORK.
- B has a DVMRP route 10.10.10.0 with a metric of 10 and an upstream 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.
- S sends data that is received by both A and B through the interface 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.

PIM-SSM over SMLT/RSMLT

Fast failover for multicast traffic in a PIM-SSM network can be achieved using SMLT/RSMLT. PIM-SSM is supported in triangle, square, and full mesh SMLT/RSMLT topologies.

The following figures show some examples of the supported topologies of PIM-SSM over SMLT.

The following figure shows a triangle topology in which all the Ethernet Routing Switch 8800/8600s are running PIM-SSM at the core, and the Ethernet Routing Switch 8300 and the stackable Ethernet Routing Switches (5xxx/4500/2500) are also running PIM-SSM at the edge.

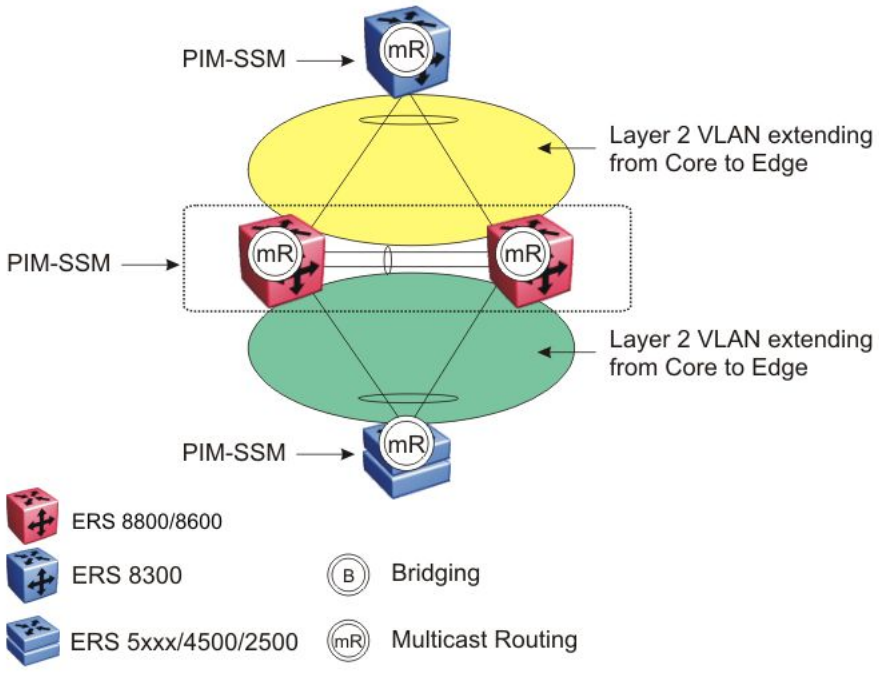


Figure 108: Triangle topology with PIM-SSM in the core and at the edge

The following figure shows a similar triangle topology, with the Ethernet Routing Switch 8300 and the stackable Ethernet Routing Switches (5xxx/4500/2500) running PIM-SSM at the edge. In this case, however, the Ethernet Routing Switch 8800/8600s are running PIM-SM in the core.

With the extended VLANs from the SSM edge to the SM core, the operating version of the interfaces in the core must be IGMPv2. Hence the querier for that VLAN sends out IGMPv2 queries. If there are receivers in the same extended VLAN from the edge to the core, they must only send IGMPv1/v2 reports because IGMPV3 reports are dropped by IGMPv2 interfaces.

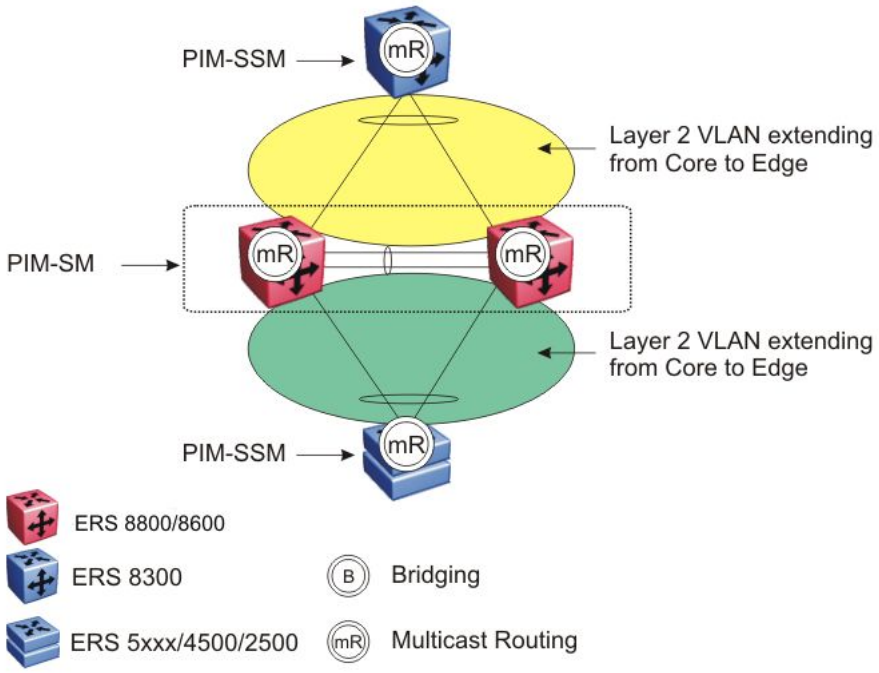


Figure 109: Triangle topology with PIM-SM in the core and PIM-SSM at the edge

The following figure shows a square or full mesh topology in which one Ethernet Routing Switch 8800/8600 IST pair is running PIM-SSM in the core, and the other IST pair is running Layer 2 IGMP. The Ethernet Routing Switch 8300 and the stackable Ethernet Routing Switches (5xxx/4500/2500) are also running Layer 2 IGMP at the edge.

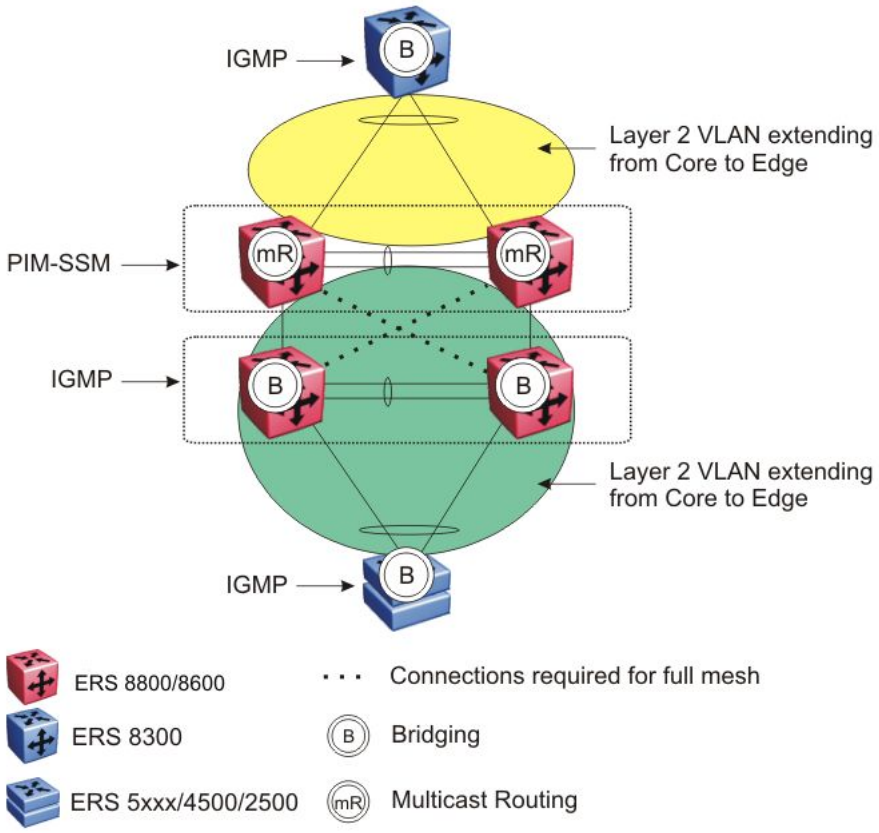


Figure 110: Square/full mesh topology with Layer 2 IGMP edge and PIM-SSM core

The following figure shows a square or full mesh topology in which both Ethernet Routing Switch 8800/8600 IST pairs are running PIM-SSM and RSMLT in the core. The Ethernet Routing Switch 8300 and the stackable Ethernet Routing Switches (5xxx/4500/2500) are running Layer 2 IGMP at the edge.

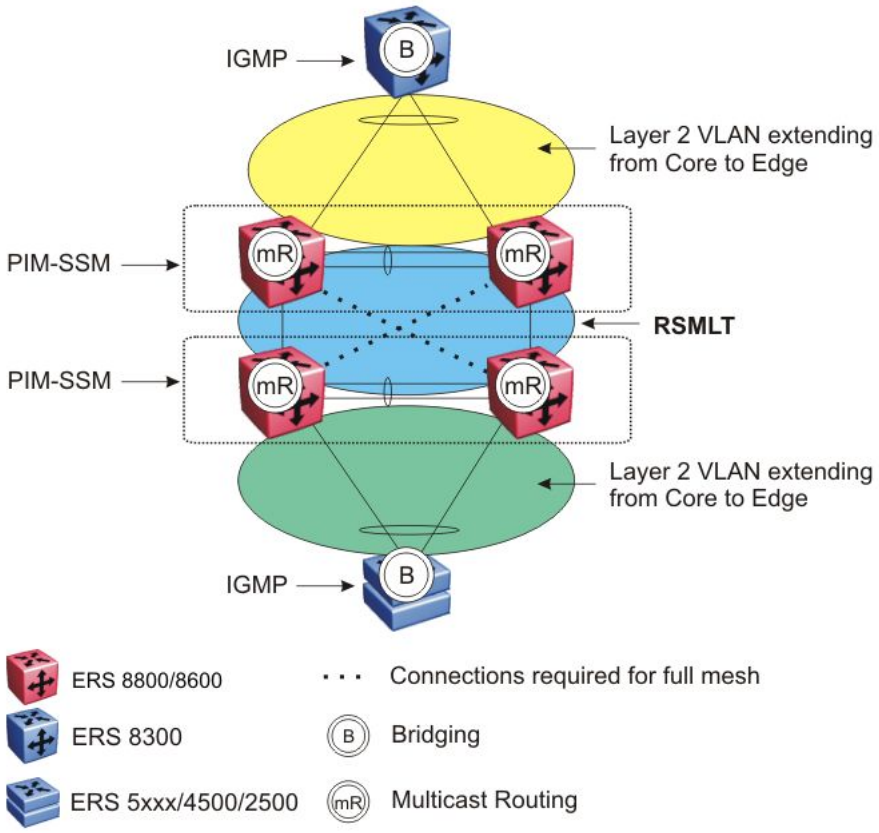


Figure 111: Square/full mesh topology with Layer 2 IGMP edge and PIM-SSM core with RSMLT

The following figure shows a square or full mesh topology in which both Ethernet Routing Switch 8800/8600 IST pairs are running PIM-SSM and RSMLT in the core. The Ethernet Routing Switch 8300 and the stackable Ethernet Routing Switches (5xxx/4500/2500) are running PIM-SSM at the edge.

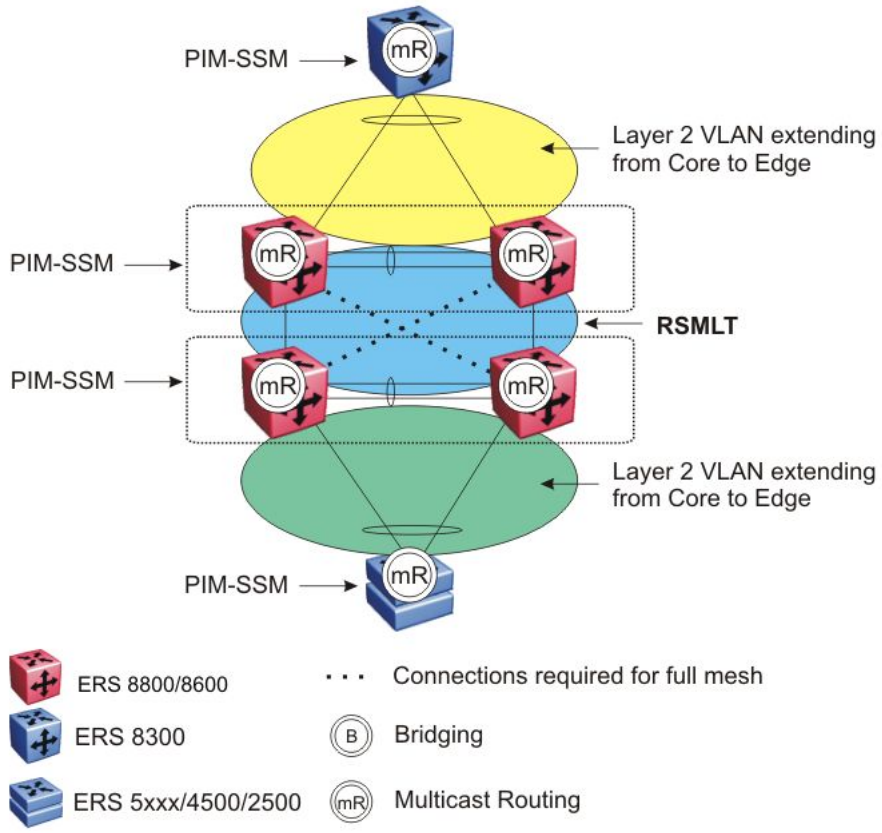


Figure 112: Square/full mesh topology with PIM-SSM edge and PIM-SSM core with RSMLT

Static-RP in SMLT using the same CLIP address

In a normal PIM SMLT network, in the event of a failed or unreachable RP, all (S,G) entries are deleted from the network because of the unreachable RP.

To provide faster failover in a switch cluster, you can configure each switch in an IST pair as a static RP using the same CLIP address, as shown in the following figure. In this case, (S,G) entries are not flushed out in the event of a failed RP, and therefore this configuration provides faster failover.

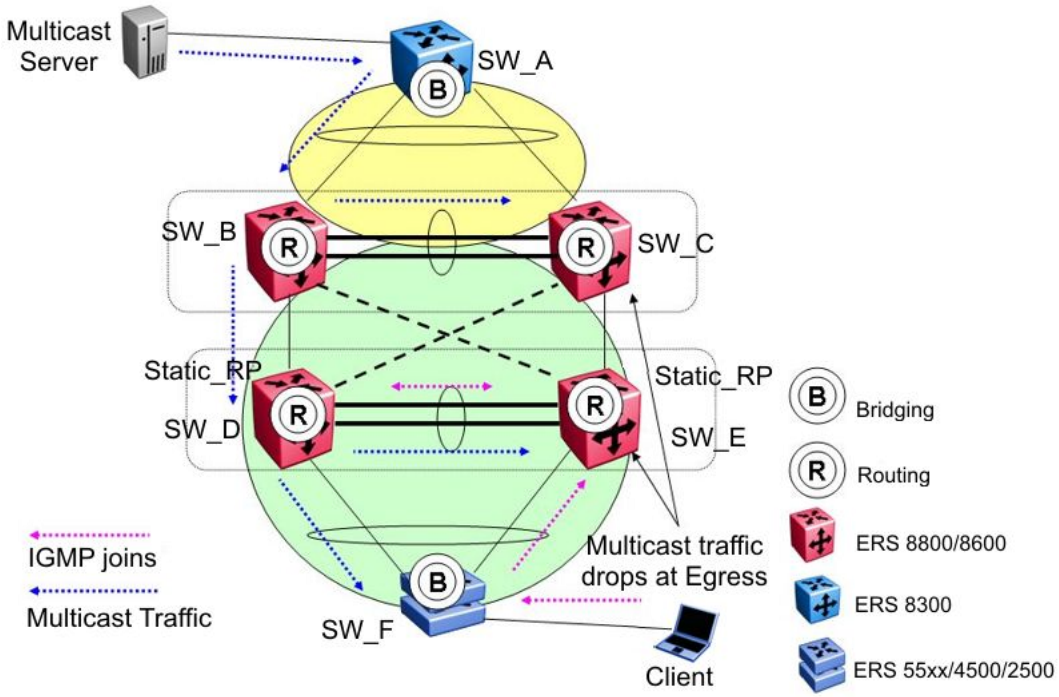


Figure 113: Static-RP in SMLT using the same CLIP address

In the preceding figure, the multicast traffic flows as follows:

1. The multicast server sends multicast data towards the Source DR (SDR) SW_A.
2. The SDR sends register messages with encapsulated multicast data towards the RP.
3. Once the client sends IGMP membership reports towards the multicast router, the multicast router creates a (*,G) entry
4. The RP sends joins towards the source on the reverse path.
5. When the SDR receives the joins, it sends native multicast traffic.
6. When SW_B or SW_D receive multicast traffic from upstream, they forward the traffic on the IST as well as on the SMLT link. However, on the other aggregation switches (SW_C and SW_E) multicast traffic is dropped at the egress side. (This is in accordance with the SMLT/RSMLT design to provide fast failover for multicast traffic.) Both aggregation switches SW_D and SW_E have similar (S,G) records.
7. In the case of SW_D (RP) failure, SW_B changes only the next-hop interface towards SW_E, and since the RP address is the same, it does not flush any (S,G) entries. In this way, this configuration achieves faster failover.

Multicast for multimedia

The Avaya Ethernet Routing Switch 8800/8600 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

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 Avaya Ethernet Routing Switch 8800/8600. 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 Avaya Ethernet Routing Switch 8800/8600 supports more than a thousand Joins/Leaves per second, which is well adapted to TV applications.

Important:

For IGMPv3, Avaya recommends that you ensure a Join rate of 250 per second or less. If the Avaya Ethernet Routing Switch 8800/8600 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 Avaya Ethernet Routing Switch 8800/8600. 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.

Important:

In networks that have only one user connected to each port, Avaya 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 Avaya Ethernet Routing Switch 8800/8600 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 8800/8600 supports the Single User and Multiple User Fast Leave modes for IGAP.

For more information about IGAP, see *Avaya 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 Avaya Ethernet Routing Switch 8800/8600 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.

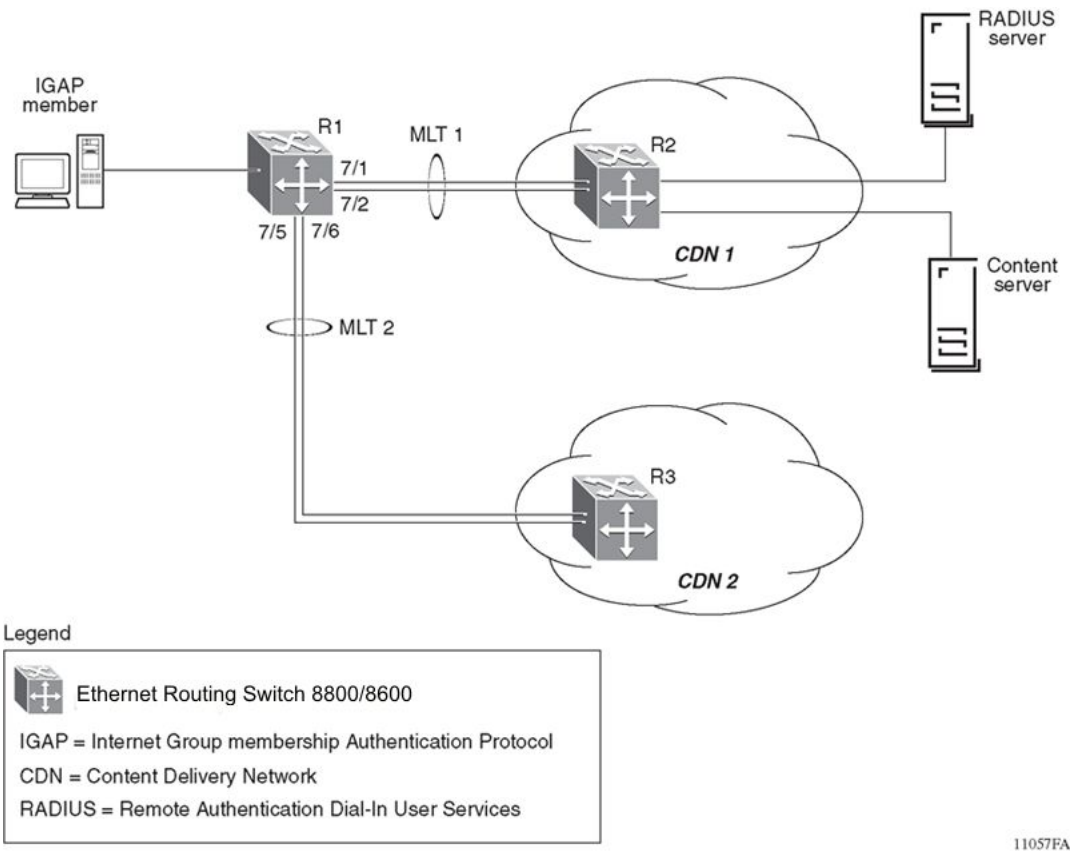


Figure 114: Avoiding an interruption of IGAP traffic

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 you must disable an MLT link, 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.

Chapter 13: MPLS IP VPN and IP VPN Lite

The Avaya Ethernet Routing Switch 8800/8600 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 8800/8600 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.

MPLS IP VPN

Beginning with Release 5.0, the Avaya 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.

Avaya 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 through 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).

The following figure shows a sample MPLS network.

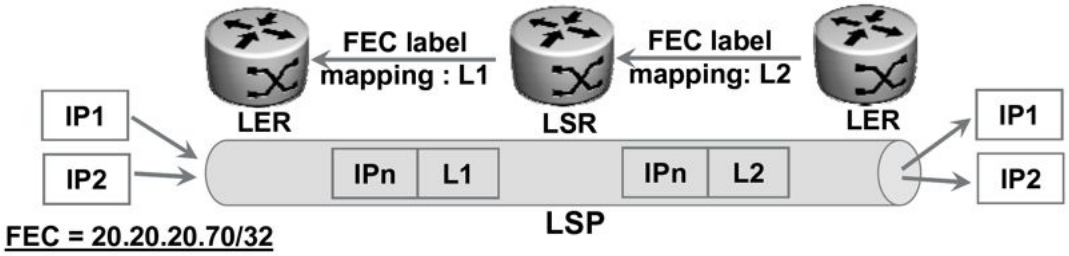
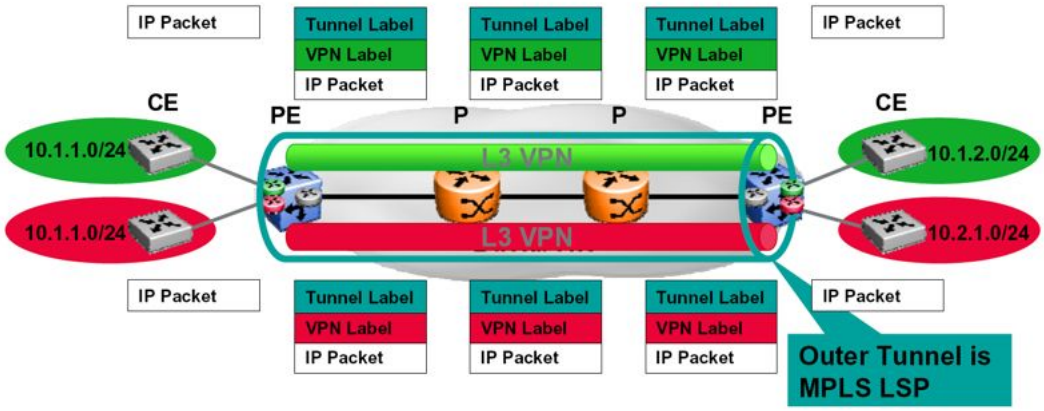


Figure 115: Label Switched Path and Forwarding Equivalent Class

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 Avaya Ethernet Routing Switch 8800/8600 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.



- Tunnel Label is MPLS outer label (changes at every hop)
- VPN Label is MPLS inner label (assigns packet to correct VRF at egress PE)
- P nodes are MPLS Label switch Routers (LSR)
- PE nodes are Label Edge Routers (LER)

Figure 116: 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 8800/8600 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.

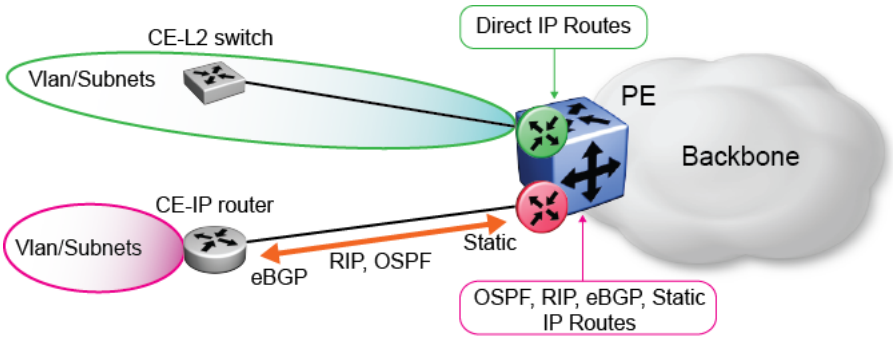


Figure 117: CE to PE connectivity

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 through 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.

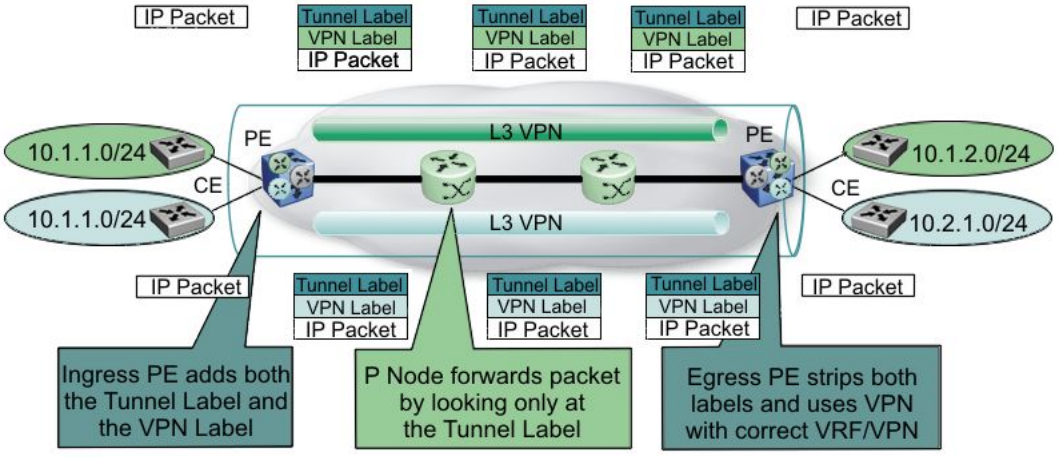


Figure 118: PE router label switching

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.

The following figure shows the VPN-IPv4 address.

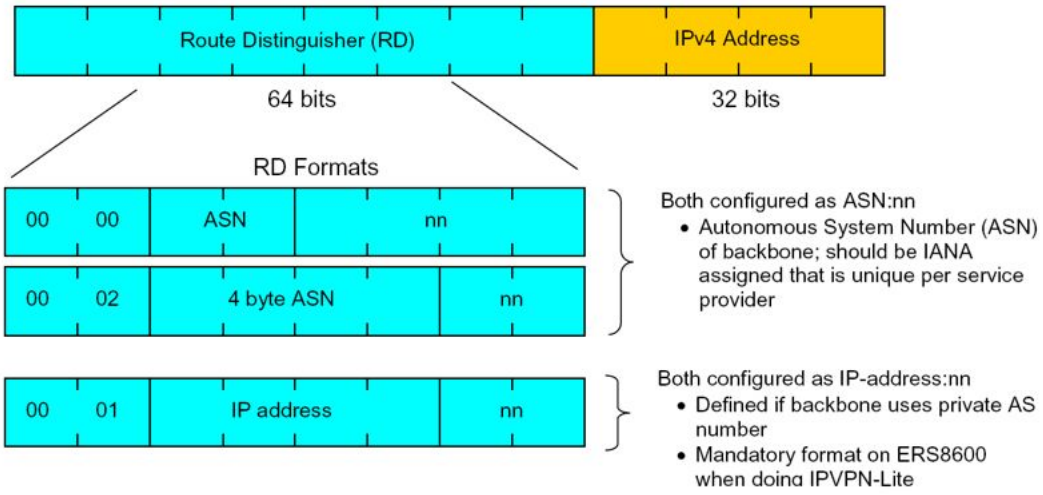


Figure 119: VPN-IPv4 Address

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.

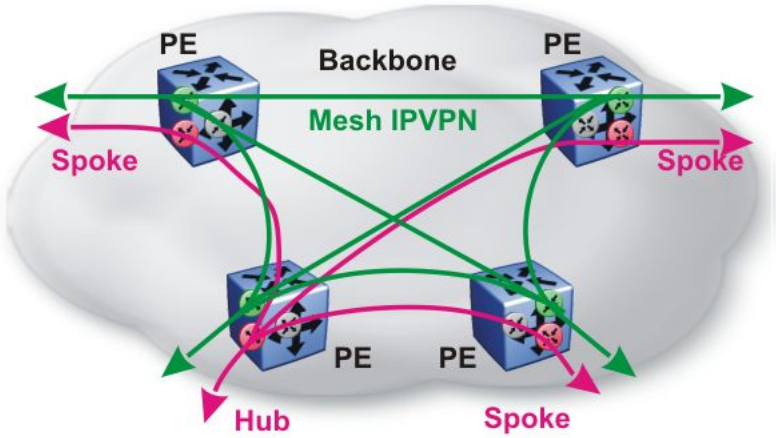


Figure 120: IP VPN hub and spoke

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 119: VPN-IPv4 Address on page 257.

IP VPN requirements and recommendations

To use IP VPN, you require R or RS modules, as well as the 8692 SF/CPU with SuperMezz or 8895 SF/CPU. You also require the appropriate license.

The Avaya Ethernet Routing Switch 8800/8600 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 the 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 runtime state information for the module from the previous Master CPU. If a module 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 the most recent configuration, the failover time is improved compared to a single SF/CPU restart in non-HA mode.

VPN tunnel dampening is not supported.

The Ethernet Routing Switch 8800/8600 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 more information about IP VPN scalability, see Table 5: Supported scaling capabilities on page 28 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 Avaya Ethernet Routing Switch 8800/8600 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 8800/8600 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 8800/86000 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 8800/8600 Technical Configuration Guide, NN48500-570*.

MPLS interoperability

The Avaya Ethernet Routing Switch 8800/8600 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 8800/8600)
- 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 Avaya Ethernet Routing Switch 8800/8600, 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 Avaya IP VPN-Lite, the Avaya Ethernet Routing Switch 8800/8600 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, Avaya 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 Avaya 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 figure, 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.

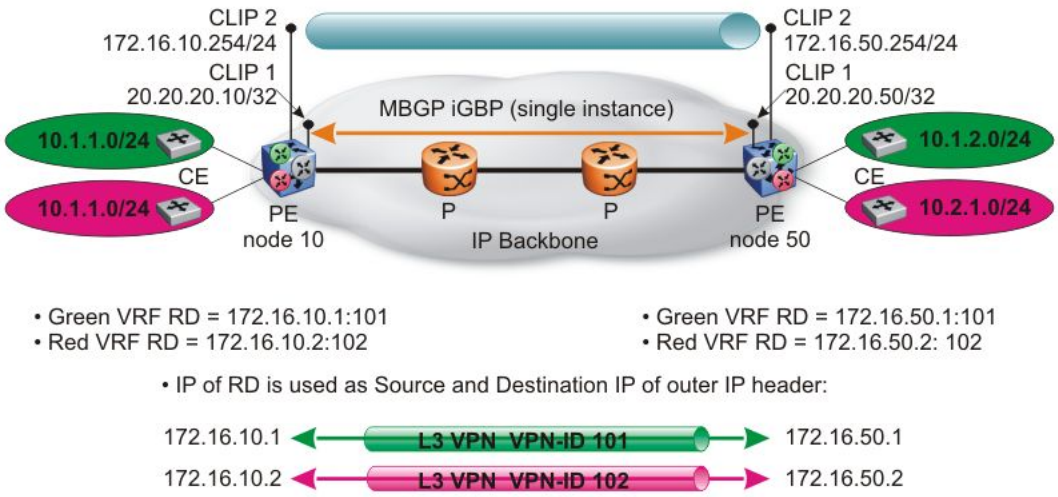


Figure 121: IP VPN Lite

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, Avaya 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 Avaya 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).

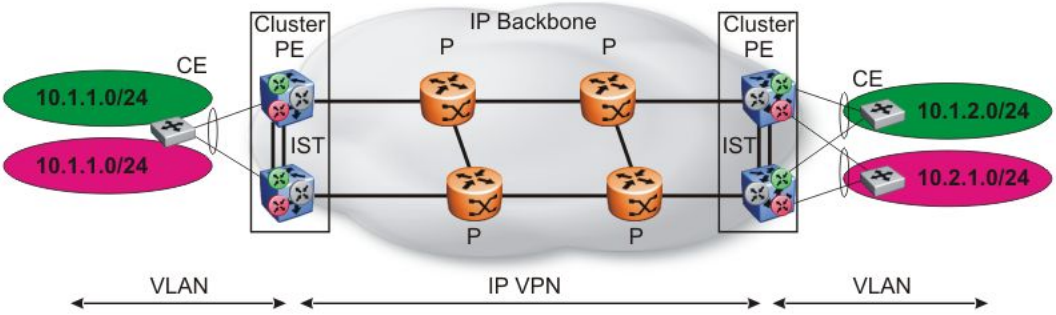


Figure 122: IP VPN Lite with SMLT

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 Avaya Ethernet Routing Switch 8800/8600 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, nonblocking design where all traffic is hardware switched

For detailed configuration steps for these examples, see *IP VPN-Lite for Avaya Ethernet Routing Switch 8800/8600 Technical Configuration Guide , NN48500-562*.

SMLT design

To meet the design requirements, an Avaya Ethernet Routing Switch 8800/8600 is deployed at each site. As shown in the following figure, the five Ethernet Routing Switch 8800/8600s are interconnected using 10 gigabit Ethernet links in an SMLT cluster configuration. The Ethernet Routing Switch 8800/8600s 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.

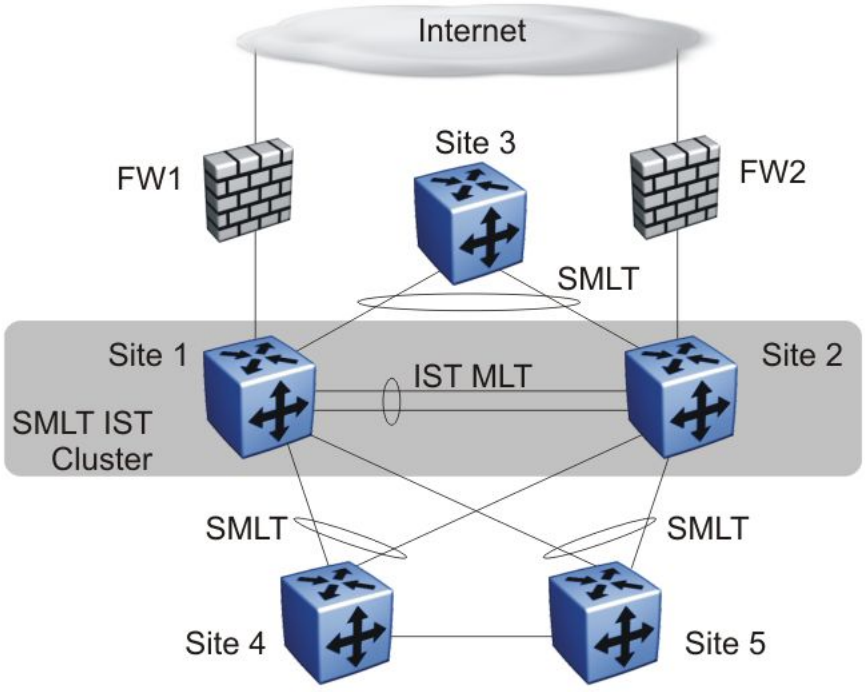


Figure 123: SMLT design

With Avaya's advanced packet processor architecture, the Avaya Ethernet Routing Switch 8800/8600 always hardware switches all traffic flows including IP VPN traffic used in this design. This means that if a nonblocking 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.

Important:

To support the VRF and IP VPN functionalities used in this design, you must equip the Avaya Ethernet Routing Switch 8800/8600 with R or RS I/O Modules, 8692 SF/CPU with Super-Mezzanine daughter card or 8895 SF/CPU, and 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 Avaya Ethernet

Routing Switch 8800/8600 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.

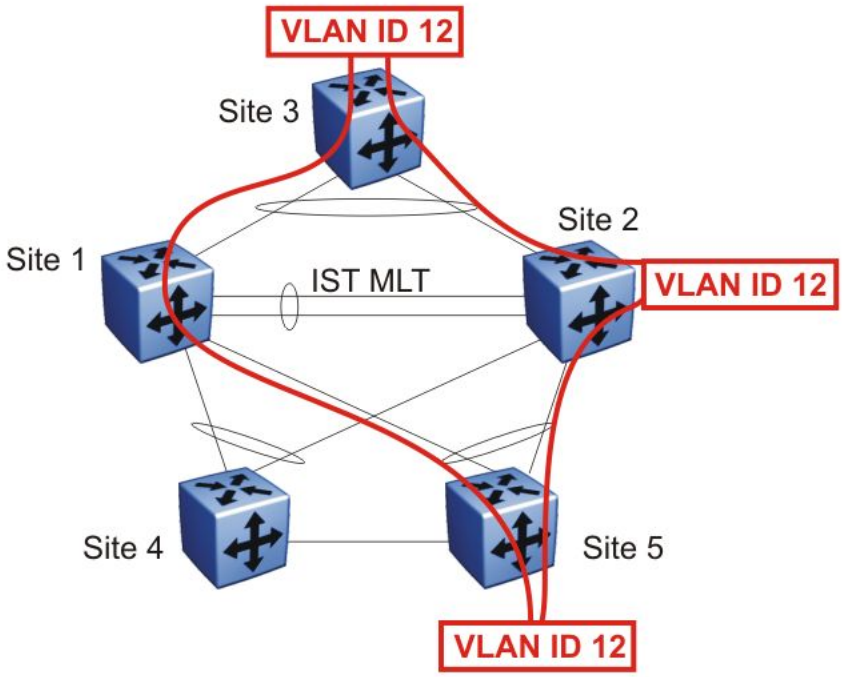


Figure 124: Layer 2 VPN example: VLAN ID 12 dropped off at sites 2, 3, and 5

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 Avaya Ethernet Routing Switch 8800/8600s. This routing instance constitutes the default routing instance of the Avaya Ethernet Routing Switch 8800/8600 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 8800/8600.

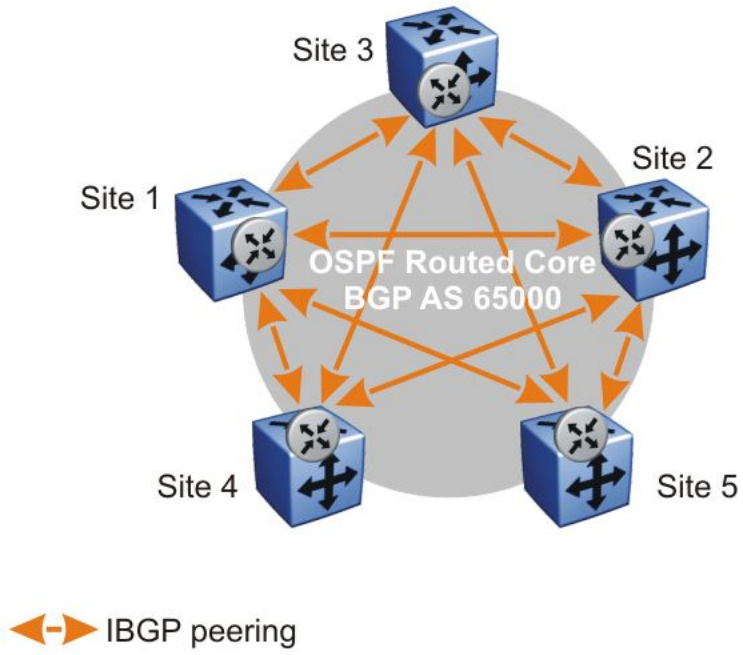


Figure 125: GRT IGP VLAN running OSPF and full mesh of IBGP peerings

Each Avaya Ethernet Routing Switch 8800/8600 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 8800/8600s 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, Avaya recommends the use of BGP Route-Reflectors. This can be accomplished by making the Ethernet Routing Switch 8800/8600 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 Avaya IP VPN Lite.

To provide address space for the IPinIP encapsulation, each Avaya Ethernet Routing Switch 8800/8600 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.

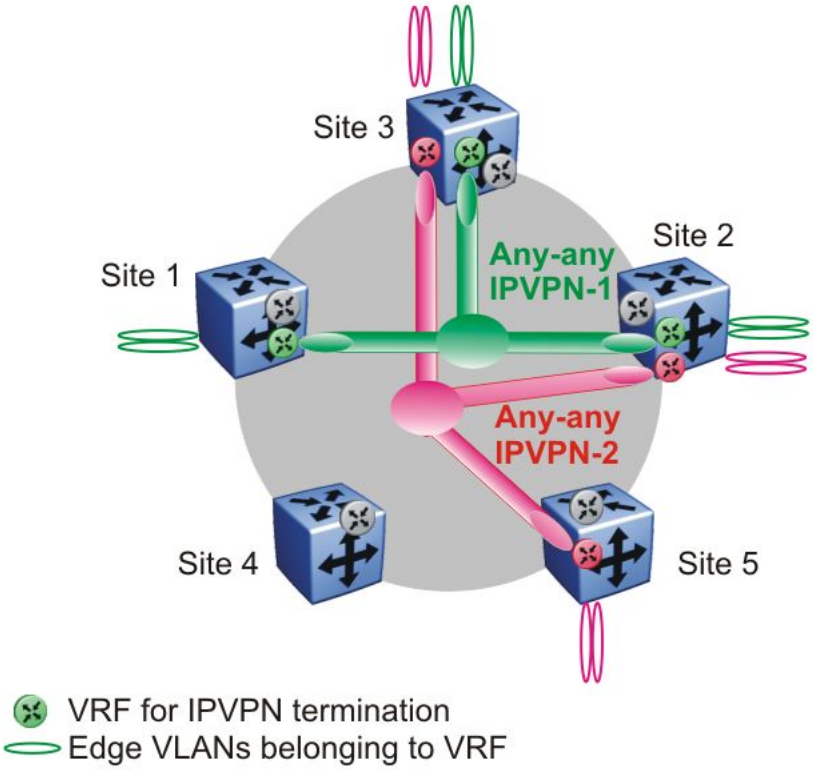
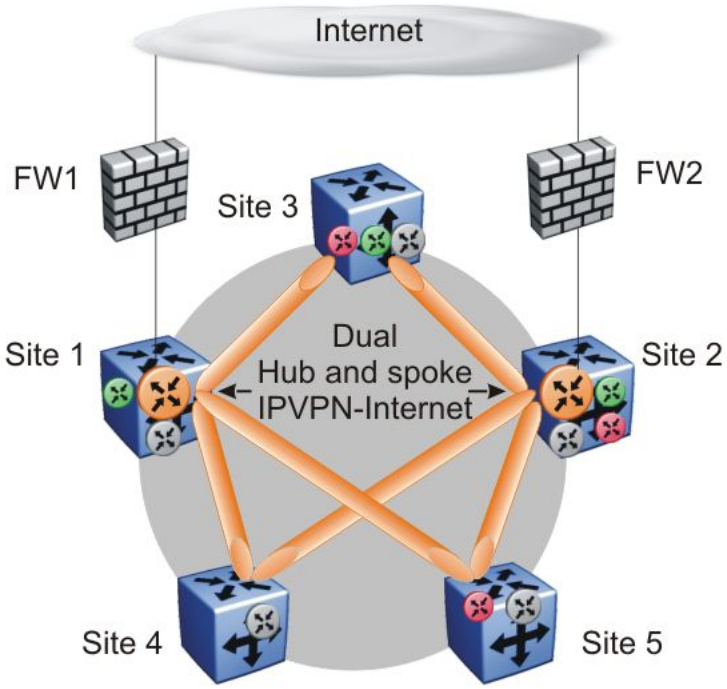


Figure 126: Example of two separate Layer 3 VPNs

Internet Layer 3 VPN design

The two Avaya Ethernet Routing Switch 8800/8600s 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 8800/8600s 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 8800/8600 Technical Configuration Guide , NN48500-562*.

Chapter 14: 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 Avaya recommends to optimize IP for network redundancy.

Layer 1 examples

The following figures are a series of Layer 1 examples that illustrate the physical network layout.

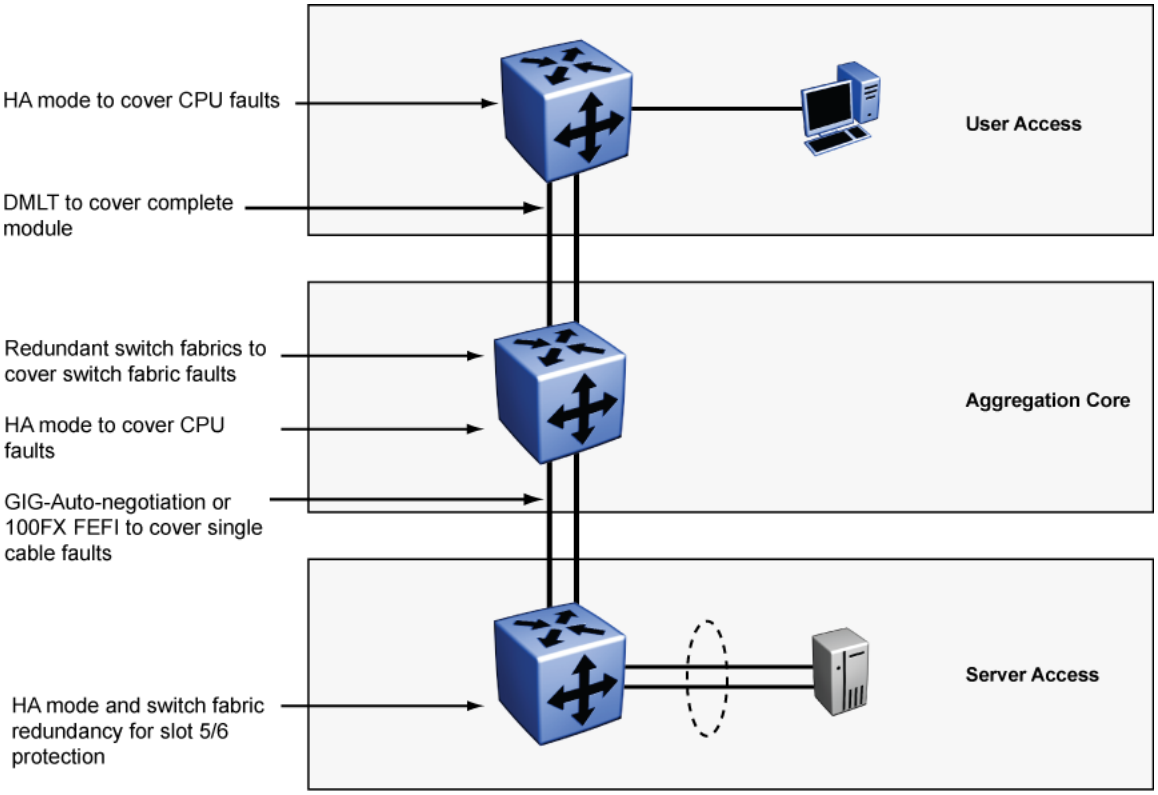


Figure 127: Layer 1 design example 1

All the Layer 1 redundancy mechanisms are described in example 2.

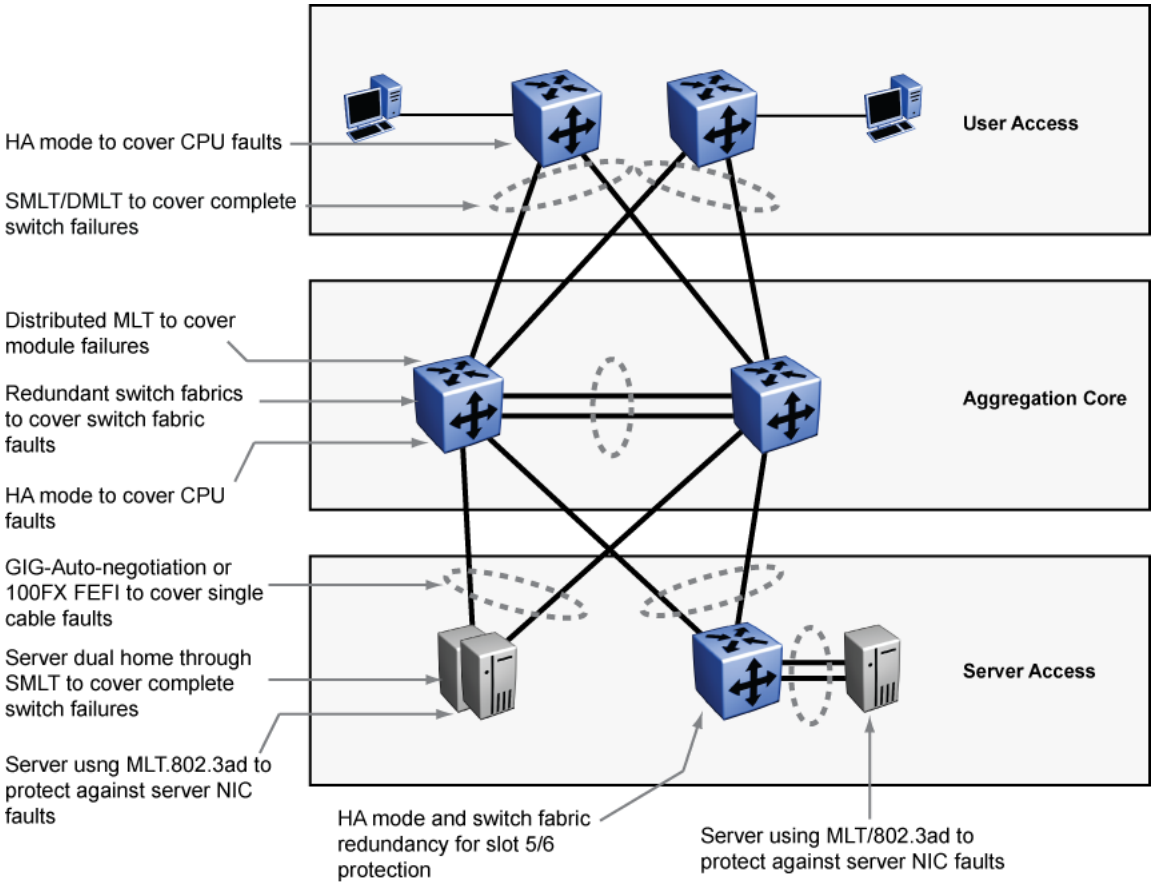


Figure 128: Layer 1 design example 2

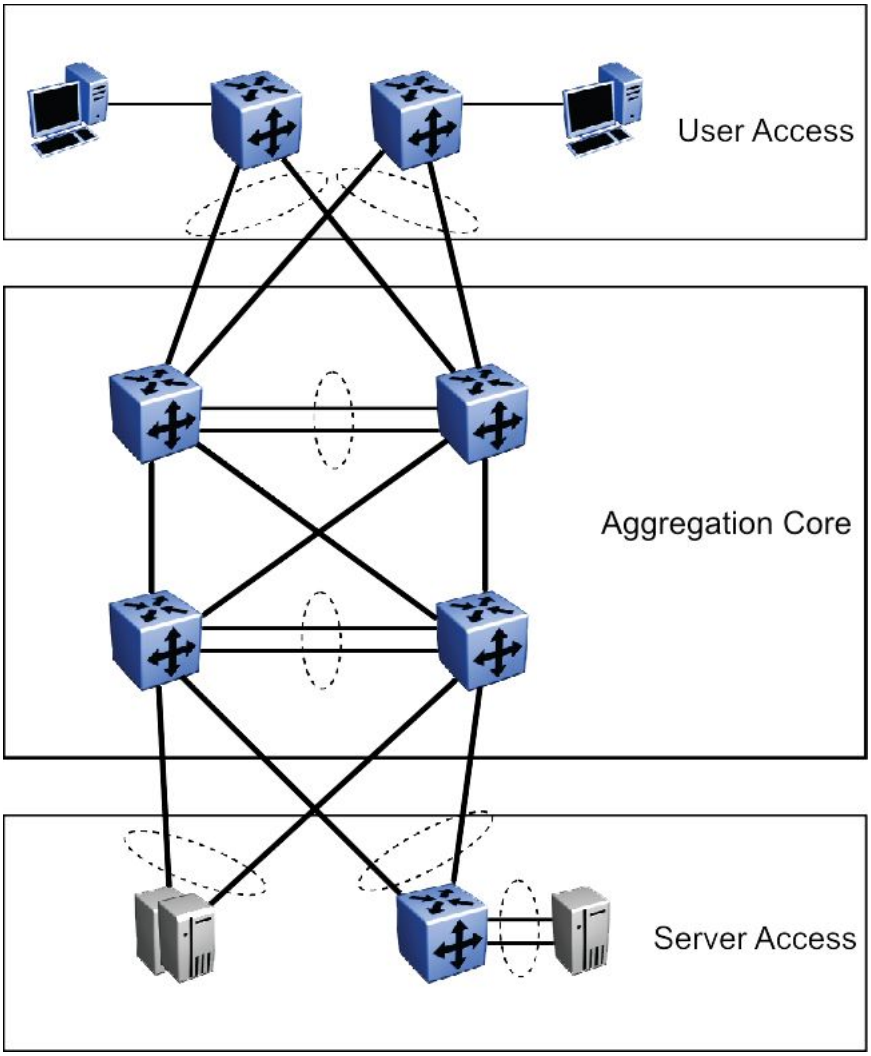


Figure 129: Layer 1 design example 3

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.

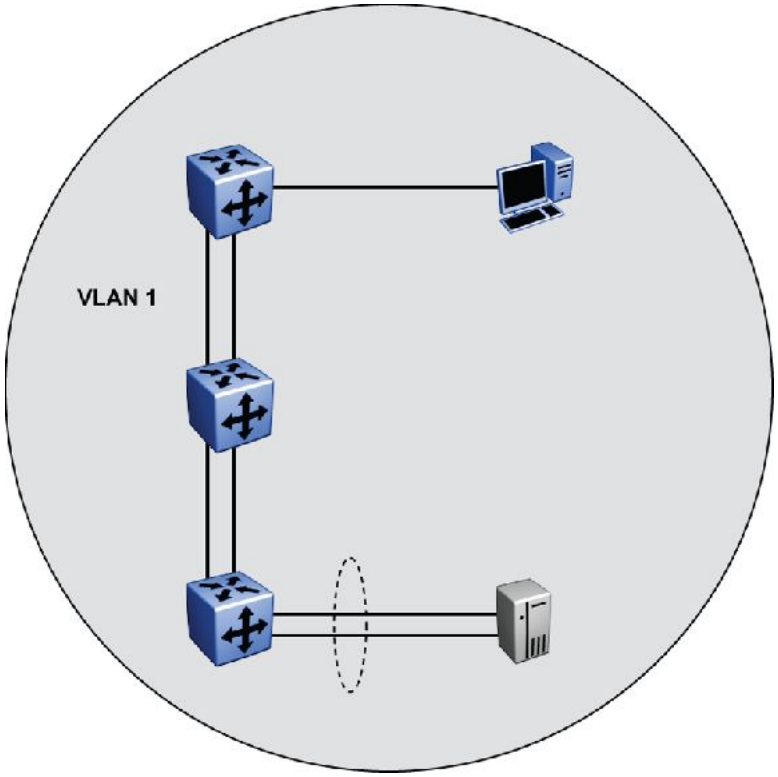


Figure 130: Layer 2 design example 1

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.

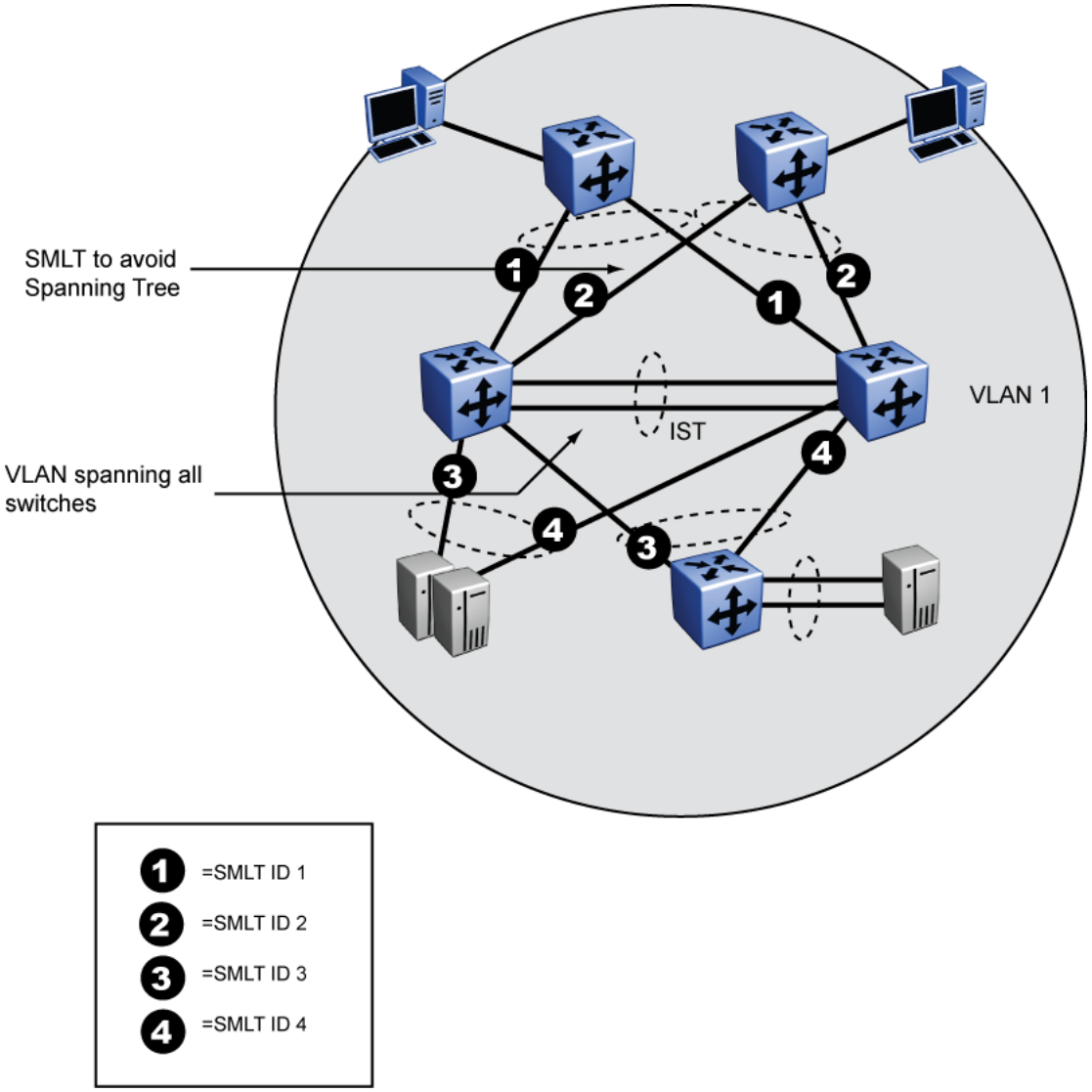


Figure 131: Layer 2 design example 2

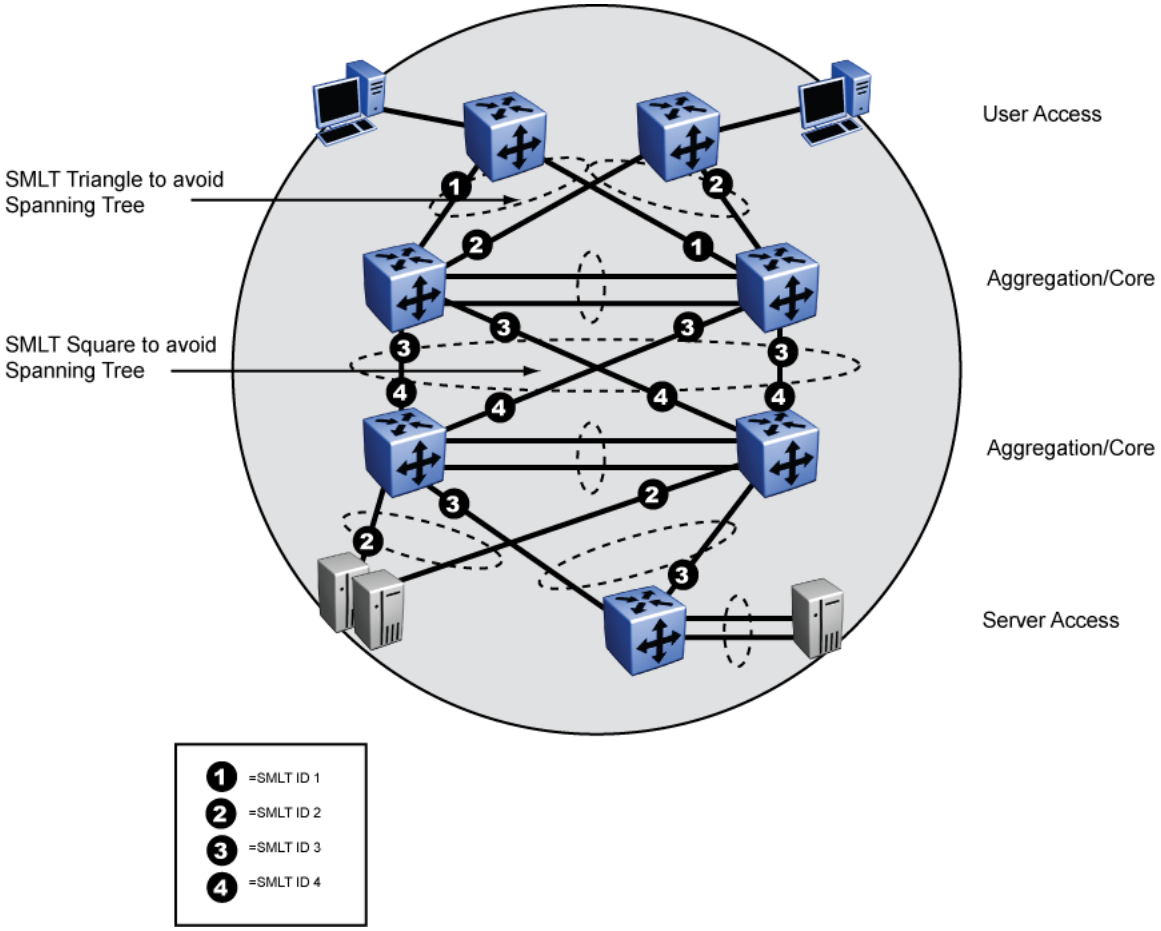


Figure 132: Layer 2 design example 3

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.

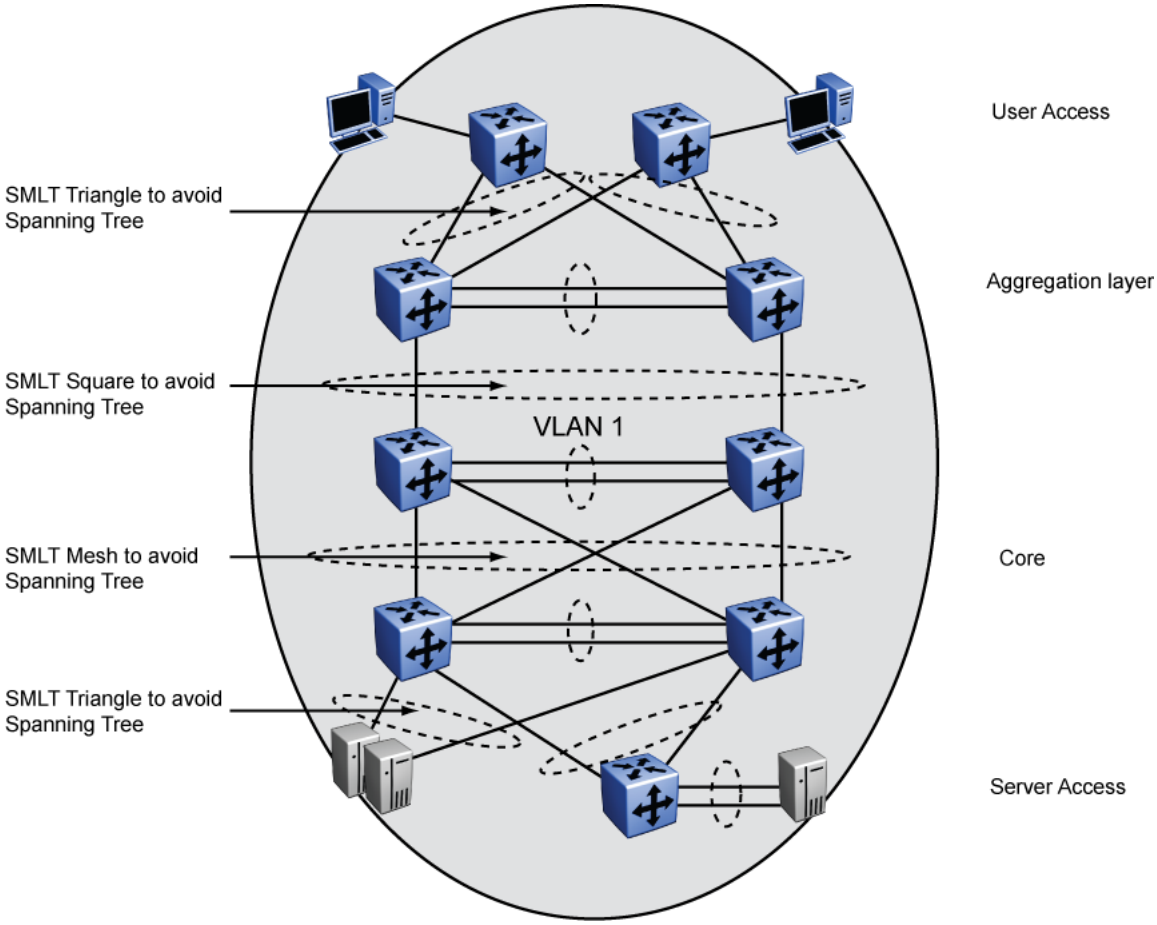


Figure 133: Layer 2 design example 4

Layer 3 examples

The following figures are a series of Layer 3 network design examples that show the routing instances that Avaya recommends you use to optimize IP for network redundancy.

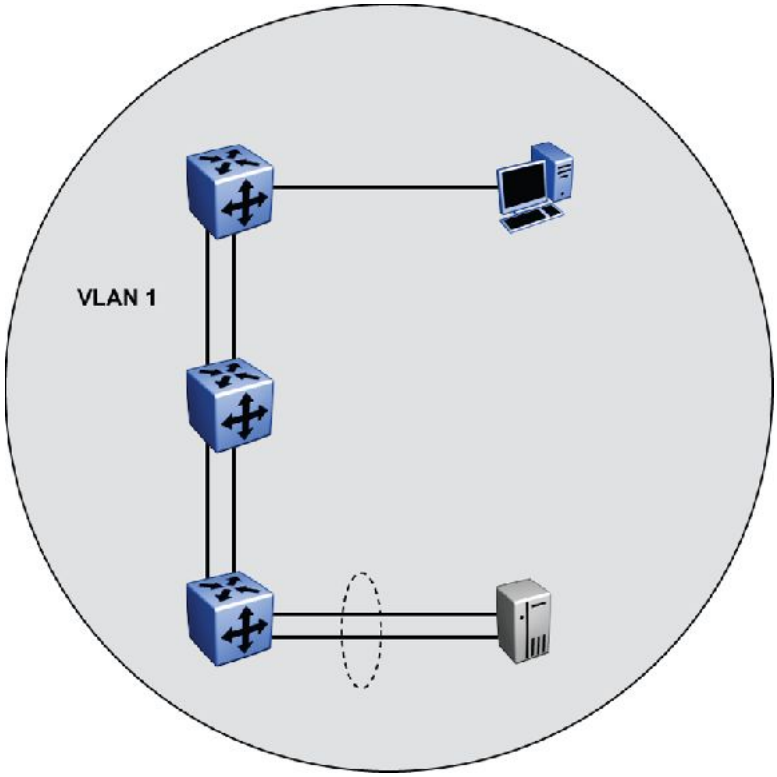


Figure 134: Layer 3 design example 1

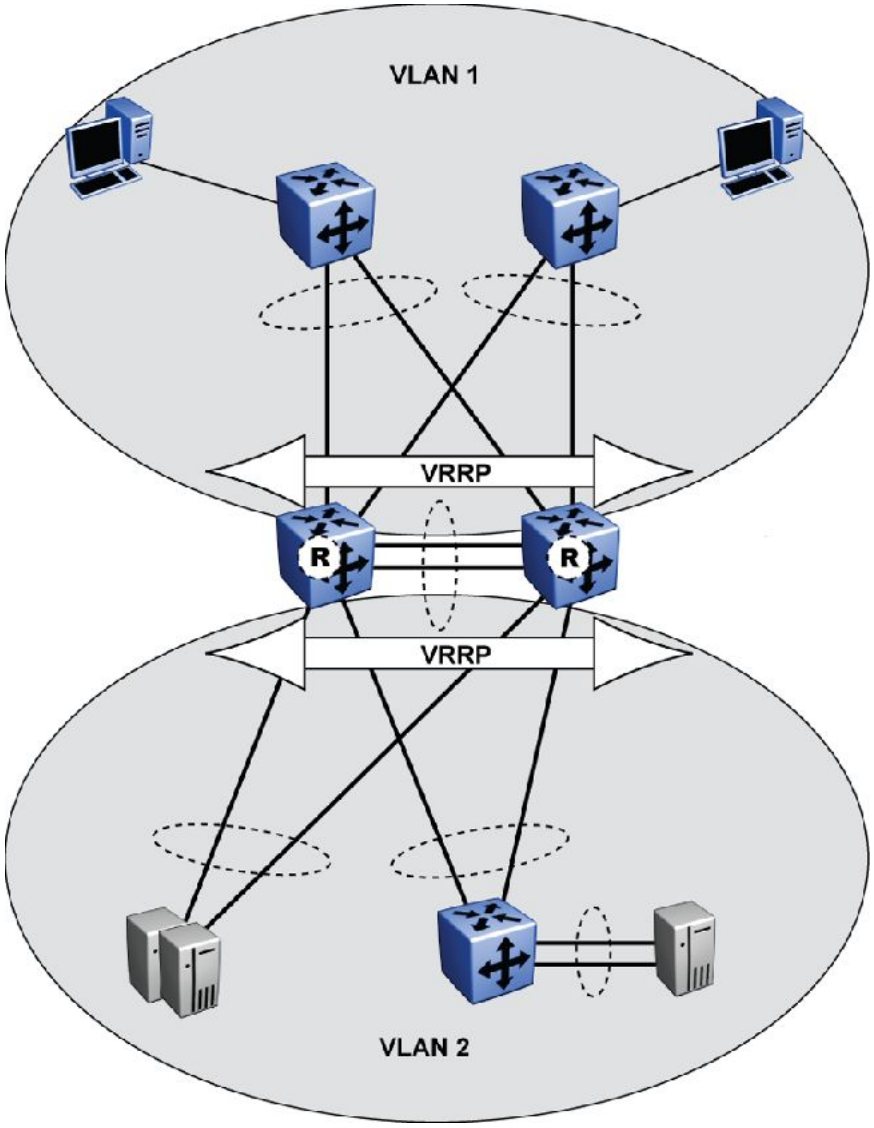


Figure 135: Layer 3 design example 2

In the following figures, DGW denotes Data GateWay.

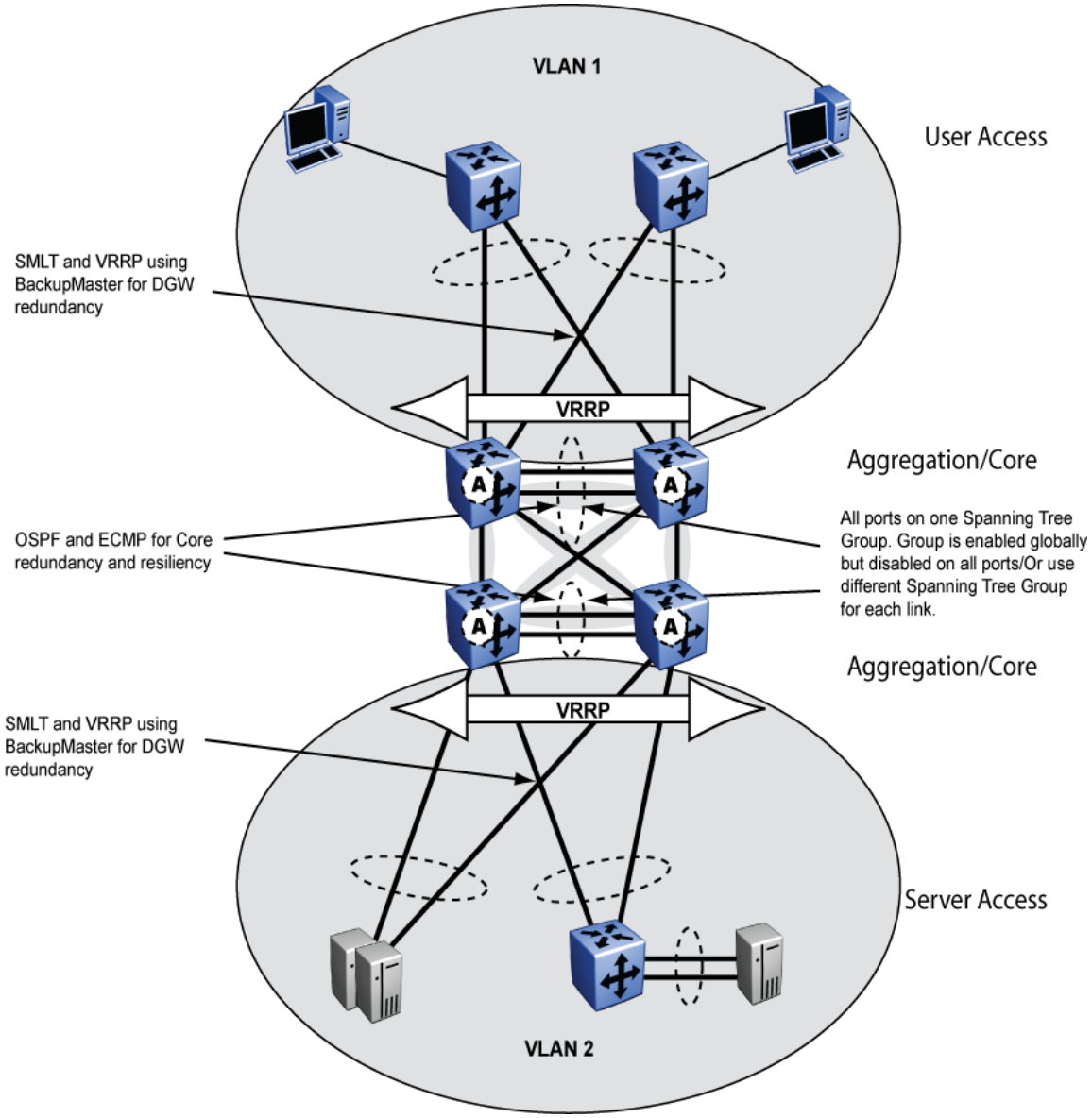


Figure 136: Layer 3 design example 3

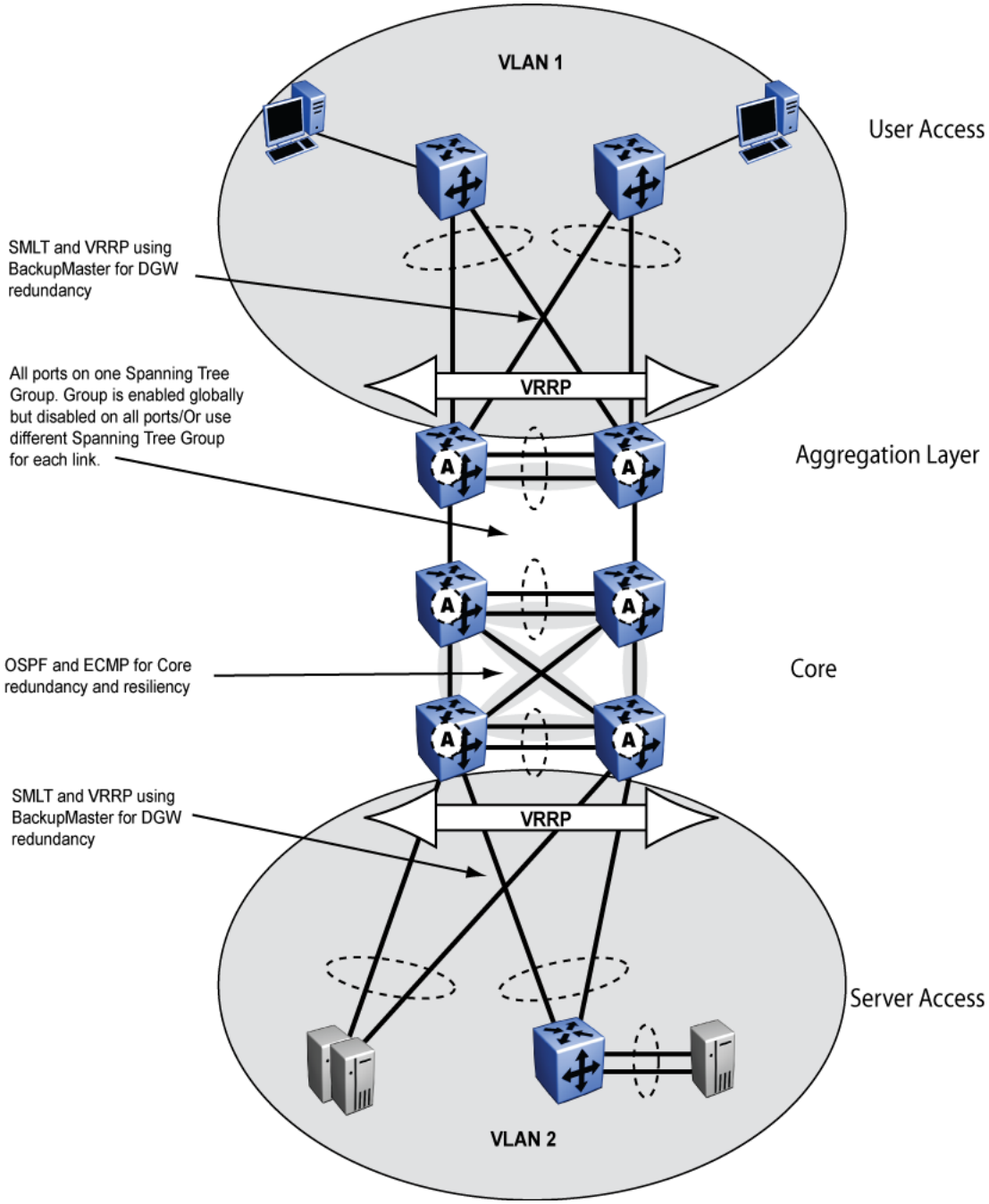


Figure 137: Layer 3 design example 4

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.

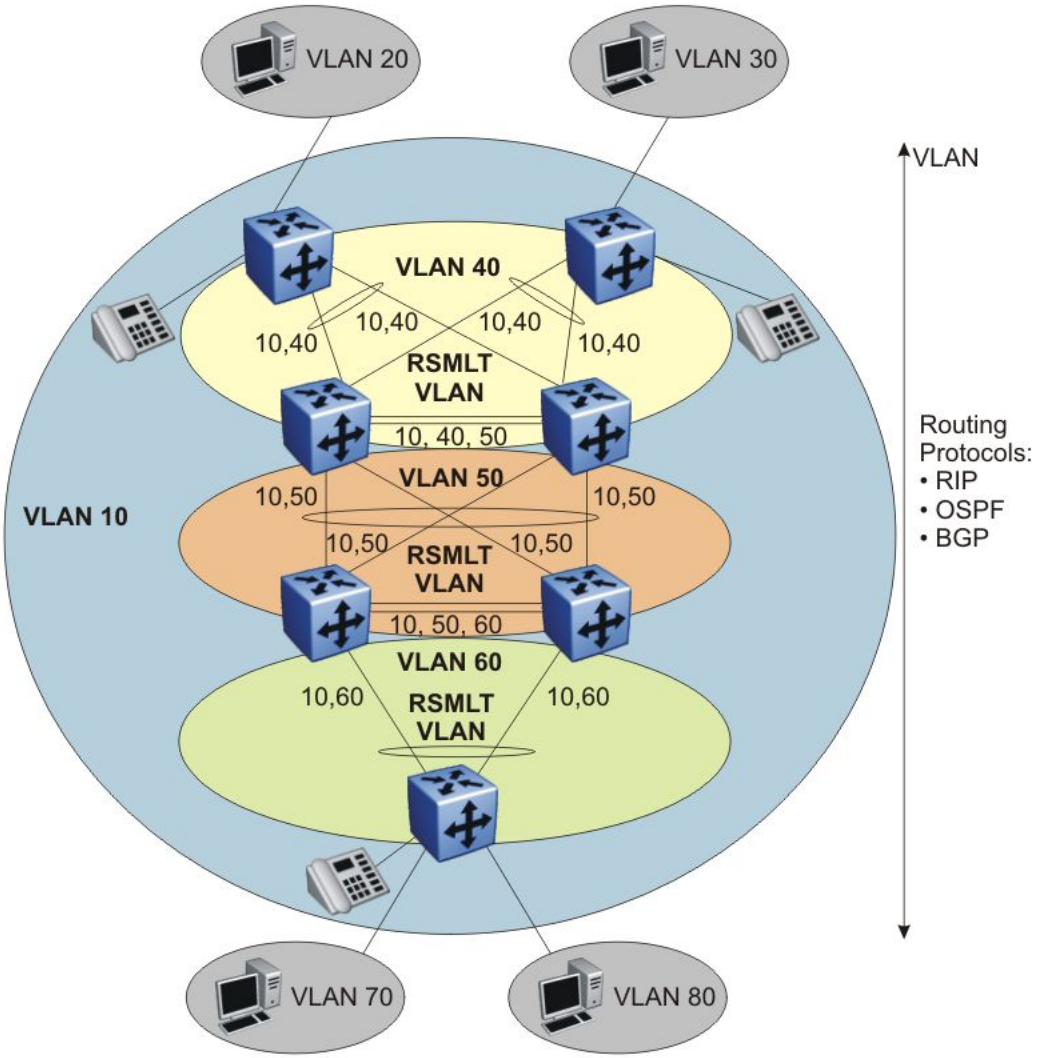


Figure 138: Redundant network design

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, Avaya 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 138: Redundant network design on page 282 shows how to enable routing at the access layer while keeping the routing protocol design robust and simple.

Chapter 15: 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. For more information, see Redundant network design on page 57. Without redundancy, all services can be brought down.

To provide additional network security, you can use the Avaya Contivity VPN product suite 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.

For more information about setting the rate limits for broadcast or multicast packets on a port, see *Avaya Ethernet Routing Switch 8800/8600 Configuration — Ethernet Modules, NN46205-503*.

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 Avaya Ethernet Routing Switch 8800/8600 is protected by directed broadcast suppression. This feature is enabled by default. Avaya recommends that you not disable it.

For more information about directed broadcast suppression, see *Avaya Ethernet Routing Switch 8800/8600 Security, NN46205-601*.

Prioritization of control traffic

The Avaya Ethernet Routing Switch 8800/8600 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 29: CP Limit recommended values

	CP Limit Values when using the 8895 SF/CPU		CP Limit Values when using the 8692 SF/CPU with SuperMezz	
	Broadcast	**Multicast**	**Broadcast**	**Multicast**
Severe				
Workstation (PC)	1000	1000	1000	1000
Server	2500	2500	2500	2500
NonIST Interconnection	7500	6000	3000	3000
Moderate				
Workstation (PC)	2500	2500	2500	2500
Server	5000	5000	3000	3000
NonIST Interconnection	9000	9000	3000	3000
Relaxed				
Workstation (PC)	4000	4000	3000	3000
Server	7000	7000	3000	3000
NonIST Interconnection	10 000	10 000	3000	3000

Important:

The 8692 SF/CPU with SuperMezz requires additional processing to send control packets from the CP to the SuperMezz and to program any hardware records in the I/O modules. Both operations now require an additional hop because they require CP involvement. To accommodate this additional processing, you must use the **`cp-limit broadcast-limit <value>`** and **`cp-limit multicast-limit <value>`** commands to lower the broadcast and multicast thresholds to 3000 packets per second.

ARP request threshold recommendations

The Address Resolution Protocol (ARP) request threshold limits the ability of the Avaya Ethernet Routing Switch 8800/8600 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), Avaya 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

From Release 3.5.0 and later, you can access the ARP request threshold feature through the CLI. For more information about the **`config ip arp arpreqthreshold`** command, see *Avaya Ethernet Routing Switch 8800/8600 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 *Avaya Ethernet Routing Switch 8800/8600 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:

- Prevent IP spoofing.

You can use the spoof-detect feature.

- Prevent your network from being used as a broadcast amplification site.
- Enable the **hsecure** flag (High Secure mode) to block illegal IP addresses.

  For more information, see High Secure mode on page 290 or *Avaya Ethernet Routing Switch 8800/8600 Security, NN46205-601*.

Damage prevention navigation

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 to filter.

Table 30: 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
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 *Avaya Ethernet Routing Switch 8800/8600 Configuration — VLANs and Spanning Tree, NN46205-517*.

You can also use the predefined Access Control Template (ACT) for ARP spoof detection. For more information about this ACT, see *Avaya Ethernet Routing Switch 8800/8600 Configuration — QoS and IP Filtering for R and RS Modules, NN46205-507*.

High Secure mode

To ensure that the Avaya Ethernet Routing Switch 8800/8600 does not route packets with an illegal source address of 255.255.255.255 (in accordance with 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].

For more information about hsecure, see *Avaya Ethernet Routing Switch 8800/8600 Security, NN46205-601*.

Spanning Tree BPDU filtering

To prevent unknown devices from influencing the Spanning Tree topology, the Avaya Ethernet Routing Switch 8800/8600 supports Bridge Protocol Data Unit (BPDU) Filtering for Avaya

Spanning Tree Groups (STPG), Rapid Spanning Tree Protocol (RSTP), and Multiple Spanning Tree Protocol (MSTP).

With BPDU Filtering, the network administrator can achieve the following:

- Block an unwanted root selection process when an edge device (for example, a laptop running Linux and enabled with STP) is added to the network. This prevents unknown devices from influencing an existing spanning tree topology.
- Block the flooding of BPDUs from an unknown device.

When a port has BPDU Filtering enabled and the port receives an STP BPDU, the following actions take place:

- The port is immediately put in the operational disabled state.
- A trap is generated and the following log message is written to the log: `Ethernet <x> is shut down by BPDU Filter`
- The port timer starts.
- The port stays in the operational disabled state until the port timer expires.

If you disable the timer or reset the switch before the timer expires, the port remains in the disabled state. If you disable BPDU Filtering while the timer is running, the timer stops and the port remains in the disabled state. You must then manually enable the port to return it to the normal mode.

The STP BPDU Filtering feature is not supported on MLT/IST/SMLT/RSMLT ports.

For more information about BPDU Filtering, Avaya Ethernet Routing Switch 8800/8600 Configuration — VLANs and Spanning Tree (NN46205-517).

Security and redundancy

Redundancy in hardware and software is one of the key security features of the Avaya Ethernet Routing Switch 8800/8600. High availability is achieved by eliminating single points of failure in the network and by using the unique features of the Avaya Ethernet Routing Switch 8800/8600 including:

- a complete, redundant hardware architecture (switching fabrics in 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
- a list of software features that allow high availability including:
    - 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 Avaya Technical Support Web site in the Ethernet Routing Switch 8800/8600 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

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 Avaya 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.

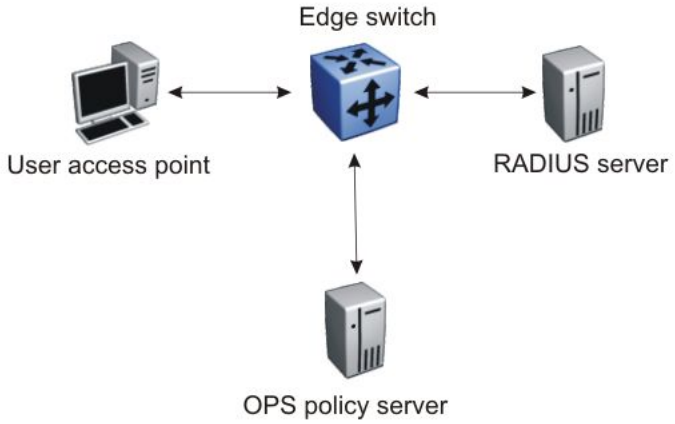


Figure 139: 802.1x and OPS interaction

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 Avaya Ethernet Routing Switch 8800/8600. For more information, contact your Avaya representative.

802.1x and the LAN Enforcer or VPN TunnelGuard

The Sygate LAN Enforcer or the Avaya VPN TunnelGuard enables the Avaya Ethernet Routing Switch 8800/8600 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 Avaya Ethernet Routing Switch 8800/8600 to build secure VLANs. When you configure port-based VLANs, each VLAN is completely separated from the others.

The Avaya Ethernet Routing Switch 8800/8600 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 *Avaya Ethernet Routing Switch 8800/8600 Configuration — VLANs and Spanning Tree, NN46205-517*.

DHCP snooping

Dynamic Host Configuration Protocol (DHCP) snooping provides security to the network by preventing DHCP spoofing. DHCP spoofing refers to an attacker's ability to respond to DHCP requests with false IP information. DHCP snooping acts like a firewall between untrusted hosts and the DHCP servers so that DHCP spoofing cannot occur.

The following figure shows a simplified DHCP snooping topology.

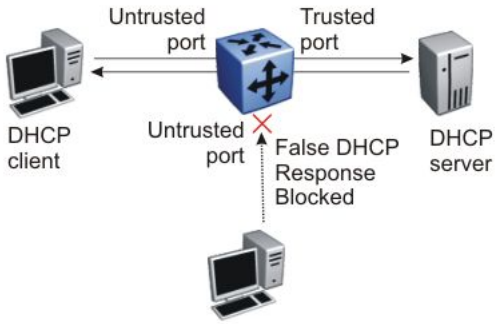


Figure 140: DHCP snooping

DHCP snooping classifies ports into two types:

- Untrusted: ports that are configured to receive messages from outside the network or firewall. Only DHCP requests are allowed.
- Trusted: ports, such as switch-to-switch and DHCP server ports, that are configured to receive messages only from within the network. All types of DHCP messages are allowed.

To eliminate the capability to set up rogue DHCP servers on untrusted ports, the untrusted ports allow DHCP request packets only. DHCP replies and all other types of DHCP messages from untrusted ports are dropped.

DHCP snooping verifies the source of DHCP packets as follows:

- When the switch receives a DHCP request on an untrusted port, DHCP snooping compares the source MAC address and the DHCP client hardware address. If the addresses match, the switch forwards the packet. If the addresses do not match, the switch drops the packet.
- When the switch receives a DHCP release or DHCP decline broadcast message from a client, DHCP snooping verifies that the port on which the message was received matches the port information for the client MAC address in the DHCP binding table. If the port information matches, the switch forwards the DHCP packet.

DHCP snooping supports MLT/SMLT ports as trusted ports only.

DHCP binding table

DHCP snooping dynamically creates and maintains an IP-to-MAC binding table. The DHCP binding table includes the following information about DHCP leases on untrusted interfaces:

- source MAC address
- IP address
- lease duration
- time to expiry
- VLAN ID
- port

You can also configure static DHCP binding entries. Dynamic binding entries are lost after a restart.

For more information about DHCP snooping, see Avaya Ethernet Routing Switch 8800/8600 Security (NN46205-601).

Dynamic ARP Inspection (DAI)

Dynamic ARP Inspection (DAI) is a security feature that validates ARP packets in the network. It intercepts, discards, and logs ARP packets with invalid IP-to-MAC address bindings.

Without Dynamic ARP inspection, a malicious user can attack hosts, switches, and routers connected to the Layer 2 network by poisoning the ARP caches of systems connected to the subnet and by intercepting traffic intended for other hosts on the subnet (man-in-the-middle attacks). Dynamic ARP Inspection prevents this type of attack.

Important:

For Dynamic ARP inspection to function, you must enable DHCP snooping globally and on the VLAN. For information on DHCP snooping, see DHCP snooping on page 294.

DHCP snooping dynamically creates and maintains a binding table gathered from DHCP requests and replies. The MAC address from the DHCP request is paired with the IP address from the DHCP reply to create an entry in the DHCP binding table.

When you enable Dynamic ARP inspection, ARP packets on untrusted ports are filtered based on the source MAC and IP addresses. The switch forwards an ARP packet when the source MAC and IP addresses match an entry in the address binding table. Otherwise, the ARP packet is dropped.

Like DHCP snooping, Dynamic ARP Inspection supports MLT/SMLT ports as trusted ports only.

For more information about Dynamic ARP Inspection, see Avaya Ethernet Routing Switch 8800/8600 Security (NN46205-601).

IP Source Guard

IP Source Guard is a security feature that validates IP packets by intercepting IP packets with invalid IP-to-MAC bindings.

IP Source Guard works closely with DHCP snooping and prevents IP spoofing by allowing only IP addresses that are obtained through DHCP on a particular port. Initially, all IP traffic on the port is blocked except for the DHCP packets that are captured by DHCP snooping. When a client receives a valid IP address from the DHCP server, traffic on the port is permitted when the source IP and MAC addresses match a DCHP binding table entry for the port. Any IP traffic that does not match an entry in the DHCP binding table is filtered out. This filtering limits the ability of a host to attack the network by claiming a neighbor host's IP address.

Important:

For IP Source Guard to function, you must enable DHCP snooping and Dynamic ARP Inspection globally and at the VLAN level. To enable IP Source Guard on a port, the port must be configured as untrusted for DHCP snooping and untrusted for Dynamic ARP Inspection.

IP Source Guard cannot be enabled on MLT/SMLT ports.

For more information about IP Source Guard, see Avaya Ethernet Routing Switch 8800/8600 Security (NN46205-601).

Security at layer 2

At Layer 2, the Avaya Ethernet Routing Switch 8800/8600 provides the following security mechanisms:

- **Filters**

The Avaya Ethernet Routing Switch 8800/8600 provides Layer 2 filtering based on the MAC destination and source addresses. This is available per-VLAN.

- **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 8800/8600 Release 3.5.2*.

  For more information about the Layer 2 MAC filter, see *Avaya Ethernet Routing Switch 8800/8600 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 8800/8600 Release 3.5.2*.

Security at Layer 3: filtering

At Layer 3 and above, the Avaya Ethernet Routing Switch 8800/8600 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 Avaya Ethernet Routing Switch 8800/8600: 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 Avaya 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 on page 203.

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 *Avaya Ethernet Routing Switch 8800/8600 Configuration — OSPF and RIP, NN46205-522* and *Avaya Ethernet Routing Switch 8800/8600 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 on page 299
- Management access control on page 300
- High Secure mode on page 290
- Security and access policies on page 301

Management port

The Avaya Ethernet Routing Switch 8800/8600 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.

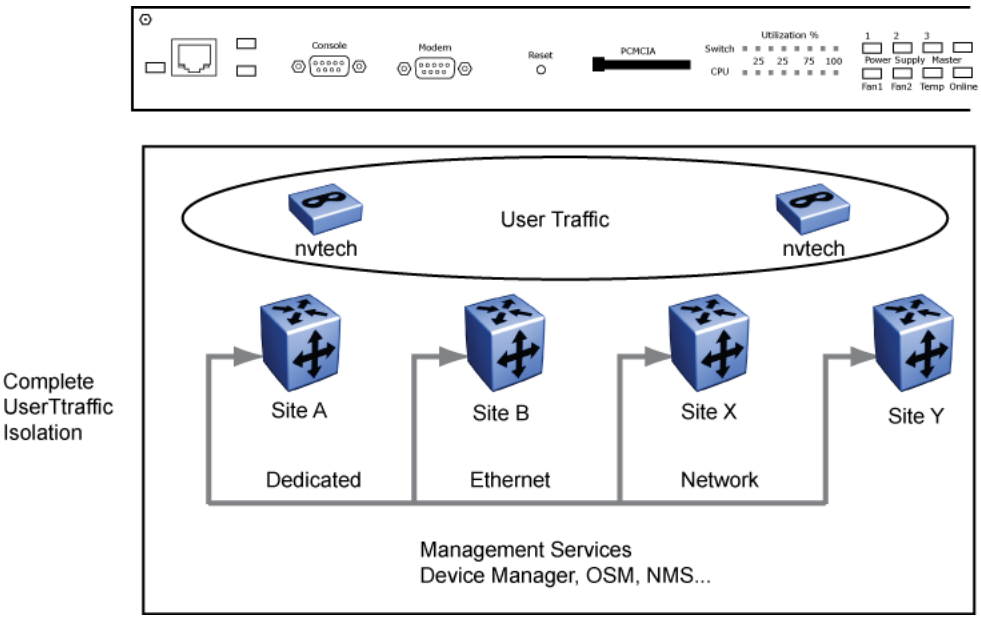


Figure 141: Dedicated Ethernet management link

You can also use the terminal servers/modems to access the console/modems ports on the switch (see the following figure).

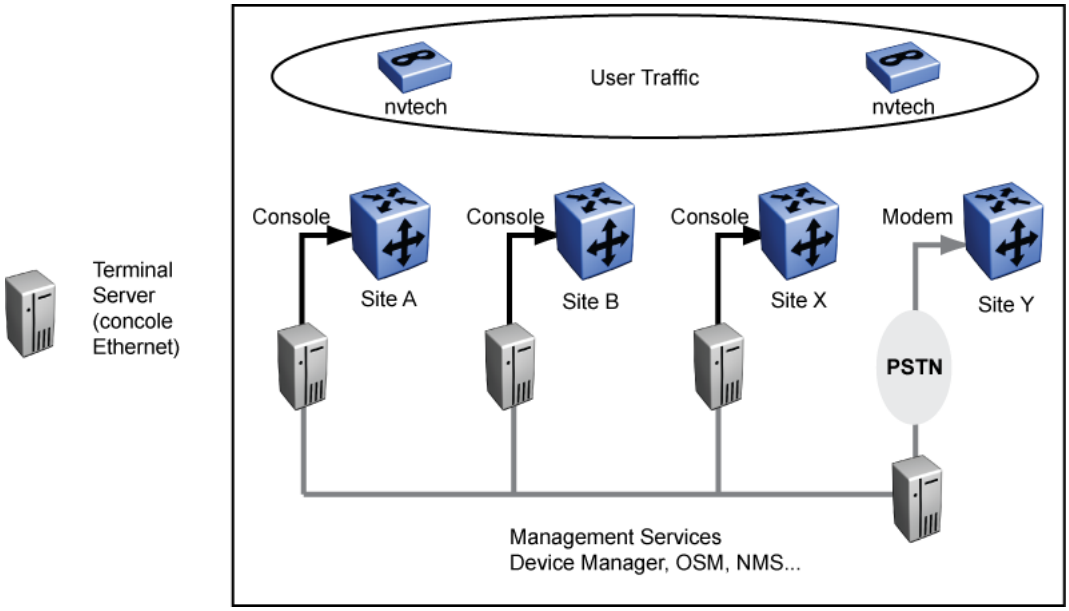


Figure 142: Terminal servers/modem access

When it is an absolute necessity for you to access the switch, Avaya recommends that you use this configuration. The switch is always reachable, even if an issue occurs with the in-band network management interface.

Important:

Connection of the Out-of-Band (OOB) Ethernet Management port to an In-Band I/O port is not recommended, as erroneous behavior on the network, such as a loop, can cause issues with the operation of the SF/CPU module. The most common issues seen are a loss of file management and inability to access the /pcmcia directory. To clear the condition, you must reboot or reset the SF/CPU.

To maintain a true OOB management network, do not include the switch In-Band I/O ports as part of the management network design. Rather than connect the OOB port to an In-band I/O port, you can achieve the same desired functionality by creating a management VLAN and assigning a management IP address to the VLAN.

Management access control

The following table shows management access levels. For more information, see *Avaya Ethernet Routing Switch 8800/8600 Security, NN46205-601*.

Table 31: Avaya Ethernet Routing Switch 8800/8600 management access levels

Access level	Description
Read only	Use this level to view the device settings. You cannot change any of the settings.

Access level	Description
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. 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 8800/8600 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, and rlogin. Avaya recommends that you not use any unsecured protocols. For more information, see High Secure mode.

Use Secure Copy (SCP) rather than FTP or TFTP. For more information, see SSHv1/v2 on page 306.

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).

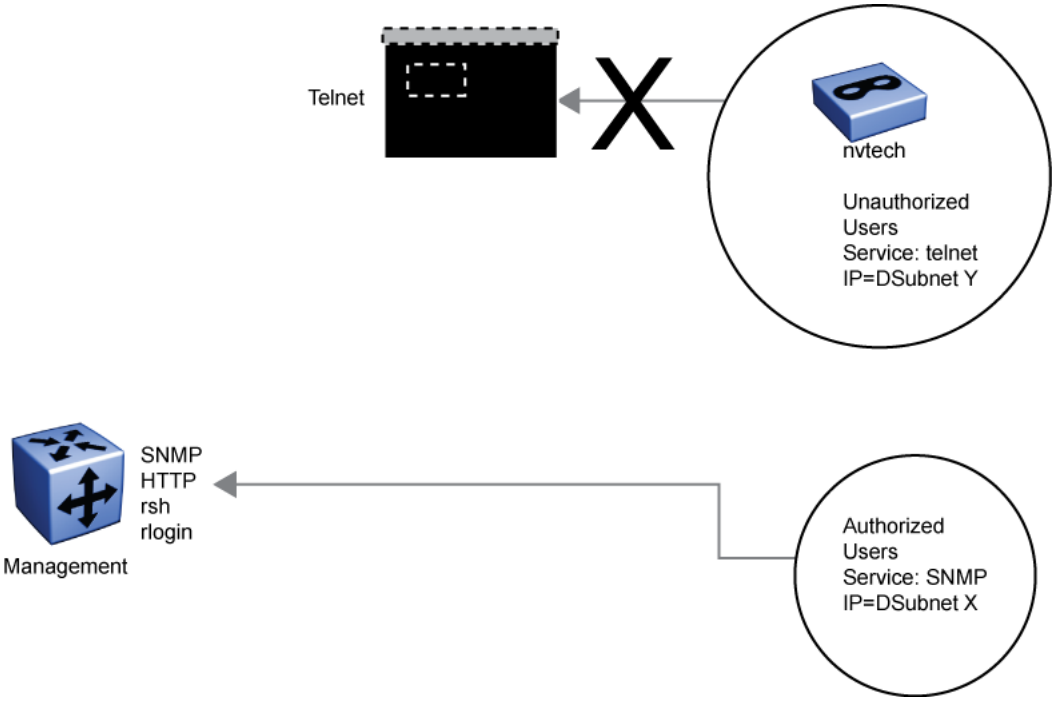


Figure 143: Access levels

Avaya 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 Avaya Ethernet Routing Switch 8800/8600 tries the second, and then attempts each server in sequence 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+

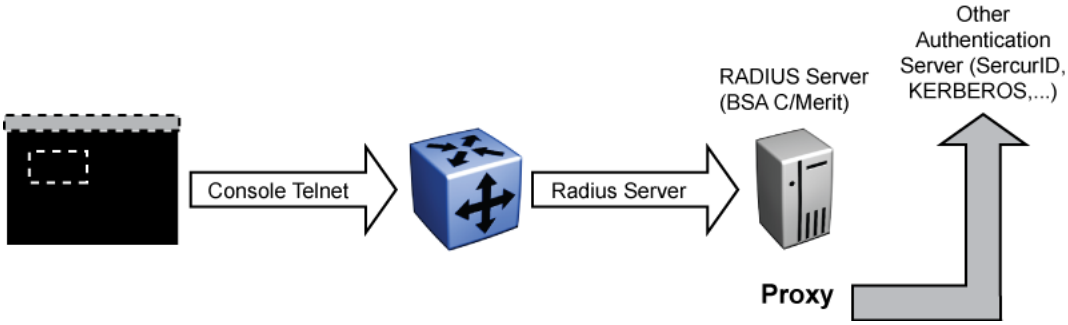


Figure 144: RADIUS server as proxy for stronger authentication

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.

Avaya 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 *Avaya Ethernet Routing Switch 8800/8600 Security, NN46205-601*.

RADIUS over IPv6

The Avaya Ethernet Routing Switch 8800/8600 supports RADIUS over IPv6 networks to provide security against unauthorized access.

For more information about RADIUS over IPv6, see *Avaya Ethernet Routing Switch 8800/8600 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).

Important:

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.

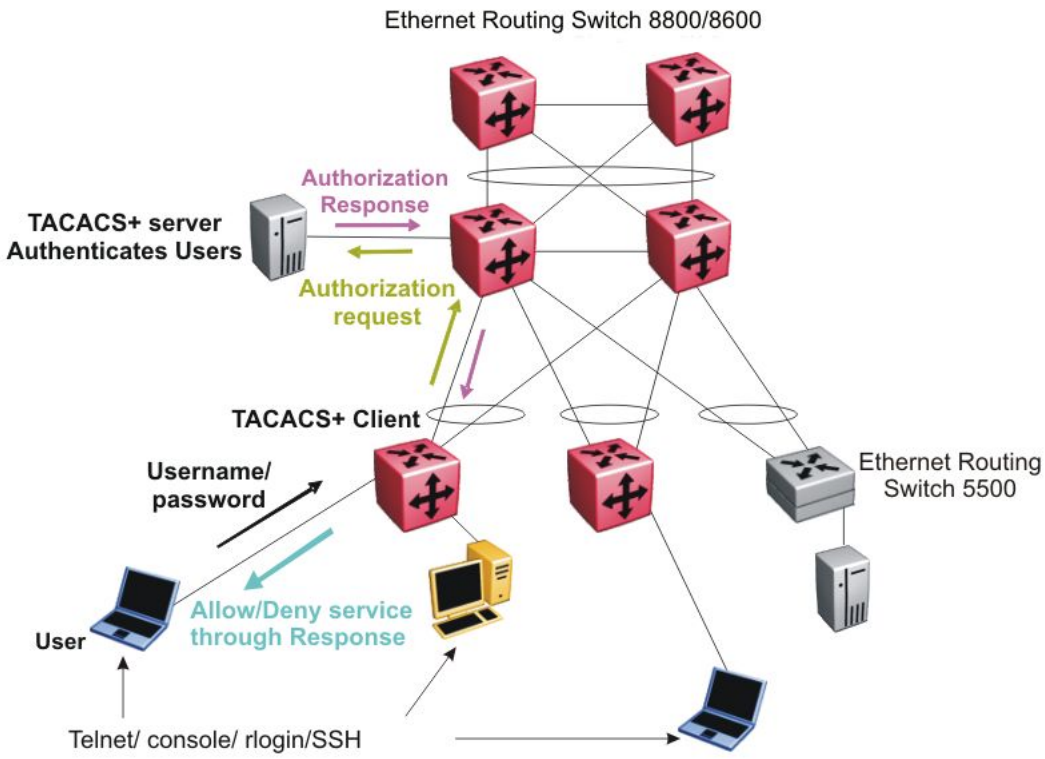


Figure 145: TACACS+

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 Avaya Ethernet Routing Switch 8800/8600 supports eight users logged in to the chassis simultaneously with TACACS+.

For more information about TACACS+, see *Avaya Ethernet Routing Switch 8800/8600 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 Avaya Ethernet Routing Switch 8800/8600 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.

Avaya recommends that you use SNMP version 3. SNMPv3 provides stronger authentication services and the encryption of data traffic for network management.

Other security equipment

Avaya 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.

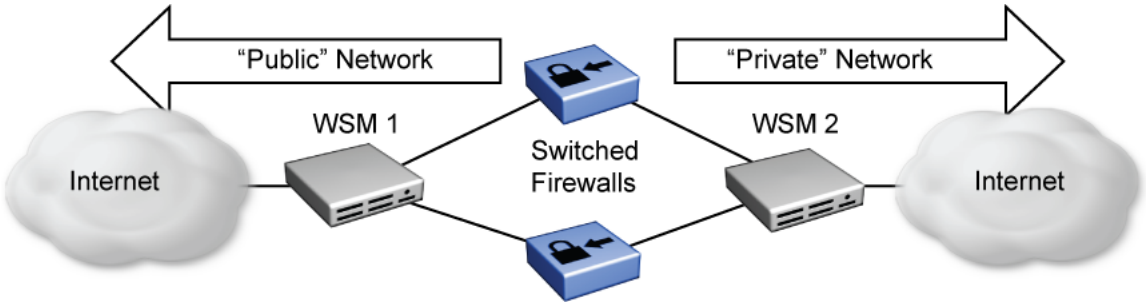


Figure 146: Firewall load balancing configuration

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 Avaya 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, Avaya 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)

Chapter 16: 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 *Avaya Ethernet Routing Switch 8800/8600 Configuration — QoS and IP Filtering for R and RS Modules, NN46205-507*.

QoS mechanisms

The Avaya Ethernet Routing Switch 8800/8600 has a solid, well-defined architecture to handle QoS in an efficient and effective manner. Several QoS mechanisms used by the Ethernet Routing Switch 8800/8600 are briefly described in the sections that follow.

QoS mechanisms navigation

- QoS classification and mapping on page 309
- QoS and queues on page 311
- QoS and filters on page 312
- Policing and shaping on page 314

QoS classification and mapping

The Avaya Ethernet Routing Switch 8800/8600 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 8800/8600 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 8800/8600 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 8800/8600 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 32: ADSSC, DSCP, and 802.1p-bit mappings on page 310 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, Avaya recommends that you do not change the mappings.

Table 32: ADSSC, DSCP, and 802.1p-bit mappings

ADSSC	DSCP	802.1p
Critical	CS7	7
Network	CS6	7
Premium	EF, CS5	6
Platinum	AF4x, CS4	5
Gold	AF3x, CS3	4
Silver	AF2x, CS2	3
Bronze	AF1x, CS1	2
Standard	DE, CS0	0
Custom/best effort	User Defined	1

In this table, ADSSC denotes Avaya Data Solutions Service Class, CS denotes Class Selector, EF denotes Expedited Forwarding, AF denotes Assured Forwarding, and DE denotes Default forwarding.

Important:

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 Avaya 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 Avaya

Data Solutions Service Classes (ADSSC). The following table provides a summary of the mappings and their typical traffic types.

Table 33: Traffic categories and ADSSC mappings

Traffic category		Application example	ADSSC
Network Control		Alarms and heartbeats	Critical
		Routing table updates	Network
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 ADSSC 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 Avaya Ethernet Routing Switch 8800/8600 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 *Avaya Ethernet Routing Switch 8800/8600 Configuration — QoS and IP Filtering for R and RS Modules, NN46205-507*.

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.

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 remarked 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 Avaya Ethernet Routing Switch 8800/8600, 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 **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.

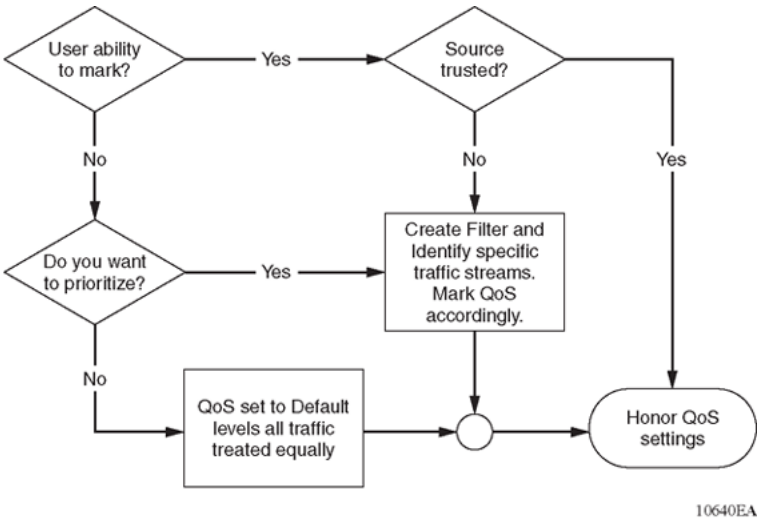


Figure 147: Filter decision-making process

Advanced filters

Advanced filters are provided 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 ACEsper port for ingress and egress. The Avaya Ethernet Routing Switch 8800/8600 supports a maximum of 4000 ACEs. For each ACL, a maximum of 500 ACEsare supported.

When you configure Advanced filters, keep the following scaling limits in mind.

Table 34: 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 Advanced 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 Avaya Ethernet Routing Switch 8800/8600, 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
- remark 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 8800/8600 supports traffic shaping at the port level and at the per-transmit-queue level for outgoing traffic.

Provisioning QoS networks using Advanced filters

You can use Advanced filters (ACLs) to provision the network.

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.

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 on page 316
- Bridged and routed traffic on page 317
- 802.1p and 802.1Q recommendations on page 317

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.

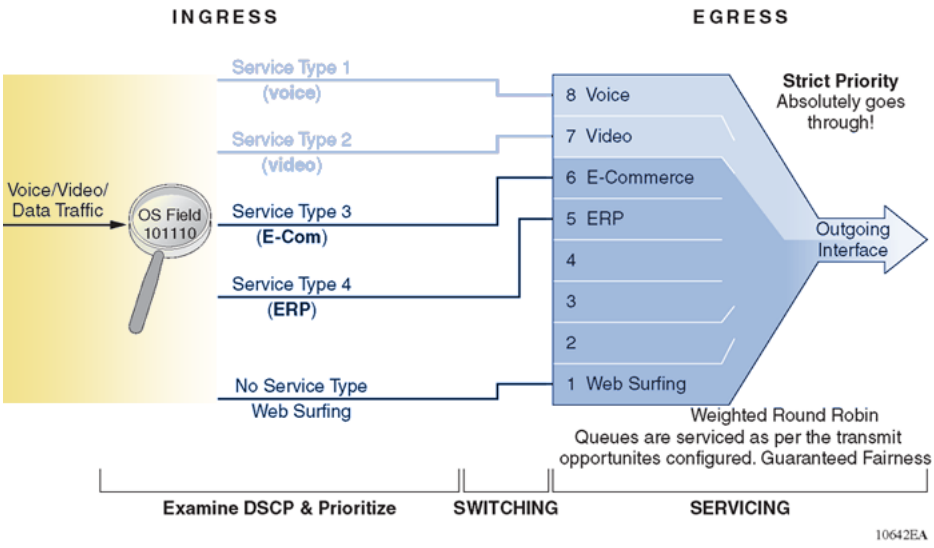


Figure 148: Core port QoS actions

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 remark 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 remarked with the appropriate DSCP and 802.1p markings, and given an internal QoS level. This remarking is done based on the filters and traffic policies that you configure. The following figure shows access port actions.

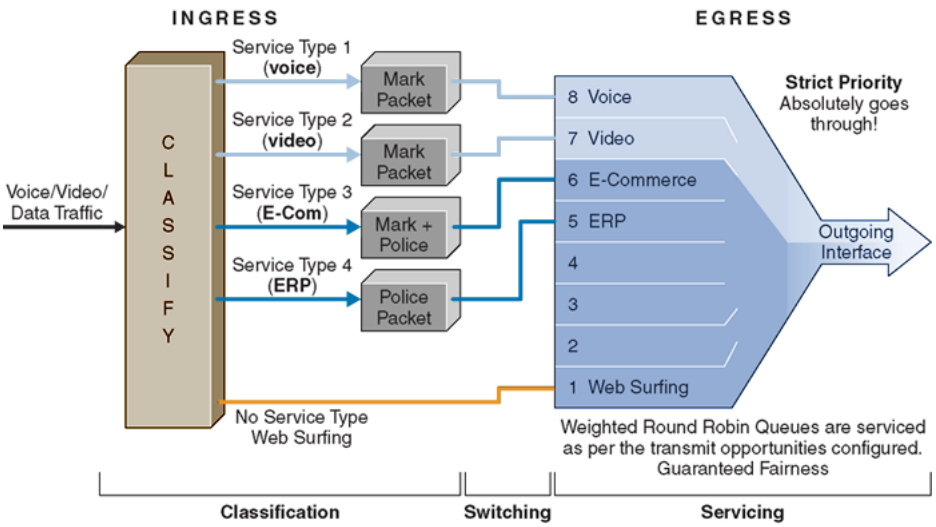


Figure 149: Access port QoS actions

Bridged and routed traffic

In a service provider network, access nodes use the Avaya Ethernet Routing Switch 8800/8600 configured for bridging. In this case, the Ethernet Routing Switch 8800/8600 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 8800/8600 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 8800/8600 configured for bridging, and core nodes use the Ethernet Routing Switch 8800/8600 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 in accordance with Table 32: ADSSC, DSCP, and 802.1p-bit mappings on page 310.

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 Avaya Ethernet Routing Switch 8800/8600 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 8800/8600, 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 8800/8600

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 remarked 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 Avaya Ethernet Routing Switch 8800/8600 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 8800/8600 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 Avaya Ethernet Routing Switch 8800/8600 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.

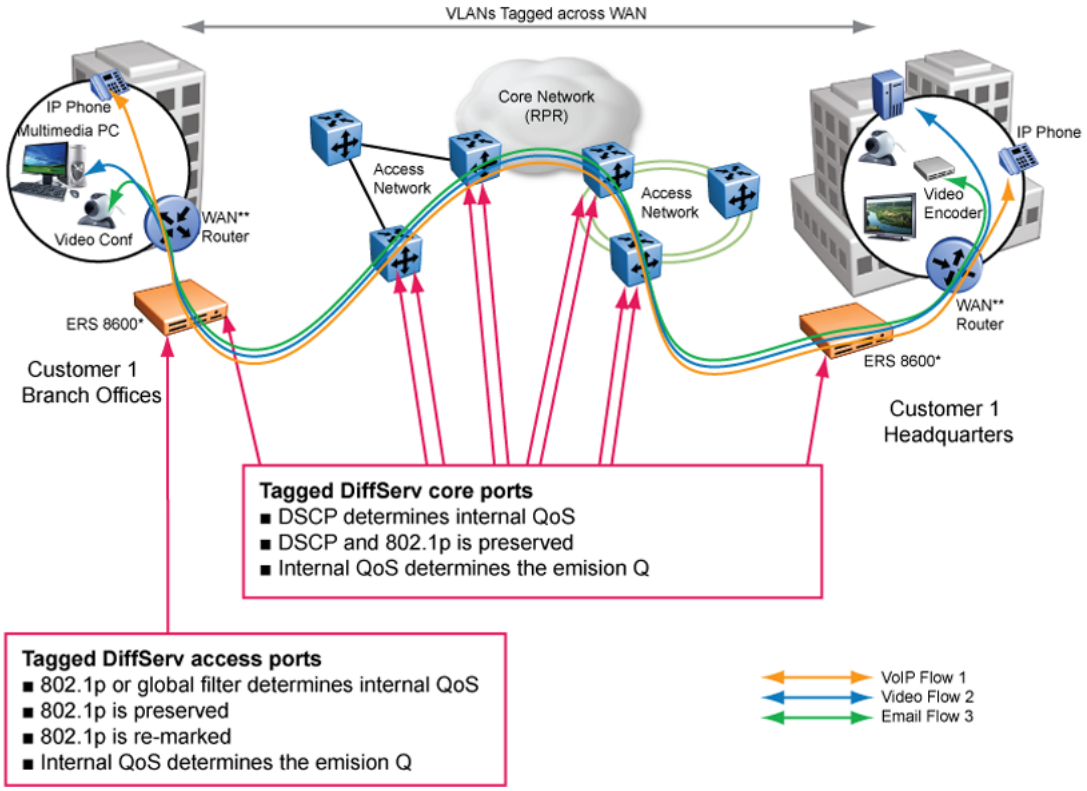


Figure 150: Trusted bridged traffic

The following figure shows what happens inside an Ethernet Routing Switch 8800/8600 access node. Packets enter through a tagged or untagged access port, and exit through a tagged or untagged core port.

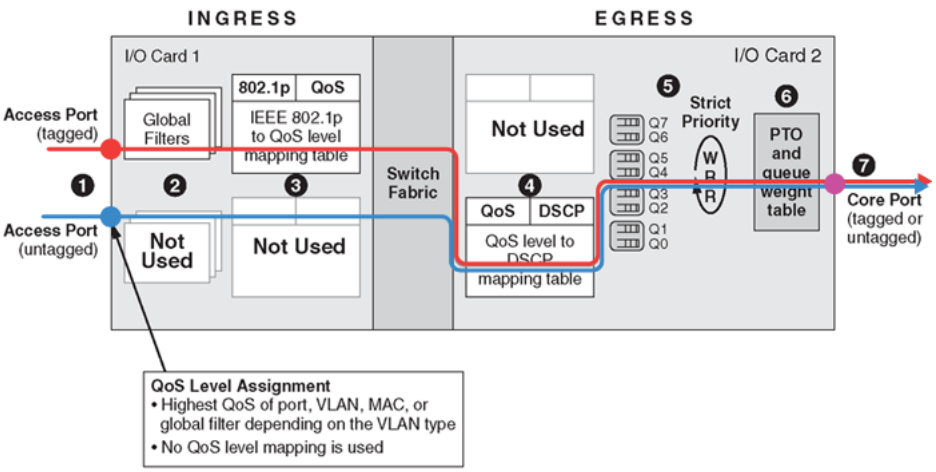


Figure 151: QoS actions on bridged access ports

The following figure shows what happens inside an Ethernet Routing Switch 8800/8600 core node. Packets enter through a tagged or untagged core port, and exit through a tagged or untagged core port.

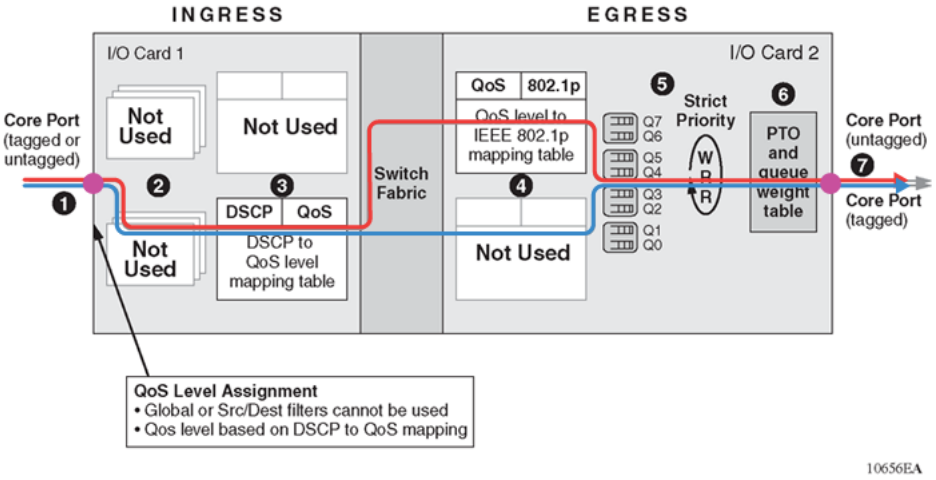


Figure 152: QoS actions on bridged or routed core ports

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.

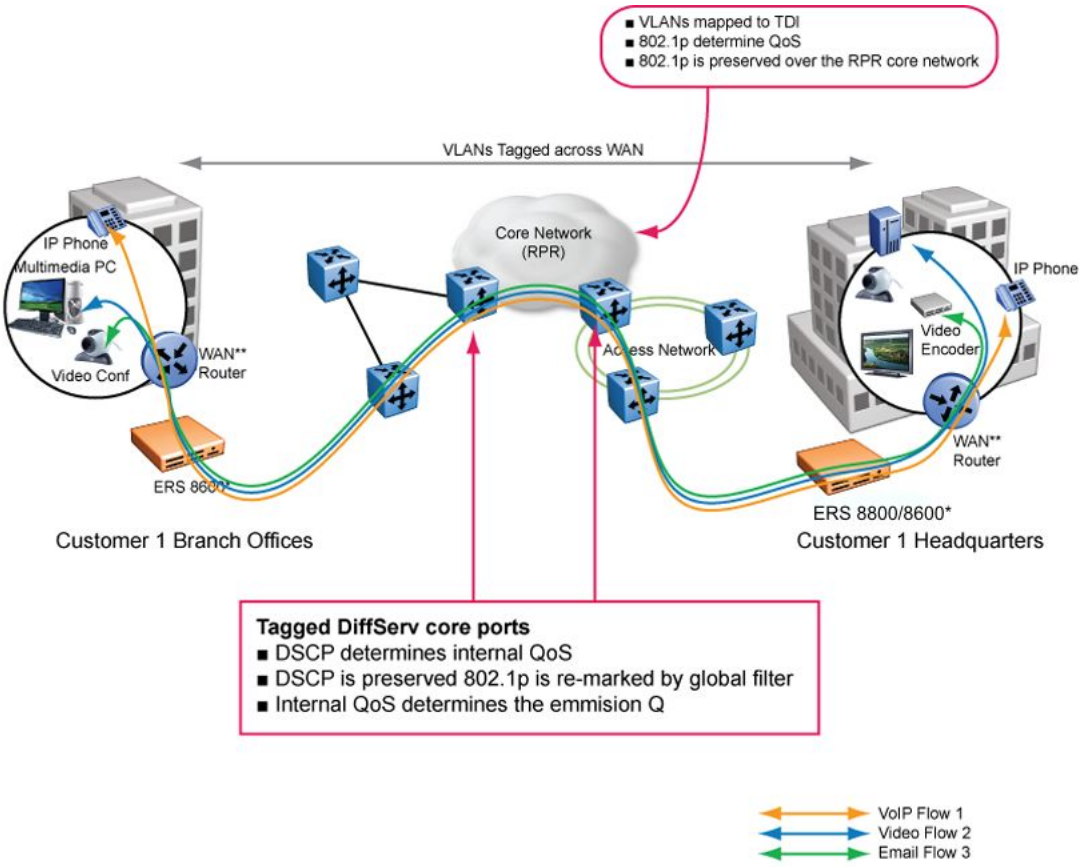


Figure 153: RPR QoS internetworking

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 Avaya Ethernet Routing Switch 8800/8600 network.

Routed trusted traffic

When you set the port to core, you assumethat, for all incoming traffic, the QoS setting is properly marked.All core switch ports simply read and forward packets. The packetsare not re-marked or reclassifed from the switch. All initial QoSmarkings 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.

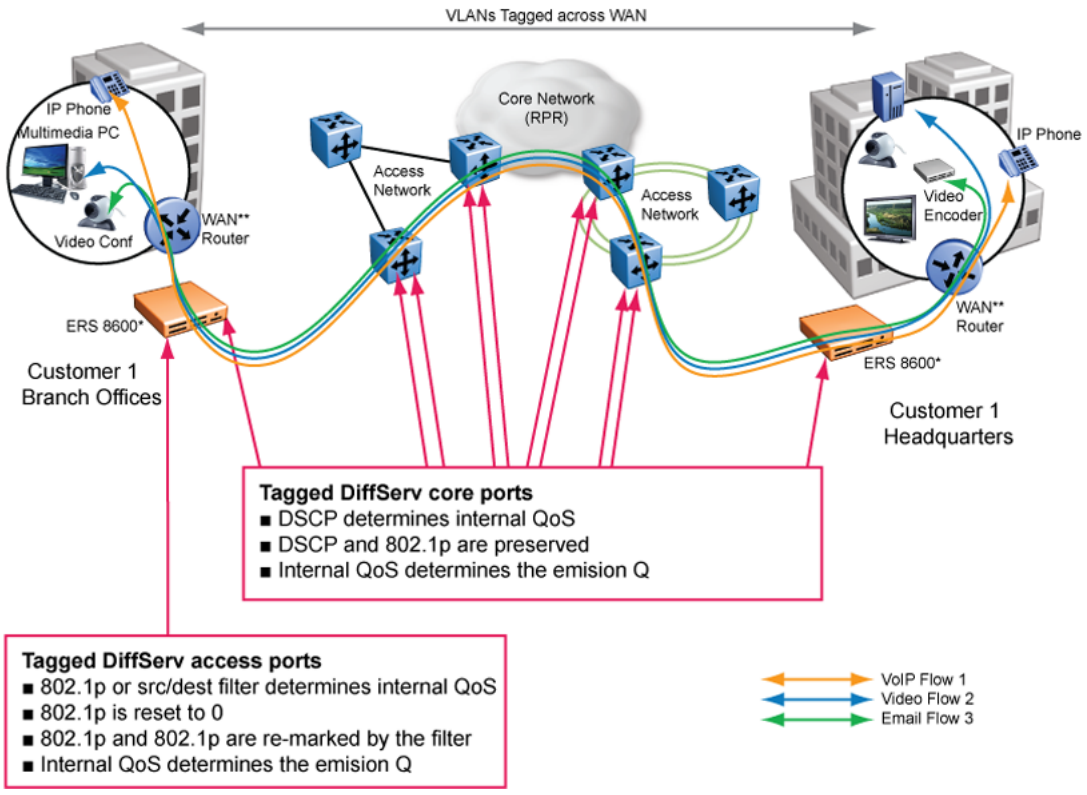


Figure 154: Trusted routed traffic

Routed untrusted traffic

The following figure shows what happens inside an Avaya Ethernet Routing Switch 8800/8600 access node. Packets enter through a tagged or untagged access port and exit through a tagged or untagged core port.

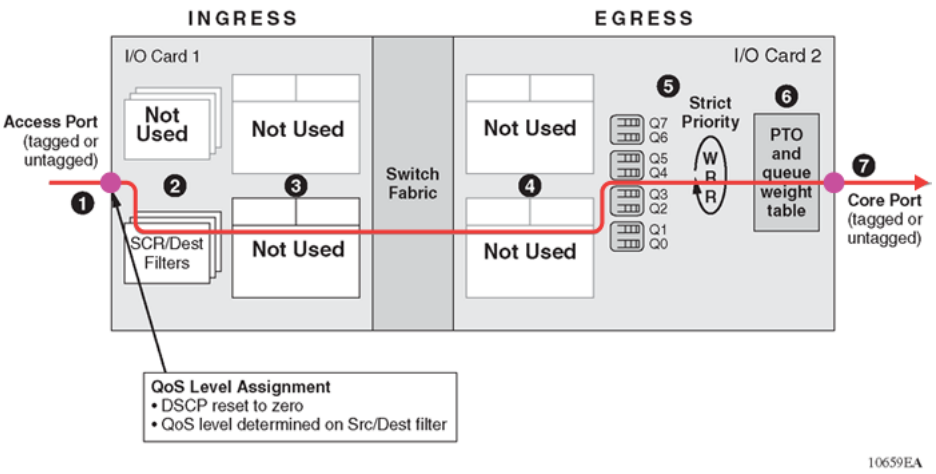


Figure 155: QoS actions on routed access ports

Chapter 17: Customer service

Visit the Avaya Web site to access the complete range of services and support that Avaya provides. Go to www.avaya.com or go to one of the pages listed in the following sections.

Getting technical documentation

To download and print selected technical publications and release notes directly from the Internet, go to www.avaya.com/support.

Getting Product training

Ongoing product training is available. For more information or to register, you can access the Web site at www.avaya.com/support. From this Web site, you can locate the Training contacts link on the left-hand navigation pane.

Getting help from a distributor or reseller

If you purchased a service contract for your Avaya product from a distributor or authorized reseller, contact the technical support staff for that distributor or reseller for assistance.

Getting technical support from the Avaya Web site

The easiest and most effective way to get technical support for Avaya products is from the Avaya Technical Support Web site at www.avaya.com/support.

Appendix A: Hardware and supporting software compatibility

The following table describes Avaya Ethernet Routing Switch 8800/8600 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 35: Ethernet Routing Switch 8800/8600 chassis and SF/CPUs

Item		Minimum software version	Part number
Chassis			
8010 chassis	10-slot chassis	3.0.0	DS1402001-E5 DS1402001-E5GS
8006 chassis	6-slot chassis	3.0.0	DS1402002-E5 DS1402002-E5GS
8010co chassis	10-slot chassis	3.1.2	DS1402004-E5 DS1402004-E5GS
8003-R chassis	3-slot chassis	7.0.0	DS1402011- E5
Switch fabric/CPU			
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
8895 SF/CPU	Switch fabric	7.0.0	DS1404120-E5
Power supplies			
8004AC	850 W AC	3.1.2	DS1405x08
8004DC	850 W DC	3.1.2	DS1405007
8005AC	1462 W AC	4.0.0	DS1405012

Item		Minimum software version	Part number
8005DI AC	1462 W Dual input AC	5.0	DS1405018-E6
8005DI DC	1462 W Dual input DC	5.1	DS1405017-E5
8005DC	1462 W DC	4.0.x	DS1405011

Table 36: Ethernet Routing Switch 8800/8600 modules and components

Module or component		Minimum software version	Part number
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
8800 series modules			
8848GT	48-port 10Base-T/100Base-TX/1000Base-T	7.1	DS1404124-E6
8848GB	48-port 1000Base-X SFP	7.1	DS1404122-E6
8834XG	2-port XFP, 24-port 1000Base-X SFP, 8-port RJ-45	7.1	DS1404123-E6
8812XL	12-port 10 Gbit/s LAN, supports SFP+	7.1.3	DS1404121-E6
SFPs			

Module or component		Minimum software version	Part number
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
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-BX	Up to 40 km, 1310 nm	7.0	AA1419076- E6
1000BASE-BX	Up to 40 km, 1490 nm	7.0	AA1419077- E6
1000BASE-EX	Up to 120 km, 1550 nm DDI	5.0.0	AA1419071-E6
SFP+s			
10GBASE-LR	1310 nm single-mode fiber (SMF). The range is up to 10 km.	7.1.3	AA1403011–E6
10GBASE-ER	1550 nm SMF. The range is up to 40 km.	7.1.3	AA1403013–E6
10GBASE-SR	850 nanometers (nm). The range is up to: • 22 m using 62.5 micrometer (µm), 160 megaHertz times km (MHz-km) MMF. • 33 m using 62.5 µm, 200 MHz-km MMF. • 66 m using 62.5 µm, 500 MHz-km MMF.	7.1.3	AA1403015–E6

Module or component		Minimum software version	Part number
	• 82 m using 50 µm, 500 MHz-km MMF. • 300 m using 50 µm, 2000 MHz-km MMF.		
10GBASE-LRM	1310 nm. Up to 220 m reach over Fiber Distributed Data Interface (FDDI)-grade 62.5 µm multimode fiber. Suited for campus LANs.	7.1.3	AA1403017–E6
10GBASE-CX	4-pair direct attach twinaxial copper cable to connect 10 Gb ports. The maximum range is 10 m.	7.1.3	AA1403018–E6
10GBASE-CX	4-pair direct attach twinaxial copper cable to connect 10 Gb ports. The maximum range is 3 m.	7.1.3	AA1403019–E6
10GBASE-CX	4-pair direct attach twinaxial copper cable to connect 10 Gb ports. The maximum range is 5 m.	7.1.3	AA1403020–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

Module or component		Minimum software version	Part number
10GBASE DWDM	1534.25 nm (195.40 THz)	5.1.0	NTK587AQE5
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

Appendix B: 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 37: Supported IEEE standards

Supported standard	Description
EEE 802.1D (2001 standard)	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	Autonegotiation 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.

IPv4 Layer 3/Layer 4 Intelligence

The following table describes the supported IETF RFCs for IPv4 Layer 3/Layer 4 Intelligence.

Table 38: IPv4 Layer 3/Layer 4 Intelligence RFCs

Supported standard	Description
RFC 768	UDP Protocol
RFC 783	TFTP Protocol
RFC 791	IP Protocol
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

Supported standard	Description
RFC 951 / RFC 2131	BootP/DHCP
RFC 1027	Using ARP to implement transparent subnet gateways/ Avaya 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 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
RFC 1966	BGP-4 Route Reflectors
RFC 1998	An Application of the BGP Community Attribute in Multihome 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

Supported standard	Description
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
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

Supported standard	Description
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 39: 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)
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 40: 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

Supported standard	Description
RFC 3587	IPv6 Global Unicast Address Format
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, EDM support for IPv6

Platform

The following table describes the supported IETF platform RFCs.

Table 41: 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 42: 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 43: Network Management RFCs

Supported standard	Description
RFC 1155	SMI

Supported standard	Description
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
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 and 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 Avaya Ethernet Routing Switch 8800/8600 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 Avaya 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 44: Standard IEEE MIBs

Protocol	IEEE standard	File name
LACP	802.3ad	ieee802-lag.mib
EAPoL	802.1x	ieee8021x.mib

Table 45: 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 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 number	MIB name
RFC 1658	Definitions of Managed Objects for Character Stream Devices 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 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 number	MIB name
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 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 46: 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
Avaya RSTP/MSTP proprietary MIBs	nnrst000.mib, nnmst000.mib
Avaya IGMP MIB	rfc_igmp.mib
Avaya IP Multicast MIB	ipmroute_rcc.mib
Avaya DVMRP MIB	dvmrp_rcc.mib
Avaya PIM MIB	pim-rcc.mib
Avaya MIB definitions	wf_com.mib

Proprietary MIB name	File name
Avaya PGM MIB	wf_pgm.mib
The Definitions of Managed Objects for the Link Control Protocol of the Point-to-Point Protocol – Avaya Proprietary	rfc1471rcc.mib
The Definitions of Managed Objects for the IP Network Control Protocol of the Point-to-Point Protocol – Avaya 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 – Avaya Proprietary	rfc1595rcc.mib
OSPF Version 2 Management Information Base – Avaya proprietary extensions	rfc1850t_rcc.mib
Avaya IPv6 proprietary MIB definitions	rfc_ipv6_tc.mib, inet_address_tc.mib, ipv6_flow_label.mib

Glossary

add/drop multiplexer (ADM)	A network element in which facilities are added, dropped, or passed directly through for transmission to other network elements.
bit error rate (BER)	The ratio of the number of bit errors to the total number of bits transmitted in a given time interval.
coarse wavelength division multiplexing (CWDM)	A technology that uses multiple optical signals with different wavelengths to simultaneously transmit in the same direction over one fiber, and then separates by wavelength at the distant end.
Custom AutoNegotiation Advertisement (CANA)	An enhancement of the IEEE 802.3 autonegotiation process on the 10/100/1000 copper ports. Custom AutoNegotiation Advertisement offers improved control over the autonegotiation process. The system advertises all port capabilities that include, for tri-speed ports, 10 Mb/s, 100 Mb/s, 1000 Mb/s speeds, and duplex and half-duplex modes of operation. This advertisement results in autonegotiation between the local and remote end that settles on the highest common denominator. Custom AutoNegotiation Advertisement can advertise a user-defined subset of the capabilities that settle on a lower or particular capability.
forwarding database (FDB)	A database that maps a port for every MAC address. If a packet is sent to a specific MAC address, the switch refers to the forwarding database for the corresponding port number and sends the data packet through that port.
MultiLink Trunking (MLT)	A method of link aggregation that uses multiple Ethernet trunks aggregated to provide a single logical trunk. A multilink trunk provides the combined bandwidth of multiple links and the physical layer protection against the failure of a single link.
Open Shortest Path First (OSPF)	A link-state routing protocol used as an Interior Gateway Protocol (IGP).
small form factor pluggable (SFP)	A hot-swappable input and output enhancement component used with Avaya products to allow gigabit Ethernet ports to link with other gigabit Ethernet ports over various media types.
Split MultiLink Trunking (SMLT)	An Avaya extension to IEEE 802.1AX (link aggregation), provides nodal and link failure protection and flexible bandwidth scaling to improve on the level of Layer 2 resiliency.